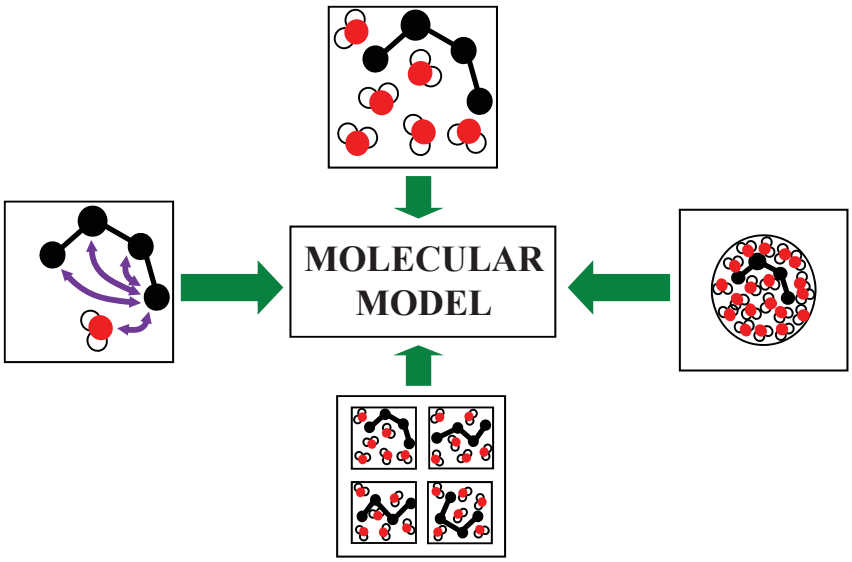
CSBMS

Classical Simulation of (Bio)Molecular Systems

Prof. P.H. Hünenberger



STUDENT LECTURE SCRIPT HS19

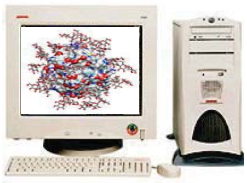
(version August 2019)

http://www.csms.ethz.ch/education/CSBMS
(downloads: use your nethz-password)

Computer Simulation in Chemistry, Biology and Physics

P.H. Hünenberger

Classical Simulation of (Bio)Molecular Systems



Lecture 529-0004-00

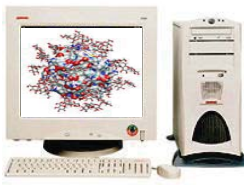
www.csms.ethz.ch/education/CSBMS

Herbstsemester 2019
Tuesday 9:45-11:30 a.m.
HCI D2

LECTURE 1 (WEEK 1):
Course Information
Introduction
GROMOS

The Hünenberger group

Prof. P.H. Hünenberger
Lab. of physical chemistry
HCI G233
[www.csms.ethz.ch,
phil@igc.phys.chem.ethz.ch]



Computer simulation of molecular systems (CSMS)

methodology and applications

Salomé Walthard

Philippe Hünenberger

Sadra Gheta

Marina Pereira

Alzbeta Kubincová

Course Organization: Lecture

- Web page: **www.csms.ethz.ch/education/CSBMS**
  → All basic information on the lecture / exercises
  → Updated regularly over the semester
  → Worth a good read this afternoon / evening !

- Updates

check this one out to see what has changed recently (so you don't have to re-read the full site every time...)

CSBMS (anciently CSCBP)

Classical Simulation of (Bio)Molecular Systems (CSBMS)

Prof. Philippe H. Hünenberger / HS19

Recent updates

date	update
22.08.2019	Update of the web pages in progress for HS19
-	Content is not yet final for HS19 and may fluctuate!!!

Recent updates (last few weeks) in the CSCBP web pages

Course Organization: Lecture

- Lecture

Lecture
Tuesdays, 9.45-11.30 hrs, HCI → D2 → , ETH Hönggerberg (14x2 hours); lecture N° 529-0004-00; note that ***the break is usually reduced to 5 minutes***, so that the actual time is 9.45-11:20 instead

- Lecture material

Course material
- A booklet with the ***lecture slides*** is distributed at the start of the semester (the corresponding pdf file can also be found in the documents → page)
- The ***slides*** corresponding to the lecture as given (which may also slightly differ from those in the booklet) will also be posted on this site after each lecture (pdf files in the documents → page)

slides marked with ▽

- Requirements

Requirements
- To take advantage of the lectures, a good ***general knowledge*** of ***mathematics, physics, chemistry and biology*** is required
- To take advantage of the exercises, a basic knowledge of ***computer science*** and a good practical knowledge of the ***UNIX operating system*** are required

Talk to me if you have doubts!

Course Organization: Lecture

- Lecture schedule

Lecture schedule

lecture	week	date	theme	exercise
1	38	17.09.2019	Introduction / Molecular models / GROMOS	-
2	39	24.09.2019	Force-fields	1
3	40	01.10.2019	Force-fields / Sampling	1
4	41	08.10.2019	Sampling	2
5	42	15.10.2019	Boundary conditions	2
6	43	22.10.2019	Electrostatic interactions	3
7	44	29.10.2019	Analysis of simulations	3
8	45	05.11.2019	Free-energy calculations I	4
9	46	12.11.2019	Free-energy calculations II	4
10	47	19.11.2019	Enhanced sampling	5
11	48	26.11.2019	Structure refinement	5
12	49	03.12.2019	Special topics (assistant presentations)	6
13	50	10.12.2019	Answer to thinking questions	6
14	51	17.12.2019	Concluding remarks	-

Lecture number, calendar week, date, theme, and exercise number of the exercise in progress during this week

(may be slightly readjusted in the course of the semester)

Course Organization: Lecture

- Assessment

Assessment

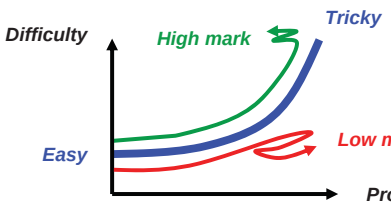
- The assessment of the course consists of an ***oral examination*** of 30 minutes duration, probing the entire content of the lecture
- Since the practical exercises do convey different skills as those being conveyed during the lectures, ***the performance in the exercises are taken into account*** in the final exam mark
- ***Bachelor and Master students*** must do the practical exercises (and take the exam)
- ***Ph.D. students and postdocs*** need not do the practical exercises to get their ETH credit points or Zulassungsprüng (but are very welcome to join if they wish)

- Exam: January or August

→ Oral exam, 30', no preparation, Phil + Beisitzer

→ Atmosphere is friendly

→ Basic principle of an oral exam



Corollary: the number of good answers is not immediately related to the mark...

- Most important

→ Understanding of the **working principles**

→ It is a good idea to know some of the **key equations** "by heart" (there is no time to derive everything); but knowing "by heart" won't help at all if you don't know the working principles

PHIL'S TIPS FOR ORAL EXAMS:
(1) leave your emotions behind on entrance (e.g. nervosity) and take them back on exit (e.g. happiness, disappointment)
(2) never self-assess in the course of the exam (e.g. little voice that whispers you are doing badly): stay focused and give all you can!

Course Organization: Exercises

Exercises

The practical part of the course consists of ***six hands-on exercises***, each involving the set-up, execution and analysis of molecular simulations using the GROMOS program and force field. The optimal way of carrying out these exercises is to participate a ***two-hours weekly session*** supervised by two assistants in our computer room HCI → D267 →

Organisation

- There are ***six exercises*** to be carried out. The exercises are formally independent from each other, but skills acquired at one exercise are assumed to be available for the next. The exercises ***start on the second semester week***. In the first semester week, there will be a short introduction to GROMOS at the lecture.
- Each exercise lasts ***two weeks*** (see table "Exercise Schedule" below). Typically, week 1 is planned for the set-up and week 2 for the analysis and discussion. The actual computation (simulation) is carried out in-between, using our group PC cluster "beaver"
- Each exercise is documented in a ***detailed script*** (distributed in principle at the preceding Tuesday lecture; pdf also deposited in the documents → page at this point). The document labelled exercise "zero" (and the associated quickref sheet) describes the computational setup, and is associated with the first real exercise
- The optimal way of carrying out these exercises is to ***participate on both week 1 and week 2*** in a ***two-hours session*** supervised by ***two assistants*** in our ***computer room*** HCI → D267 → . We offer ***two options*** for this two-hours block on different days of the week (see paragraph "Exercise Sessions" below) and you can chose the session that fits best your schedule. It is in principle allowed to do the exercises on your own instead or to only come for part of the two-hours blocks, but not recommended (you will spend more time and learn less). In any case, the assistant responsible for a given exercise will be available for you over the two weeks, even outside the two-hours blocks if needed (email, personal meeting)
- For each exercise, you are expected to hand in a ***short report*** (if possible, maximum 2-3 pages of text, *i.e.* excluding the space taken by possible graphs/tables) summarizing briefly your findings and answering some "thinking questions" asked in the script. The ***deadline for handing in the report*** is one week after closing of the specific exercise (see table "Exercise Schedule" below)
- The exercises can be made ***individually or in groups of two students*** (recommendation: work individually at the computer and, if you want to make a pair, discuss your findings during the course of the session and make the report together)
- The ***enrolment*** for the exercises is at the first lecture

HS18,HS19: distributed at start (booklet)

Course Organization: Exercises

- Principle

→ **Six** exercise series, lasting **two weeks** each

starting from the second to the before-last semester week

→ Progressive **build up** of skills in **setting up**, **executing** and **analyzing** simulations

→ must be done in sequence + don't miss one (especially of ex 1-3)

→ **Hands-on**, using **GROMOS**, workstations in **HCI D267.4**, computations on **beaver** cluster

→ later run your own simulations using GROMOS

→ or with other packages (GROMACS, CHARMM, AMBER, NAMD, ...) as the basic principles underlying all condensed-phase (bio)molecular simulation programs are the same

→ Documented in a **computer-setup** + **six exercise scripts** (distributed as booklet at semester start)

→ Short (max. 2-3 page excl. graphs/tables) **reports** to be handed in after each exercise

the 2-3 pages max is rather to protect you from "overwork" !!!

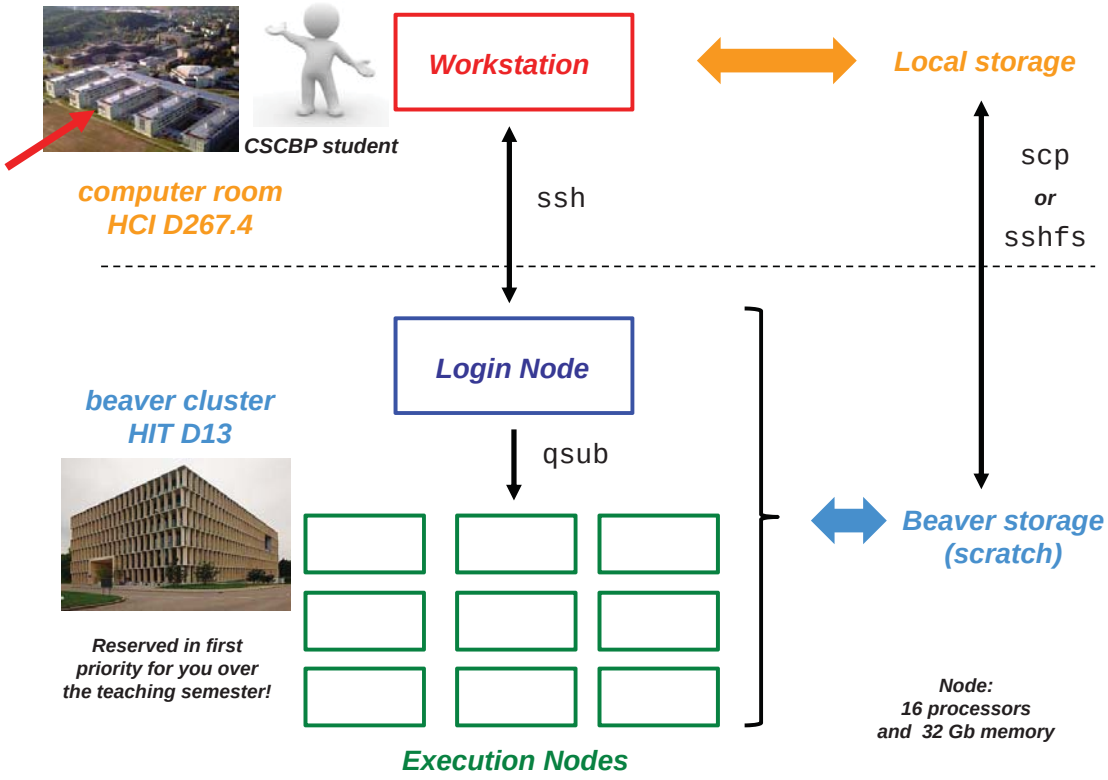
the scripts end with a number of "thinking questions" – try to answer them, but don't worry too much if you cannot answer them all (it is about thinking, not about completeness)

→ **Assessment**: the corrected reports will be returned to you; I'll also keep some rough record (for mark rounding at the exam)

the main goal of the reports is <u>not</u> assessment: it is about considering / interpreting / questioning your raw computer outputs

Course Organization: Exercises

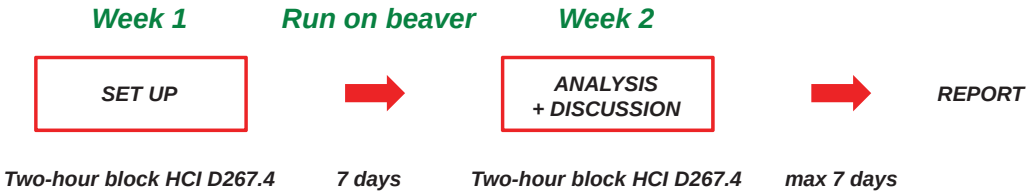
- Setup



Course Organization: Exercises

- Principle (continued)

→ **Two-week** rhythm *General principle (there may be variations)*



→ **Individually** or by **groups of two** *Advice: make exercises individually at the computer, even if you make a group of two for the report*

- Note on the downloads before starting each exercise:
  - → You should copy the directory **/usr/local/CSCBP/ex*** (with *=1..6) only on the **first day** of a given exercise
  - → Reason: the assistants may modify the content until the last minute and if you are out of sync with the latest files, things may not work as expected

Course Organization: Exercises

- Beaver
  - → You will receive an **e-mail** about your account
  - → Please **change the password** as soon as possible
  - → To **login** to beaver from the outside `ssh user@realbeaver.ethz.ch`
- D267.4 computers
  - → **Login** from outside is only possible *via* first a "ssh" to the ETHZ ID login machines `ssh user@slab[1-4].ethz.ch`
  - → Then try another "ssh" to one of the D267.4 machines (but these machines are not always switched on!) `ssh user@slabhcib[002-041].ethz.ch`
  - → For testing purposes, it may be sufficient to work on the login machines (which are always switched on)
  - → This does not seem to require "vpn"
  - → Use "ssh -X" or "ssh -Y" to enable X11 forwarding (*e.g.* "xmgrace")
- Usage of **GROMOS**
  - → GROMOS is ***entirely free !*** – you are just asked to register (free license) on the web site `www.gromos.net`
  - → In principle, GROMOS should even work **on your own desktop/laptop** if you have a C++ compiler

... how easy it is in practice depends a bit on your computer setup (and the GROMOS team does not provide support for that – so, try & see)

Course Organization: Exercises

- Assistants

Assistant Schedule

first name	last name	room	tel	cde	1st	2nd	3rd
Stephanie	Linker →	HCI G239	34590	SL	1	2	4
Thomas	Stadelmann →	HCI G238/E314	34320/23144	TS	2	4	3
Sadra	Gheta →	HCI G243	26861	SG	3	6	1
Carmen	Esposito →	HCI G235	38898	CE	4	5	6
Alzbeta	Kubincova →	HCI G227	34593	AK	5	1	2
Shuzhe	Wang →	HSI G238	22347	SW	6	3	5

Philippe Hünenberger

Stephanie Linker

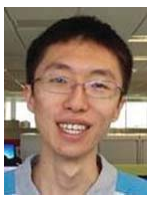
Thomas Stadelmann

Sadra Gheta

Carmen Esposito

Alzbeta Kubincová

Shuzhe Wang



Course Organization: Exercises

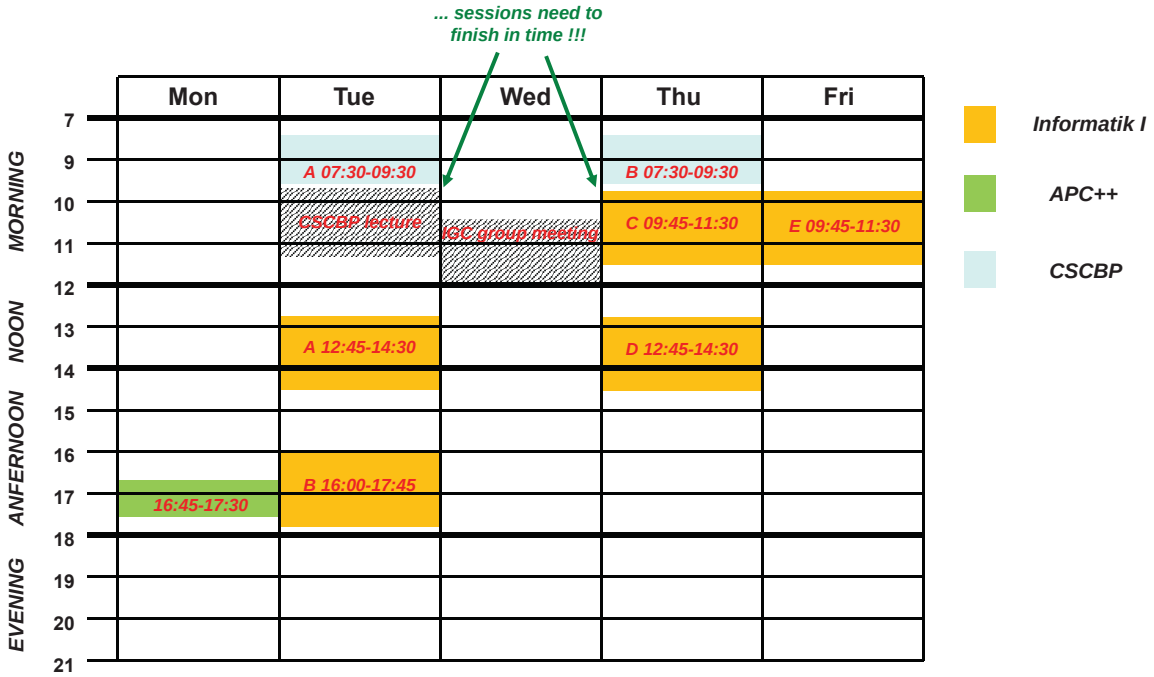
- Exercise schedule

Exercise schedule

exe	week	start	end	1st	2nd	3rd	theme	deadline
-	38	17.09	22.09	-	-	-	No exercises on the first week	-
1	39	23.09	29.09	SL	AK	SG	Topology building & first simulation	-
id.	40	30.09	06.10	id.	id.	id.	id.	13.10
2	41	07.10	13.10	TS	SL	AK	Peptide simulation & properties	-
id.	42	14.10	20.10	id.	id.	id.	id.	27.10
3	43	21.10	27.10	SG	SW	TS	Protein simulation & properties	-
id.	44	28.10	03.11	id.	id.	id.	id.	10.11
4	45	04.11	10.11	CE	TS	SL	Liquid simulation & properties	-
id.	46	11.11	17.11	id.	id.	id.	id.	24.11
5	47	19.11	25.11	AK	CE	SW	Free energy calculations	-
id.	48	25.11	01.12	id.	id.	id.	id.	08.12
6	49	02.12	08.12	SW	SG	CE	Structure refinement	-
id.	50	09.12	15.12	id.	id.	id.	id.	17.12
-	51	16.12	22.12	-	-	-	No exercises on the last week	-

Course Organization: Exercises

- Group sessions
  - → **two-hour block** with **two assistants**, computer room **HCI D267.4**
  - → we offer *two* **sessions per week**
  - → but you can still do (part of) the work on your own at other times or/and from home



Course Organization: Exercises

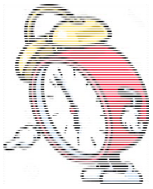
- In HS14, we have made all the exercise series **from scratch** !

~200 pages of «IKEA manual»

... was a lot of work!

→ Some things may still not go perfectly as expected and we'll have to **improvise a bit**

→ Your **understanding** (for what does not work) and constructive **feed-back** (for improving things) will be highly appreciated !



- Please

→ Be **ready to start** at the **indicated time** for the beginning of the session

The assistants cannot lead their sessions efficiently if everyone comes at a random time-point within the first 20 minutes !!!

→ **Take a look** at the **exercise scripts** before the session

Booklet distributed at semester start (+individual pdf's on web page [same content])

The time to read during the exercise sessions is relatively short, an a quick browsing in advance will already give you an idea of the overall exercise principle (this is especially important for the first two series!)

Course Organization: Documents

«locked» documents: use your n-ethz password

- Documents **(old page for HS18 – the page is by now [almost] updated for HS19!)**

after each lecture

coming soon

HS18, HS19: booklet from the start!

Documents

Computer Simulation in Chemistry, Biology and Physics (CSCBP)

Prof. Philippe H. Hünenberger / HS17

(documents with a "lock" require a n-ethz password)

document	version	upload	link
(none available yet)	HS19	22.08.2019	dummy.pdf

Lecture slides (lecture as given, added after each lecture)

Exercise Scripts

document	version	upload	link
(none available yet)	HS19	22.08.2019	dummy.pdf

Exercise scripts (unless stated explicitly, these scripts are rigorously identical to the material you can find in the exercise-script booklet distributed at the lecture start)

Miscellaneous Documents

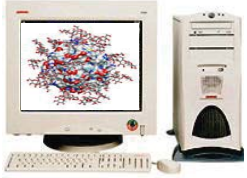
document	version	upload	link
Lecture script (student version for HS18)	HS18	29.08.2018	CSCBP_scr_ini_HS18.pdf
Exercise script (student version for HS18)	HS18	29.08.2018	CSCBP_exe_ini_HS18.pdf
Intro to UNIX (slides from Info I)	HS18	29.08.2018	CSCBP_unx_sld_HS18.pdf
Intro to UNIX (exercise from Info I)	HS18	29.08.2018	CSCBP_unx_exe_HS18.pdf
Simulation guide	HS13	30.08.2013	CSCBP_sim_gde_HS13.pdf
Simulation validation	HS13	30.08.2013	CSCBP_sim_vld_HS13.pdf
Wilfred van Gunsteren and GROMOS	HS13	26.09.2013	CSCBP_wfg_edi_HS13.pdf
Exsikkator article May 2014	2014	13.08.2014	CSCBP_exs_art_may_2014.pdf
Bibliometry essay Dec 2016	2016	12.12.2016	CSCBP_bib_ess_dec_2016.pdf
(Full infozine S1 issue)	2016	12.12.2016	CSCBP_bib_infozine_dec_2016.pdf
Computer vs brain essay Dec 2018	2018	04.12.2018	CSCBP_cvb_ess_nov_2018.pdf
(Full infozine S2 issue)	2018	04.12.2018	CSCBP_cvb_infozine_nov_2018.pdf
GROMOS manual (complete)	HS17	28.09.2017	CSCBP_gro_man_HS17.pdf
GROMOS manual (vol 1; overview)	HS17	28.09.2017	CSCBP_gro_man_v1_HS17.pdf
GROMOS manual (vol 2; algorithms)	HS17	28.09.2017	CSCBP_gro_man_v2_HS17.pdf
GROMOS manual (vol 3; force field)	HS17	28.09.2017	CSCBP_gro_man_v3_HS17.pdf
GROMOS manual (vol 4; files & formats)	HS17	28.09.2017	CSCBP_gro_man_v4_HS17.pdf
GROMOS manual (vol 5; program library)	HS17	28.09.2017	CSCBP_gro_man_v5_HS17.pdf
GROMOS manual (vol 6; technical details)	HS17	28.09.2017	CSCBP_gro_man_v6_HS17.pdf
GROMOS manual (vol 7; tutorial)	HS17	28.09.2017	CSCBP_gro_man_v7_HS17.pdf
GROMOS manual (vol 8; installation guide)	HS17	28.09.2017	CSCBP_gro_man_v8_HS17.pdf

QUESTIONS?

Computer Simulation in Chemistry, Biology and Physics

P.H. Hünenberger

COMPUTER SIMULATION OF MOLECULAR SYSTEMS



Herbstsemester 2019
Tuesday 9:45-11:30 a.m.
HCI D2

Lecture 529-0004-00
www.csms.ethz.ch/education/CSCBP

LECTURE 1 (WEEK 1):
Introduction

Why simulation ?

Every attempt to employ mathematical methods in the study of chemical questions must be considered profoundly irrational and contrary to the spirit of chemistry. If mathematical analysis should ever hold a prominent place in chemistry - an aberration which is happily almost impossible - it would occasion a rapid and widespread degeneration of that science.

Auguste Comte, “Cours de philosophie positive”,1830, volume I

For those who want to leave the room now, it is still time... for the others, let’s prove him wrong...

(although... in the big-data era... I am sometimes wondering if there is not still a little bit of truth in there...)

Why simulation ?

The advantages of simulation are learned early in life…

EXPERIMENTAL ROUTE

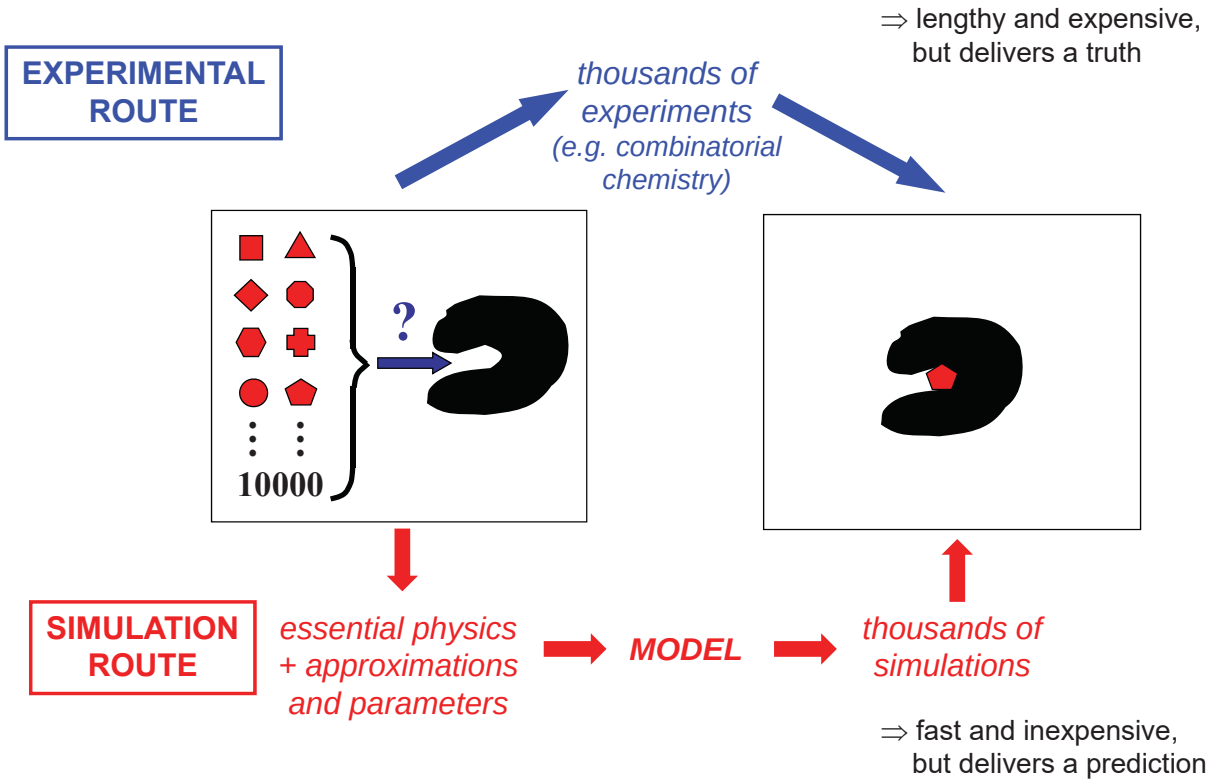
be hungry → *look unhappy* → *cry a bit* → *cry really loud*

SIMULATION ROUTE

crying loud is the key → *don’t wait till you are hungry…* → *cry really loud*

(1) make a model **(2) simulate**

Why simulation ?



★ = key slides ▽

When is simulation useful ?

The equations governing the model may be

but: a lot of chemistry deals with liquids, solutions & (bio)polymers !

- very simple ⇒ analytical treatment (*e.g.* ideal gas, harmonic crystal)
- moderately complex but numerous ⇒ computer simulation (numerical solution)
- not known or too complex ⇒ small-scale simulation (*e.g.* avalanches)

Simulation is used *instead* of experiment when

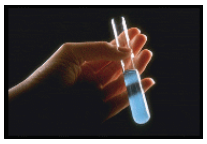
- the process **cannot** be studied experimentally
  e.g. interior of a star, weather forecast (experiment is too late !)
- the process is **dangerous** or **unethical** to study experimentally
  e.g. spread of an epidemy, flight simulators, explosion of a nuclear bomb, fighting ability of the Swiss army
- the process is **expensive** to study experimentally
  e.g. volcanism on Venus, aerodynamics in aircraft design

e.g. Führungssimulator in Kriens

Simulation is used *in complement to* experiment when

- approximate simulations may **reduce** the number of experiments to be performed or/and **increase** their likelihood of success
  e.g. modeling in industry: drug design, protein engineering, stock market predictions (banks), risk assessment (insurances)
- a simulation reproducing an experiment provides additional **insight**
  e.g. modeling in academia: quantum chemistry, molecular simulations

Why do molecular simulations provide insight ?

	Experiment	*Classical simulations*	
			
Typical resolution*			
Length :	10^{23} molecules	1 atom	
Time :	1 second	10^{-15} second	*femto-second (fs)*

*: single molecule / femtosecond is also possible (but not simultaneously in condensed phase)

Typical system sizes		*with current computers*	
Length :	10^{-3} meter	10^{-9} meter	*nano-meter (nm)*
Time :	10^{3} seconds	10^{-6} second	*micro-second (μs)*

low resolution / averaged	***high resolution / instantaneous***
large scale	***small scale***
complex physics	***elementary physics***
true	***approximate***

⇒ simulation and experiment are complementary methods !

Advantages of molecular simulations

- **Experimentalist:**

  Someone who knows how to operate the Natural computer...

- **Theoretician (classical simulator):**

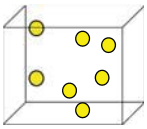
  ... someone who did not show any skill at that during undergraduate study, and decided to go instead for a silicon-based Ersatz (prosthesis ?), with a primitive chipset (classical force field) and an outdated operating system (Newton™, release 1.0)

- But there are two **key advantages** of simulations over experiment:

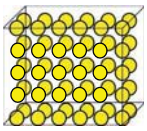
total transparency (of the output)	***absolute freedom (in the procedure)***
→ single-atom spatial resolution	*→ unphysical equations of motion OK*
→ femtosecond time resolution	*→ weird statistical mechanical ensembles OK*
→ direct access to the instantaneous atomic coordinates/velocities/energies	*→ inclusion of artificial forces OK*
→ no implicit averaging over molecules and time	*→ paths relying on alchemical processes OK*
	→ ...
⇒ permits (in favorable circumstances) the detailed interpretation of experimental observations at atomic and quasi-instantaneous resolution	*⇒ permits to carry out "impossible" experiments*
	⇒ gives room to "improve" on Nature's way (e.g. in terms of sampling speed)

What are (bio)chemists interested in ?

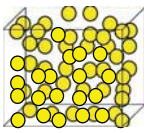
(*) important for industry



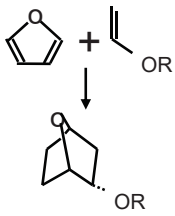
molecules in the gas phase:
- thermodynamics (real gases)
- molecular structure
- fragments (ions, radicals)
- spectroscopic properties
- reactions*



crystals:
- thermodynamics, phases*
- molecular structure
- packing forces
- disorder



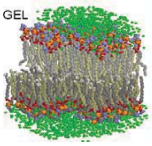
liquids/polymers:
- thermodynamics, phases*
- rheological properties*
- transport properties
- dielectric properties
- mixtures



molecules in solution:
- molecular structure and solvation
- spectroscopic properties
- conformational equilibria
- complexation equilibria
- acido-basic and redox properties
- chemical reactions*
  (thermodynamics & mechanism)

proteins:
- structure (1°,2°,3°,4°) and solvation*
- folding, assembly and binding*
- dynamics and function*
- catalysis*
  (thermodynamics & mechanism)
- effect of mutations*

nucleic acids:
- structure and solvation (ions)
- interaction with proteins/ligands*
- dynamics
- packaging
- expression*

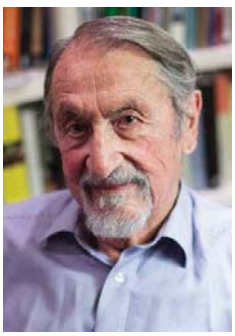


lipids:
- structure and solvation
- dynamics and fluidity
- permeation and diffusion*
- interaction with proteins*

carbohydrates:
- structure and solvation
- dynamics
- rheological properties*
- interaction with proteins*

Classical atomistic simulations

- As a result, since October 2013, our field has its Nobel prize



Question:
Who is the brains of the gang ?

Martin Karplus
Université de Strasbourg, France and Harvard University, Cambridge, MA, USA

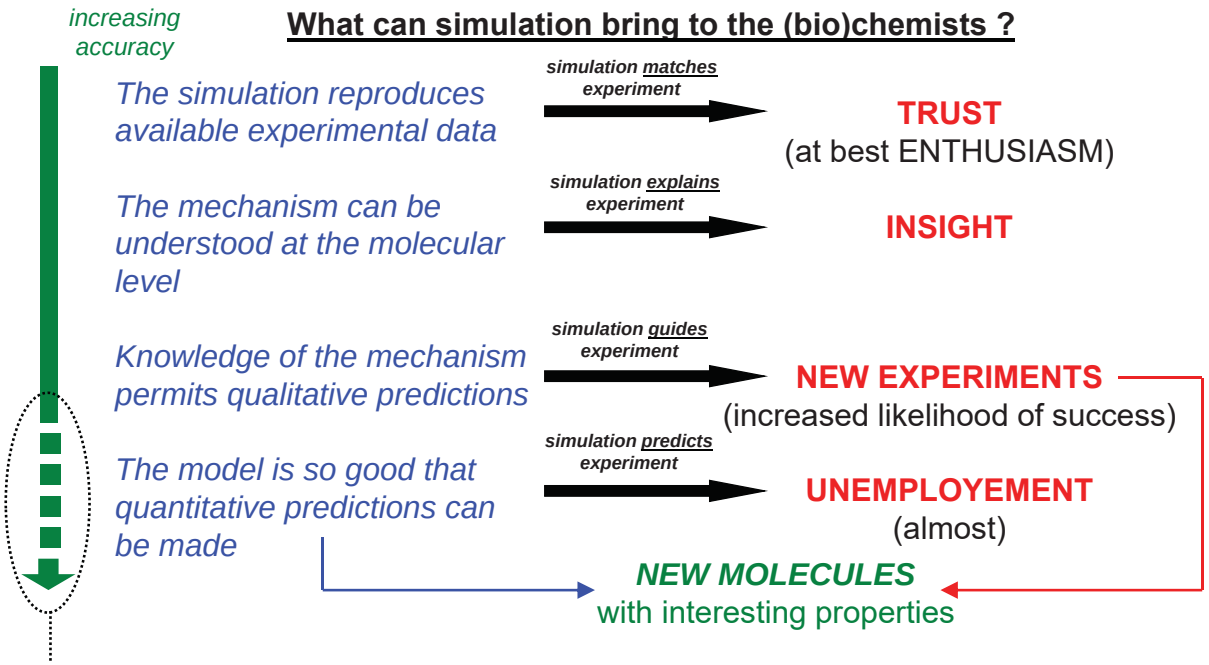
Michael Levitt
Stanford University School of Medicine, CA, USA

Arieh Warshel
University of Southern California, Los Angeles, CA, USA

"För utvecklandet av flerskalemodeller för komplexa kemiska system."

"For the development of multiscale models for complex chemical systems."

What can simulation bring to the (bio)chemists ?



Will this ever happen for molecular simulations ?

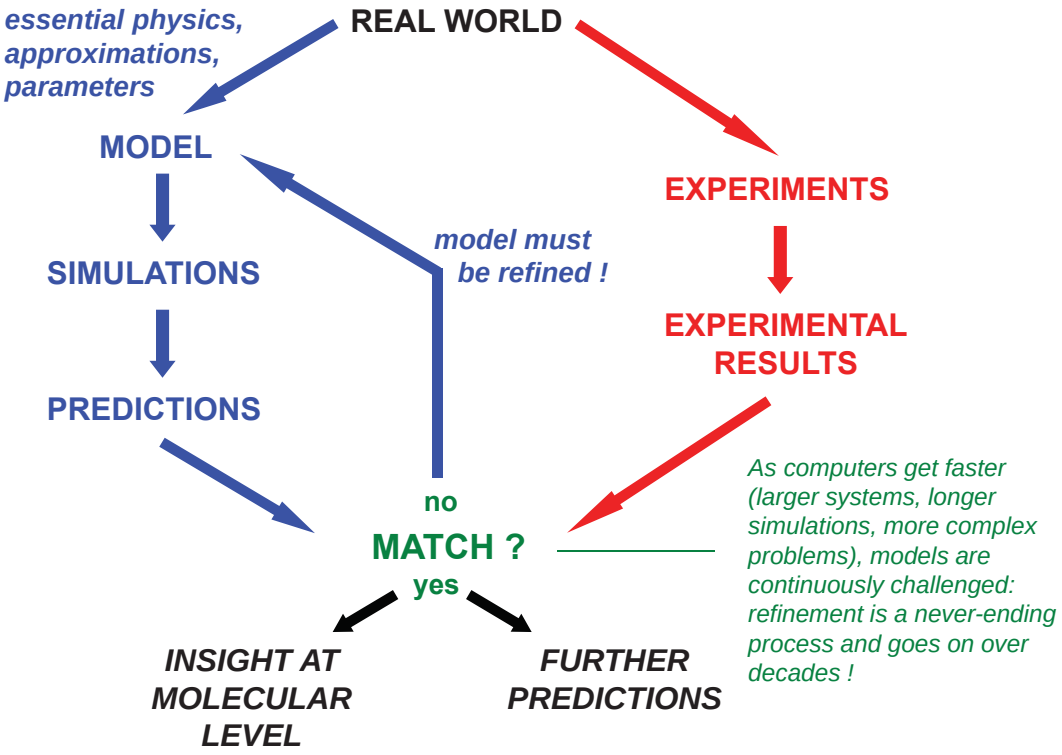
Possibly: the power of computers steadily grows, leading to increased resolution, system sizes, and complexity (accuracy) of the models

Possibly not: numerical solutions will remain approximate, and new ideas nearly always come from experimental observations

Note: in the "big-data" era, more emphasis is often placed on purely predicting relative to explaining/guiding (not a healthy trend in my opinion!)

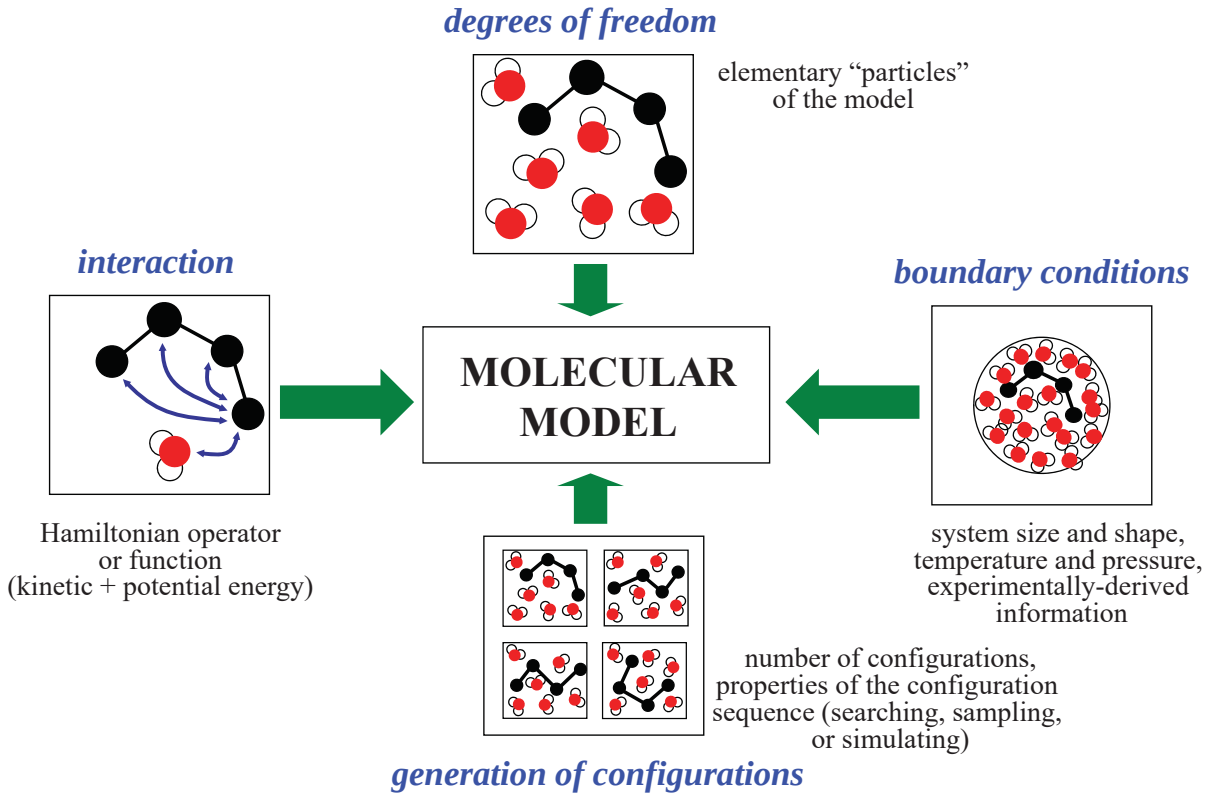
Definition and refinement of a molecular model

Generally a slow, iterative, and highly non-trivial task !



As computers get faster (larger systems, longer simulations, more complex problems), models are continuously challenged: refinement is a never-ending process and goes on over decades !

Four basic choices defining a molecular model



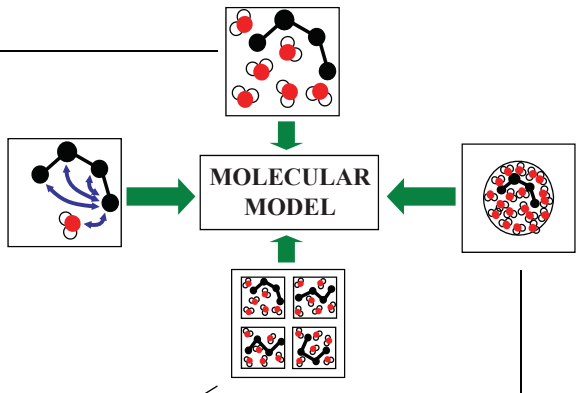
Scope of this course

1) INTRODUCTION
- what is simulation
- basic choices defining a model
- *choice of the degrees of freedom*
- computational limitations
- brief overview of quantum chemistry
- classical atomistic simulations (→rest of the course)

2+3) INTERACTION (FORCE FIELDS)
- basis of the classical description
- molecular topology
- covalent force-field terms
- non-bonded force-field terms
- calculating atomic forces
- force-field parameterization

3+4) GENERATING CONFIGURATIONS
- searching (incl. energy minimization)
- sampling (incl. Monte-Carlo sampling)
- simulating (incl. molecular and stochastic dynamics)

5) BOUNDARY CONDITIONS
- spatial boundary conditions
- thermodynamic boundary conditions (incl. temperature, pressure)
- [experimentally-derived boundary conditions (incl. X-ray, NMR)]



Scope of this course

6) ELECTROSTATIC INTERACTIONS
- the long-range problem
- methods to handle electrostatic interactions in simulations
- finite-size effects in simulations

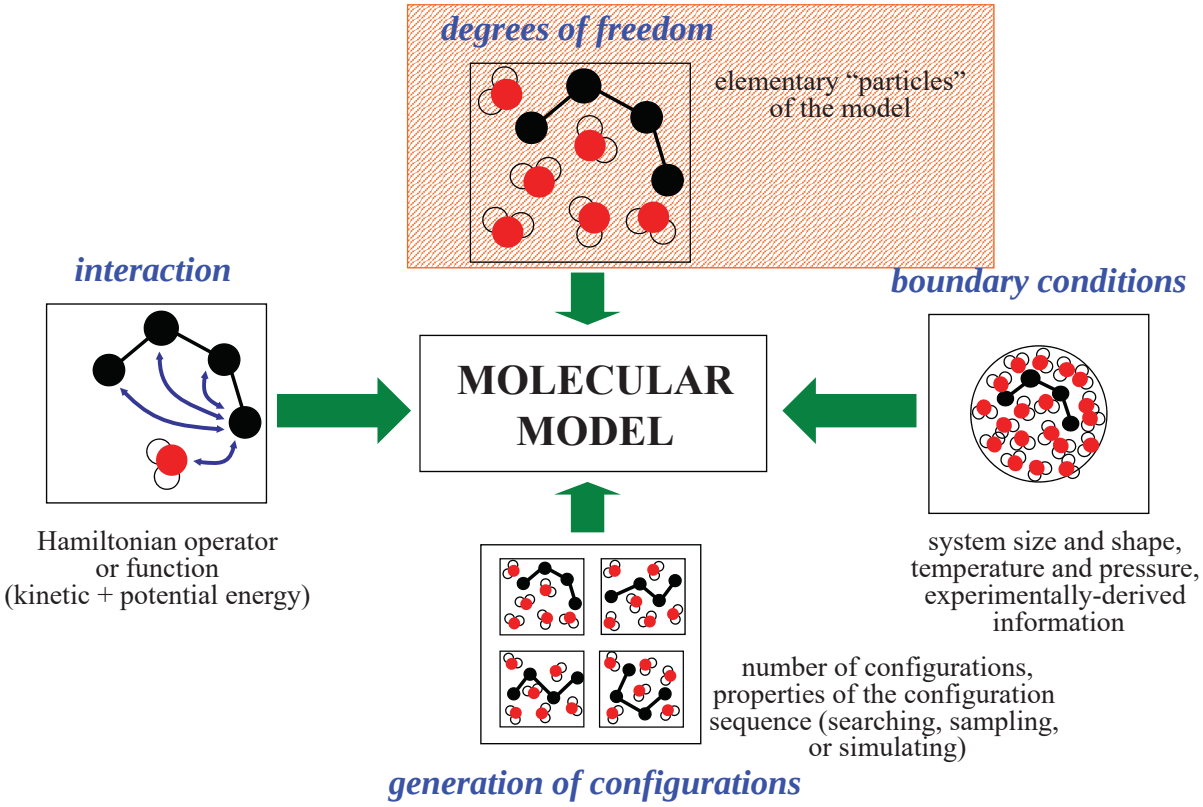
7) ANALYSIS OF SIMULATIONS
- liquid simulations
- biomolecular simulations
- examples

8-9) FREE ENERGY CALCULATIONS
- introduction to statistical mechanics
- determining free energy and entropy: methodology
- determining free energy and entropy: practical issues

10-14) SPECIAL TOPICS
- efficient methods for searching configuration space
- structure refinement based on X-ray or NMR data
- comparison between simulation and experiment

one of these (possibly)
- (treatment of quantum effects)
- (coarse-graining)
- (calculation of kinetic properties)
- (inclusion of polarization)
- (time-saving techniques)

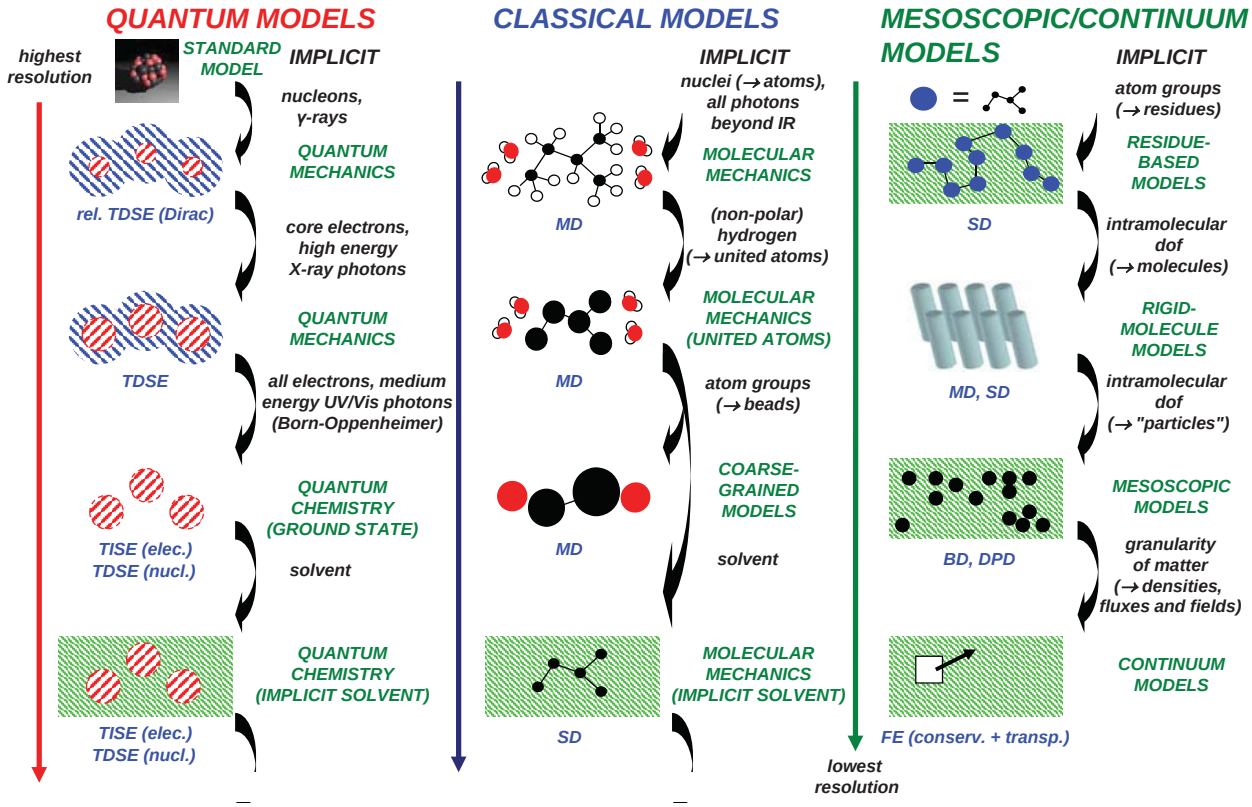
Four basic choices defining a molecular model



Choice of the degrees of freedom

TDSE / TISE: Time-(in)dependent Schrödinger equation

- A **hierarchy** of models with progressively decreasing levels of resolution



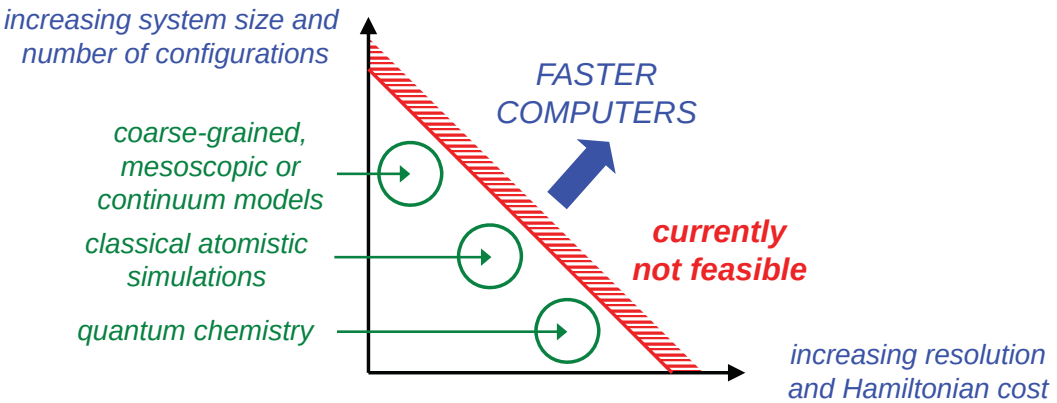
Computational limitations

The **choice of degrees of freedom** of the model largely determines:

- the ***resolution*** that can be achieved
- the ***types of phenomena*** that can(not) be described by the model
- the ***Hamiltonian (operator or function)*** describing the inter-particle interactions and its ***intrinsic computational cost*** (cost increases with resolution)

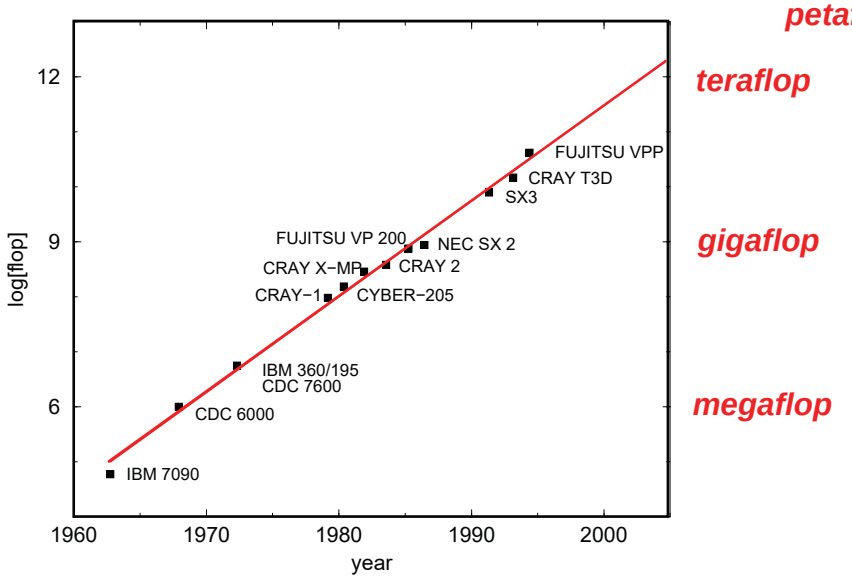
The **computational cost** is then determined by:

- the ***intrinsic computational cost*** of evaluating the Hamiltonian (operator or function) for a given system size and for one configuration
- the ***required system size*** to represent the property of interest
- the ***required number of configurations*** to evaluate the property of interest



Moore's law

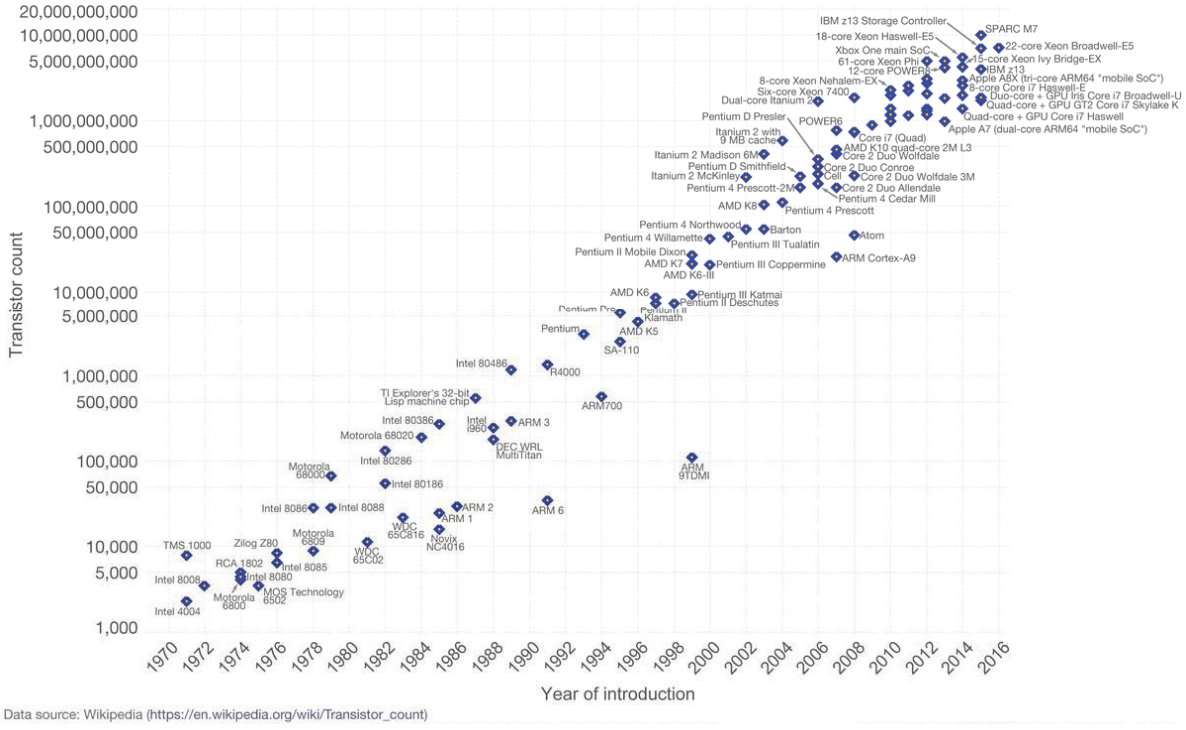
Now reached: petaflop



1 flop = 1 floating-point operation (14 digit precision) per second
good human brain (~5 minutes) ≅ 3 milliflop…

Moore's law

The computing power increases on average *by a factor 10 every 6 years*
[verified over the past four decades – future ?]



A super-quick overview of quantum chemistry

- Broadly speaking

→ ***refreshing overview if you already know, or sketchy introduction if you have no clue***

→ ***not exam material !***

→ ***understand the flow, forget about the equations...***

molecular calculations/simulations

quantum-chemical calculations | *classical atomistic simulations* | *coarse-grained, mesoscopic or continuum models*

Not the goal of this course but good to have minimal knowledge

→ ***brief (qualitative) overview***

Goal of the course

Mentioned at the end

From next lecture onward...

... towards last lecture

For more info

Quantum Chemistry

Advanced Quantum Chemistry

Prof. Reiher

- *In principle*, we should all do quantum calculations, because **molecules** obey the laws of **quantum** (*not classical!*) **mechanics**
- *In practice*, the quantum description involves many **approximations** and computational **limitations**

→ The *classical description* is *more appropriate* for <u>many problems</u> but may also be *clearly inappropriate* for <u>other problems</u>

Goal of this overview:

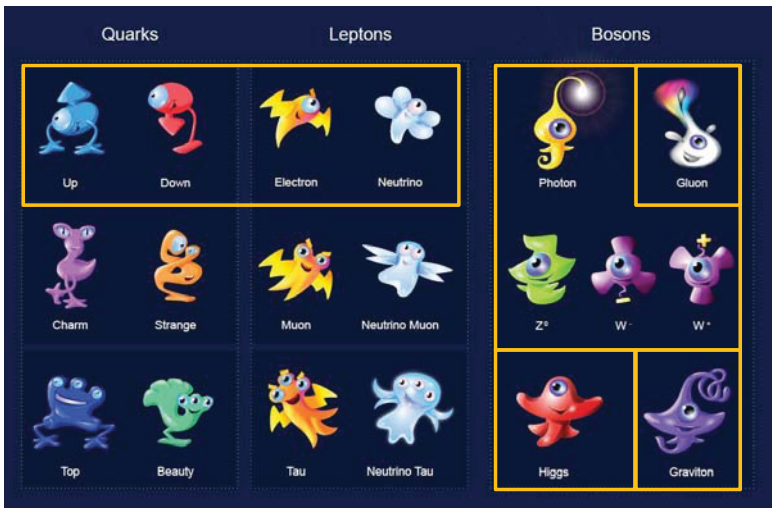
WHICH ONES AND WHY

To the best of our knowledge

- The **standard model** of physics: 30 **elementary particles** + a **quantum-field theory** (QFT)

Fermions ***spin 1/2*** | *Bosons* ***spin 0 (g,γ,Z⁰,H⁰), 1 (W⁺,W⁻) or 2 (G)***

~1960-1975 Glashow, Salam & Weinberg (following work of Dirac, Feynman, Higgs, ...)



Generation

I (usual)

II

exotic

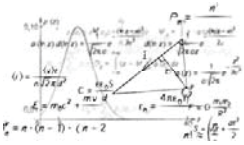
III

strong

electroweak

+ 6 anti-quarks | *+ 6 anti-leptons* | *mass-giver* ***(existence proved in 2012 in CERN/LHC)*** | *gravitation* — ***actually not in the standard model***

→ Good for



→ Not good for

Major problems of the model

→ remarkable account of experiments

→ cannot be the end of the story: incompatible with general relativity, which also gives remarkable account of other experiments

Approximation 1: simplified particle model

- Under the action of the **strong interaction** (*gluon*) — *the strongest but shortest-ranged (~fm) interaction in nature*

  → Quarks assemble into hadrons { mesons ($q\bar{q}$) / baryons (qqq) } → *e.g.* proton (uud) and neutron (udd) — *u: charge 2/3, d: charge -1/3*

  → Baryons assemble into nuclei 

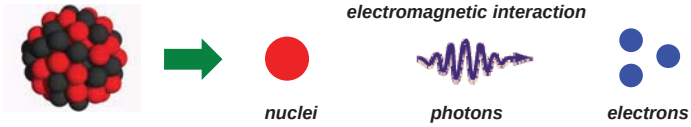
- The **weak interaction** (Z^o,W^- *and* W^+ *bosons*) can be omitted — *the weak interaction is short-ranged (~fm) as well, but much weaker; it plays a role in radioactive decay processes*

- The **gravitational interaction** (*graviton [?]*) can be omitted or treated macroscopically — *longest-ranged ($\sim r^{-1}$) but weakest interaction*; *its contribution is entirely negligible between particles/atoms/molecules – only worth considering between macroscopic assemblies !*

- The **neutrinos** can be omitted — *only sensitive to weak and gravitational interactions*

- The **2nd and 3rd generation particles** can be omitted — *exotic and unstable*

- We are left with **nuclei** and **electrons** subject to the **electromagnetic interaction** (*photon*) and we can change to traditional (but still relativistic) **quantum mechanics** (QM)



Approximation 1: simplified particle model

- The **coordinates** to describe a system of particles are

  M nuclei ● ● ...

  $\boldsymbol{\nu} = \{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, ..., \boldsymbol{\nu}_M\}$ — 4M-dimensional nuclear coordinate+spin vector

  $\boldsymbol{\nu}_\alpha = (x_\alpha, y_\alpha, z_\alpha, \sigma_\alpha)$ — Cartesian and spin ($\in \mathbb{Z}$ or $\frac{1}{2}\mathbb{Z}$) coordinates of nucleus α

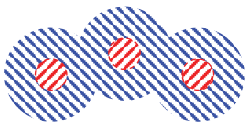
  N electrons ● ● ● ...

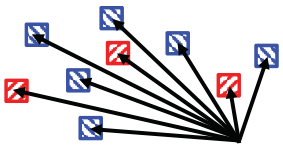
  $\boldsymbol{\tau} = \{\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, ..., \boldsymbol{\tau}_N\}$ — 4N-dimensional electronic coordinate+spin vector

  $\boldsymbol{\tau}_i = (x_i, y_i, z_i, \sigma_i)$ — Cartesian and spin ($\in \{-\frac{1}{2}, \frac{1}{2}\}$) coordinates of electron *i*

- The system **wavefunction** is

  $\tilde{\Psi}(\boldsymbol{\tau}, \boldsymbol{\nu}, t)$ — *complex, normalized*

  → it is defined by

  $|\tilde{\Psi}(\boldsymbol{\tau}, \boldsymbol{\nu}, t)|^2 \, d\boldsymbol{\tau} \, d\boldsymbol{\nu}$ = Probability of finding the nuclei and electrons in the 4(M+N)-dimensional coordinate/spin volume at time *t*

- The system **Hamiltonian** is

  $\hat{\mathrm{H}}(t)$ — *with terms for*
  - kinetic energy (nuclei and electrons)
  - electric and magnetic interactions (charges and spins)
  - interaction with (time-dependent) external fields (*e.g.* photons)

  particles in fast motion relative to the speed of light → relativistic

  → main property

  $\int d\boldsymbol{\tau} \, d\boldsymbol{\nu} \, \tilde{\Psi}^*(\boldsymbol{\tau}, \boldsymbol{\nu}, t) \hat{\mathrm{H}}(t) \tilde{\Psi}(\boldsymbol{\tau}, \boldsymbol{\nu}, t)$ = Total energy of the system at time *t*

- The system wavefunction should satisfy the **time-dependent Schrödinger equation** (TDSE)

$$\hat{\mathrm{H}}(t)\tilde{\Psi}(\boldsymbol{\tau}, \boldsymbol{\nu}, t) = i\hbar \frac{\partial \tilde{\Psi}(\boldsymbol{\tau}, \boldsymbol{\nu}, t)}{\partial t}$$

Approximation 2: time-independent Hamiltonian

- If the **Hamiltonian** does not depend explicitly on **time** $\hat{H}(t) = \hat{H}$

  → no interaction with a fluctuating external field, *e.g.* photons

  → the wavefunction becomes **separable** in the coordinate and time variables

$$\tilde{\Psi}(\boldsymbol{\tau}, \boldsymbol{\nu}, t) = \Psi(\boldsymbol{\tau}, \boldsymbol{\nu})\, T(t)$$

- Inserting into the TDSE gives

$$\hat{H}\Psi(\boldsymbol{\tau}, \boldsymbol{\nu})T(t) = i\hbar\Psi(\boldsymbol{\tau}, \boldsymbol{\nu})\frac{dT(t)}{dt}$$

coordinate domain ↙ ↘ *time domain*

$$\hat{H}\Psi(\boldsymbol{\tau}, \boldsymbol{\nu}) = E\Psi(\boldsymbol{\tau}, \boldsymbol{\nu}) \qquad \frac{dT(t)}{dt} = \frac{E}{i\hbar}T(t) = -i\omega T(t)$$

solution ↓

$$T(t) = e^{-i\omega t} \qquad \omega = E/\hbar$$

- This is the **time-independent Schrödinger equation** (TISE)

  → where

$\Psi(\boldsymbol{\tau}, \boldsymbol{\nu})$	**Stationary wavefunction (complex, normalized, defined within a phase factor $e^{i\phi}$ - arbitrary time origin)**
$\lvert\Psi(\boldsymbol{\tau}, \boldsymbol{\nu})\rvert^2\, d\boldsymbol{\tau}\, d\boldsymbol{\nu}$	**Probability of finding the nuclei and electrons in the 4(M+N)-dimensional coordinate volume (time independent)**

Approximation 3: isolated system / pure electrostatic interactions

- As a good approximation, the Hamiltonian of an **isolated molecular system** can be written in terms of **purely electrostatic** (no magnetic) interactions within the system

kinetic energy of the nuclei — *kinetic energy of the electrons*

$$\hat{H} = -\frac{\hbar^2}{2}\sum_{\alpha}^{M}\frac{1}{m_\alpha}\nabla_\alpha^2 - \frac{\hbar^2}{2m_e}\sum_{i}^{N}\nabla_i^2$$

Coulombic interaction nuclei-electrons

$$-\frac{e}{4\pi\varepsilon_o}\sum_{i}^{N}\sum_{\alpha}^{M}\frac{Q_\alpha}{r_{\alpha i}}$$

$$+\frac{1}{4\pi\varepsilon_o}\sum_{\alpha}^{M}\sum_{\beta>\alpha}^{M}\frac{Q_\alpha Q_\beta}{r_{\alpha\beta}} + \frac{e^2}{4\pi\varepsilon_o}\sum_{i}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}}$$

Coulombic interaction of the nuclei — *Coulombic interaction of the electrons*

- This neglects
  - → Relativistic effects (except spin [Fermi correlation only – see later])
  - → Interaction with (static) external fields (system is isolated)
  - → Magnetic interactions (spin-spin, spin-orbit and orbit-orbit)
  - → Non-electromagnetic forces (*e.g.* strong, weak and gravitational interactions)

already from previous level

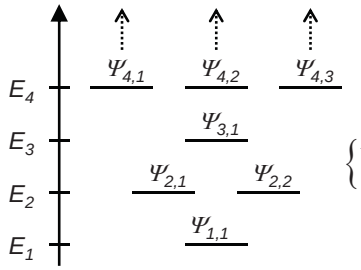
Approximation 3: pure electrostatic interactions

- The TISE + normalization read

$$\hat{H}\Psi(\tau,\nu) = E\Psi(\tau,\nu) \qquad \textit{with} \qquad \int d\tau\, d\nu\, |\Psi(\tau,\nu)|^2 = 1$$

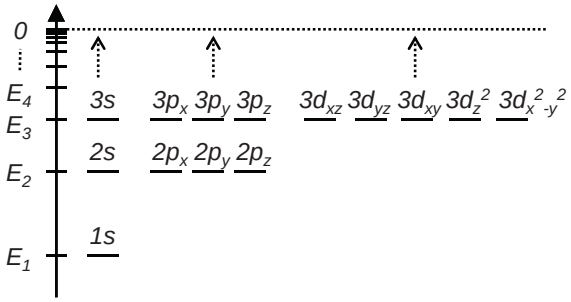
- This is a second-order differential **eigenvalue equation**

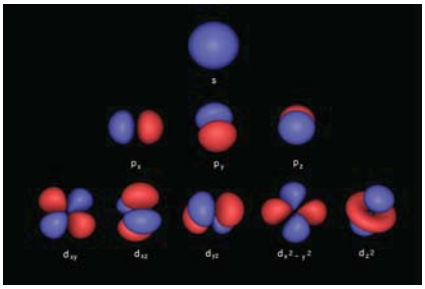
→ solutions $\{E_n\}$ $\{\Psi_{n,m}\}$



infinite set of discrete eigenvalues — *associated (possibly degenerate) eigenvectors*

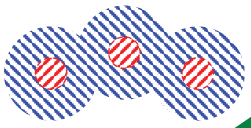
- At this level, we can solve **two-particle problems** (*e.g.* H, He^+, Li^{++}, ...) exactly *First solution: Pauli (1926)*





Approximation 4: Born-Oppenheimer approximation

- Electrons are >10^4 times **lighter** than nuclei so that they move >10^4 times **faster**

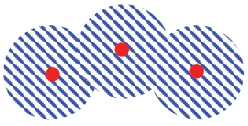


$$\hat{H}\Psi(\tau,\nu) = E\Psi(\tau,\nu)$$

Born & Oppenheimer (1927)

Electronic problem: ***nuclei are quasi motionless***

Nuclear problem (for a given *k*): ***electrons relax instantaneously***



$$\hat{H}_e\Psi_e(\tau;\nu) = V(\nu)\Psi_e(\tau;\nu)$$

⇒ ***set of solutions for the electronic energy levels $\{V_k(\nu)\}$ corresponding to a given nuclear configuration***

for a given k $\hat{H}_n\Psi_n(\nu) = E\Psi_n(\nu)$

⇒ ***set of solutions for the rotational/vibrational energy levels $\{E_{kl}\}$ corresponding to a given electronic energy level k***

$$\Psi(\tau,\nu) = \Psi_{e,k}(\tau;\nu)\Psi_{n,kl}(\nu)$$

(valid within the Born-Oppenheimer approximation)

$\Psi_{e,k}(\tau;\nu)$ *further noted* $\psi(\tau)$

- At this level, we can solve **polynuclear/monoelectronic problems** exactly (*e.g.* dihydrogen cation H_2^+) *Øyvind Burrau (1927)*

Approximation 5: Neglect of electron correlation

- If **electron correlation** is **neglected**, each electron sees the **mean effect** of the charge density generated by the other electrons

etc…

→ then the polyelectronic wavefunction, satisfying the parity constraint — *Pauli (1927)*
(sign of the wavefunction is inverted upon interchange of two electrons)
can be written as a **Slater determinant** — *Hartree, Fock, Slater (1930-1935)*

$$\psi(\tau) = (N!)^{-1/2} \begin{Vmatrix} \chi_1(\tau_1) & \chi_2(\tau_1) & \cdots & \chi_N(\tau_1) \\ \chi_1(\tau_2) & \chi_2(\tau_2) & \cdots & \chi_N(\tau_2) \\ \vdots & \vdots & & \vdots \\ \chi_1(\tau_N) & \chi_2(\tau_N) & \cdots & \chi_N(\tau_N) \end{Vmatrix}$$

→ the single-electron χ functions are called **spin-orbitals**

Approximation 5: Neglect of electron correlation

- Using the **variational principle**, one finds that the spin-orbitals obey the **Hartree-Fock equation**

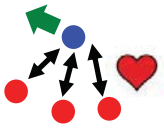
$$\hat{F}\chi_i(\tau_1) = \varepsilon_i \chi_i(\tau_1), \quad \forall i$$

Fock operator — optimal spin-orbital — orbital energy

→ Form of the **Fock operator**

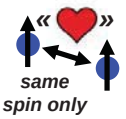
$$\hat{F} = \hat{h} + \sum_{j\ occupied}^{N} (\hat{J}_j - \hat{K}_j)$$

- $\hat{h}$ — ***Core operator:*** *Kinetic energy + interaction with nuclei*



- $\hat{J}_j$ — ***Coulomb operator:*** *Coulombic interaction with electron in occupied spin-orbitals χ_j*

- $\hat{K}_j$ — ***Exchange operator:*** *Reduction of the Coulombic repulsion with electron in occupied spin-orbital χ_j (only if same spin as χ_i)*



rather:

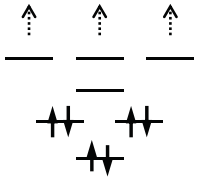
→ Must be solved **self-consistently** — ***the Fock operator depends on all occupied spin-orbitals $\{\chi_{j\neq i}\}$***

→ Relies on a definition of a specific **electronic configuration**

i.e. we have to say which spin-orbitals are occupied

e.g. closed-shell with doubly-occupied molecular orbitals

$$\phi_i(r_1)$$



Approximation 6: Basis-set expansion

- Computers are of no direct help in solving problems of *continuous* vector calculus...
  But: they are really good at solving *discrete* linear algebra problems !

- To transform the Hartree-Fock equation (for a given electronic configuration → rewritten in terms of molecular orbitals) into a computationally-tractable problem, *molecular orbitals* are expressed as a linear combination of **atomic orbitals**

$$\phi_i(\mathbf{r}_1) = \sum_\nu c_{\nu i}\varphi_\nu(\mathbf{r}_1)$$

→ I this basis set, the Hartree-Fock equation becomes the **Roothaan-Hall equation**

$$\mathbb{FC} = \mathbb{SCE} \quad \text{with} \quad \mathbb{F} = \mathbb{F}(\mathbb{C})$$

Roothan & Hall (1951)

*→ the birth of **COMPUTATIONAL QUANTUM CHEMISTRY!***

Fock matrix $\{F_{\mu\nu}\}$ — *coefficient matrix* $\{c_{\nu i}\}$ — *overlap matrix* $\{S_{\mu\nu}\}$ — *energy matrix* $\{\varepsilon_i\delta_{ij}\}$

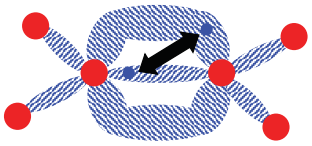
$$F_{\mu\nu} = \int d\tau_1\, \varphi_\mu^*(\tau_1)\, \hat{F}\, \varphi_\nu(\tau_1) \qquad S_{\mu\nu} = \int d\tau_1\, \varphi_\mu^*(\tau_1)\, \varphi_\nu(\tau_1)$$

→ Can be solved iteratively on a computer until **self-consistency**

→ The atomic orbitals are usually sets of **atom-centered Gaussians**

Beyond approximation 5: Electron correlation

- Electrons do not simply move in the *average field* produced by the other electrons



⇒ ***electrons tend to avoid each other dynamically***

- the corresponding electron-correlation energy is negative
- it is neglected in Hartree-Fock theory (except for Fermi correlation)
- it is important to correctly account for dispersion interactions

⇒ ***electron correlation can be partly reintroduced as a post-Hartree-Fock correction***

- configuration interaction CI (a variational approach)
- many-body perturbation theory PT (a perturbational approach)

See also: coupled cluster (CC)

⇒ ***computationally expensive, unfavorable scaling with system size, but (unfortunately!) absolutely necessary for many problems !***

Two other routes

- **Semi-empirical methods**
  - → Start from **Hartree-Fock**
  - → Introduce **approximations and parameters** for the one- and two-electron integrals
  - → **Calibrate** the parameters against experimental or against accurate quantum-chemical data
  - → The method becomes **computationally less expensive** and the parametrization against experiment reintroduces **some effective electron correlation** for the systems considered
- **Density-functional theory**
  - → It can be shown that the electron density entirely determines the energy; this simplifies the problem considerably! *Hohenberg & Kohn (1964)*

$$\psi(\tau) \quad \Rightarrow \quad \rho(\boldsymbol{r})$$

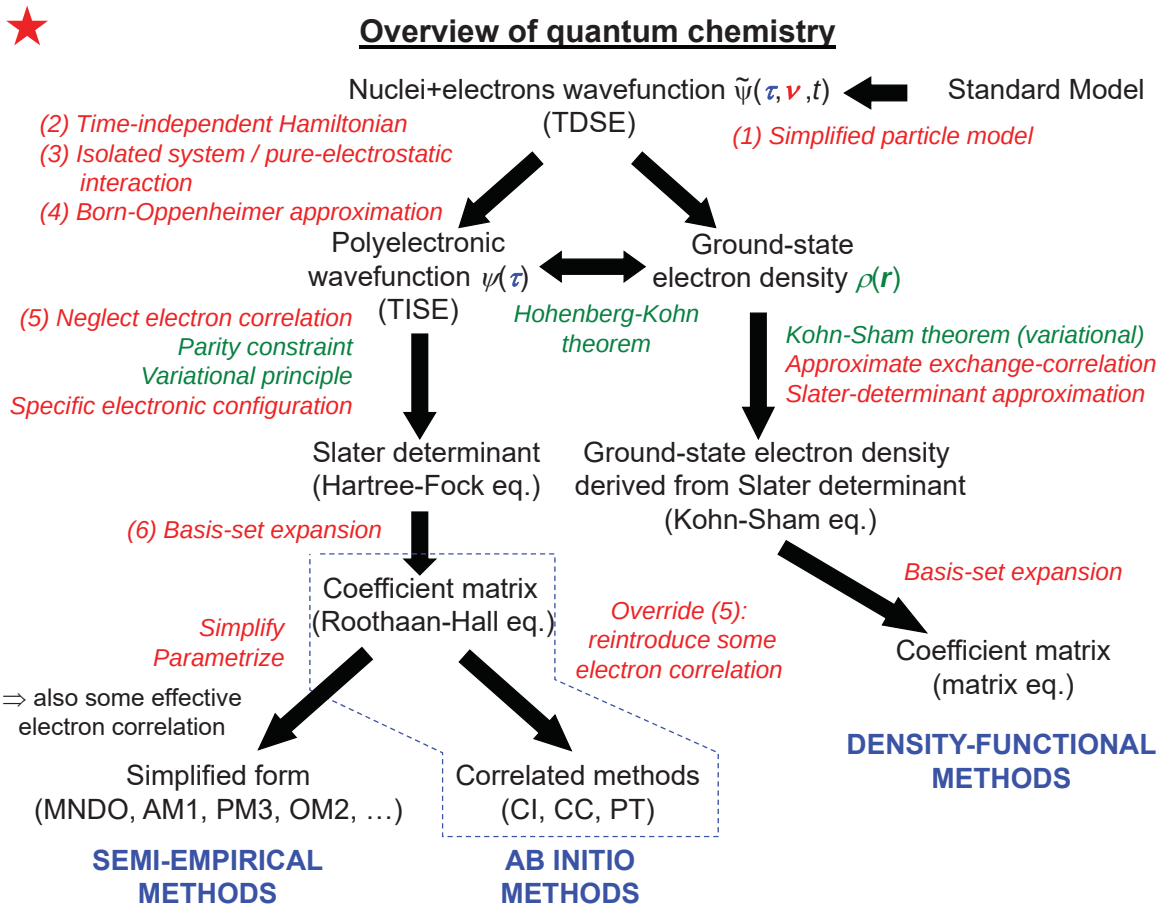
4N-dimensional ***3-dimensional***

  - → Introduce an (approximate) **energy functional**

$$V[\rho] = T[\rho] + E_n[\rho] + E_c[\rho] + E_{xc}[\rho]$$

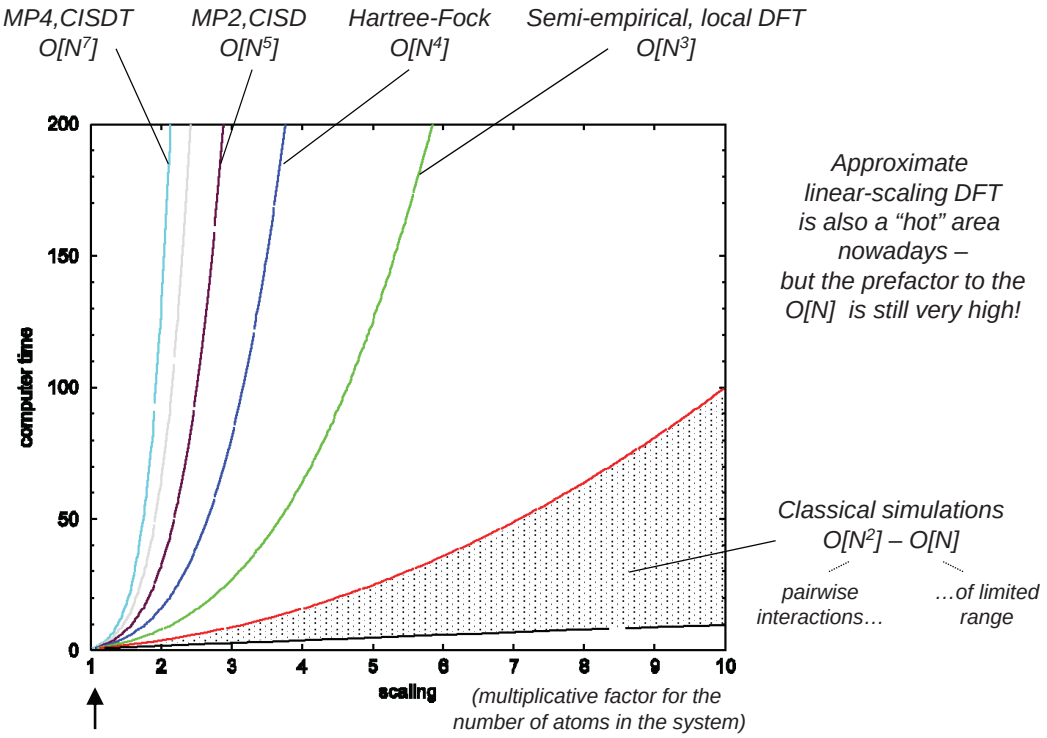
X[ρ] is a functional of ρ if X is fully determined by the knowledge of ρ(r) over all space

  - → Solve the problem using the **variational principle** and a **basis-set expansion** *Kohn & Scham (1965)*
  - → The method is in principle exact (including electron correlation) but the **design of good functionals** (especially exchange-correlation) is difficult in practice (→ often poor description of dispersion)



Computational scaling of quantum-chemical methods

- Computational **scaling**



all methods assumed as fast at this point

In short

- The field of **quantum chemistry** is about:

the art of mastering approximations to achieve a good trade-off between accuracy and computer time for a given molecular system (and its relevant properties)

- As you will see in this lecture, the field of classical simulations is completely different...

- The field of **classical simulations** is about:

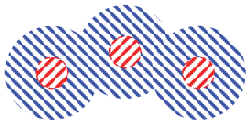
the art of mastering approximations to achieve a good trade-off between accuracy and computer time for a given molecular system (and its relevant properties)

- Just the **types of systems** (and properties) we are interested in, and the nature of the **approximations** involved are very different!

  → Quantum mechanics is conceptually more **complicated /abstract** (the classical representation is in many respects more *intuitive* !)

  → But classical simulations also involves a lot of **tricky methodological issues**... (the devil is in the details)

Quantum-chemical methods



- Major **advantages**

  → Sound theoretical basis — *the chemical world is described by quantum mechanics → hope to obtain exact solutions*

  → Few (or no) empirical parameters — *ab initio methods → full specification needs only:*
  - *number and types of nuclei (mass, charge, spin)*
  - *nuclear coordinates*
  - *number of electrons and electronic configuration*
  - *physical constants*

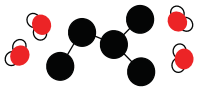
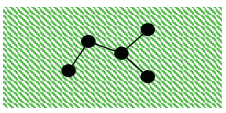
- Major **limitations**

  → **Numerous approximations** are invoked in practice

  → Accurate calculations are computationally **expensive** and scale very unfavorably — *require including electron correlation and using large basis sets*

  ⇒ small systems only (typically <<100 atoms)
  ⇒ few configurations (no statistical mechanics)
  ⇒ difficult to handle solvation (microsolvation or continuum)

Classical atomistic simulations

- Major **advantages**

  → Suited for the study of ***condensed-phase systems*** — *i.e. most of (bio)chemistry*

  → Sufficient timescales
  - *→ calculation of thermodynamic properties through statistical mechanics*
  - *→ bridge with experimentally-accessible timescales (now or near future)*

  → Sufficient system sizes — *→ access to (bio-)macromolecules in solution*

  → Sufficient resolution — *→ almost correct dynamics at the atomic level*

  → Complementary to experiment
  - *→ X-ray diffraction (crystals)*
  - *→ NMR (solutions)*

  (structure and dynamics at atomic resolution)

- Major **limitations**

  → **Empirical**, **numerous parameters**, **parameter-sensitive**

  → Unable to account (accurately) for **quantum effects**, mainly

  ⇒ proton and electron transfers
  ⇒ chemical reactions
  ⇒ high-frequency vibrations
  ⇒ low-temperature properties

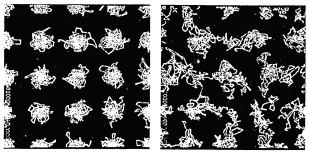
  → need for hybrid methods

History of classical atomistic simulations

1957 First molecular dynamics simulation (hard disks in 2D)

Adler & Wainwright

solid



liquid

Monte Carlo is older...

1964	Atomic liquid (argon)	10 ps
1971	Molecular liquid (water)	5 ps
1975	Simple short polymer (no solvent)	10 ps
1977	Protein (no solvent)	20 ps
1982	Model membrane (no solvent)	200 ps
1983	Protein in water	20 ps
1986	Nucleic acid in water	100 ps
1989	Protein/nucleic acid complex in water	100 ps
1996	Protein/membrane system in water	100 ps
1997	Peptide folding in solution (ETHZ)	100 ns
1998	Protein(?) folding(?) in water (UCSF)	1 μs
2000	Spontaneous micelle/membrane formation in water	50 ns
2002	Membrane fusion in water	200 ns

2018 Standard : biomolecule in water (10000 atoms) ~1 μs

~ 10 CPU days → ~10^{12} times slower than nature…

Future of classical atomistic simulations

using Moore's law, one may speculate about the future…

Standard classical simulations:

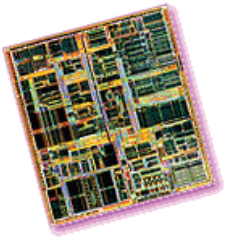
2018	Biomolecule in water (~10^4 atoms)	1 μs	
2029	Biomolecule in water	1 ms	***protein folding (?)***
2034	E-Coli (~10^{11} atoms)	1 ns ⟵	***my retirement…***
2056	Mammalian cell (~10^{15} atoms)	1 ns	
2080	Biomolecule in water	10^6 s	***as fast as nature…***
2180	Human body (~10^{27} atoms)	1 s	

But :

- ***Scientists are impatient – they want answers now !***
- ***There may be a limit to the speed of computers ...***
- ***How to get the starting configurations ?***
- ***Are the classical models sufficiently accurate at all ?***

The partners in the simulation business

Chip designers



Faster computers
→ larger systems,
longer timescales

Experimentator

Finer resolution
→ Starting structures
Data for model refinement
New problems …

Simulator

Always better models !
New methods !
Faster algorithms !

Computer Simulation in Chemistry, Biology and Physics

P.H. Hünenberger

COMPUTER SIMULATION OF MOLECULAR SYSTEMS

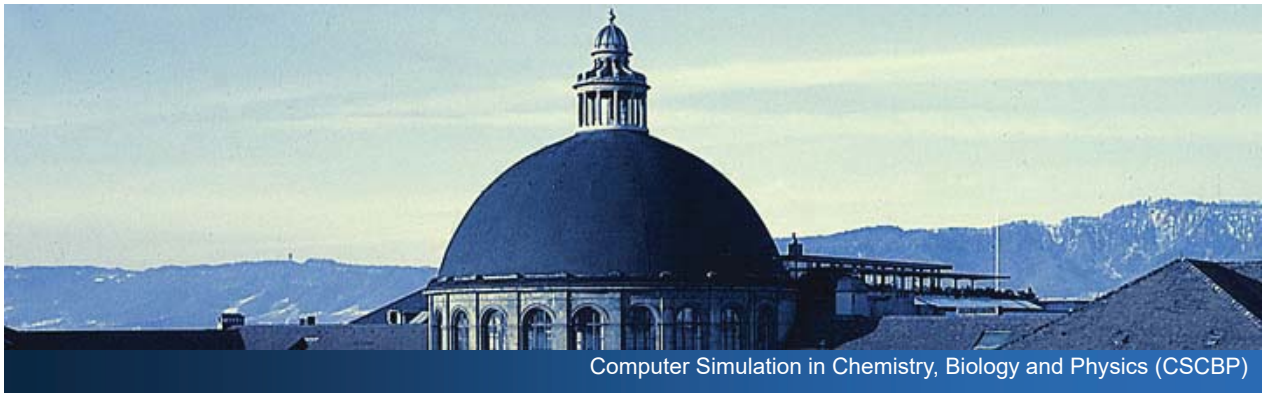
Herbstsemester 2019
Tuesday 9:45-11:30 a.m.
HCI D2

Lecture 529-0004-00
www.csms.ethz.ch/education/CSCBP

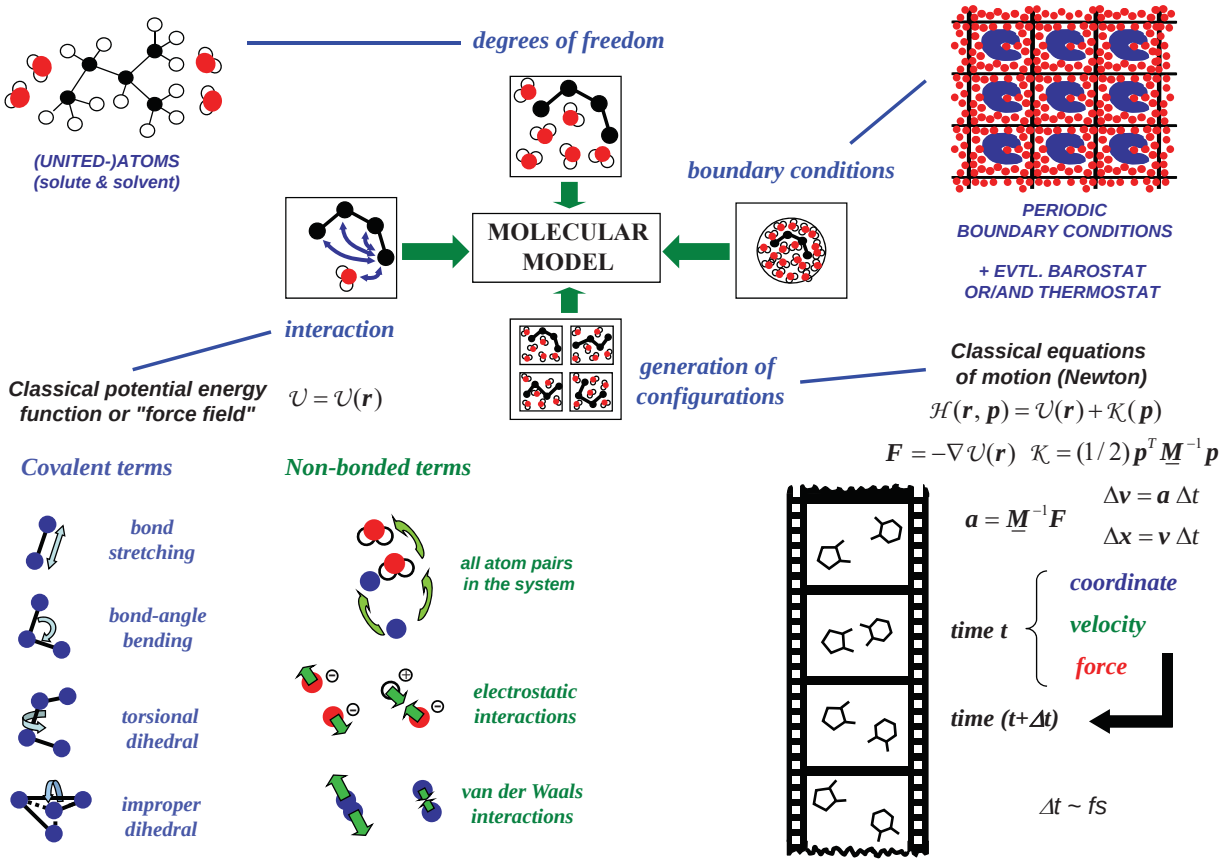
LECTURE 1' (WEEK 1): Simulating using GROMOS

How to simulate using GROMOS

An introduction to the
GROningen ***MO***lecular ***S***imulation package



A one-slide crash course in Molecular Dynamics (MD) simulation



The GROMOS program

http://www.gromos.net

a simulation program

set-up and analysis tools

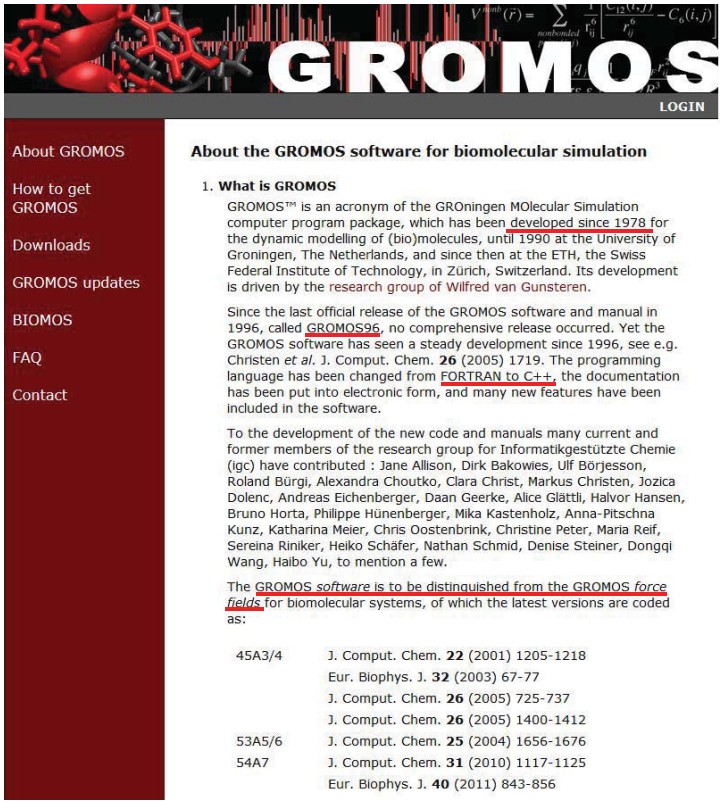
a force field

distributed with source

free software (just register for a no-cost license)

very reliable, but at present, not the fastest program around

(e.g. GROMACS is faster, but also significatly more "buggy"...)



The GROMOS program

http://www.csms.ethz.ch/education/CSCBP

see «Wilfred van Gunsteren and GROMOS» under downloads

JCTC Journal of Chemical Theory and Computation — Editorial
pubs.acs.org/JCTC

Wilfred van Gunsteren: 35 Years of Biomolecular Simulation

Photograph of Wilfred van Gunsteren: Giulia Marthaler/ETH Zürich

This special issue of the *Journal of Chemical Theory and Computation* is dedicated to one of the founders of the field of biomolecular simulation, Prof. Wilfred F. van Gunsteren, in honor of his 65th birthday and 35 years of research in this field.

cum laude in 1976. In addition to studying physics, Wilfred also prepared (without actually following the lectures!) a master's degree in law ("meester in de rechten"), which he obtained in 1974. While he never formally worked as a lawyer, the unusual combination of law and physics has been visible throughout his later career, in the form of a keen interest in matters of correctness, integrity, and justice, and in the careful formulation of arguments.

Despite having just completed a highly successful Ph. D., Wilfred was unsure whether he wanted to pursue a career in nuclear physics. It was at this point that he met Prof. Herman J. C. Berendsen, professor of physical chemistry at the University of Groningen. Herman convinced Wilfred that his skills in physics and computation might enable him to address fundamental questions in biology, a new and radical idea for the time. Wilfred worked with Herman at the University of Groningen as a postdoctoral fellow from 1976 to 1978, developing the basic algorithms and programs needed to efficiently simulate (bio)-molecular systems. This period was followed by a second postdoctoral stay from 1978 to 1980 in the group of Prof. Martin Karplus at Harvard University, another leading center in this

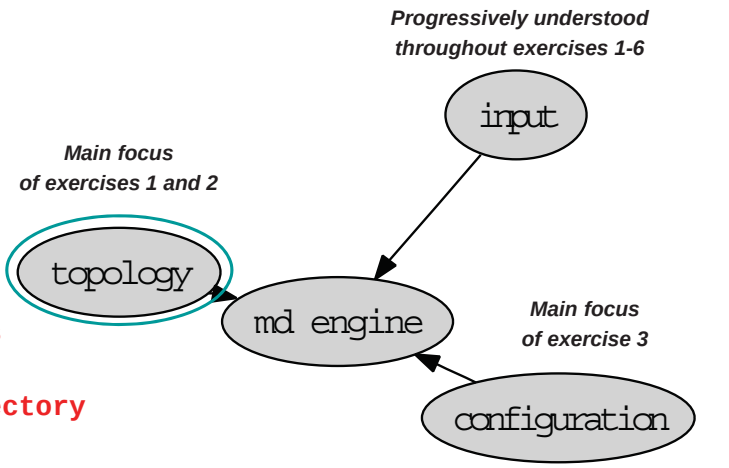
...

Philippe H. Hünenberger*,†
Alan E. Mark‡
Herman J.C. Berendsen§

As a CSCBP tribute, the following slides are still the original "vintage" slides of Wilfred... (only with a few added comments)

Usage of md – the molecular dynamics engine of GROMOS

```
leia:~> md
# topology data
@topo         filename
# coordinates
@conf         filename
# input parameter
@input        filename
# output final coordinates
@fin          filename
# output coordinates trajectory
@trc          filename
# output energy trajectory
# @tre        filename
# position restraints specification
# @posresspec filename
```



Legend: `# input files` `# output files`

The GROMOS topology

- Contains the topological and the force field data for the molecular system
- Written with the help of programs (see tutorial)
- Contains information in blocks

Prominently: make_top (not used in exercise 1 but afterwards)

```
TITLE
MAKETOP topology, using:
mtb54a7.dat
ifp54a7.dat
Force-field code: 54A7
END
```

Program which generated topology

Forcefield used

The GROMOS topology

```
PHYSICALCONSTANTS
# FPEPSI: 1.0/(4.0*PI*EPS0) (EPS0 is the permittivity of vacuum)
138.9354
# HBAR: Planck's constant HBAR = H/(2* PI)
0.0635078
# SPDL: Speed of light (nm/ps)
299792.458
# BOLTZ: Boltzmann's constant kB
0.00831441
END
[...]
ATOMTYPENAME
# NRATT: number of van der Waals atom types
54
# TYPE: atom type names
O
[...]
NUrea
CH3p
END
```

Physical constants used by program md

Determine the units... standard:

Length:	*nm*
Time:	*ps*
Mass:	*g/mol*
Charge:	*e*
Energy:	*kJ/mol*
Temperature:	*K*
Pressure:	*kJ/(nm³ mol) ≈ 16.6 bar*
Angles:	*degrees (converted internally to rad)*

Different atom types used in this topology

carbonyl oxygen (C=O)

nitrogen in urea

The GROMOS topology

```
RESNAME
# NRAA2: number of residues in a solute molecule
5
# AANM: residue names
VAL
TYR
ARG
LYSH
GLN
END
```

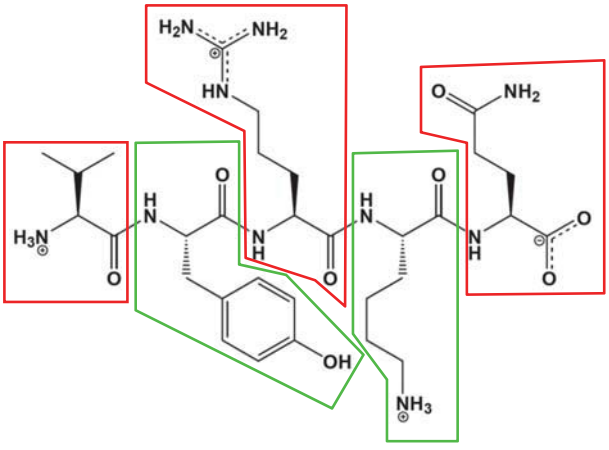
Valine

Tyrosine

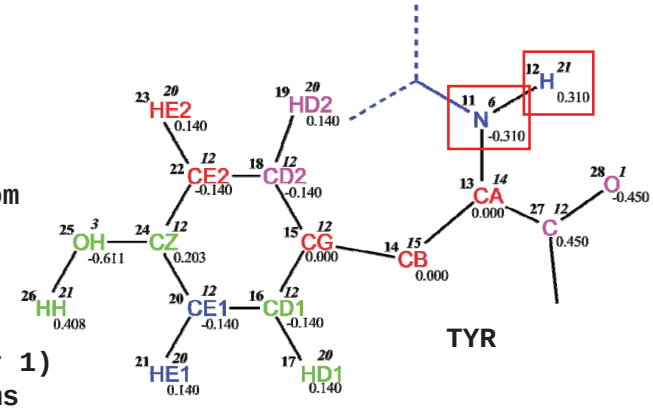
Arginine (protonated; charge +1)

Lysine (protonated; charge +1)

Glutamine

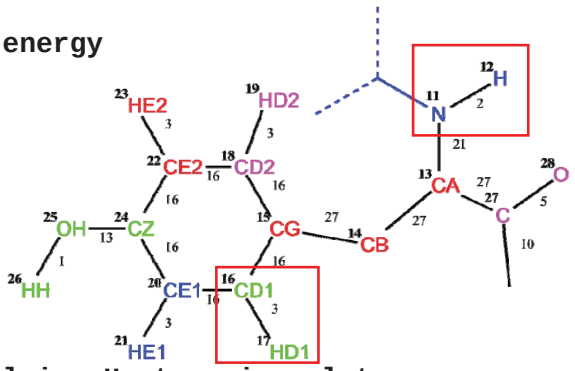


```
SOLUTEATOM
#    NRP: number of solute atoms
    71
#   ATNM: atom number
#   MRES: residue number
#   PANM: atom name of solute atom
#    IAC: integer (van der Waals)
#         atom type code
#   MASS: mass of solute atom
#     CG: charge of solute atom
#    CGC: charge group code (0 or 1)
#    INE: number of excluded atoms
# INE14: number of 1-4 interactions
# ATNM MRES PANM IAC     MASS       CG   CGC INE
#                                        INE14
[...]
     11    2    N    6 14.00670 -0.31000   0   4     12     13     14     27
                                              3     15     28     29
     12    2    H   21  1.00800  0.31000   1   1     13
                                              2     14     27
[...]
END
```



(comes directly from .ifp file, i.e. not system specific !)

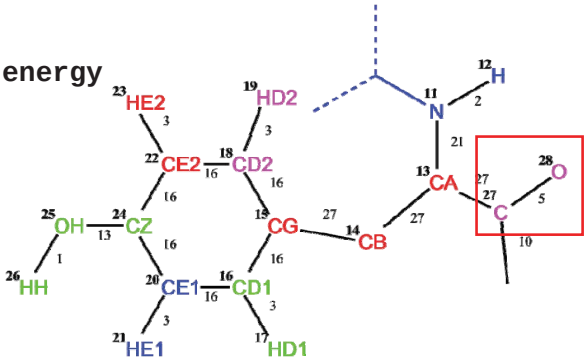
```
BONDSTRETCHTYPE
#  NBTY: number of covalent bond types
52
#  CB: force constant
#  B0: bond length at minimum energy
#         CB          B0
 1.57000e+07 1.00000e-01
 1.87000e+07 1.00000e-01
 1.23000e+07 1.09000e-01
 3.70000e+07 1.12000e-01
 1.66000e+07 1.23000e-01
[...]
END
```




```
BONDH
#  NBONH: number of bonds involving H atoms in solute
22
#  IBH, JBH: atom sequence numbers of atoms forming a bond
#  ICBH: bond type code
#    IBH    JBH ICBH
[...]
     11     12     2
     16     17     3
[...]
END
```

Bonds involving hydrogen atoms

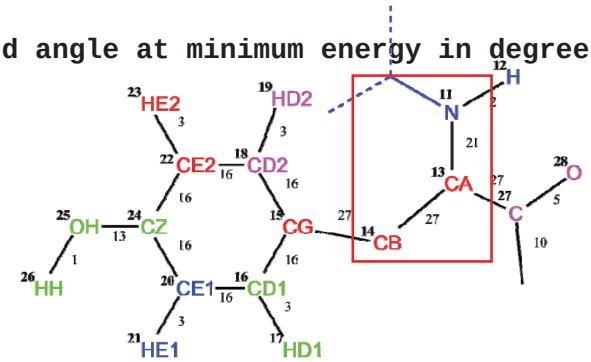
```
BONDSTRETCHTYPE      (same block as on last slide)
#  NBTY: number of covalent bond types
52
#  CB: force constant
#  B0: bond length at minimum energy
#         CB           B0
 1.57000e+07 1.00000e-01
 1.87000e+07 1.00000e-01
 1.23000e+07 1.09000e-01
 3.70000e+07 1.12000e-01
 1.66000e+07 1.23000e-01
[...]
END
BOND
#  NBON: number of bonds NOT involving H atoms in solute
49
#  IB, JB: atom sequence numbers of atoms forming a bond
#  ICB: bond type code
#    IB     JB  ICB
[...]
      27     28    5
[...]
END
```



Bonds not involving hydrogen atoms

(comes directly from .ifp file, i.e. not system specific !)

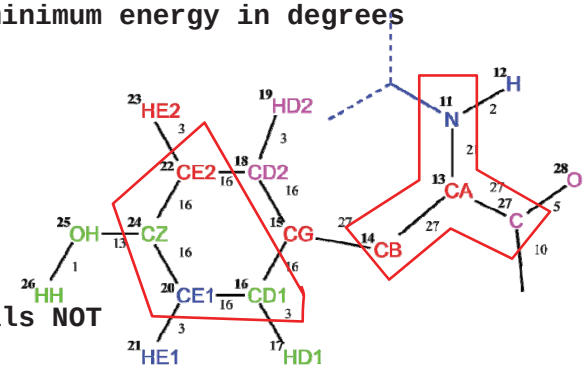
```
BONDANGLEBENDTYPE
#  NTTY: number of bond angle types
54
#  CT: force constant,T0: bond angle at minimum energy in degrees
#         CT           T0
[...]
# 10
 4.25000e+02 1.09500e+02
 4.50000e+02 1.09500e+02
 5.20000e+02 1.09500e+02
[...]
END
BONDANGLE
#  NTHE: number of bond angles NOT involving H atoms in solute
64
#  IT, JT, KT: atom sequence numbers of atoms forming a bond angle
#  ICT: bond angle type code
#    IT     JT     KT   ICT
[...]
# 10
      11     13     14    13
[...]
END
```



Angles not involving hydrogen atoms (BONDANGLEH: involving hydrogen atoms)

(comes directly from .ifp file, i.e. not system specific !)

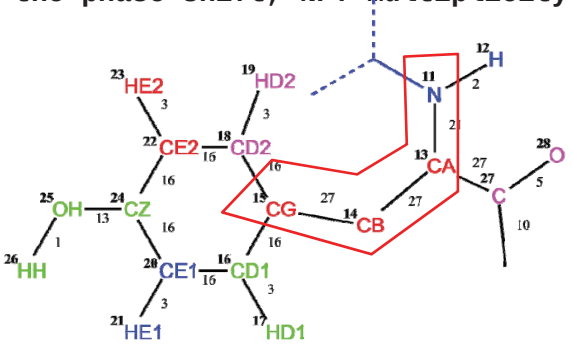
```
IMPDIHEDRALTYPE
#  NQTY: number of improper dihedrals
5
#  CQ: force constant of improper dihedral per degrees square
#  Q0: improper dihedral angle at minimum energy in degrees
#         CQ          Q0
 5.10000e-02 0.00000e+00
 1.02000e-01 3.52644e+01
 2.04000e-01 0.00000e+00
 5.10000e-02 1.80000e+02
 1.02000e-01 -3.52644e+01
END
IMPDIHEDRAL
#  NQHI: number of improper dihedrals NOT
#    involving H atoms in solute
21
#  IQ,JQ,KQ,LQ: atom seq. numbers of atoms forming an improper dihedral
#  ICQ: improper dihedral type code
#    IQ     JQ     KQ     LQ  ICQ
[...]
     16     20     24     22    1
     13     11     27     14    2
[...]
END
```



Improper dihedrals not involving hydrogen atoms (IMPDIHEDRALH: involving hydrogen atoms)

(comes directly from .ifp file, i.e. not system specific !)

```
TORSDIHEDRALTYPE
#  NPTY: number of dihedral types
45
#  CP: force const.,PD: cosine of the phase shift, NP: multiplicity
#       CP         PD   NP
[...]
# 30
    4.18000    1.00000    3
    4.69000    1.00000    3
    5.44000    1.00000    3
    5.92000    1.00000    3
[...]
END
DIHEDRAL
#  NPHI: number of dihedrals NOT involving H atoms in solute
28
#  IP, JP, KP, LP: atom sequence numbers
#     of atoms forming a dihedral
#  ICP: dihedral type code
#    IP     JP     KP     LP   ICP
[...]
     11     13     14     15    34
END
```



Torsional dihedrals not involving hydrogen atoms (DIHEDRALH: involving hydrogen atoms)

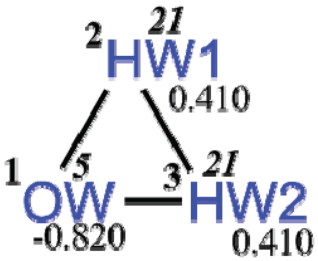
(comes directly from .ifp file, i.e. not system specific !)

```
LJPARAMETERS
#  NRATT2: number of LJ interaction types = NRATT*(NRATT+1)/2
1485
#  IAC,JAC: integer (van der Waals) atom type code
#  C12: r**(-12) term in nonbonded interactions
#   C6: r**(-6) term in nonbonded interactions
# CS12: r**(-12) term in 1-4 nonbonded interactions
#  CS6: r**(-6) term in 1-4 nonbonded interactions
# IAC  JAC           C12            C6          CS12           CS6
    1    1  1.000000e-06  2.261954e-03  7.414932e-07  2.261954e-03
[...]
END

SOLUTEMOLECULES
# NSPM: number of separate molecules in solute block
# NSP[1...NSPM]: atom sequence number of last atom
#                of the successive submolecules
#      NSPM  NSP[1...NSPM]
       1     71
END
```

These are just used for classification (or not used at all). For use in temperature or/and pressure control, see blocks TEMPERATUREGROUPS and PRESSUREGROUPS

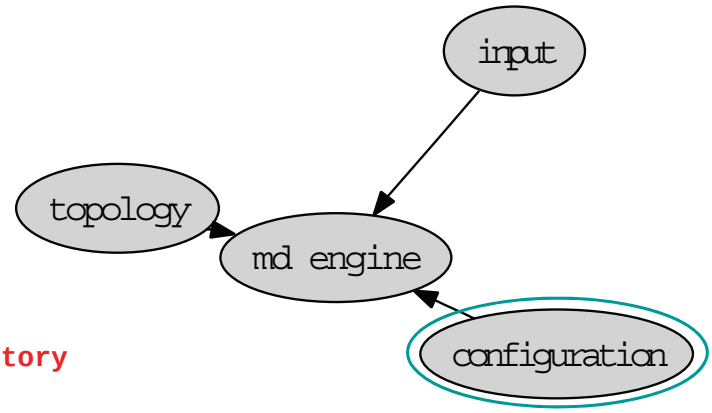
```
SOLVENTATOM
#  NRAM: number of atoms per solvent molecule
3
#     I: solvent atom sequence number
#  IACS: integer (van der Waals) atom type code
#  ANMS: atom name of solvent atom
#  MASS: mass of solvent atom
#   CGS: charge of solvent atom
#  I  ANMS IACS      MASS        CGS
   1    OW   5   15.99940   -0.82000
   2   HW1  21    1.00800    0.41000
   3   HW2  21    1.00800    0.41000
END
```




```
SOLVENTCONSTR
#  NCONS: number of constraints
3
#  ICONS, JCONS: atom sequence numbers forming constraint
#   CONS constraint length
#ICONS JCONS         CONS
    1    2      0.1000000
    1    3      0.1000000
    2    3      0.1632990
END
```

Usage of md – the molecular dynamics engine of GROMOS

```
leia:~> md
# topology data
@topo         filename
# coordinates
@conf         filename
# input parameter
@input        filename
# output final coordinates
@fin          filename
# output coordinates trajectory
@trc          filename
# output energy trajectory
# @tre        filename
# position restraints specification
# @posresspec filename
```




```
# input files
# output files
```

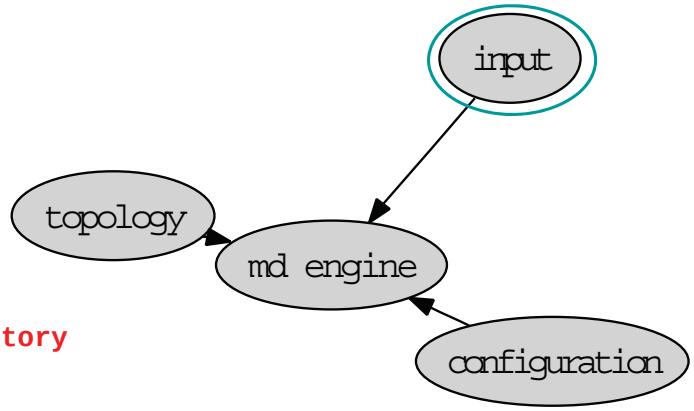
```
TITLE
Solvating ../coord/peptide.cnf in spc.cnf
Box dimensions (cubic) were calculated from maximum
solute atom-atom distance (not rotated):
       1.41551 between atoms 1:42 and 1:65
Added 865 solvent molecules
END
POSITION
    1 VAL   H1         1   -0.344305342   -0.135185949    0.124481007
    1 VAL   H2         2   -0.324119246   -0.007525893    0.024623263
[...]
    1 SOLV  OW        72   -1.375019730   -0.116880247   -1.118649662
    1 SOLV  HW1       73   -1.349965361   -0.020069275   -1.119966747
    1 SOLV  HW2       74   -1.331102156   -0.161598224   -1.040718404
[...]
  865 SOLV  HW2     2666   -0.263513130    0.913675353   -1.521003964
END
GENBOX
    1
    3.015511503    3.015511503    3.015511503
   90.000000000   90.000000000   90.000000000
    0.000000000    0.000000000    0.000000000
    0.000000000    0.000000000    0.000000000
END
```

Atom/residue names are redundant with topology file and will be ignored !

Initial velocities are also needed for a continuation run !

Usage of md – the molecular dynamics engine of GROMOS

```
leia:~> md
# topology data
@topo         filename
# coordinates
@conf         filename
# input parameter
@input        filename
# output final coordinates
@fin          filename
# output coordinates trajectory
@trc          filename
# output energy trajectory
# @tre        filename
# position restraints specification
# @posresspec filename
```




```
# input files
# output files
```

Overview over the GROMOS input file

```
SYSTEM
# NPM NSM
    1 865
END
INITIALISE
# Default values for NTI values: 0
# NTIVEL    NTISHK    NTINHT    NTINHB
        1         3         0         0
# NTISHI    NTIRTC    NTICOM
        1         0         0
# NTISTI
        0
#     IG   TEMPI
  210185     298
END
STEP
# NSTLIM T   DT
  10000  0.0 0.002
END
```

(initial time is just to have correct times in the output for continuation runs)

Number of solute molecules

Number of solvent molecules

Read coordinates and velocities from configuration file

initial SHAKE for coordinates and velocities

random number seed

Temperature to generate initial velocities (if generated!)

Number of simulation steps

Time at simulation start

Integration time step

```
MULTIBATH
# ALGORITHM:
#       0:      use weak-coupling scheme
#       1:        use Nose Hoover scheme
#       2: use Nose Hoover chains scheme
# ALGORITHM
        0
# NBATHS
        2
# TEMP0(1 ... NBATHS)  TAU(1 ... NBATHS)
     293    0.1       293    0.1
# DOFSET: number of distinguishable sets of d.o.f.
          2
# LAST(1 ... DOFSET)  COMBATH(1 ... DOFSET)  IRBATH(1 ... DOFSET)
      71      1       1      2666      2       2
END
BOUNDCOND
#      NTB     NDFMIN
        1          0
END
COMTRANSROT
#      NSCM
       1000
END
```

Temperature and coupling time

Solute internal, rotational, and com translational motion coupled to one bath

(Baths 1 and 2)

Solvent coupled to a different bath

(solute coupled to bath 1)

(solvent coupled to bath 2)

Rectangular, periodic boundary condition

Remove com motion every NSCM steps

```
WRITETRAJ
# NTWX NTWSE NTWV NTWE NTWG NTWB
  100  0     0    100  0    0
END
PRINTOUT
# NTPR: print out energies, etc. every NTPR steps
# NTPP: =1 perform dihedral angle transition monitoring
# NTPR NTPP
  100     0
END
CONSTRAINT
# NTC
    3
# NTCP  NTCP0(1)
     1    0.0001
# NTCS  NTCS0(1)
     1    0.0001
END
```

Write position trajectory every NTWX steps

Write energy trajectory every NTWE steps

Print energies etc. to output file every NTPR steps

Constrain all bonds

```
FORCE
# NTF(1)    NTF(2)     NTF(3)    NTF(4)    NTF(5)         NTF(6)
# bonds     angles     improper  dihedral  electrostatic  vdW
  0         1          1         1         1              1
# NEGR NRE(1) NRE(2) ... NRE(NEGR)
  4    69     70     71  2668
END
PRESSURESCALE
# COUPLE   SCALE    COMP     TAUP  VIRIAL
       2       1 0.000917      0.5       2
# SEMIANISOTROPIC COUPLINGS(X, Y, Z)
       1        1        1
# PRES0(1...3,1...3)
 0.06102        0        0
       0  0.06102        0
       0        0  0.06102
END
PAIRLIST
# algorithm     NSNB    RCUTP    RCUTL     SIZE     TYPE
          1        5      0.8      1.4      0.4        0
END
```

Do not calculate bond forces (constraints).

Definition of 4 energy groups: Energies between these groups will be calculated and will appear in the energy trajectory.

(pressure in GROMOS units!)

Use fast grid pairlist algorithm to search for pairs of atom that interact via nonbonded interactions.

update pairlist every 5th step

Do not include pairs that are further apart than these cutoff distances

```
NONBONDED
# NLRELE
       1
#  APPAK     RCRF    EPSRF
       0      1.4     66.6
# NSHAPE   ASHAPE   NA2CLC    TOLA2    EPSLS
       3      1.4        2    1e-10        0
#     NKX      NKY      NKZ    KCUT
      10       10       10     100
#     NGX      NGY      NGZ  NASORD  NFDORD  NALIAS  NSPORD
      32       32       32       3       2       3       4
# NQEVAL   FACCUR  NRDGRD  NWRGRD
  100000      1.6       0       0
#  NLRLJ   SLVDNS
       0    33.3
END
```

Beyond this cutoff replace electrostatic interactions by a static reaction field.

Dielectric permittivity of reaction field.

There are many more input blocks! E.g.

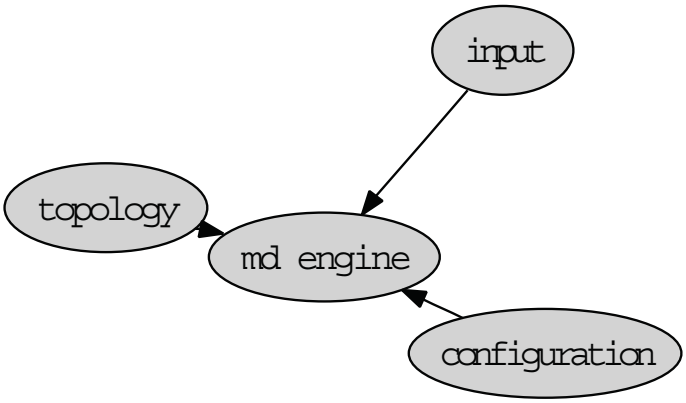
```
POSITIONRES
#    NTPOR    NTPORB   NTPORS     CPOR
         1         1        0      300
END
```

Do position restraining

Force constant used in restraining

Most basic usage of the program

```
leia:~> md @topo topo @conf conf @input input \
          @fin fin @trc trc @tre tre > out
```



The *gromos*++ program mkscript generates shell scripts for running more complicated simulations. E.g. minimization followed by thermalization and equilibration (see tutorial).

Documentation

GROMOS manual

will be accessible to you as pdf in the computer room D267.4 (check in the «computer setup» document)

The GROMOS Software for (Bio)Molecular Simulation

Volume 1: About the GROMOS package: Overview

June 14, 2013

Contents

Documentation

GROMOS manual

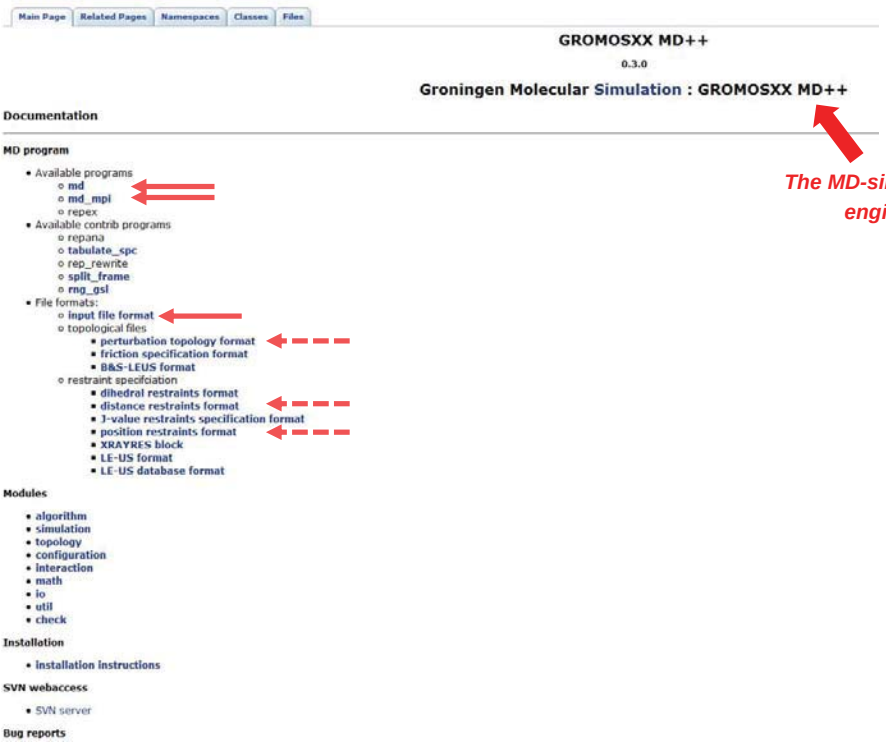
The current GROMOS manual and user guide exists of 9 volumes:

The GROMOS Software for (Bio)Molecular Simulation
Volume 1: About the GROMOS Package: Overview
Volume 2: Algorithms and Formulae for Modelling of Molecular Systems
Volume 3: Force Field and Topology Data Set
Volume 4: Data Structures and Formats
Volume 5: Program Library Manual
Volume 6: Technical Details
Volume 7: Tutorial with Examples
Volume 8: Installation Guide
Volume 9: Index

Documentation

doxygen html-based documentation

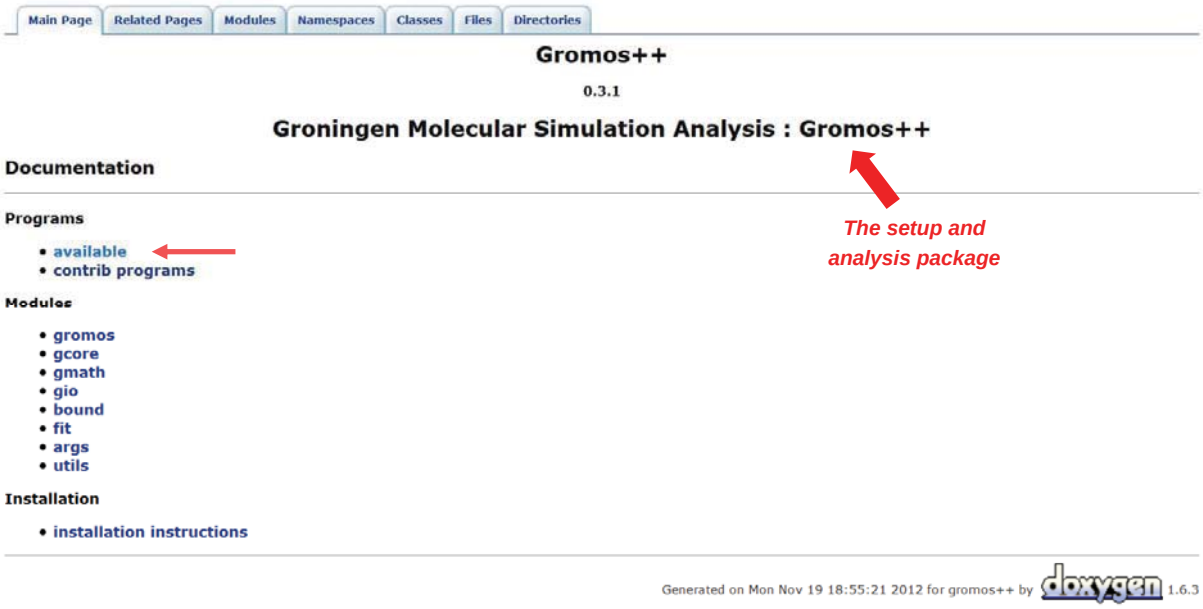
will be accessible to you as web-document in the computer room D267.4 (check in the «computer setup» document)



The MD-simulation engine

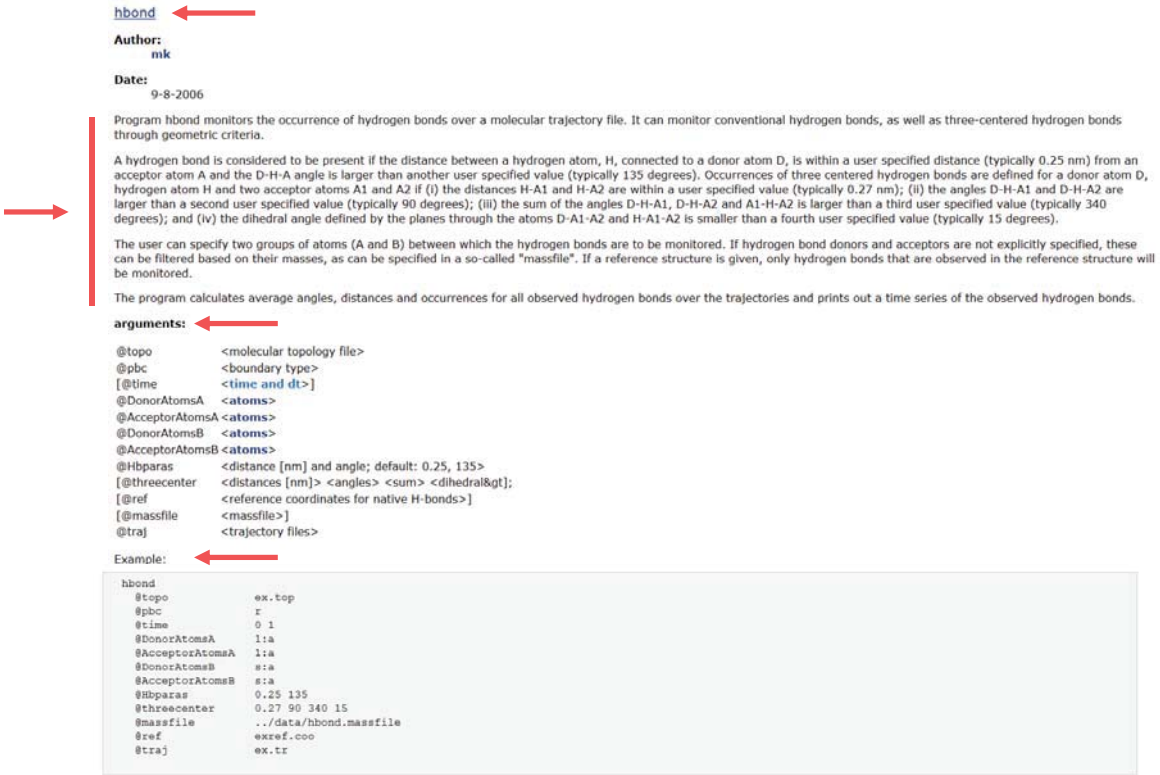
Documentation

doxygen html-based documentation



Documentation

Example: analysis program "hbond"



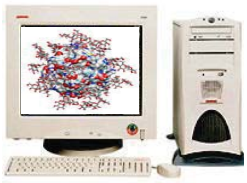
That's all folks

→ I wish you a happy GROMOS start next week!

The best way to learn swimming is...
...to swim !

P.H. Hünenberger

COMPUTER SIMULATION OF MOLECULAR SYSTEMS

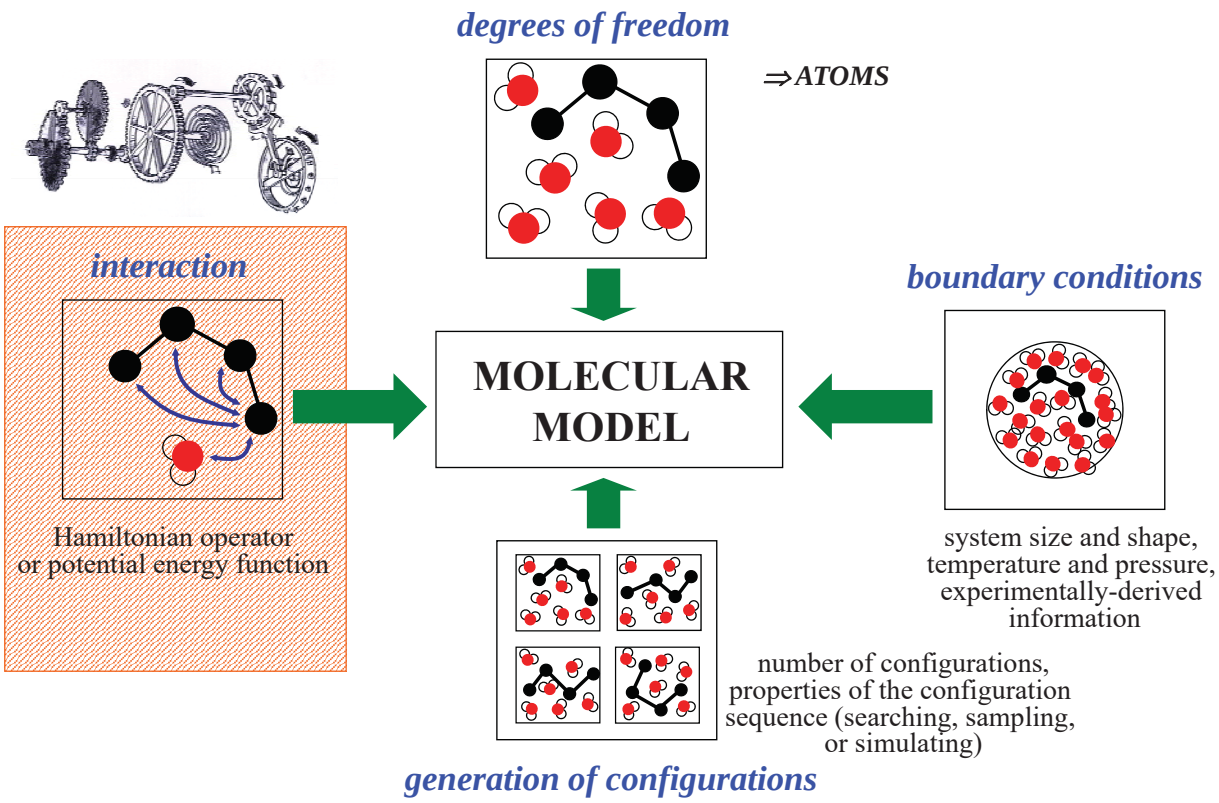


Herbstsemester 2019
Tuesday 9:45 a.m. – 11:30 p.m.
HCI D2

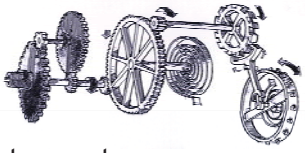
Lecture 529-0004-00
www.csms.ethz.ch/education/CSCBP

LECTURE 2 (WEEKS 2+3):
Force fields

Four basic choices defining a molecular model



Force field



- Force fields are the **machinery** of classical molecular simulations

- Given a set of elementary **particles** (atoms, atom groups, coarse-grained beads,...) assumed to behave **classically**, a **force field** is the **potential-energy function**, *i.e.* is the function returning the potential energy of the system in a given configuration

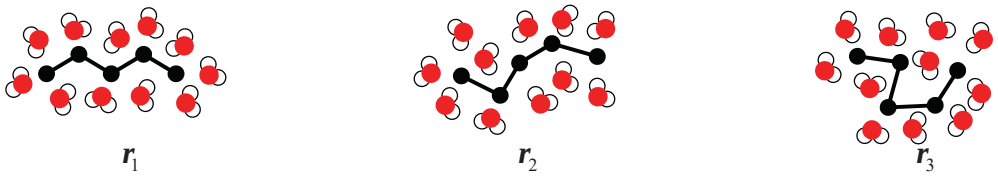
3N-dimensional coordinate vector — *3D coordinates of the N particles* ⇒ configuration

roughly: minus V tells how «happy» the system is in this configuration

$$V(\boldsymbol{r}) = V(\{\boldsymbol{r}_i \mid i = 1,2,..N\})$$

⇒ potential energy of the system in the given configuration

- *E.g.* united-atom pentane in water (● = CH_2 or CH_3)



- The **classical Hamiltonian** is a function that gives the **total energy** of the system

coordinates *momenta*

$$H(\boldsymbol{r},\boldsymbol{p}) = V(\boldsymbol{r}) + K(\boldsymbol{p})$$ *Hamiltonian function*

$$K(\boldsymbol{p}) = \sum_{i=1}^{N} \frac{\boldsymbol{p}_i^2}{2m_i}$$ *Kinetic energy*

(will be discussed later)

Potential-energy function

- In **all force fields**, the potential energy function is represented as a **sum** of force-field **terms**

- Each force-field term has a given **functional form** and depends on
  - → one or more **internal coordinates** of the system
  - → **parameters** specific to the term

internal (generalized) coordinate: any function of the Cartesian coordinates of a all particles (generally: of a small subset of atoms) e.g. a distance, an angle, a dihedral angle...

- A detailed (but tedious) way of writing this is

number of terms (there may easily be millions!) — *due to pairwise atom-atom interactions e.g. 10^4 atoms → ~$0.5 \cdot 10^8$ pairs*

$$V(\boldsymbol{r}) = \sum_{\alpha=1}^{N_{terms}} {}^{(n_\alpha)}V^{(t_\alpha)}(q_{\alpha,1}, q_{\alpha,2}, \ldots; s_{\alpha,1}, s_{\alpha,2}, \ldots)$$

order of term α *(number of particles involved)*	***type of term α*** *(functional form)*	***internal coordinates*** *(derived from **r**)*	***parameters*** *(belong to force-field definition)*

→ *E.g.* term α corresponding to a harmonic bond between atoms 4 and 5 in a molecule

$${}^{(2)}V^{(b)}(b_{45}; b^o_{45}, k^b_{45}) = \underbrace{(1/2)k^b_{45}(b_{45} - b^o_{45})^2}_{\text{harmonic spring}}$$

2-body — *bonded (harmonic form)*

- b_{45} — *actual distance 4-5 in given configuration r*
- b^o_{45} — *reference distance (parameter)*
- k^b_{45} — *harmonic force constant (parameter)*

Categories of force-field terms

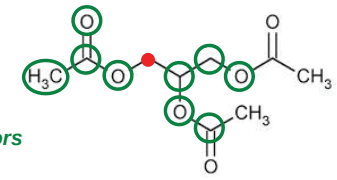
- Force-field terms belong to one of the **three** following categories

→ **physical** terms / **covalent** : between atoms within the **same molecule** that are **close covalent neighbors**, *i.e.* separated by 1, 2 or 3 bonds

here, the peculiarity of the covalent bonding must be taken into account

Example:
Glycerol triacetate

Reference atom
Close covalent neighbors



⇒ the electron density between the nuclei drives them to adopt specific bond distances, angles, point geometries (e.g. planar or tetrahedral) and torsional preferences

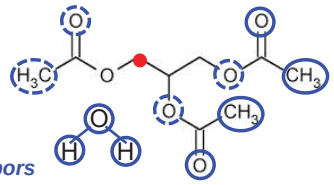
⇒ quantum mechanics of covalent interactions

→ **physical** terms / **non-bonded**: between atoms within **different molecules** or atoms within the **same molecule** that are **not close covalent neighbors**, *i.e.* separated by 3 or more bonds

here, we can consider a generic form of closed-shell interatomic interactions

Example:
Glycerol triacetate

Reference atom
Non-bonded neighbors
Non-bonded third neighbors



⇒ at these covalent distances, the interaction between atoms within the same molecule is about the same as if they were in different molecules

⇒ quantum mechanics of closed-shell interactions

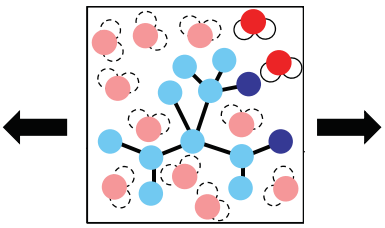
⇒ third neighbors are special: they are non-bonded as well as covalent neighbors (discussed later)

→ **unphysical** (artificial) terms

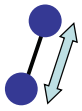
they do not correspond to interactions found in nature, and are used for various "engineering" purposes in simulations → when including such terms, the dynamics will be biased (non-natural)

Physical force-field terms

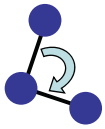
Covalent terms



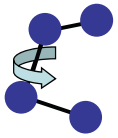
Non-bonded terms



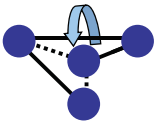
bond stretching



bond-angle bending

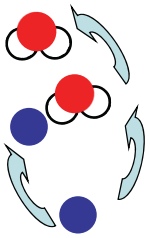


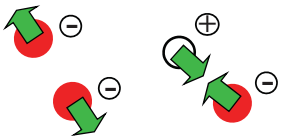
torsional dihedral



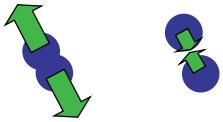
improper dihedral

all atom pairs in the system except close covalent neighbors



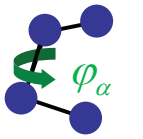
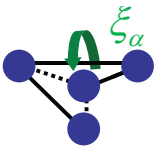
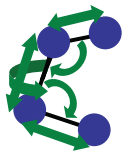


electrostatic interactions

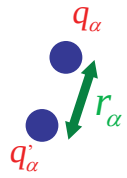
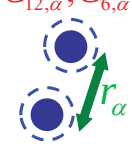
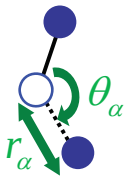


van der Waals interactions

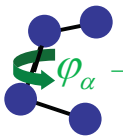
Covalent force-field terms

TYPE	GENERAL FORM	EXAMPLE
 bond stretching	$^{(2)}V^{(b)}(b_\alpha;\ldots)$	$^{(2)}V^{(b)}(b_\alpha;b_\alpha^o,k_\alpha^b) = (1/2)k_\alpha^b(b_\alpha - b_\alpha^o)^2$
 bond-angle bending	$^{(3)}V^{(\theta)}(\theta_\alpha;\ldots)$	$^{(3)}V^{(\theta)}(\theta_\alpha;\theta_\alpha^o,k_\alpha^\theta) = (1/2)k_\alpha^\theta(\theta_\alpha - \theta_\alpha^o)^2$
 torsional dihedral	$^{(4)}V^{(\varphi)}(\varphi_\alpha;\ldots)$	$^{(4)}V^{(\varphi)}(\varphi_\alpha;\{{}^nk_\alpha^\varphi,{}^n\delta_\alpha\}) = \sum_n {}^nk_\alpha^\varphi(1+\cos{}^n\delta_\alpha\cos n\,\varphi_\alpha)$
 out-of-plane (or tetrahedron) distortion *e.g.* improper dihedral	$^{(4)}V^{(\xi)}(\xi_\alpha;\ldots)$	$^{(4)}V^{(\xi)}(\xi_\alpha;\xi_\alpha^o,k_\alpha^\xi) = k_\alpha^\xi(\xi_\alpha - \xi_\alpha^o)^2$
 covalent cross-terms	$^{(3\,or\,4)}V^{(ct)}(\{b_\alpha\},\{\theta_\alpha\},\varphi_\alpha;\ldots)$	*various terms*

Non-bonded force-field terms

TYPE	GENERAL FORM	EXAMPLE
 electrostatic (all atom pairs)	$^{(2)}V^{(el)}(r_\alpha;q_\alpha,q_\alpha',\ldots)$	$^{(2)}V^{(el)}(r_\alpha;q_\alpha,q_\alpha') = (4\pi\varepsilon_o)^{-1}q_\alpha q_\alpha' r_\alpha^{-1}$ ***permittivity of vacuum (physical constant)***
 van der Waals (all atom pairs)	$^{(2)}V^{(vdW)}(r_\alpha;\ldots)$	$^{(2)}V^{(vdW)}(r_\alpha;C_{12,\alpha},C_{6,\alpha}) = C_{12,\alpha}r_\alpha^{-12} - C_{6,\alpha}r_\alpha^{-6}$ ***e.g. here, a Lennard-Jones potential***
 H-bond	$^{(3\,or\,4)}V^{(hb)}(r_\alpha,\theta_\alpha;\ldots)$	*various forms*

Unphysical force-field terms

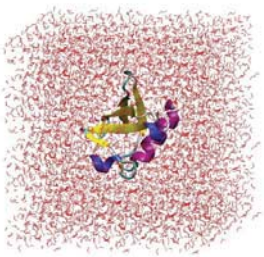
TYPE	GENERAL FORM	EXAMPLE
 position restraint	$^{(1)}V^{(pr)}(l_\alpha;\ldots)$	$^{(1)}V^{(pr)}(l_\alpha;k_\alpha^{pr}) = (1/2)k_\alpha^{pr}l_\alpha^2$
 distance restraint	$^{(2)}V^{(dr)}(d_\alpha;\ldots)$	$^{(2)}V^{(dr)}(d_\alpha;d_\alpha^o,k_\alpha^{dr}) = (1/2)k_\alpha^{dr}(d_\alpha-d_\alpha^o)^2 h(d_\alpha-d_\alpha^o)$ *e.g. half-harmonic (attractive) restraints* *Heaviside function*
 ^{3}J-value restraint (*Karplus equation*)	$^{(4)}V^{(jr)}(J_\alpha;\ldots)$	$^{(4)}V^{(jr)}(J_\alpha;J_\alpha^o,k_\alpha^{jr}) = (1/2)k_\alpha^{jr}(J_\alpha-J_\alpha^o)^2$

Used *e.g.* to:
- *apply perturbation (e.g. external field)*
- *restrain/confine system to prevent conformational drifts*
- *bias/enhance sampling or/and force processes to occur*
- *enforce (on average) agreement with experimental data*
- *...*

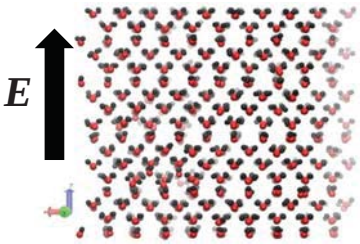
Remember from lecture 1: one key advantage of simulation

absolute freedom (in the procedure)

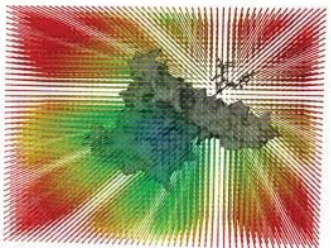
Use of unphysical force-field terms



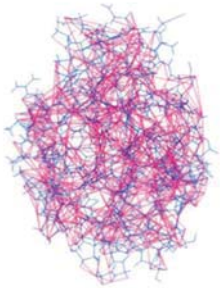
A protein was (roughly) solvated in a water box → equilibrate using MD with position restraints to the atom coordinates in the X-ray structure (avoid structural distortions)



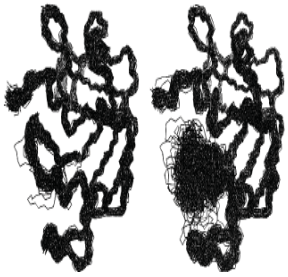
Water is simulated in the presence of a strong electric field (5 V nm^{-1}, along the z-axis); this is essentially impossible to achieve experimentally (electrical double-layer at electrodes!) → it freezes to hexagonal ice!



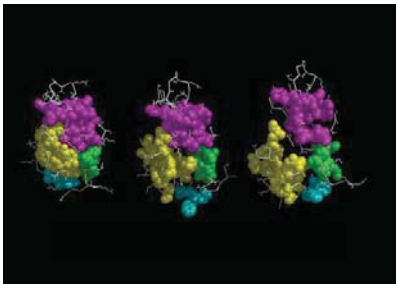
A ligand is artificially «pulled» into the active-site of a protein using a potential decreasing a «topological» distance (distancefield)



A protein is simulated with distance restraints between close protons pairs (based on NOE cross-peak intensities) → the generated ensemble is compatible with the NOE data



The sampling of backbone conformations in a protein loop is enhanced by applying a potential that «flattens» the torsional barriers in the loop



A protein is artificially unfolded by application of a potential that forces its radius of gyration to increase

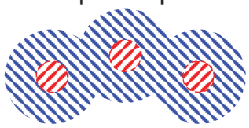
→We will use (and abuse) of these from Lecture 6 onward; but for now, let us look at ***physical terms***!

Approximations behind the force-field description

- The **microscopic world** is only accurately described by **quantum mechanics**.
  Why is the force-field description a reasonable approximation ?

Approximation 1 : simplified particle model (nuclei + electrons + electromagnetic interactions)

M nuclei
N electrons



$$\hat{\mathrm{H}}(t)\tilde{\Psi}(\boldsymbol{\tau},\boldsymbol{\nu},t) = i\hbar\frac{\partial\tilde{\Psi}(\boldsymbol{\tau},\boldsymbol{\nu},t)}{\partial t}$$

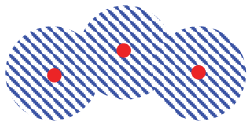
time-dependent Schrödinger equation

Approximation 2 : time-independent Hamiltonian [time-independent Schrödinger equation]
⇒ constant environment, no interaction with light

Approximation 3 : isolated system / pure-electrostatic interaction [form of the Hamiltonian]
⇒ no external field, no magnetic interactions in system, no relativistic effects

Approximation 4 : decoupling of electronic and nuclear motion (Born-Oppenheimer)

Electronic problem:
nuclei are quasi motionless



(compared to speed of electrons)

Nuclear problem:
electrons relax instantaneously

$$\hat{\mathrm{H}}_e\Psi_e(\boldsymbol{\tau};\boldsymbol{\nu}) = V_e(\boldsymbol{\nu})\Psi_e(\boldsymbol{\tau};\boldsymbol{\nu})$$

→ solutions for electronic energy levels $\{V_k(\boldsymbol{\nu})\}$

for a given k

$$\hat{\mathrm{H}}_n\Psi_n(\boldsymbol{\nu}) = E\Psi_n(\boldsymbol{\nu})$$

→ solutions for rotational/vibrational energies $\{E_{kl}\}$

Approximation 5' : electronic ground state
⇒ no electronic excitation

⇒ ground-state mean electronic energy $V_e(\boldsymbol{\nu})$ (kinetic e and Coulombic n-e and e-e)

⇒ set of solutions for the electronic/ rotational/vibrational energy levels E_{kl}

(prime: approximations required for classical simulations)

Approximation 6': Classical nuclei approximation

- At this point, nuclei are particles moving **quantum-mechanically** under the influence of a potential energy function $V(\boldsymbol{\nu})$ including
  → the **mean effect** of the **electrons** in the **ground electronic state**
  → the **internuclear repulsion**

$$\hat{\mathrm{H}}_n\Psi_n(\boldsymbol{\nu}) = E\Psi_n(\boldsymbol{\nu})$$

time-independent Schrödinger equation for the nuclei

operator with energy as eigenvalue

$$\hat{\mathrm{H}}_n = -\frac{1}{2}\sum_{\alpha}^{M}\frac{1}{\eta_\alpha}\nabla_\alpha^2 + V(\boldsymbol{\nu})$$

$$V(\boldsymbol{\nu}) = V_e(\boldsymbol{\nu}) + \sum_{\alpha}^{M}\sum_{\beta>\alpha}^{M}\frac{Z_\alpha Z_\beta}{r_{\alpha\beta}}$$

spin becomes irrelevant

Assumption 6' : nuclei behave as **classical particles**, named **atoms** (notation: $\boldsymbol{\nu}\rightarrow\boldsymbol{r}$)

Cartesian coordinates

Lagrangian

$$\frac{\partial \mathrm{H}(\boldsymbol{r},\boldsymbol{p})}{\partial \boldsymbol{r}} = -\dot{\boldsymbol{p}} \quad \textit{and} \quad \frac{\partial \mathrm{H}(\boldsymbol{r},\boldsymbol{p})}{\partial \boldsymbol{p}} = \dot{\boldsymbol{r}} \quad \textit{with} \quad \mathrm{H}(\boldsymbol{r},\boldsymbol{p}) = V(\boldsymbol{r}) + K(\boldsymbol{p})$$

Cartesian momenta i.e. $\boldsymbol{p} = \frac{\partial[K(\dot{\boldsymbol{r}}) - V(\boldsymbol{r})]}{\partial\dot{\boldsymbol{r}}}$

= total energy of the system

function with energy as value

classical Hamiltonian equations of motion (here, Cartesian)

→ can be generalized to arbitrary coordinate systems (see later)

Equivalent to

$$-\boldsymbol{F}(\boldsymbol{r}) = -\dot{\boldsymbol{p}} \quad \textit{and} \quad \underline{\boldsymbol{M}}^{-1}\boldsymbol{p} = \dot{\boldsymbol{r}}$$

classical Newtonian equations of motion

***Restrictions** (thermodynamics and dynamics):*

- *no electronic processes (≠chemical reactions)*
- *heavy-enough nuclei (≠protons)*
- *low-frequency oscillations (≠bond stretching)*
- *high-enough temperature (classical stat. mech.)*

Approximation 7': Analytical function approximation

- The dynamics is fully determined if we know the **potential-energy function** *V(**r**)*

→ In general, this function would be a very complex (non-analytical) function of the atomic coordinates, obtained by solving the electronic time-independent Schrödinger equation (electronic ground state) for different nuclear configurations

Assumption 7' : the classical potential energy function may be well *approximated* by a **simple analytical function**, expressed as a **sum of terms**.

$$V(\boldsymbol{r}) = \sum_{\alpha=1}^{N_{terms}} {}^{(n_\alpha)}V^{(t_\alpha)}(q_{\alpha,1}, q_{\alpha,2}, \ldots; s_{\alpha,1}, s_{\alpha,2}, \ldots)$$

Justification :

- *chemical intuition (e.g. entities such as "bonds" exist)*
- *work on simple systems (e.g. vibrations in small molecules, real monoatomic gases)*
- *with good parameters, agreement with experiment may be reached*

V(r) is then fully specified by:

- *The functional form of the terms of type t_α (for all t_α)*
- *The definition of the internal coordinate(s) involved in the terms of type t_α (for all t_α)*

together with (for a given molecular system):

- *The number of terms α of a given type t_α, and for each term α of type t_α :*
  - *The list of atoms involved in the definition of the corresponding internal coordinate(s)*
  - *The parameters involved in the corresponding term*

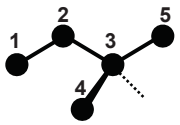
belongs to the force-field definition (hard-coded in the program)

belongs to the molecular topology (stored in a file for a given system)

★

Molecular topology file

E.g. united-atom isopentane



type	*#terms*	*atom i*	*atom j*	*atom k*	*atom l*	*parameters (example)*
bonds	4	1	2			b^o_{12}, k^b_{12}
		2	3			b^o_{23}, k^b_{23}
		3	4			b^o_{34}, k^b_{34}
		3	5			b^o_{35}, k^b_{35}
bond angles	4	1	2	3		$\theta^o_{123}, k^\theta_{123}$
		2	3	4		$\theta^o_{234}, k^\theta_{234}$
		2	3	5		$\theta^o_{235}, k^\theta_{235}$
		4	3	5		$\theta^o_{435}, k^\theta_{435}$
torsional dihedral	2	1	2	3	4	$\{{}^n\tilde{k}^\varphi_{1234}, {}^n\delta_{1234}\}$
		1	2	3	5	$\{{}^n k^\varphi_{1235}, {}^n\delta_{1235}\}$
improper dihedral	1	2	3	4	5	$\xi^o_{2345}, k^\xi_{2345}$
non-bonded pairs	10	1	2			$q_1q_2, C_{12,12}, C_{6,12}$
		1	3			$q_1q_3, C_{12,13}, C_{6,13}$
		...	...			...

N(N-1)/2 →TOO MANY

atoms	*#atoms*	*atom*	*charge*	*vdW-code*	*...*
	5	1	+0.1	1	
		2	- 0.1	2	
		3	- 0.2	3	
		4	+0.1	1	
		5	+0.1	1	

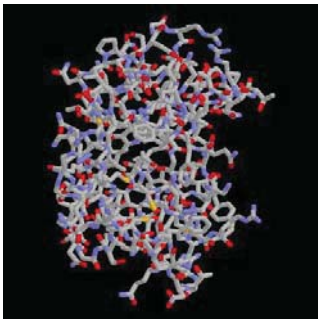
[exclusions, 3rd neighbours]

not included, generated on the flight from atomic parameters

Molecular topology file

E.g. the protein hen-egg-white lysozyme, within GROMOS96

129 residues
1321 atoms



type	*#terms*	*#parameters*
bonds	1345	2690
bond angles	1970	3940
torsional dihedrals	698	2094
improper dihedrals	687	1374
non-bonded pairs	871860	2615580 (or 408)
total:	876560	2625678 (or 10506)

[16 atom types, 3 parameters per type →3×(16×17)/2

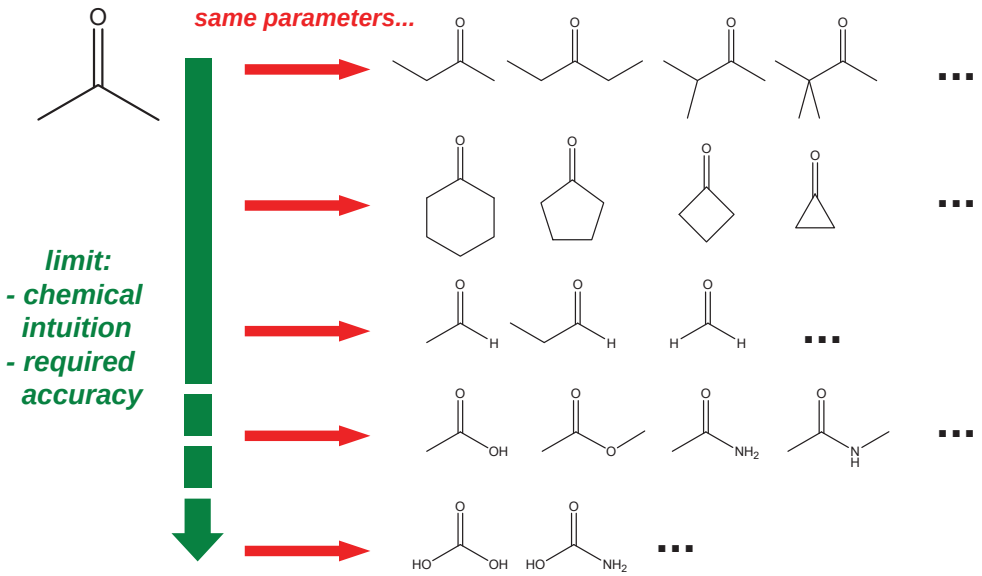
Without further simplifications, defining so many parameters is a mission impossible .....

Approximation 8': Transferability

One needs to reduce the number of parameters to be defined and calibrated...

Assumption 8' : a force-field term describing the interaction of a **given set of atoms** in a **given chemical environment** can be used for any ***set of the same atoms*** in a ***similar chemical environment***

E.g. carbonyl bond-stretching parameters:



Using transferability

Atom type:
"particles" considered by the the force field, corresponding to a *specific (united-)atom* in a *specific chemical environment*.

+

Combination rule:
rule determining the *value of a force-field parameter* given the *atom types* of the atoms involved.

or...

Topology building block:
List of *atom types* with a defined *connectivity* (+*parameters* for terms where no combination rule is used), characterizing a *molecule* or a *monomer in a polymer*.

Example:

type	*code*	*description*
...	...	...
C	1	aliphatic carbon
CR	2	aromatic carbon
CO	3	carbonyl carbon
...	...	...
OH	4	hydroxyl oxygen
O	5	ether oxygen
OC	6	carbonyl oxygen
...	...	...

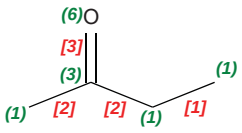
table

type i	*type j*	*bond parameters*
...	...	...
3	6	$r^{o}_{(CO,OC)}, k^{b}_{(CO,OC)}$
...	...	...

formula

$$C_{6,ij} = (C_{6,i} \times C_{6,j})^{1/2} \quad , \quad C_{12,ij} = (C_{12,i} \times C_{12,j})^{1/2}$$

type i	*atomic parameters*
...	...
3	$C_{12,(CO)}, C_{6,(CO)}$
6	$C_{12,(OC)}, C_{6,(OC)}$
...	...



(#) : atom type

[#] : bond type
(if no combination rule)

Example: The GROMOS 54A7 atom types

integer atom code	atom type		description
IAC[N]	TYPE[N]		
1	O		carbonyl oxygen (C=O)
2	OM		carboxyl oxygen (CO^-)
3	OA	***OXYGEN***	hydroxyl or sugar oxygen
4	OE	***types***	ether or ester oxygen
5	OW		water oxygen
6	N		peptide nitrogen (NH)
7	NT		terminal nitrogen (NH2)
8	NL	***NITROGEN***	terminal nitrogen (NH3)
9	NR	***types***	aromatic nitrogen
10	NZ		Arg NH (NH2)
11	NE		Arg NE (NH)
12	C		bare carbon
13	CH0		bare sp3 carbon, 4 bound heavy atoms
14	CH1	***(UNITED)***	aliphatic or sugar CH-group
15	CH2	***CARBON***	aliphatic or sugar CH2-group
16	CH3	***types***	aliphatic CH3-group
17	CH4		methane
18	CH2r		aliphatic or sugar CH2 group in ring
19	CR1		aromatic CH-group
20	HC	***HYDROGEN***	hydrogen bound to carbon
21	H	***types***	hydrogen not bound to carbon
22	DUM		dummy atom
23	S		sulphur
24	CU1+		copper (charge 1+)
25	CU2+		copper (charge 2+)
26	FE		iron (heme)
27	ZN2+		zinc (charge 2+)
28	MG2+		magnesium (charge 2+)
29	CA2+		calcium (charge 2+)
30	P, SI		phosphor or silicon
31	AR		argon
32	F		fluor (non-ionic)
33	CL		chlorine (non-ionic)
34	BR		bromine (non-ionic)

35	CMet	CH3-group in methanol (solvent)
36	OMet	oxygen in methanol (solvent)
37	NA+	sodium (charge 1+)
38	CL-	chloride (charge 1-)
39	CChl	carbon in chloroform (solvent)
40	CLChl	chloride in chloroform (solvent)
41	HChl	hydrogen in chloroform (solvent)
42	SDmso	sulphur in DMSO (solvent)
43	CDmso	CH3-group in DMSO (solvent)
44	ODmso	oxygen in DMSO (solvent)
45	CCl4	carbon in carbontetrachloride (solvent)
46	CLCl4	chloride in carbontetrachloride (solvent)
47	FTfe	fluor in trifluorethanol
48	CTfe	carbon in trifluorethanol
49	CHTfe	CH2-group in trifluorethanol
50	OTfe	oxygen in trifluorethanol
51	CUrea	carbon in urea
52	OUrea	oxygen in urea
53	NUrea	nitrogen in urea
54	CH3p	positively charged methyl

GROMOS manual Vol 3 (table 3.21) / file 54A7.ifp

Uses united atoms CH_1, CH_2, CH_3, CH_4
- *treated as single particles*
- *reduces the number of solute atoms by ~50% (proteins) or ~30% (nucleic acids) [but: most expensive is generally the solvent!]*
- *removes high-frequency C-H bond vibration*
- *NOT applied to aromatic CH and polar XH (X=N,O,S,...)*

Example: The GROMOS 54A7 building block for alanine

>6 atoms refer to the next residue (or end-cap) in a protein

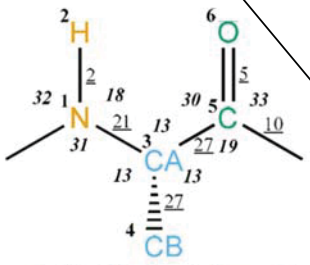


FIGURE 4.2. ALA bonded parameters.

I	J	Type
1	2	2
1	3	21
3	4	27
3	5	27
5	6	5
5	7	10

Table 4.2: Bonds

negative atoms refer to the previous residue (or end-cap) in a protein

I	J	K	Type
-1	1	2	32
-1	1	3	31
2	1	3	18
1	3	4	13
1	3	5	13
4	3	5	13
3	5	6	30
3	5	7	19
6	5	7	33

Table 4.3: Bond-angles

I	J	K	L	Type
-2	-1	1	3	14
-1	1	3	5	43
-1	1	3	5	44
1	3	5	7	42
1	3	5	7	45

Table 4.4: Dihedral angles

I	J	K	L	Type
1	-1	3	2	1
3	1	5	4	2
5	3	7	6	1

Table 4.5: Improper dihedral-angles

GROMOS manual Vol 4 / file 54A7.mtb

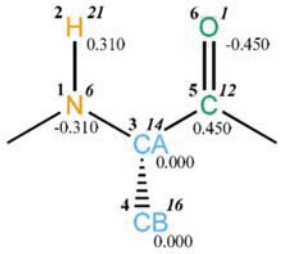


FIGURE 4.1. ALA non-bonded parameters.

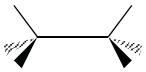
Seq.	Name	IAC	Mass	Charge	Exclusions
-1					0 1 2 3
0					1
1	N	6	14	-0.31000	2 3 4 5
2	H	21	1	0.31000	3
3	CA	14	3	0.00000	4 5 6 7
4	CB	16	5	0.00000	5
5	C	12	12	0.45000	
6	O	1	16	-0.45000	

TABLE 4.1. Atoms of building block ALA.

Types → table (in ifp file). No explicit combination rules for covalent terms (they are still used implicitly, but not explicitly enforced)

Only a subset of the possible torsional and improper dihedrals is considered (difficult without a building block approach...)

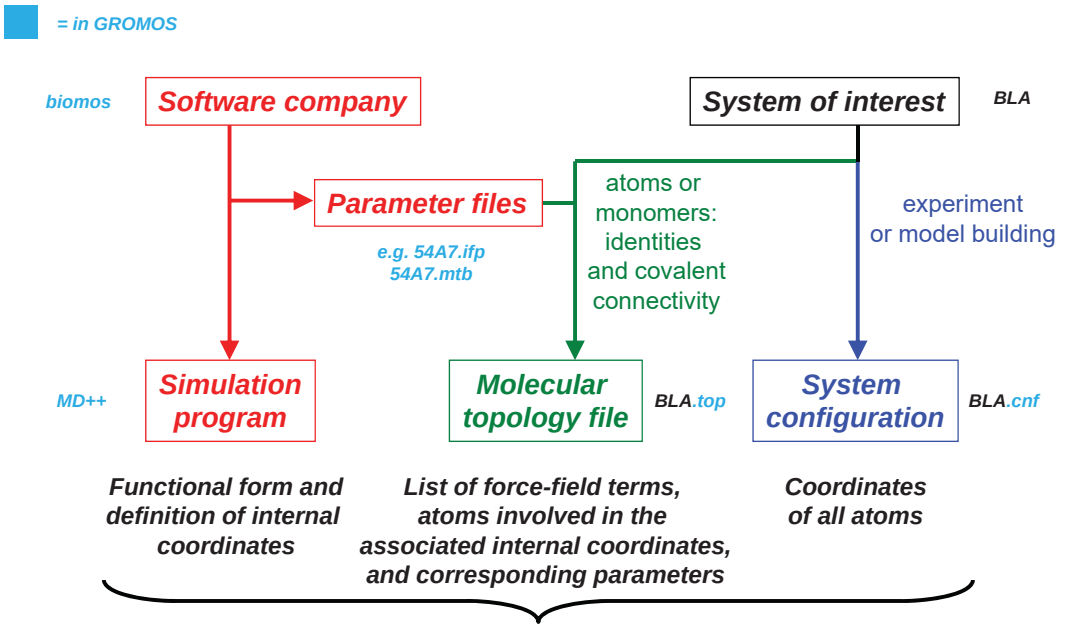
one out of nine possible dihedrals



Combination rule for van der Waals interaction IAC_i , IAC_j → table (in ifp file) → interaction parameters for ij (a geometric-mean formula is implicitly used to construct the table, but not explicitly enforced)

Combination rule for electrostatic interaction q_i , q_j → $q_i q_j$ (i.e. a formula)

Overview of a potential-energy calculation



$$V(\boldsymbol{r})$$

Main force-fields for condensed-phase biomolecular simulations

- **CHARMM**: Chemistry at HARvard Molecular Mechanics (www.charmm.org)

  → developed since the early 80's under the leadership of Martin Karplus, Harvard

- **AMBER**: Assisted Model Building with Energy Refinement (amber.scripps.edu)

  → Developed since the early 80's under the leadership of Peter Kollman, UCSF

- **OPLS**: Optimized Potentials for Liquid Simulation (zarbi.chem.yale.edu)

  → Developed since mid 80's under the leadership of William Jorgensen, Yale

- **GROMOS**: GROningen MOlecular Simulation (www.gromos.net)

  → Developed since the early 80's under the leadership of Wilfred van Gunsteren, ETHZ

  → Do not confuse with GROMACS (simulation program only, no own force-field)

- **TraPPE**: Transferable Potentials for Phase Equilibria
  (http://chem-siepmann.oit.umn.edu/siepmann/trappe)

  → Developed since 1998 under the leadership of J. Ilja Siepmann, Uni Minnesota

- ...

➡ **BUT**: don't ask which one is the best – it is *very difficult* to compare force-fields in a fair manner (what system, what conditions?)

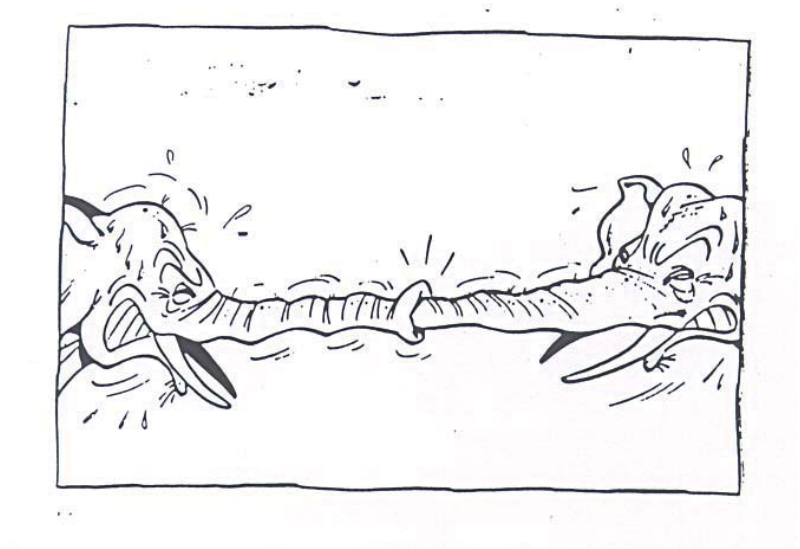
Overview of the GROMOS development

	PROGRAM	*FORCE FIELD*	*REFERENCES*
1984		26C1	HE84.1 (VA82.1)
1987	PROMD87 (Fortran77)		VA87.3
1987		37C4	VA87.3
1996	PROMD96 (Fortran 77)		VA96.1, SC99.1 (SC95.4, BO00.8)
1996		43A1	VA96.1, DA98.3, VA98.3
2000		43A2	SC00.1
2001		45A3	SC01.4 (SO04.1)
2005		45A4	LI05.1 (CH03.1, BO04.2, SO05.1)
2004		53A5	OO04.1
2004		53A6	OO04.1 (OO05.1)
2005	GROMOS05 (C++)		CH05.1
2011		56A6@CARBO	HA11.1
2011		53A6@OXY	HO11.1
2012		53A6@OXY+N	HO12.1
2012		53A6@OXY+D	FU12.1
2011	GROMOS11/MD++ (C++)		SC12.1, RI11.2, SC11.3, EI11.1, KU12.1 (SC10.1)
2011		54A7	SC11.1 (PO10.7, HU11.2)
2012		54A8	RE12.2 (RE13.1)
2016		2016H66	HO16.1

Overview of the GROMOS development

HE84.1 : [Hermans/Postma] A consistent empirical potential for water-protein interactions.
VA82.1 : [van Gunsteren/Karplus] Effect of constraints on the dynamics of macromolecules.
VA87.3 : [van Gunsteren/Berendsen] (book) Groningen molecular simulation (GROMOS) library manual.
VA96.1 : [van Gunsteren/Tironi] (book) Biomolecular simulation: The GROMOS96 manual and user guide.
SC99.1 : [Scott/vanGunsteren] The GROMOS biomolecular simulation program package.
SC95.4 : [Scott/vanGunsteren] NO TITLE IN RECORD
BO00.8 : [Bonvin/vanGunsteren] The GROMOS96 benchmarks for molecular simulation.
VA96.1 : [van Gunsteren/Tironi] (book) Biomolecular simulation: The GROMOS96 manual and user guide.
DA98.3 : [Daura/vanGunsteren] Parametrization of aliphatic CHn united atoms of GROMOS96 force field.
VA98.3 : [van Gunsteren/Mark] NO TITLE IN RECORD
SC00.1 : [Schuler/vanGunsteren] On the choice of dihedral angle potential energy functions for n-alkanes.
SC01.4 : [Schuler/vanGunsteren] An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase.
SO04.1 : [Soares/vanGunsteren] Validation of the GROMOS force-field parameter set 45A3 against nuclear magnetic resonance data of hen egg lysozyme.
LI05.1 : [Lins/Hunenberger] A new GROMOS force field for hexopyranose-based carbohydrates.
CH03.1 : [Chandrasekhar/vanGunsteren] A consistent potential energy parameter set for lipids: dipalmitoylphosphatidylcholine as a benchmark of the GROMOS96 45A3 force field.
BO04.2 : [Borjesson/Hunenberger] pH-dependent stability of a decalysine alpha-helix studied by explicit-solvent molecular dynamics simulations at constant pH.
SO05.1 : [Soares/vanGunsteren] An improved nucleic-acid parameter set for the GROMOS force field.
OO04.1 : [Oostenbrink/vanGunsteren] A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6.
OO05.1 : [Oostenbrink/vanGunsteren] Validation of the 53A6 GROMOS force field.
CH05.1 : [Christen/vanGunsteren] The GROMOS software for biomolecular simulation: GROMOS05.
HA11.1 : [Hansen/Hunenberger] A reoptimized GROMOS force field for hexopyranose-based carbohydrates accounting for the relative free energies of ring conformers, anomers, epimers, hydroxymethyl rotamers and glycosidic linkage conformers.
HO11.1 : [Horta/Hunenberger] New interaction parameters for oxygen compounds in the GROMOS force field: Improved pure-liquid and solvation properties for alcohols, ethers, aldehydes, ketones, carboxylic acids and esters.
HO12.1 : [Horta/Hunenberger] Reoptimized interaction parameters for the peptide-backbone model compound N-methylacetamide in the GROMOS force field: Influence on the folding properties of two beta-peptides in methanol.
FU12.1 : [Fuchs/Horta] A GROMOS parameter set for vicinal diether functions: propertiespolyethyleneoxide and polyethyleneglycol.
SC12.1 : [Schmid/vanGunsteren] Architecture, implementation and parallelisation of the GROMOS software for biomolecular simulation.
RI11.2 : [Riniker/vanGunsteren] Calculation of relative free energies for ligand-protein binding, solvation and conformational transitions using the GROMOS biomolecular simulation software.
SC11.3 : [Schmid/vanGunsteren] Biomolecular structure refinement using the GROMOS simulation software.
EI11.1 : [Eichenberger/vanGunsteren] The GROMOS++ software for the analysis of biomolecular simulation trajectories.
KU12.1 : [Kunz/vanGunsteren] New functionalities in the GROMOS biomolecular simulation software.
SC10.1 : [Schmid/{vanGunsteren}] A GPU solvent-solvent interaction calculation accelerator for biomolecular simulations using the GROMOS software.
SC11.1 : [Schmid/vanGunsteren] Definition and testing of the GROMOS force-field versions 54A7 and 54B7.
PO10.7 : [Poger/Mark] A new force field for simulating phosphatidylcholine bilayers.
HU11.2 : [Huang/vanGunsteren] Validation of the GROMOS 54A7 force field with respect to beta-peptide folding.
RE12.2 : [Reif/Oostenbrink] New interaction parameters for charged amino acid side chains in the GROMOS force field.
RE13.1 : [Reif/Oostenbrink] Testing of the GROMOS force-field parameter set 54A8: Structural properties of electrolyte solutions, lipid bilayers, and proteins.
HO16.1 : [Horta/Hunenberger] A GROMOS-compatible force field for small organic molecules in the condensed phase: The 2016H66 parameter set.

Covalent force-field terms



Two elephants kindly accepted to illustrate the concept of "hard" degrees of freedom...

Covalent terms: Bond stretching

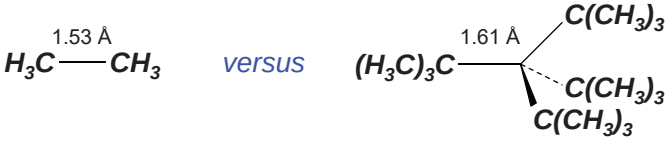
bond stretching $\quad {}^{(2)}V^{(b)}(b_\alpha;\ldots)$

Most common form: harmonic (GROMOS87,AMBER,CHARMM,...)

$$ {}^{(2)}V^{(b)}(b_\alpha;b_\alpha^o,k_\alpha^b)=(1/2)k_\alpha^b(b_\alpha-b_\alpha^o)^2 $$

reference bond length *harmonic force constant*

- No first-order term (ensure a zero force when $b_\alpha = b_\alpha^o$)
- The *reference bond length* b_α^o is not necessarily the *equilibrium bond length* in any molecule (strain), *e.g.*



- No terms of higher order than two (reasonable for moderately strained molecules and at low enough temperature)
- Does not allow for bond dissociation
- Bond lengths are *"hard"* degrees of freedom, *i.e.* small deviations away from the reference length induce a large energy increase compared to k_BT (average thermal energy per d.o.f.)

Covalent terms: Bond stretching

Example: harmonic reference lengths and force constants from the MM2 force field

Bond	b^o [nm]	k^b [kJ·mol^{-1}·nm^{-2}]
C(sp^3)-C(sp^3)	0.1523	1.326·10^5
C(sp^3)-C(sp^2)	0.1497	1.326·10^5
C(sp^2)=C(sp^2)	0.1337	2.887·10^5
C(sp^2)=O	0.1208	3.251·10^5
C(sp^3)-N(sp^3)	0.1438	1.536·10^5
C(sp^2)-N(sp^2) [amide]	0.1345	3.008·10^5

- Follow chemical intuition
- A C(sp^3)-C(sp^3) bond stretched by 0.01 nm already corresponds to an energy increase by 13.3 kJ·mol^{-1} (~5 k_BT at 300K)

Covalent terms: Bond stretching

Computationally cheaper form: quadratic in b^2 (GROMOS96)

$$^{(2)}V^{(b)}(b_\alpha; b_\alpha^o, k_\alpha^b) = (1/4)k_\alpha^b[b_\alpha^2 - (b_\alpha^o)^2]^2$$

- Avoids a square-root calculation (b_α not required)
- Saves very little in practice (covalent terms are cheap to compute)

Covalent terms: Bond stretching

More complex forms: Taylor expansion (MM2,MM3,CFF93,...)

$$^{(2)}V^{(b)}(b_\alpha; b_\alpha^o, {}^2k_\alpha^b, {}^3k_\alpha^b, ...) = (1/2)\,{}^2k_\alpha^b(b_\alpha - b_\alpha^o)^2 + (1/6)\,{}^3k_\alpha^b(b_\alpha - b_\alpha^o)^3 + ...$$

reference bond length *force constants*

- Still no first-order term (ensure a zero force when $b_\alpha = b_\alpha^o$)
- Terms up to order 3 (MM2; bad choice !) or 4 (MM3,CFF93)
- Better structural description of strained molecule, of vibrational frequencies (includes anharmonicities) and of molecules at elevated temperatures
- Useful if accurate structures, vibrational frequencies or heats of formation in the gas phase are wanted
- Does not allow for bond dissociation (at least not on purpose, see MM2 !)
- Typically an unnecessary complication for condensed-phase simulations of macromolecules, where bond description is not critical
  - high frequency bond-stretching vibrations have little influence on the dynamics
  - their statistical mechanics at the classical level is incorrect anyway
  - low frequency motions (solvent relaxation, conformational changes) have the largest impact on the thermodynamic properties
- More parameters to calibrate (compared to simple harmonic) !

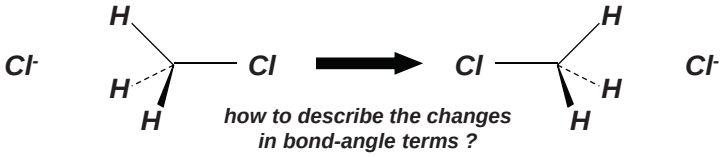
Covalent terms: Bond stretching

Dissociative forms: e.g. Morse function

$$^{(2)}V^{(b)}(b_\alpha; b_\alpha^o, D_\alpha, a_\alpha, ...) = D_\alpha[\exp(-a_\alpha(b_\alpha - b_\alpha^o)) - 1]^2$$

well depth *inverse well width*

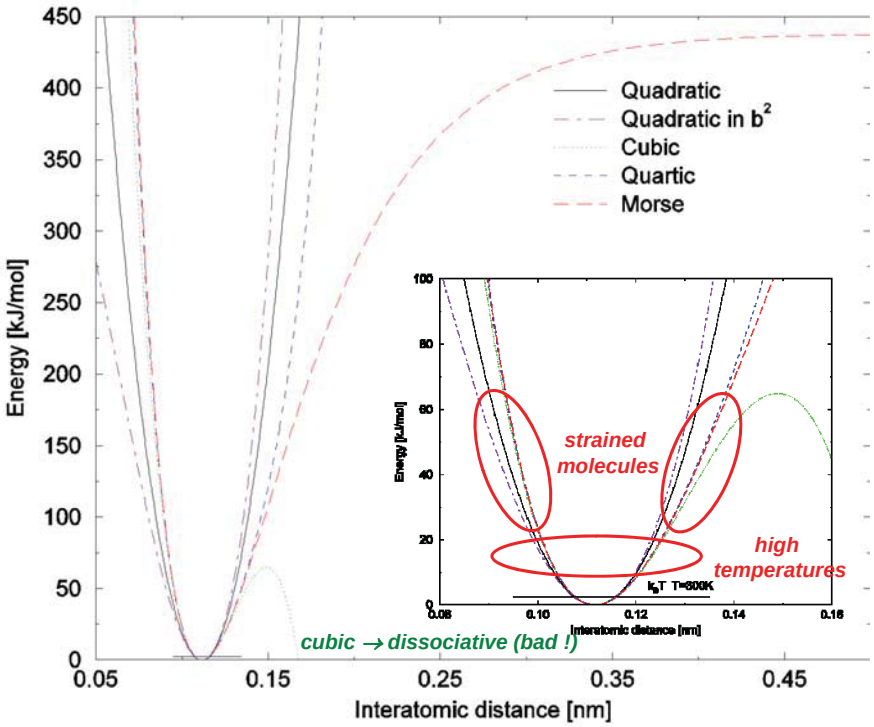
- Zero force when $b_\alpha = b_\alpha^o$
- Better structural description of strained molecule, of vibrational frequencies (includes anharmonicities) and of molecules at elevated temperatures
- Allows for bond dissociation, but seldom useful beyond diatomic molecules, because of changes involved in other terms



- Twice more parameters to calibrate (compared to simple harmonic) !
- Computationally expensive (exponential function)
- There are many other choice for three-parameter dissociative functions (generally developed for diatomic molecules)

Covalent terms: Bond stretching

C-H bond, Morse function ($D_\alpha = 437.8\, kJ \cdot mol^{-1}, a_\alpha = 17.87\, nm^{-1}, b_\alpha^o = 0.117\, nm$) and alternatives with identical second (and higher) derivatives at the minimum



Covalent terms: Bond stretching

Use of bond-length constraints

- Bond distance is rigorously fixed to the reference vaue ($b_\alpha = b_\alpha^o$)
  e.g. using SHAKE (equilibrium distance would be better - so-called *soft constraints*)

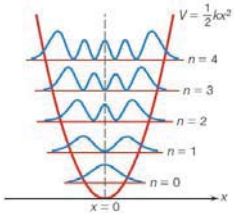
- No potential energy term / no kinetic energy contribution

- Justifications:

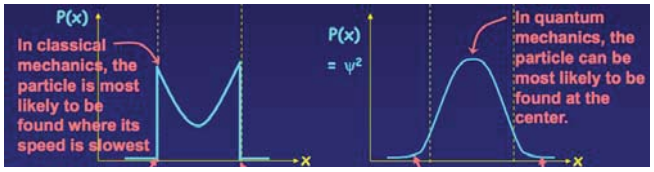
  - bond-stretching vibrations are often uninteresting

  - frequencies (≥1000 cm^{-1}) are above k_BT (200 cm^{-1}) and these vibrations (which should be treated quantum-mechanically) are not excited at room temperature (zero-point energy, no heat capacity/entropy contribution)
    → a constraint may be a better approximation than a classical oscillator in terms of statistical mechanics

QM
(energy levels, probability)



classical limit QM ground state



- allows for longer timesteps in MD (typically: 0.5 fs → 2.0 fs)

- narrower frequency spectrum → better exchange of kinetic energy

Covalent terms: Bond stretching

Combination rules (if applied):

- Generally in the form of tables:

  atom type of *i*
  atom type of *j* ➡ parameters for bond *i-j*

Exclusions:

- Bonded atoms (*first neighbors*) are excluded from any non-bonded interactions
  - this interaction would be huge
  - it should be encompassed in the bond-stretching term
  - non-bonded (van der Waals) interactions are only appropriate to account for non-bonding interactions (between closed-shell atoms)

Covalent terms: Bond-angle bending

bond-angle bending

$^{(3)}V^{(\theta)}(\theta_\alpha;...)$

Most common form: harmonic (GROMOS87,AMBER,CHARMM,...)

$$^{(3)}V^{(\theta)}(\theta_\alpha;\theta_\alpha^o,k_\alpha^\theta)=(1/2)k_\alpha^\theta(\theta_\alpha-\theta_\alpha^o)^2$$

reference bond angle *harmonic force constant*

- Most considerations made for bond stretching also apply here...
- Bond angles are also *"hard"* degrees of freedom, *i.e.* small deviations away from the reference angle induce a large energy increase compared to k_BT (but they are less hard than bonds)

Covalent terms: Bond-angle bending

Example: harmonic reference angles and force constants from the MM2 force field

Angle	θ^o [deg]	k^θ [kJ·mol^{-1}·deg^{-2}]
C(sp^3)-C(sp^3)-C(sp^3)	109.47	4.142·10^{-2}
C(sp^3)-C(sp^3)-H	109.47	3.305·10^{-2}
H-C(sp^3)-H	109.47	2.929·10^{-2}
C(sp^3)-C(sp^2)-C(sp^3)	117.20	4.142·10^{-2}
C(sp^3)-C(sp^2)=C(sp^2)	121.40	5.063·10^{-2}
C(sp^3)-C(sp^2)=O	122.50	4.226·10^{-2}

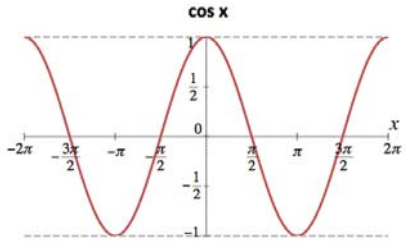
- Follow chemical intuition
- A C(sp^3)-C(sp^3)-C(sp^3) bond-angle bent by 10 degrees already corresponds to an energy increase by 4.1 kJ·mol^{-1} (~1.5 k_BT at 300K)

Covalent terms: Bond-angle bending

Computationally cheaper form: harmonic in cos(θ) (GROMOS96,DREIDING)

$$^{(3)}V^{(\theta)}(\theta_\alpha;\theta_\alpha^o,k_\alpha^\theta)=(1/2)k_\alpha^\theta(\cos\theta_\alpha-\cos\theta_\alpha^o)^2$$

- Avoids an "arccos" calculation (θ_α not required)
- Saves very little in practice (covalent terms are cheap to compute)
- A poor choice for molecules involving linear groups (because the potential is "flat" at 0 and π - zero derivative) !!!



Covalent terms: Bond-angle bending

More complex forms: Taylor expansion (MM2,MM3,CFF93,...)

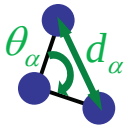
$$^{(3)}V^{(\theta)}(\theta_\alpha;\theta_\alpha^o,{}^2k_\alpha^\theta,{}^3k_\alpha^\theta,\ldots)=(1/2)\,{}^2k_\alpha^\theta(\theta_\alpha-\theta_\alpha^o)^2+(1/6)\,{}^3k_\alpha^\theta(\theta_\alpha-\theta_\alpha^o)^3+\ldots$$

reference bond angle *force constants*

- Most considerations made for bond stretching also apply here...
- Terms up to order 4 (CFF93) or 6 (MM2,MM3)
- Better structural description of strained molecule, of vibrational frequencies (includes anharmonicities) and of molecules at elevated temperatures
- Useful if accurate structures, vibrational frequencies or heats of formation in the gas phase are wanted
- Typically an unnecessary complication for condensed-phase simulations of macromolecules, where bond-angle description is not critical (+their statistical mechanics at the classical level is incorrect anyway)
- More parameters to calibrate (compared to simple harmonic) !

More complex forms: Urey-Bradley (e.g. CHARMM/DNA)

$$^{(3)}V^{(\theta)}(\theta_\alpha,d_\alpha;\theta_\alpha^o,d_\alpha^o,{}^2k_\alpha^\theta,{}^1k_\alpha^d,{}^2k_\alpha^d)=$$
$$(1/2)\,{}^2k_\alpha^\theta(\theta_\alpha-\theta_\alpha^o)^2+{}^1k_\alpha^d(d_\alpha-d_\alpha^o)+{}^2k_\alpha^d(d_\alpha-d_\alpha^o)^2$$



- Includes some anharmonicity and bond to bond-angle coupling
- If *V* is defined within a constant and d_α^o is replaced by an effective distance, ${}^1k_\alpha^d$ can be omitted

Covalent terms: Bond-angle bending

Use of bond-angle constraints

- The use of bond-angle constraints (in addition to bond-length constraints) would seem reasonable
  - frequencies (≥700 cm^{-1}) are above k_BT (200 cm^{-1}) and these vibrations (which should be treated quantum-mechanically) are not excited at room temperature (zero-point energy, no heat capacity / entropy contribution)
    → a constraint may be a better approximation than a classical oscillator

But: it is strongly disadvised in practice ! *except for fully-rigid molecules*

⇒ dynamic (hindered dihedral-angle rotations around bonds) and thermodynamic (incorrect dihedral-angle distributions related to metric tensor effects in MD) artifacts

Combination rules (if applied):

- Generally in the form of tables:

atom type of *i*
atom type of *j* ➡ parameters for bond-angle *i-j-k*
atom type of *k*

Exclusions:

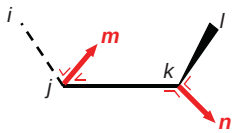
- Atoms separated by two covalent bonds (*second neighbors*) are excluded from any non-bonded interactions (for similar reasons as first neighbors)

Covalent terms: Torsional-dihedral rotation

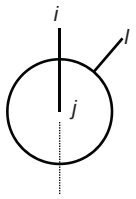
torsional dihedral

$^{(4)}V^{(\varphi)}(\varphi_\alpha;\ldots)$

Geometrical definition



$|\varphi_\alpha| = ang(\boldsymbol{m},\boldsymbol{n})$



Note: dihedral(i,j,k,l)=dihedral(l,k,j,i)

Most common form: cosine series

$$^{(4)}V^{(\varphi)}(\varphi_\alpha;\{^{n}k_\alpha^{\varphi}\}) = \sum_{n} {}^{n}k_\alpha^{\varphi}(1-\cos n\,\varphi_\alpha)$$

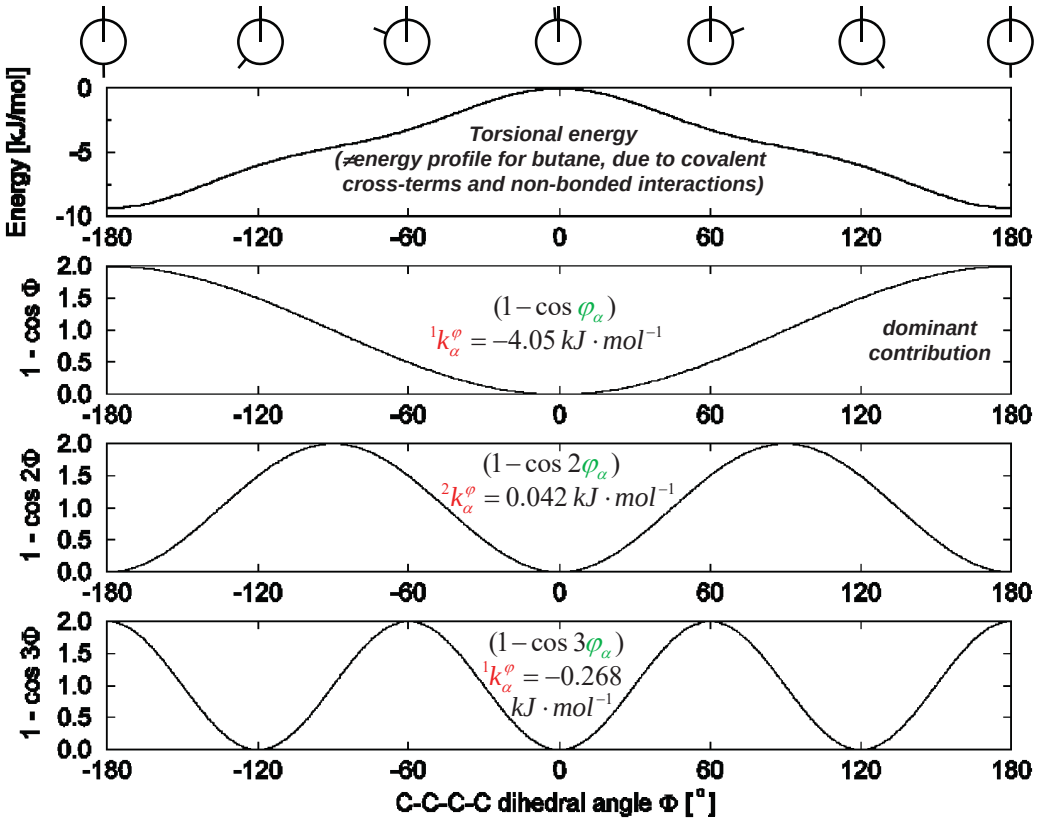
force constants — *n: multiplicity*

$$= {}^{1}k_\alpha^{\varphi}(1-\cos\varphi_\alpha) + {}^{2}k_\alpha^{\varphi}(1-\cos 2\varphi_\alpha) + {}^{3}k_\alpha^{\varphi}(1-\cos 3\,\varphi_\alpha) + \ldots$$

- Term is 2π-periodic and symmetric (minimum or maximum) at 0 and 2π
- In contrast to bonds and bond-angles, torsional dihedrals are *"soft"* degrees of freedom, *i.e.* the whole range [0;2π[is generally sampled (for this reason, Taylor series expansion does not make sense here – except if the dihedral angle is oscillating around some equilibrium value, *e.g.* at very low temperature)

Covalent terms: Torsional-dihedral rotation

C-C-C-C dihedral, CFF93 force field

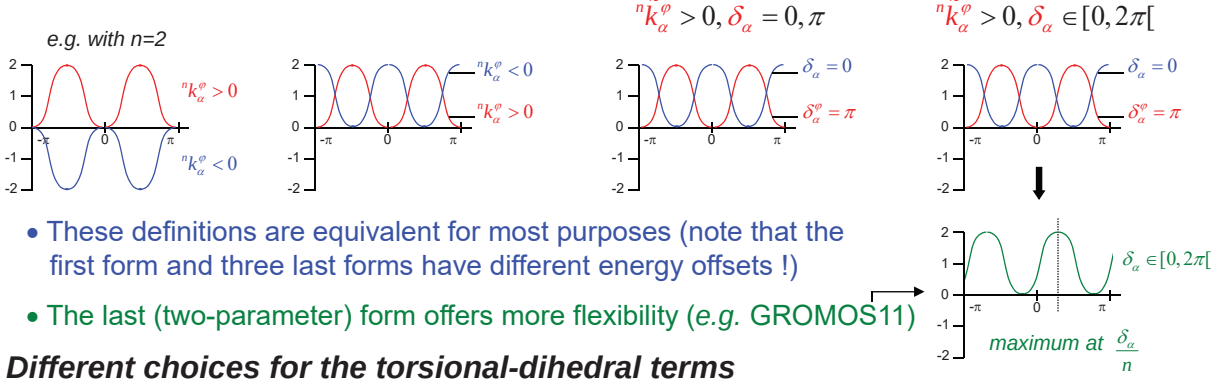


Covalent terms: Torsional-dihedral rotation

Definition variants for terms of multiplicity n:

$\cos(a-b) = \cos a\cos b + \sin a\sin b$

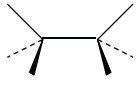
$^nk_\alpha^\varphi(1-\cos n\,\varphi_\alpha)$ | $^nk_\alpha^\varphi$ | $-{}^nk_\alpha^\varphi\cos n\,\varphi_\alpha$ | $^n\tilde{k}_\alpha^\varphi[1+\cos(\delta_\alpha)\cos(n\,\varphi_\alpha)]$ $^n\tilde{k}_\alpha^\varphi>0, \delta_\alpha = 0,\pi$ | $^n\tilde{k}_\alpha^\varphi[1+\cos(n\,\varphi_\alpha-\delta_\alpha)]$ $^n\tilde{k}_\alpha^\varphi>0, \delta_\alpha\in[0,2\pi[$



- These definitions are equivalent for most purposes (note that the first form and three last forms have different energy offsets !)
- The last (two-parameter) form offers more flexibility (*e.g.* GROMOS11)

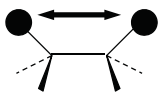
Different choices for the torsional-dihedral terms

- Terms retained
  - → all terms up to *n=3* (CFF93,MM3)
  - → generally one (evtl. 2) term with n≤6 (GROMOS,CHARMM)
- Dihedral angles retained
  - → all possible
  - → only one



9 possible dihedrals

- Third-neighbor non-bonded interactions
  - → excluded
  - → reduced (GROMOS) or scaled (OPLS)
  - → included



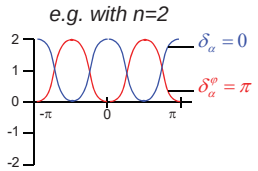
Torsional parameters are generally not transferable from one force field to another !

Covalent terms: Torsional-dihedral rotation

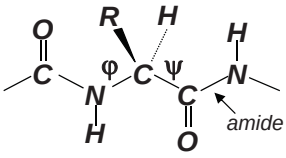
Example: multiplicities, phase shifts and force constants from the GROMOS96 force field

$$ {}^{n}\tilde{k}_{\alpha}^{\varphi}[1+\cos(\delta_{\alpha})\cos(n\,\varphi_{\alpha})] $$

$$ {}^{n}\tilde{k}_{\alpha}^{\varphi} > 0, \delta_{\alpha} = 0, \pi $$



Dihedral angle	*n*	${}^{n}\delta$ [rad]	${}^{n}\tilde{k}^{\theta}$ [kJ·mol^{-1}]
X-CH_n-CH_n-Y	3	0	5.86
X-CH_n-N(sp^3)-Y	3	0	3.77
X-CH_n-O(sp^3)-Y	3	0	1.26
X-C(sp^2)-N(sp^2)-Y [amide]	2	π	33.50
X- CH-N(sp^2)-Y [peptide φ]	6	π	1.00
X- CH-C(sp^2)-Y [peptide ψ]	6	0	1.00



- Barrier height is $2\,{}^{n}\tilde{k}^{\theta}$ (generally small - except planar amide group)
  → multiple rotamers are generally populated

Covalent terms: Torsional-dihedral rotation

Combination rules (if applied):

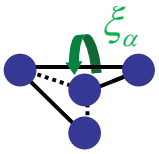
- Generally in the form of tables:

in some force fields: "any" (wildcard) → (atom type of *i*) atom type of *j* atom type of *k* (atom type of *l*) ➡ parameters for torsional-dihedral rotation *i-j-k-l*

Exclusions:

- Atoms separated by three covalent bonds (*third neighbors*) are either
  - excluded from any non-bonded interactions
  - interacting with scaled interactions (OPLS; different scaling factors for electrostatic and van der Waals interactions)
  - interacting with a special set or parameters (GROMOS96; normal electrostatics, special set of van der Waals parameters)
  - interacting through normal non-bonded interactions (usually a bad idea !)

Covalent terms: Geometrical distortion



out-of-plane
(or tetrahedron)
distortion

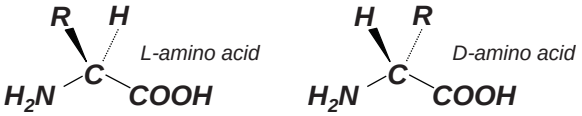
e.g. improper dihedral

$$^{(4)}V^{(\xi)}(\xi_\alpha;\ldots)$$

Usage

- The previous covalent terms are in principle sufficient to define the entire covalent contribution in a force field (*e.g.* CFF93 for alkanes)
- Need for out-of-plane (or out-of-tetrahedron) distortion terms:

→ For enforcing absolute chirality (unphysical but... convenient !)

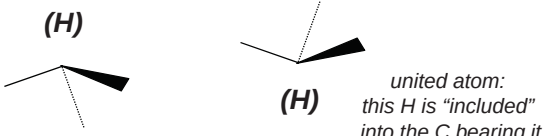


Internal coordinates defined in terms of scalar products cannot distinguish (→ requires correct initial coordinates and high barriers)

→ For enforcing geometry around planar or tetrahedral centers (*e.g.* carbonyl, aromatic rings, amine or sulfoxide groups)

Enforcing planarity (and, to a lesser extent, tetrahedral geometry) via bond angles requires excessively stiff bond-angle potentials

→ For enforcing geometry and chirality around CHR_3 united atoms



united atom: this H is "included" into the C bearing it

- The term describes how difficult it is to distort from the specified (planar or tetrahedral) geometry

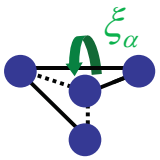
Covalent terms: Geometrical distortion

Functional form: harmonic

$$^{(4)}V^{(\xi)}(\xi_\alpha;\xi^o_\alpha,k^\xi_\alpha)=k^\xi_\alpha(\xi_\alpha-\xi^o_\alpha)^2$$

reference value *harmonic force constant*

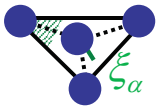
Geometrical definition



- Improper dihedral angle (six possible choices), *the most common one (e.g. GROMOS) !*

Recall: dihedral(i,j,k,l)=dihedral(l,k,j,i)

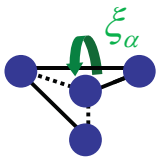
- Bond/plane angle (three possible choices)



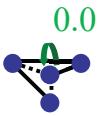
- Pyramid height (one possible choice)

Covalent terms: Geometrical distortion

Example: harmonic reference angles and force constants from the GROMOS96 force field



improper dihedral angle definition



planar group

tetrahedral group

Improper dihedral angle	ξ^o [deg]	k^ξ [kJ·mol^{-1}·deg^{-2}]
planar groups	0.0	0.051
tetrahedral centers	35.26	0.102
heme iron	0.0	0.204

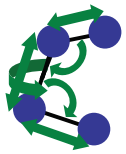
watch out: order of the atoms i-j-k-l (vs e.g. i-k-j-l) defines chirality !

Combination rules (if applied):

- Generally in the form of tables:

in some force fields: "any" (wildcard) → atom type of *i* / (atom type of *j*) / (atom type of *k*) / atom type of *l* → parameters for improper-dihedral distortion *i-j-k-l*

Covalent terms: Cross terms

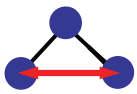


covalent cross-terms

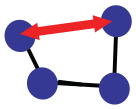
$$^{(3\,or\,4)}V^{(ct)}(\{b_\alpha\},\{\theta_\alpha\},\varphi_\alpha;\ldots)$$

Usage

- Improve the description of gas-phase properties (structures, heats of formation, vibrational frequencies) - compared to *ab initio* calculations and experiment
- Urey-Bradley term and third-neighbor non-bonded interactions implicitly encompass valence-coordinate cross terms (not very accurate)



Urey-Bradley: bond/bond-angle coupling



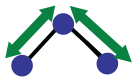
Third neighbor non-bonded: bond/bond-angle/torsional-dihedral coupling

- No such terms (GROMOS,CHARMM,AMBER), some terms (MM2,CVFF), or many terms (MM3,CFF93)
- Typically an unnecessary complication for condensed-phase simulations of macromolecules
  - increases the force-field complexity and number of parameters
  - prevents the use of bond-length constraints
  - not very relevant: other types of motion (solvent relaxation, conformational changes) have the largest impact on the thermodynamic and dynamic properties

Covalent terms: Cross terms

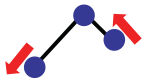
Example: cross-terms included in the CFF93 force field

- Bond/bond coupling



$$^{(3)}V^{(bb')}(b_\alpha, b_\alpha'; b_\alpha^o, b_\alpha^{o'}, k_\alpha^{bb'}) = k_\alpha^{bb'}(b_\alpha - b_\alpha^o)(b_\alpha' - b_\alpha^{o'})$$

$k_\alpha^{bb'} > 0$



- Bond/angle coupling

$$^{(3)}V^{(\theta b)}(\theta_\alpha, b_\alpha; \theta_\alpha^o, b_\alpha^o, k_\alpha^{\theta b}) = k_\alpha^{\theta b}(\theta_\alpha - \theta_\alpha^o)(b_\alpha - b_\alpha^o)$$

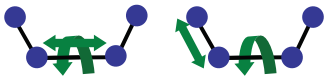
$k_\alpha^{\theta b} > 0$

- Angle/angle coupling

$$^{(4)}V^{(\theta\theta')}(\theta_\alpha, \theta_\alpha'; \theta_\alpha^o, \theta_\alpha^{o'}, k_\alpha^{\theta\theta'}) = k_\alpha^{\theta\theta'}(\theta_\alpha - \theta_\alpha^o)(\theta_\alpha' - \theta_\alpha^{o'})$$

not easily interpreted

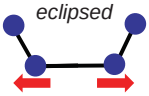
- Dihedral/bond coupling



for the central bond:

$$^{(4)}V^{(\phi b)}(\phi_\alpha, b_\alpha; b_\alpha^o, {}^1k_\alpha^{\phi b}, {}^2k_\alpha^{\phi b}, \ldots) = (b_\alpha - b_\alpha^o)({}^1k_\alpha^{\phi b}\cos\phi_\alpha + {}^2k_\alpha^{\phi b}\cos 2\phi_\alpha + \ldots)$$

$^1k_\alpha^{\phi b} < 0$ *eclipsed*



- Dihedral/angle coupling

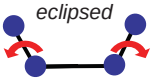
$$^{(4)}V^{(\phi\theta)}(\phi_\alpha, \theta_\alpha; \theta_\alpha^o, {}^1k_\alpha^{\phi\theta}, {}^2k_\alpha^{\phi\theta}, \ldots) = (b_\alpha - b_\alpha^o)({}^1k_\alpha^{\phi\theta}\cos\phi_\alpha + {}^2k_\alpha^{\phi\theta}\cos 2\phi_\alpha + \ldots)$$

not easily interpreted

- Dihedral/angle/angle coupling

$$^{(4)}V^{(\phi\theta\theta')}(\phi_\alpha, \theta_\alpha, \theta_\alpha'; \theta_\alpha^o, \theta_\alpha^{o'}, k_\alpha^{\phi\theta\theta'}) = k_\alpha^{\phi\theta\theta'}(\theta_\alpha - \theta_\alpha^o)(\theta_\alpha' - \theta_\alpha^{o'})\cos\phi_\alpha$$

$k_\alpha^{\phi\theta\theta'} < 0$ *eclipsed*



Biomolecular vs spectrocopic force fields

Class I (biomolecular): e.g. AMBER

similar for GROMOS, CHARMM, OPLS, …

$$E_{total} = \sum_{bonds} K_R(R - R_o)^2 + \sum_{angles} K_\theta(\theta - \theta_o)^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right] + \sum_{H\text{-}bonds}\left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}}\right]$$

+ Simple energy expression

+ Few parameters to calibrate

+ Focus on ***intermolecular*** interactions **(+torsions in molecules)**
(critical to reproduce condensed-phase properties)

+ Parameters are more intuitive and transferable

- Moderately accurate (if not just bad) gas-phase properties and molecular geometries (incorrect vibrational frequencies and heats of formation !)

- Many atom types

Class II (spectroscopic): e.g. CFF93

$$E = \sum_b [{}^2K_b(b - b_0)^2 + {}^3K_b(b - b_0)^3 + {}^4K_b(b - b_0)^4] + \sum_\theta [{}^2K_\theta(\theta - \theta_0)^2 + {}^3K_\theta(\theta - \theta_0)^3 + {}^4K_\theta(\theta - \theta_0)^4] + \sum_\phi [{}^1K_\phi(1 - \cos\phi) + {}^2K_\phi(1 - \cos 2\phi) + {}^3K_\phi(1 - \cos 3\phi)] + \sum_b K_s\chi^2 + \sum_{i>j}\frac{q_i q_j}{r_{ij}} + \sum_{i>j}\epsilon\left[2\left(\frac{r^*}{r_{ij}}\right)^9 - 3\left(\frac{r^*}{r_{ij}}\right)^6\right] + \sum_b\sum_{b'} K_{bb'}(b - b_0)(b' - b_0') + \sum_\theta\sum_{\theta'} K_{\theta\theta'}(\theta - \theta_0)\times(\theta' - \theta_0') + \sum_b\sum_\theta K_{b\theta}(b - b_0)(\theta - \theta_0) + \sum_\phi\sum_b (b - b_0)[{}^1K_{b\phi}\cos\phi + {}^2K_{b\phi}\cos 2\phi + {}^3K_{b\phi}\cos 3\phi] + \sum_\phi\sum_{b'} (b' - b_0')[{}^1K_{b'\phi}\cos\phi + {}^2K_{b'\phi}\cos 2\phi + {}^3K_{b'\phi}\cos 3\phi] + \sum_\phi\sum_\theta (\theta - \theta_0)[{}^1K_{\theta\phi}\cos\phi + {}^2K_{\theta\phi}\cos 2\phi + {}^3K_{\theta\phi}\cos 3\phi] + \sum_\phi\sum_\theta\sum_{\theta'} K_{\theta\theta'\phi}(\theta - \theta_0)(\theta' - \theta_0')\cos\phi$$

+ Accurate gas-phase properties (geometries, vibrational frequencies, heats of formation)
→ cheap alternative to quantum chemistry

+ Use of quantum-mechanical data in the calibration

+ Few atom types

+ Elegant

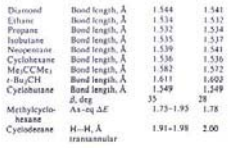
- Complicated energy expression

- Many parameters to calibrate – often needs full recalibration for adding new atom types

- Focus on ***intramolecular*** interactions
(of little help to reproduce condensed-phase properties)

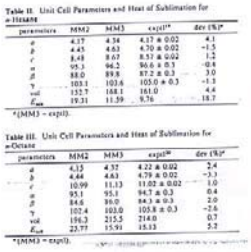
Example of spectrocopic force field: MM3 for alkanes

Geometries



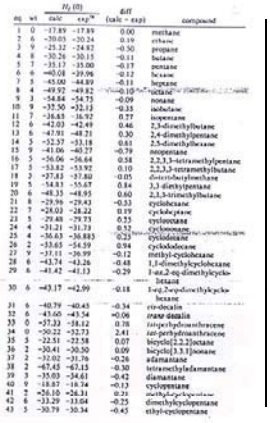
Heats of sublimation

(only data on intermolecular interactions - less accurate)

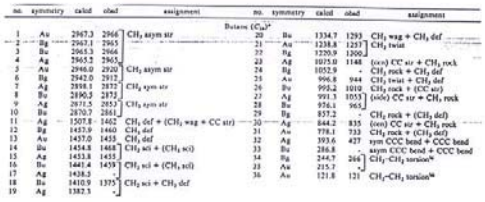


Heats of formation

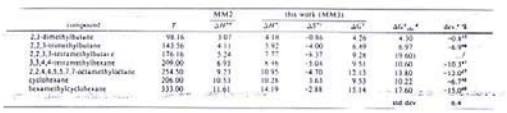
(kcal/mol)



Vibration frequencies



Torsional barriers



Interesting student question (HS16)

- When doing Ex1, HS16 students noticed that for the organic molecule they considered, the energy in vacuum (isolated molecule) is ***positive***; the same is true for the intramolecular energy in the liquid, actually

  → Now, in QM, the (intramolecular) energy of a molecule is always ***negative***

  → And this **makes sense**, otherwise the molecule would ***fall apart*** !

  → So: **what (the hell) is wrong with GROMOS** ?

- Someone has an answer to that ???

- Explanation:

  → The energy of a molecule within GROMOS is **not equal** to the QM energy of the molecule relative to the isolated atoms !

  → It only accounts for the **variations of the QM energy** upon changing the geometry of the molecule (at least approximately)

  Approximately: because in the condensed phase we don't care too much about intramolecular interactions – the intermolecular ones are much more important!

  → It is generally **positive** because the force field terms add positive contributions for *deviations from ideal covalent geometries* and *steric repulsions*

  → This also implies that the GROMOS energy differences for **chemical changes** are generally ***entirely wrong***; only **physical changes** can be handled !

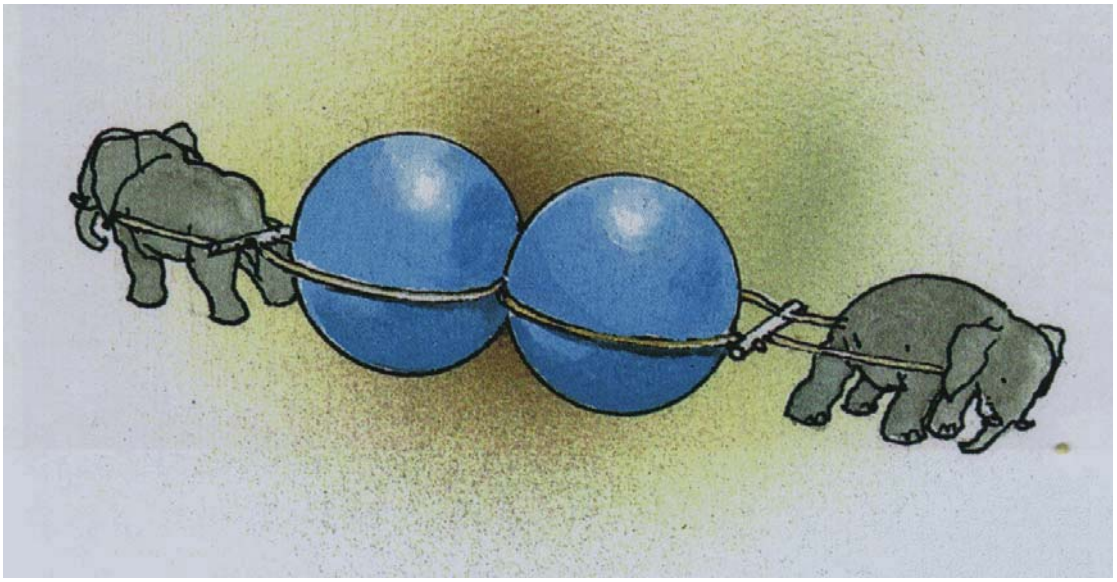
simple example: water

not OK with GROMOS | OK with GROMOS

$$H_{2,g} + \frac{1}{2}O_{2,g} \rightarrow H_2O_g \rightarrow H_2O_l$$

We would find zero !!!
(first and second covalent neighbors excluded)

Non-bonded force-field terms



Two elephants kindly accepted to illustrate how non-bonded forces can be measured accurately in a simple experiment

Non-bonded terms: Electrostatic interactions

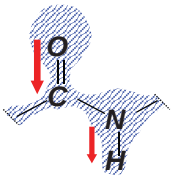
electrostatic
(all atom pairs)

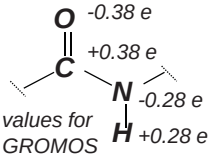
$$^{(2)}V^{(el)}(r_\alpha; q_\alpha, q'_\alpha, \ldots)$$

*usually **not** simply Coulomb's law (e.g. cutoff+correction, periodicity)*

Most common approximation: monopole (point partial charge) model

electron density
→ higher close to more electronegative atoms
→ dipoles along bonds





values for GROMOS

atomic partial charges are attributed to all (solute and solvent) atoms

- Some force fields use *off-atom partial charges* (*e.g.* middle of bonds, lone pairs, σ-hole)
- Some force fields include *higher-order point multiploles* (*e.g.* point dipoles, quadrupoles)
- Some force fields account for *electronic nduction / polarizability / charge transfer* (*e.g.* movable charge sites)

Functional form: Coulomb's law (in a perfect world !)

$$^{(2)}V^{(el)}(r_\alpha; q_\alpha, q'_\alpha) = (4\pi\varepsilon_o)^{-1} q_\alpha q'_\alpha r_\alpha^{-1}$$

partial charges

permittivity of vacuum (a physical constant)

ε_o

not to be confused with the relative permittivity ε of a medium (e.g. 80 for water)

- For a number of reasons (*discussed later*) this simple form is seldom applied in practice (→ cutoff, cutoff + long-range correction, lattice-sum, ...)

Non-bonded terms: Electrostatic interactions

Combination rules:

- Often (not always, *e.g.* GROMOS), the charge is determined by the atom type

In GROMOS, we use building blocks, and specific charges are assigned to all atoms individually (still, in practice, similar charges are used for similar functional groups – see next slide)

- Coulomb's law implicitly involves a (physically-based) combination rule

$$q_\alpha, q'_\alpha = q_\alpha \times q'_\alpha$$

- Bond increment method (used in some force fields; preserves electroneutrality)

$$q_\alpha = \sum_{\text{first neighbors } \beta}^{\leq 4} \delta(type\ \alpha, type\ \beta)$$

(δ: bond increment)

In bond-increment force-fields, one defines the δ-values for different bond types rather than charges for different atom types

Note: for a neutral non-cyclic molecule, there is a unique mapping between a set of atom charges and a set of bond dipoles (otherwise, the mapping is not unique)

Exclusions:

- First and second covalent neighbors are excluded from electrostatic interactions
- Third covalent neighbors may interact with full (*e.g.* GROMOS) or scaled (*e.g.* OPLS) electrostatic interactions

Will be discussed in more details in a separate lecture: treatment of long-range interactions & accounting for polarization (be patient !...)

Non-bonded terms: Electrostatic interactions

A GROMOS-Compatible Force Field for Small Organic Molecules in the Condensed Phase: The 2016H66 Parameter Set

Bruno A. C. Horta,*[,†,‡] Pascal T. Merz,[†] Patrick F. J. Fuchs,[‖] Jozica Dolenc,[†,⊥] Sereina Riniker,[†] and Philippe H. Hünenberger[†]

J. Chem. Theory Comput. 2016, 12, 3825−3850

Charges of functional groups in the GROMOS-compatible 2016H66 force field

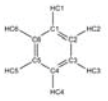
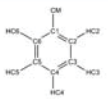
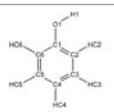
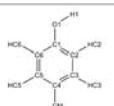
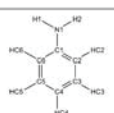
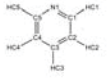
Calibrated along with the vdW parameters against pure-liquid and solvation properties for 57 organic compounds

- ρ_{liq} ***liquid density***
- ΔH_{vap} ***enthalpy of vaporization***
- ΔG_{wat} ***hydration free energy***
- ΔG_{che} ***solvation free energy in cyclohexane***

Charge signs and magnitudes follow (roughly) chemical intuition

Table 3. Atom-Type Assignments and Atomic Partial Charges of the Different Organic Chemical Functions in 2016H66[a]

group	IAC	type	q (e)
alcohol	21	H	0.410
	3	OA	−0.700
	13−18	CHn	0.290
ether	4	OE	−0.580
	13−18	CHn	0.290
aldehyde	12	C	0.375
	20	HC	0.100
	1	O	−0.475
ketone	1	O	−0.540
	12	C	0.540
carboxylic acid	12	C	0.550
	1	O	−0.550
	3	OA	−0.410
	21	H	0.410
ester	13−18	CHn	0.290
	4	OE	−0.370
	1	O	−0.550
	12	C	0.630
primary amine	13−18	CHn	0.250
	7	N	−0.980
	21	H	0.365
secondary amine	13−18	CHn	0.250
	7	N	−0.865
	21	H	0.365
tertiary amine	13−18	CHn	0.250
	7	N	−0.750
primary amide	12	C	0.540
	1	O	−0.540
	6	N	−0.760
	21	H	0.380
secondary amide	12	C	0.540
	1	O	−0.540
	6	N	−0.500
	21	H	0.310
	13−18	CHn	0.190
tertiary amide	12	C	0.540
	1	O	−0.540
	6	N	−0.380
	13−18	CHn	0.190
thiol	13−18	CHn	0.150
	23	S	−0.370
	21	H	0.220
sulfide	13−18	CHn	0.215
	23	S	−0.430
disulfide	13−18	CHn	0.150
	23	S	−0.150

group (compound)	labeling scheme	atom label	IAC	type	q (e)
benzene (BZN)		HC1-6	20	HC	0.1298
		C1-6	12	C	−0.1298
toluene (TOL)		C1	12	C	−0.1
		CM	16	CH3	0.1
		C2-6	12	C	−0.1298
		HC2-6	20	HC	0.1298
phenol (PHL)		C1	12	C	0.29
		O1	1	O	−0.7
		H1	21	H	0.41
		C2-6	12	C	−0.1298
		HC2-6	20	HC	0.1298
p-Cresol (PHT)		C1	12	C	0.29
		O1	1	O	−0.7
		H1	21	H	0.41
		C4	12	C	−0.1
		CM	16	CH3	0.1
		C2,3,5,6	12	C	−0.1298
		HC2,3,5,6	20	HC	0.1298
aniline (ANLN)		C1	12	C	0.25
		N1	7	NT	−0.98
		H1,2	21	H	0.365
		C2-6	12	C	−0.1298
		HC2-6	20	HC	0.1298
pyridine (PYRI)		N1	9	NR	−0.5
		C1,5	12	C	0.1202
		C2-4	12	C	−0.1298
		HC1-5	20	HC	0.1298

[a]For each functional group (group) and for each atom within the group (or set of equivalent atoms for the aromatic compounds), the integer atom code (IAC), atom-type name (type), and atomic partial charge (q) are reported. This table includes only the chemical functions relevant for the 57 organic compounds of Table 1 and is repeated in SI A (Table A8). The corresponding information for the nucleic-acid bases can be found in Table 4. A complete topology specification (including charge-group definitions) for the 57 organic molecules in 2016H66 can be found in the mtb files.[134]

Non-bonded terms: van der Waals interactions

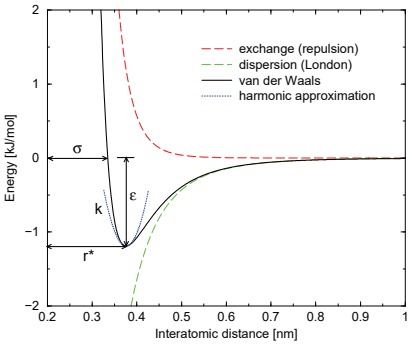
van der Waals
(all atom pairs)

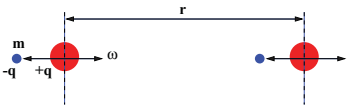
$^{(2)}V^{(vdW)}(r_\alpha;\ldots)$

Physical nature:

- Rare gas atoms have zero net charge → no "classical" electrostatic interactions
  Still, they do interact - due to so-called *van der Waals interactions*:

  - deviations from ideal-gas behavior (*e.g.* van der Waals or virial real-gas equations of state)
  - existence of liquid and solid phases (which would not be there without such interactions !)





QM treatment (with elec. correl.), 3D
(2 fixed nuclei with oscillating electrons)

Very-short-range repulsion (*exchange or Pauli term*)
→ consequence of the Pauli exclusion principle
for closed-shell atoms

→ very-short distance $\sim r^{-1}$ *repulsion between nuclei (electrons are pushed away due to Pauli exclusion)*

→ short-distance $\sim e^{-2r/a_o}$ (a_o=Bohr radius[=0.053nm for H])

Short-range attraction (*dispersion or London term*)
→ consequence of electron correlation (instantaneous dipole - induced-dipole interactions)

instantaneous fluctuation → dipole

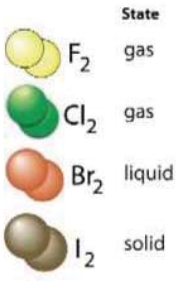
induced fluctuation → dipole

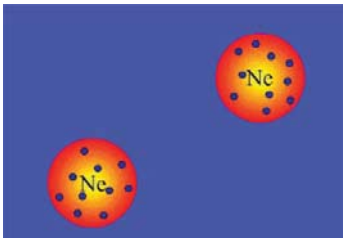
→ Drude model $E = C_6 r^{-6} + C_8 r^{-8} + C_{10} r^{-10} + \ldots \quad (C_n < 0)$

related to the atomic polarizability,
accounts for ~75% of total dispersion for liquid Ar

Dispersion interactions (London forces)

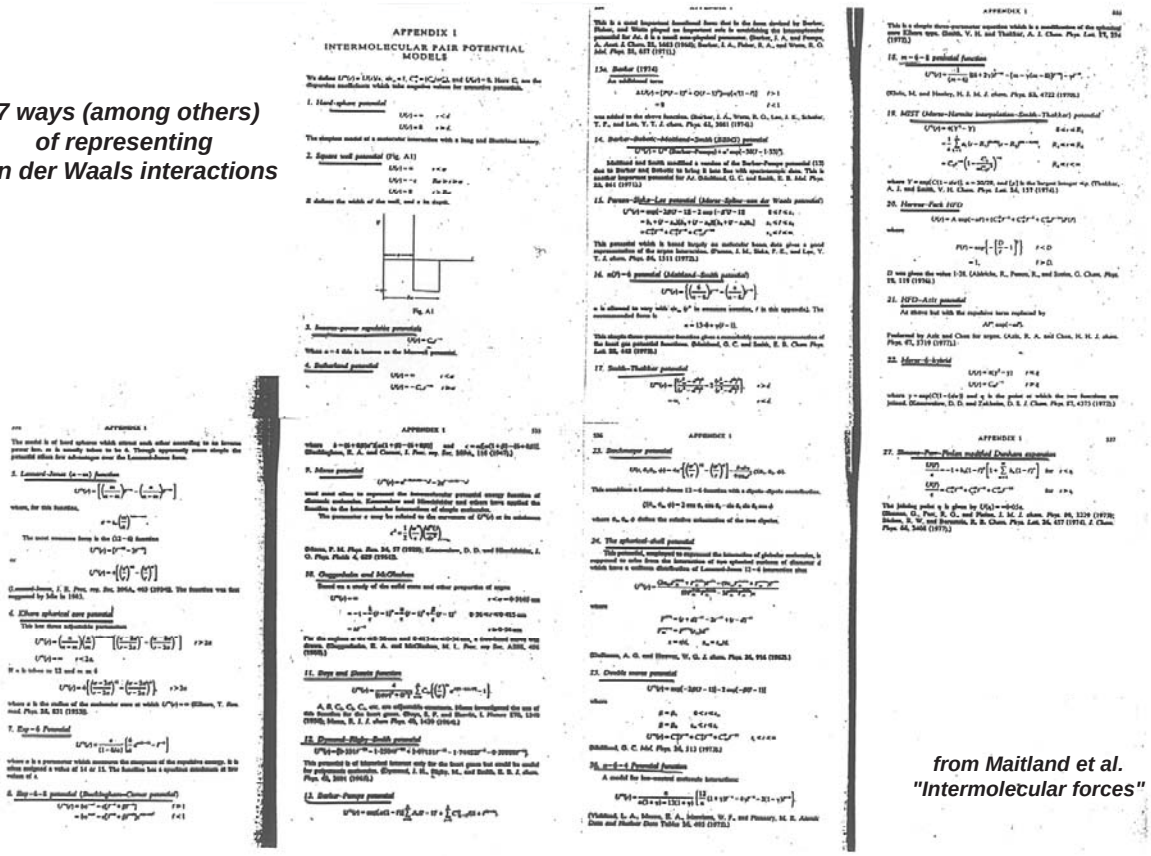
Noble gas	molar mass (g)	Boiling point (K)	Halogen	molar mass (g)	Boiling Point (K)
He	4.0	4.6	F_2	38.0	85.1
Ne	20.2	27.3	Cl_2	71.0	238.6
Ar	39.9	87.5	Br_2	159.8	332.0
Kr	83.8	120.9	I_2	253.8	457.6





Non-bonded terms: van der Waals interactions

27 ways (among others) of representing van der Waals interactions



from Maitland et al. "Intermolecular forces"

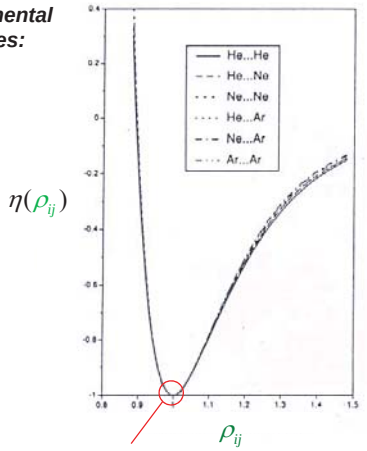
Non-bonded terms: van der Waals interactions

Reduced form:

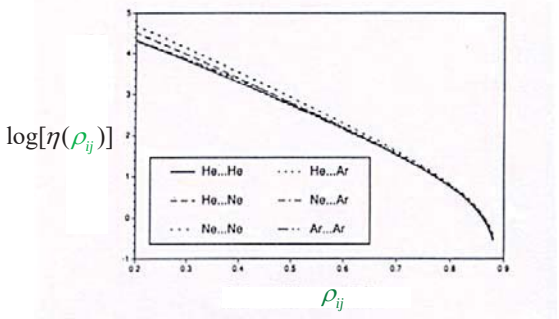
- Rare gases are the simplest training ground for van der Waals interactions (molecular beam experiments, deviations from ideality, transport properties)
- For all rare gases, van der Waals interactions follow a single *reduced form* (to a good approximation)

$$^{(2)}V_{ij}^{(vdW)}(r_{ij}) = \varepsilon(i,j)\eta(\rho_{ij}) \qquad \text{with} \qquad \rho_{ij} = \frac{r_{ij}}{r^*(i,j)}$$

experimental curves:



reference point (minimum)

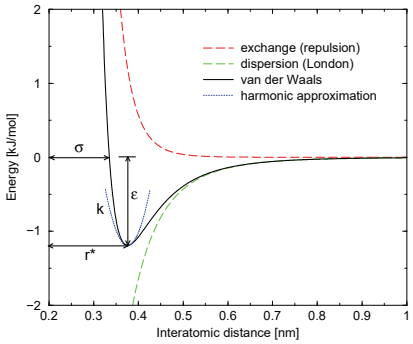


e.g.	ε [kJ/mol]	r* [nm]
He-He	0.0910	0.2963
He-Ne	0.1737	0.3036
Ne-Ne	0.3513	0.3087

Non-bonded terms: van der Waals interactions

Characteristics of the reduced form:

(non-reduced form)



$\eta(1) = -1$ *value at the minimum*

$\left.\frac{d\eta(\rho)}{d\rho}\right|_{\rho=1} = 0$ *derivative at the minimum*

$\lim_{\rho\to\infty} \eta(\rho) = 0$ *interaction is short-ranged (no interaction at infinite distance)*

$\lim_{\rho\to 0} \eta(\rho) = +\infty$ *atom overlap is forbidden*

- Collision diameter $\sigma(i,j)$

$$\eta(\xi) = 0 \quad \text{with} \quad \xi = \frac{\sigma(i,j)}{r^*(i,j)}$$

- Equivalent harmonic force constant at the minimum $k(i,j)$

$$\kappa(\xi) = \left.\frac{d^2\eta(\rho)}{d\rho^2}\right|_{\rho=1} \quad \text{with} \quad \kappa = \varepsilon^{-1}(i,j)k(i,j)[r^*(i,j)]^2$$

so that
$E = \frac{1}{2}k(i,j)[r_{ij} - r^*(i,j)]^2$
around the minimum

Non-bonded terms: van der Waals interactions

Form of the van der Waals term:

$$^{(2)}V^{(vdW)}(r_\alpha; r_\alpha^*, \varepsilon_\alpha, \ldots) = \varepsilon_\alpha \eta\left(\frac{r_\alpha}{r_\alpha^*}, \ldots\right)$$

Commonly-employed reduced form:

- *n-m* van der Waals function

$$\eta(\rho) = \frac{1}{n-m}\left(m\rho^{-n} - n\rho^{-m}\right) \qquad \xi = (m/n)^{1/(n-m)} \qquad \kappa = nm$$

Commonly used with
- *m*=6, *n*=9 : 9-6 van der Waals function (CVFF,CFF93)
- *m*=6, *n*=12 : Lennard-Jones function (GROMOS,AMBER,CHARMM; computationally advantageous [$r^{-12}=(r^{-6})^2$], *ad hoc*, may be too steep)

→ *m*=6 is correct for the leading term of dispersion

- exp-*m* van der Waals function

$$\eta(\rho;\varsigma) = \frac{1}{\varsigma - m}\left(me^{\varsigma(1-\rho)} - \varsigma\rho^{-m}\right)$$

Commonly used with m=6, ζ=pairwise adjustable parameter (Buckingham)
→ deviations from the common reduced form
→ *exp* repulsion is more realistic than power law at short distances
→ computationally expensive (*exp* function)
→ three parameters per pair to calibrate

but
$\lim_{\rho\to 0} \eta(\rho) = -\infty$
may lead to "fusion" of atoms

Non-bonded terms: van der Waals interactions

Commonly-employed reduced form (continued):

- Double-Morse van der Waals function

$$\eta(\rho;a) = e^{2a(1-\rho)} - 2e^{a(1-\rho)}$$

→ improper description of dispersion
→ computationally expensive (*exp* function)
→ three parameters per pair to calibrate

- *n-m* buffered van der Waals function

$$\eta(\rho;\gamma,\delta) = \frac{n-m}{m}\left(\frac{1+\delta}{\rho+\delta}\right)\left(\frac{1+\gamma}{\rho^m+\gamma} - \frac{m}{n-m} - 1\right)$$

→ calibrated for rare gases (and quite accurate for these !) with *n*=15 and *m*=7 (latter formally incorrect for dispersion !), leading to γ=0.12, δ=0.07

Choice of a specific function:

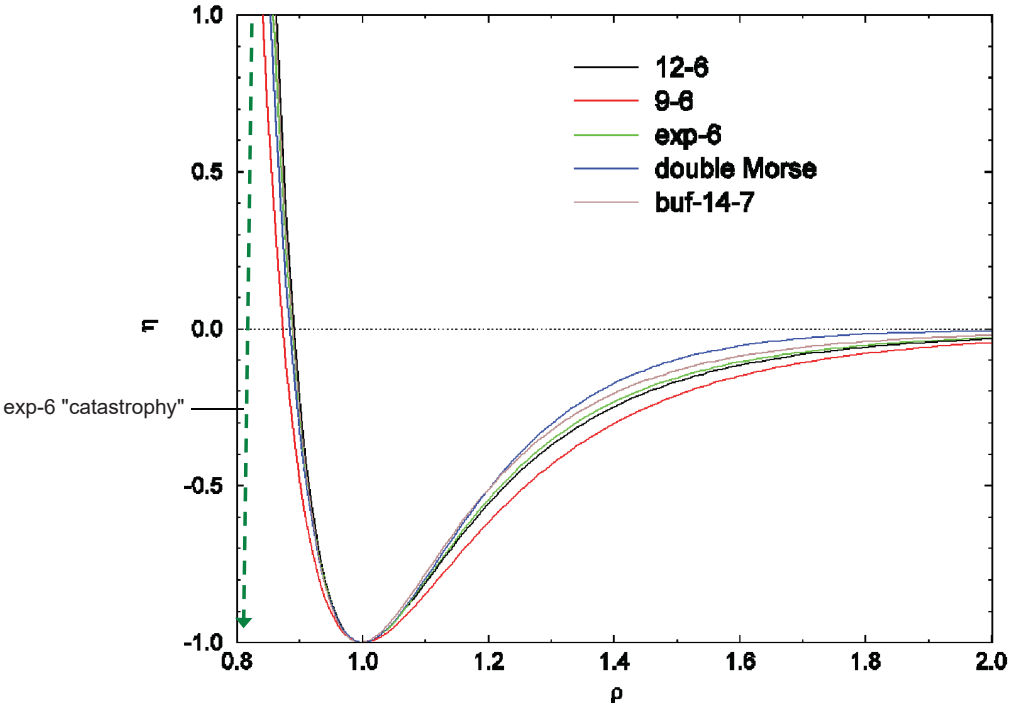
- *A priori* important because
  - the ***steepness*** of the repulsion term (below r*) determines (in balance with electrostatic interactions) the **density and compressibility** of condensed-phase systems
  - the ***dispersion*** term (above r*) makes a significant contribution to the total **potential energy** (*e.g.* comparison with experimental heats of vaporization) and **pressure** (*e.g.* comparison with experimental densities)
- Finally not so important because
  - what is ***accurate for rare gases*** may no longer be for *e.g.* water-water, water-aliphatic and aliphatic-aliphatic interactions
  - interaction parameters are ***empirical*** (adjustable) parameters

i.e. (united-)atoms in molecules are not the same as closed-shell rare gases !!!

calibration against experimental condensed-phase properties compensates for errors in the functional form

Non-bonded terms: van der Waals interactions

Graphical representation of the common reduced forms:



(three-parameter functions: curvature at the minimum chosen identical to the 12-6 function)

Non-bonded terms: van der Waals interactions

Most common choice: Lennard-Jones function

$$\eta(\rho) = \rho^{-12} - 2\rho^{-6} \qquad \xi = (1/2)^{1/6} \qquad \kappa = 72$$

- Leading to

$$^{(2)}V^{(vdW)}(r_\alpha; r_\alpha^*, \varepsilon_\alpha) = \varepsilon_\alpha \left[\left(\frac{r_\alpha}{r_\alpha^*} \right)^{-12} - 2 \left(\frac{r_\alpha}{r_\alpha^*} \right)^{-6} \right]$$

- Also sometimes written

$$^{(2)}V^{(vdW)}(r_\alpha; \sigma_\alpha, \varepsilon_\alpha) = 4\varepsilon_\alpha \left[\left(\frac{r_\alpha}{\sigma_\alpha} \right)^{-12} - \left(\frac{r_\alpha}{\sigma_\alpha} \right)^{-6} \right]$$

- Or

$$^{(2)}V^{(vdW)}(r_\alpha; C_{12,\alpha}, C_{6,\alpha}) = C_{12,\alpha} r_\alpha^{-12} - C_{6,\alpha} r_\alpha^{-6}$$

$$\begin{cases} r^* = \left(\dfrac{2C_{12}}{C_6} \right)^{1/6} \\ \sigma = \left(\dfrac{C_{12}}{C_6} \right)^{1/6} \end{cases} \qquad \varepsilon = \frac{C_6^2}{4C_{12}}$$

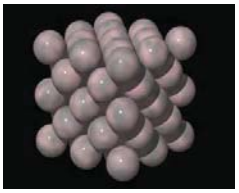
Exclusions:

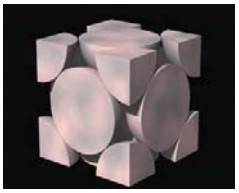
- First and second covalent neighbors are excluded from van der Waals interactions
- Third covalent neighbors may interact with full, reduced (special set; *e.g.* GROMOS) or scaled (*e.g.* OPLS) van der Waals interactions

Non-bonded terms: van der Waals interactions

- We can also play with crystals

→ Rare gases crystallize in the face-centered cubic lattice





NOTE:
I had fun playing with this when preparing Infol Ex3; but all this was already done in the 1940's !

- Assuming a Lennard-Jones potential $V(r;\sigma,\varepsilon) = 4\varepsilon \left[\left(\frac{r}{\sigma} \right)^{-12} - \left(\frac{r}{\sigma} \right)^{-6} \right]$

→ The lattice energy is *given a lattice spacing a: contact distance*

$$E_l(a) = 2\varepsilon \left[A_{12} \left(\frac{a}{\sigma} \right)^{-12} - A_6 \left(\frac{a}{\sigma} \right)^{-6} \right] \qquad A_{12} \approx 12.13 \qquad A_6 \approx 14.45$$

evaluated numerically by lattice-summation for a FCC lattice

→ Minimizing this energy, one finds

$$a = \left(\frac{2A_{12}}{A_6} \right)^{1/6} \sigma \approx 1.09\,\sigma \quad \Rightarrow \quad E = -\frac{A_6^2}{2A_{12}} \varepsilon \approx -8.61\,\varepsilon$$

→ One-to-one relationships

$\rho \leftrightarrow \sigma \qquad E \leftrightarrow \varepsilon$

related to the density → 4 atoms per unit cell → unit-cell volume $(a/2^{1/2})^3$

$$\rho = \frac{4 \cdot 2^{3/2} \cdot M}{a^3} = \left(\frac{A_{12}}{A_6} \right)^{-1/2} M\sigma^{-3} \approx 1.09 M\,\sigma^{-3}$$

M: molar mass

→ E.g. Argon *M = 39.948 [g/mol]*

$\sigma = 0.340$ *[nm]* $\qquad \varepsilon = 1.003$ *[kJ/mol]*

from gas-phase virial coefficients ⇒

crystal: $\rho = 1842$ *[kg/m³]* $\qquad E = -8.6$ *[kJ/mol]*

exp. $\rho = 1764$ *[kg/m³]* *(20K)* $\qquad E = -7.7$ *[kJ/mol]*

Non-bonded terms: van der Waals interactions

Combination rules:

- Necessary to reduce the number of interaction parameters

N atom types ➡ $2 \times [\tfrac{1}{2} \times N(N+1)]$ pair parameters

ε, r^*

- For term α involving atoms of type *i* and *j*

$$\varepsilon_\alpha = \varepsilon(i,j) = f[\varepsilon(i,i), \varepsilon(j,j)]$$

$$r^*_\alpha = r^*(i,j) = f[r^*(i,i), r^*(j,j)]$$

- Testing: rare gases (mixed systems)
- Combination rules belong to the definition of a force field !

Non-bonded terms: van der Waals interactions

Common combination rules:

- Geometric mean for r* and ε (GROMOS,OPLS)

For the Lennard-Jones potential: easily shown that this is equivalent to a geometric mean for C_6 and C_{12}

$$r^*(i,j) = [r^*(i,i) r^*(j,j)]^{1/2} \qquad \varepsilon(i,j) = [\varepsilon(i,i)\varepsilon(j,j)]^{1/2}$$

- Arithmetic mean for r* and geometric mean for ε (Lorentz-Berthelot; CHARMM,AMBER)

$$r^*(i,j) = [r^*(i,i) + r^*(j,j)]/2 \qquad \varepsilon(i,j) = [\varepsilon(i,i)\varepsilon(j,j)]^{1/2}$$

- Arithmetic mean for $(r^*)^6$ and geometric mean for $\varepsilon\,(r^*)^6$

$$r^*(i,j) = \left\{[(r^*(i,i))^6 + (r^*(j,j))^6]/2\right\}^{1/6}$$

$$\varepsilon(i,j) = \left\{[\varepsilon(i,i)(r^*(i,i))^6 \varepsilon(j,j)(r^*(j,j))^6]^{1/2}\right\}/(r^*(i,j))^6$$

- Cubic mean for r* and HHG mean (harmonic mean of harmonic and geometric means) for ε

$$r^*(i,j) = [(r^*(i,i))^3 + (r^*(j,j))^3]/[(r^*(i,i))^2 + (r^*(j,j))^2]$$

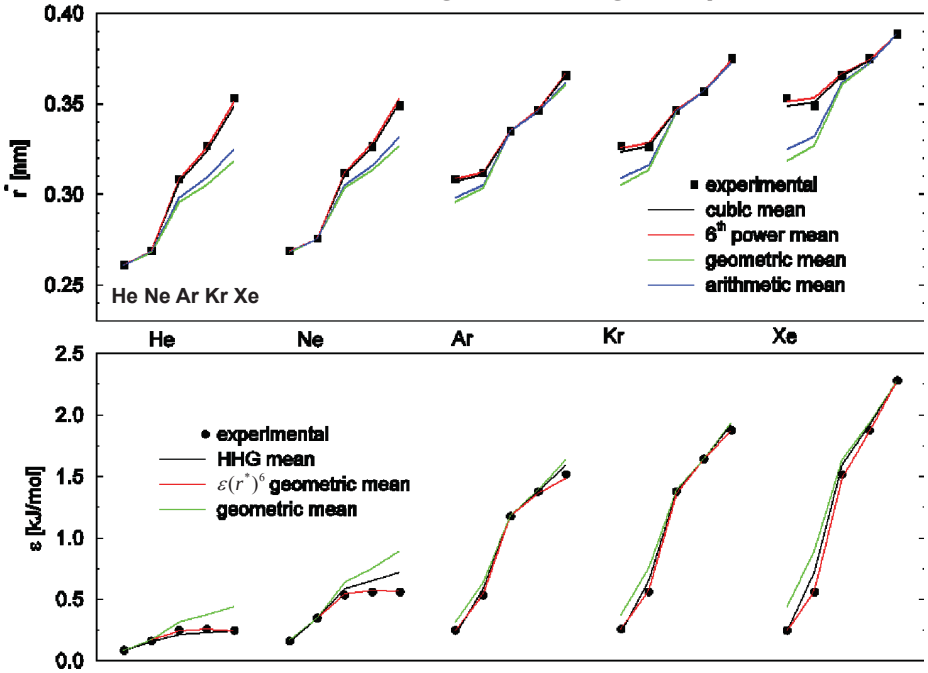
$$\varepsilon(i,j) = 4\varepsilon(i,i)\varepsilon(j,j)/[(\varepsilon(i,i))^{1/2} + (\varepsilon(j,j))^{1/2}]^2$$

Choice of a specific rule:

- *A priori* important (differences in accuracy for rare-gas mixtures)
- Finally not so important because
  - what is accurate for rare gases may no longer be for (united-)atoms in molecules
  - interaction parameters are adjustable (empirical) parameters
- But: choice must be done consistently within a force field!

Non-bonded terms: van der Waals interactions

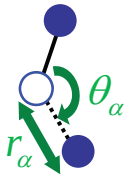
Combination rules tested against rare-gas experimental data:



- Large differences for largely different atom sizes / polarizabilities
- What is more accurate for rare gases need not be more accurate for everything (atoms in molecules are not spherical, there are other [e.g. electrostatic] non-bonded interactions, and we use of effective [calibrated] parameters anyway) !

★

Non-bonded terms: Hydrogen-bonding interactions

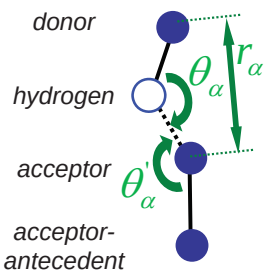


H-bond $\quad {}^{(3)}V^{(hb)}(r_\alpha, \theta_\alpha; \ldots)$

Needed because:

- Normal van der Waals interaction parameters + strong electrostatic interaction → too short H-bond distances, too strong H-bonds

One possibility: special H-bonding force-field term, e.g. CHARMM



or zero if $\theta<\pi/2$ or $\theta'<\pi/2$ or $r_\alpha>r_{cut}$

$$ {}^{(4)}V^{(hb)}(r_\alpha, \theta_\alpha, \theta'_\alpha; C_{12,\alpha}, C_{6,\alpha}, m_\alpha, n_\alpha) = $$

$$ (C_{12,\alpha} r_\alpha^{-12} - C_{6,\alpha} r_\alpha^{-6}) \cos^{m_\alpha} \theta_\alpha \cos^{n_\alpha} \theta'_\alpha $$

- With m_α=0,2,4 (depending on donor), n_α=0,2 (depending on acceptor)
- Requires some bookkeeping (list of H-bonded atoms) during the simulation

Non-bonded terms: Hydrogen-bonding interactions

Another possibility: special van der Waals parameters, e.g. GROMOS

$$^{(2)}V^{(vdW)}(r_\alpha; C_{12,\alpha}, C_{6,\alpha}) = C_{12,\alpha} r_\alpha^{-12} - C_{6,\alpha} r_\alpha^{-6}$$

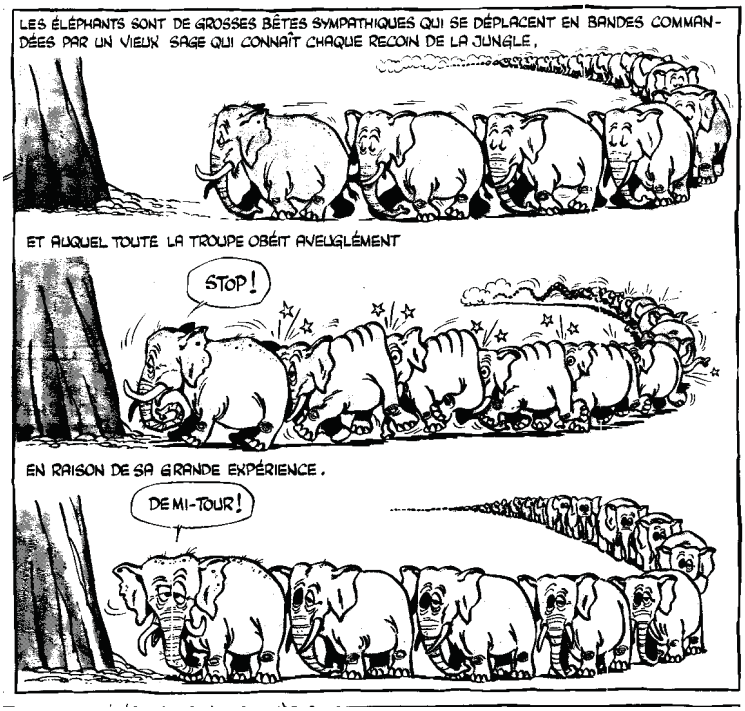
- Geometric mean rules (for ε and r* or, equivalently, for C_6 and C_{12})

$$C_{12}(i,j) = [C_{12}(i,i) C_{12}(j,j)]^{1/2} \qquad C_6(i,j) = [C_6(i,i) C_6(j,j)]^{1/2}$$

Three choices:
- non H-bonding pair (small, *e.g.* aliphatic-aliphatic)
- neutral H-bonding pair (larger, *e.g.* amine-water)
- charged H-bonding pair (even larger, *e.g.* ammonium-water)

- H-bonding determined by a balance between electrostatic and van der Waals interactions
- No bookkeeping of H-bonds
- Atom in non H-bonding orientations or distances still interact with modified parameters

Non-bonded terms: Many-body interactions

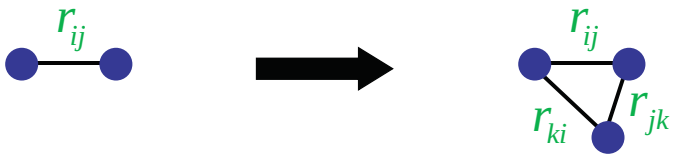


Many elephants kindly accepted to illustrate the concept of many-elephant interactions (involving more than two elephants)

Non-bonded terms: Many-body interactions

Physical origin:

- All above-listed non-bonded terms are pairwise additive. In reality, the presence of a third (fourth,...) atom affects the interaction...



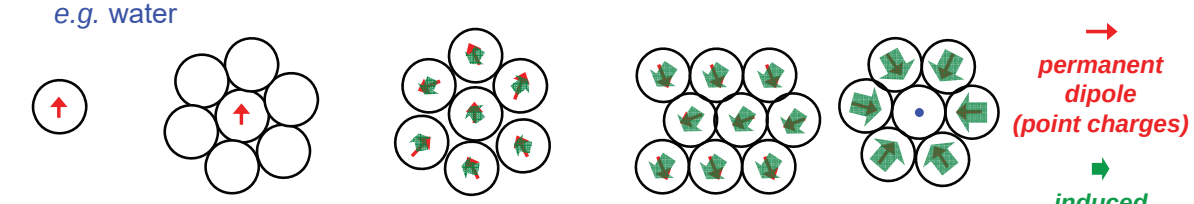
$$^{(2)}V(r_{ij};\ldots)$$

$$^{(2)}V(r_{ij};\ldots)+{}^{(2)}V(r_{jk};\ldots)+{}^{(2)}V(r_{ki};\ldots)+{}^{(3)}V(r_{ij},r_{jk},r_{ki};\ldots)$$

three-body term *(the presence of a third atom changes the way two atoms interact with each other – via electronic polarization [dispersion & electrostatics])*

In electrostatic interactions:

- Many-body effects in electrostatic interactions arise from electronic polarizability, *e.g.* water

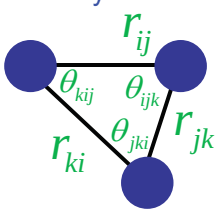


gas phase (1.85 D) | ***non-polar solution +0%*** | ***liquid phase + 25%*** | ***solid state +30%*** | ***near an ion up to +50% (?)***

Non-bonded terms: Many-body interactions

In van der Waals interactions:

- May be included through the Axilrod-Teller three-body term



$$^{(3)}V(r_{ij},r_{jk},r_{ki},\theta_{ijk},\theta_{jki},\theta_{kij};k_{ijk})=\frac{k_{ijk}\left(1+3\cos\theta_{ijk}\cos\theta_{jki}\cos\theta_{kij}\right)}{(r_{ij}r_{jk}r_{ki})^3}$$

homoatomic case: related to the atomic polarizability α

$$=3\alpha C_6/(16\pi\varepsilon_o)>0$$

- Most negative for aligned particles, most positive for equilateral-triangle configuration
- Derived from quantum-mechanics (perturbation analysis of dispersion interactions)
- About 10% of the lattice energy of crystalline argon

Many-body terms are seldom explicitly included:

- Computationally expensive:

		scaling	*e.g. N*=1000
- $^{(2)}V$	(1/2)*N*(*N*-1) terms	~O[N^2]	499’500
- $^{(3)}V$	(1/6)*N*(*N*-1)(*N*-2) terms	~O[N^3]	166’167’000
- $^{(4)}V$	(1/24)*N*(*N*-1)(*N*-2)(*N*-3) terms	~O[N^4]	41’417’125’750

- The computational scaling of classical simulations is in principle O[N^2] without many-body terms (*i.e.* only pairwise terms) – down to O[*N*] with special tricks (patience again) – the covalent terms are list-based and thus cheap
- Yet... many-body effects are not negligible... what to do ?

Challenge: avoid calculating many-body terms, but without neglecting them !

Non-bonded terms: Many-body interactions

Effective pairwise interactions:

- Incorporate the mean effect of many-body terms into "effective" pairwise terms
- The resulting effective interaction parameters are valid only for a specific environment
- *E.g.* van der Waals interactions: Barker-Fischer-Watts potential for liquid argon

loss of transferability, problems for "mixed" environments

$$\eta(\rho) = e^{a(1-\rho)} \sum_{n=0}^{5} A_n (\rho - 1)^n + \sum_{n=3}^{5} C_n (\rho^n + \delta)^{-1}$$

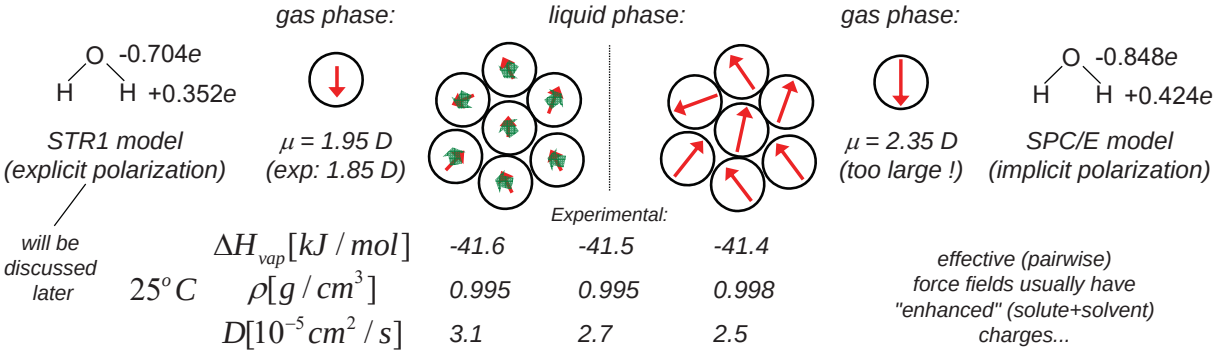
→ *11 parameters*

→ *calibrated by averaging the Axilrod-Teller term over liquid argon configurations*

→ *will no longer give a good description of gas-phase argon! (real-gas properties)*

→ *better, but still may not be good enough for solid argon...*

- *E.g.* electrostatic interactions: use of enhanced atomic charges for liquids/solution



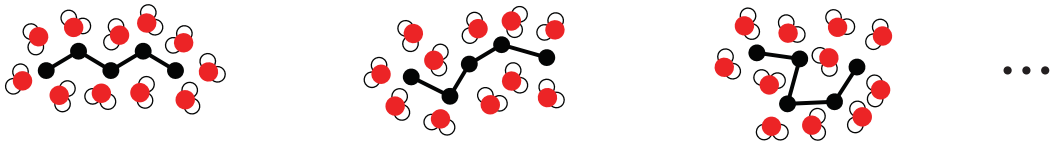
Calculating the forces

- The force field *V(r)* represents the (classical) potential energy of the system in configuration ***r***
- The corresponding force $\boldsymbol{F}_i(\boldsymbol{r})$ on each on each atom *i* is required for performing EM, MD or SD
- The force is the negative derivative of *V(r)* with respect to $\boldsymbol{r}_i$

$$\boldsymbol{F}_i(\boldsymbol{r}) \doteq -\frac{\partial V(\boldsymbol{r})}{\partial \boldsymbol{r}_i}$$ → *force on atom i in configuration* ***r***

$$\boldsymbol{F}(\boldsymbol{r}) = \{\boldsymbol{F}_i \mid i = 1,2,..N\}$$ → *3N-dimensional force vector*

- Simulation programs typically evaluate the forces analytically, simultaneously with the potential energy evaluation



$\boldsymbol{r}_1 \Rightarrow V(\boldsymbol{r}_1), \boldsymbol{F}(\boldsymbol{r}_1)$ $\boldsymbol{r}_2 \Rightarrow V(\boldsymbol{r}_2), \boldsymbol{F}(\boldsymbol{r}_2)$ $\boldsymbol{r}_3 \Rightarrow V(\boldsymbol{r}_3), \boldsymbol{F}(\boldsymbol{r}_3)$ ...

- *V(r)* is a sum over force-field terms

$$V(\boldsymbol{r}) = \sum_{\alpha=1}^{N_{terms}} {}^{(n_\alpha)}V^{(t_\alpha)}(q_{\alpha,1}, q_{\alpha,2}, \ldots; s_{\alpha,1}, s_{\alpha,2}, \ldots)$$

- ***F(r)*** is also a sum over force-field terms...

$$\boldsymbol{F}_i(\boldsymbol{r}) = \sum_{\alpha=1, i\in\alpha}^{N_{terms}} {}^{(n_\alpha)}\boldsymbol{F}_{i,\alpha}^{(t_\alpha)} \quad \text{with} \quad {}^{(n_\alpha)}\boldsymbol{F}_{i,\alpha}^{(t_\alpha)}(\{\boldsymbol{r}_j \mid j \in \alpha\}) = \sum_{n\in\alpha, i\in q_{\alpha,n}} -\frac{\partial^{(n_\alpha)}V^{(t_\alpha)}}{\partial q_{\alpha,n}} \frac{\partial q_{\alpha,n}}{\partial \boldsymbol{r}_i}$$

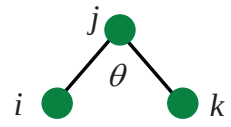
atom i is involved in term α — *internal coordinate n is involved in term α* — *atom i is involved in internal coordinate n*

Calculating the forces

- Example: forces arising from a bond-angle term with harmonic potential

Basic geometrical considerations:

$$\frac{\partial \boldsymbol{r}_{ji}}{\partial \boldsymbol{r}_i} = -\frac{\partial \boldsymbol{r}_{ji}}{\partial \boldsymbol{r}_j} = \underline{\boldsymbol{1}}$$



$$\boldsymbol{r}_{ji} \doteq \boldsymbol{r}_i - \boldsymbol{r}_j \qquad \boldsymbol{r}_{jk} \doteq \boldsymbol{r}_k - \boldsymbol{r}_j$$

$$\cos\theta = \frac{\boldsymbol{r}_{ji}\cdot\boldsymbol{r}_{jk}}{r_{ji}\cdot r_{jk}}$$

$$\frac{\partial r_{ji}}{\partial \boldsymbol{r}_i} = \frac{d\left(r_{ji}^2\right)^{1/2}}{d\boldsymbol{r}_{ji}}\frac{\partial \boldsymbol{r}_{ji}}{\partial \boldsymbol{r}_i} = (1/2)\frac{1}{r_{ji}}2\boldsymbol{r}_{ji}\underline{\boldsymbol{1}} \quad \Rightarrow \quad \frac{\partial r_{ji}}{\partial \boldsymbol{r}_i} = -\frac{\partial r_{ji}}{\partial \boldsymbol{r}_j} = \frac{\boldsymbol{r}_{ji}}{r_{ji}}$$

$$\frac{\partial \cos\theta}{\partial \boldsymbol{r}_i} = \frac{1}{r_{ji}^2 r_{jk}^2}\left((r_{ji}r_{jk})\boldsymbol{r}_{jk}\underline{\boldsymbol{1}} - (\boldsymbol{r}_{ji}\cdot\boldsymbol{r}_{jk})r_{jk}\frac{\boldsymbol{r}_{ji}}{r_{ji}}\right) \quad \Rightarrow \quad \frac{\partial \cos\theta}{\partial \boldsymbol{r}_i} = \frac{1}{r_{ji}}\left(\frac{\boldsymbol{r}_{jk}}{r_{jk}} - \cos\theta\frac{\boldsymbol{r}_{ji}}{r_{ji}}\right)$$

angle is invariant upon overall translation

$$i \leftrightarrow k \quad \Rightarrow \quad \frac{\partial \cos\theta}{\partial \boldsymbol{r}_k} = \frac{1}{r_{jk}}\left(\frac{\boldsymbol{r}_{ji}}{r_{ji}} - \cos\theta\frac{\boldsymbol{r}_{jk}}{r_{jk}}\right)$$

$$\frac{\partial \cos\theta}{\partial \boldsymbol{r}_i} + \frac{\partial \cos\theta}{\partial \boldsymbol{r}_j} + \frac{\partial \cos\theta}{\partial \boldsymbol{r}_k} = \boldsymbol{0} \quad \Rightarrow \quad \frac{\partial \cos\theta}{\partial \boldsymbol{r}_j} = -\frac{1}{r_{ji}r_{jk}}\left(\boldsymbol{r}_{jk} + \boldsymbol{r}_{ji} - \cos\theta\left(\frac{r_{jk}\boldsymbol{r}_{ji}}{r_{ji}} + \frac{r_{ji}\boldsymbol{r}_{jk}}{r_{jk}}\right)\right)$$

Forces:

$${}^{(3)}V^{\theta}(\theta;\theta_o,k_\theta) \doteq (1/2)k_\theta(\theta-\theta_o)^2$$

similar procedure for all term types (most tricky for dihedrals)

$$\boldsymbol{F}_{i,\alpha}^{\theta} = -\frac{dV^{\theta}}{d\theta}\frac{d\theta}{d\cos\theta}\frac{\partial\cos\theta}{\partial \boldsymbol{r}_i} = k^{\theta}\frac{\theta-\theta^o}{\sin\theta}\frac{1}{r_{ji}}\left(\frac{\boldsymbol{r}_{jk}}{r_{jk}} - \cos\theta\frac{\boldsymbol{r}_{ji}}{r_{ji}}\right)$$

$$\boldsymbol{F}_{k,\alpha}^{\theta} = \ldots = k^{\theta}\frac{\theta-\theta^o}{\sin\theta}\frac{1}{r_{jk}}\left(\frac{\boldsymbol{r}_{ji}}{r_{ji}} - \cos\theta\frac{\boldsymbol{r}_{jk}}{r_{jk}}\right)$$

$$\boldsymbol{F}_{j,\alpha}^{\theta} = \ldots = -(\boldsymbol{F}_{i,\alpha}^{\theta} + \boldsymbol{F}_{k,\alpha}^{\theta}) = -k^{\theta}\frac{\theta-\theta^o}{\sin\theta}\frac{1}{r_{ji}r_{jk}}\left(\boldsymbol{r}_{jk} + \boldsymbol{r}_{ji} - \cos\theta\left(\frac{r_{jk}\boldsymbol{r}_{ji}}{r_{ji}} + \frac{r_{ji}\boldsymbol{r}_{jk}}{r_{jk}}\right)\right)$$

forces sum up to zero for all physical terms (Newton 3rd law)

Force-field engineering

A highly non-trivial, time-consuming (and poorly funded) task !

Sources of data

• Sources of data

→ **Theoretical**: ***quantum-mechanical*** (QM: *ab initio*, semi-empirical, DFT) calculations on small molecules (or molecule pairs/clusters)

- Must include ***low as well as higher energy*** ("distorted") conformations
- Fit to ***energies***, ***forces***, ***Hessian*** (vibrational properties), ***torsional profiles*** (barriers)
- *Lots of information* and applicable to *any molecule*, but
  (1) **accuracy limited** by that of the QM method employed,
  (2) **little information on intermolecular interactions** (*e.g.* limited to molecule pairs in gas phase) and **bulk solvation** (*e.g.* limited to solute-solvent microclusters in gas)

→ **Experimental**:

nowadays: lots of efforts to try to get more accurate atomic charges and dispersion coefficients from QM !

- ***Spectroscopic*** (*e.g.* IR, μ-wave, UV, visible, NMR, X-ray/neutron scattering, ...) properties
- ***Equilibrium*** (*e.g.* densitiy, heat capacity, compressibility, expansivity, dielectric permittivity) and ***thermodynamic*** (free energy, enthalpy, entropy of various physical processes; *e.g.* phase transition, mixing, binding, reaction) properties
- ***Transport*** (*e.g.* diffusion, thermal/ionic conductivity, dielectric relaxation) and ***kinetic*** (rates of various physical and chemical processes) properties
- *Information more sparse* and restricted to *small molecules*. In addition:
  (1) data is also affected by **experimental errors** (!)
  (2) data may **already result from the application of a model**
  (advisable: *fit to raw measurement data* rather than to processed data
  [*e.g.* NMR NOE intensities > NOE-derived distances >> NMR-derived structure]).

nowadays: also lots of efforts to automate the parameter calibration based on experimental thermodynamic data !

Sources of data

• Tentative classification of the data

→ **Primary**: data which can (in principle) be mapped to one (or a few) force-field parameters (can be used to estimate parameters directly)

→ **Secondary**: data which can be reliably compared to the result of a (reasonably short) simulation

→ **Tertiary**: data which can be compared to the simulation results, but not reliably enough to be used for parametrization (can be used [with care] for "validation" only).

"Not reliably" means here *e.g.* :

- *Too large experimental uncertainty*
- *Underdetermination* of the experimental data (data can be reproduced by different models)
- *Full convergence not possible* within simulation timescale
- *Ambiguity* in the correspondance experimental ↔ simulated observable (interpretation)
- Data *processing already results from application of a model* (and you don't want to fit a model against another model!)

Sources of data

Table 3 *Possible source of data for force-field parametrization or validation*

Technique	Phase	Type	Property	Parameters
Spectroscopy (IR, μ-wave)	Gas	1°	Vibrational/rotational spectra, overtone analysis	$^{(n)}k_b$, $^{(n)}k_\theta$, $^{(2)}k_\xi$, CT
			Moments of inertia (small molecules)	b^o, θ^o
		3°	Rotational barriers and populations (estimates, small molecules)	$^{(n)}k_\phi$
(UV, visible, μ-wave)	Solution	3°	Time-resolved fluorescence intensities, depolarization, circular dichroism	
(NMR)	Solution/ membrane	1°	Rotational barriers	$^{(n)}k_\phi$, vdW(1,4)
		2°	Molecular structure, rotamers at equilibrium	$^{(n)}k_\phi$, vdW(1,4)
		3°	Distances (NOE, chemical shift), orientations (J-coupling), equilibrium constants, order parameters, relaxation times, diffusion constants (translation/rotation), residence times, H/D exchange rates, etc.	
Diffraction (X-ray, neutron)	Crystal	1°	Molecular structure	b^o, θ^o
			Force-length interpolation	$^{(2)}k_b$, $^{(2)}k_\theta$
		2°	Molecular structure, crystal density, packing, lattice dynamics	vdW, q, H-bond
		3°	Electron density map, B-factors, occupancy factors	
(neutron)	Liquid/ polymers	3°	Radial distribution functions	
			Static and dynamic structure factors (polymers)	$^{(n)}k_\phi$, vdW
Thermodynamic/ kinetic measurements	Gas	1°	Heats of formation	E
			Thermodynamic properties for rare gas mixtures	vdW, CR
Thermodynamic/ kinetic measurements	Liquid/ solution	2°	Density, vapour pressure, solvation free energy, heat of vaporization, heat of mixing, partition coefficients, heat capacity, compressibility, viscosity, diffusion constant, transport properties, etc.	vdW, q, H-bond
		3°	Chemical equilibrium parameters, pK_a, dielectric properties, reaction rates	
Ab initio and semiempirical calculations	Gas	1°	Equilibrium geometries	b^o, θ^o
			Vibrational analysis	$^{(n)}k_b$, $^{(n)}k_\theta$, $^{(2)}k_\xi$, CT
			Conformers analysis	$^{(n)}k_\phi$
			Population analysis or fit of the electrostatic potential	q
			Van der Waals clusters (second-order perturbation or higher)	vdW
			Energy/derivatives	E, dE, d^2E
	Solution	1°	*Idem*, supermolecule and/or reaction field approach	*Idem*

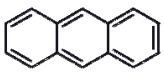
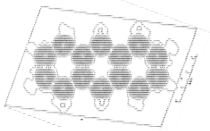
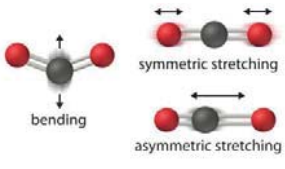
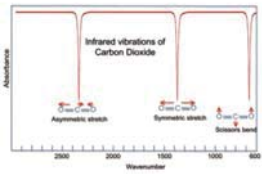
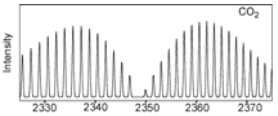
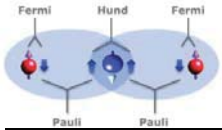
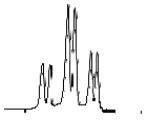
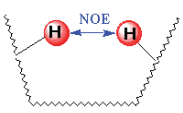
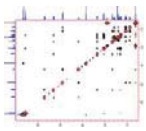
Type: 1° – primary, 2° – secondary, 3° – tertiary data, see Sec. 7.2; Parameters: parameters that can be calibrated using the corresponding data; CT: covalent coordinate cross-terms (Sec. 6.5); CR: van der Waals combination rules (Sec. 6.6.2); q: charges for Eq. 6.7.1.1; vdW: van der Waals parameters (Sec. 6.6); vdW(1,4): third-neighbour van der Waals parameters; k: force constants for Eqs. 6.1.1.1, 6.2.1.1, 6.3.1.1 and 6.4.1.1; H-bond: hydrogen-bonding interaction parameters (Sec. 6.9).

primary
secondary
tertiary

From... my thesis...
Highlighted: most important sources

Source of data: spectroscopic

- **Spectroscopy**: interaction between *matter* and *electromagnetic wave* ***absorbtion, emission, resonance, diffraction, ...***

	interacts with wavelength/wavenumber/frequency			***use for calibrating***	***special feature***
X-ray	core electrons ~0.1 nm ***(neutrons → nuclei incl. hydrogens)***			b_o, θ_o, ξ_o ***PRIMARY*** VdW, q, HB ***SECONDARY***	crystal ***(phase problem, hydrogens not visible with X-ray only)***
IR + Raman	nuclei (vibration) ~1000 cm^{-1}			k_b, k_θ, k_ξ ***PRIMARY***	solution or crystal/powder ***(normal mode analysis)***
μ-wave	nuclei (rotation) ~10 GHz			b_o, θ_o, ξ_o ***PRIMARY***	gas phase ***(easy only for small molecules)***
NMR	nuclear spins ~100 MHz ***(for magnetic field of ~10 T)***		***scalar J-coupling*** 	$k_\phi, VdW_{1,4}$ ***SECONDARY/ TERTIARY***	solution ***(Karplus equation + population model)***
			***NOE enhancement*** 	***TERTIARY***	solution ***(tumbling model + population model)***

Source of data: liquids and mixtures

- **Pure liquids** are favorable systems for the use of *secondary data*

Structural	***X-ray (SAXS) or neutron scattering***	***Radial distribution functions (RDF)***	***pair correlations***		
Thermodynamic	ρ ***density***	ΔH_{vap} ***enthalpy of vaporization***	c_P ***heat capacity***	$\left(\Delta H_f\right)$	***spectroscopic force-fields only***
		κ_T ***compressibility***	γ_P ***thermal expansion coefficient***		
Transport	D ***self-diffusion constant***	η ***viscosity***			
Dielectric	ε ***permittivity***	τ_D ***Debye relaxation time***			
Kinetic	τ_{mol} ***Molecular reorientation time***				

- So are **liquid mixtures** (mixed liquids, finite-concentration or infinitely dilute solutions)

$\Delta H_{mix}(x)$ $\Delta V_{mix}(x)$	$\gamma_{sol}(x)$ $\gamma_{slv}(x)$	ΔG_{slv}
Thermodynamic properties of mixing (composition x)	***(Limiting) activity coefficients***	***Solvation free energy***

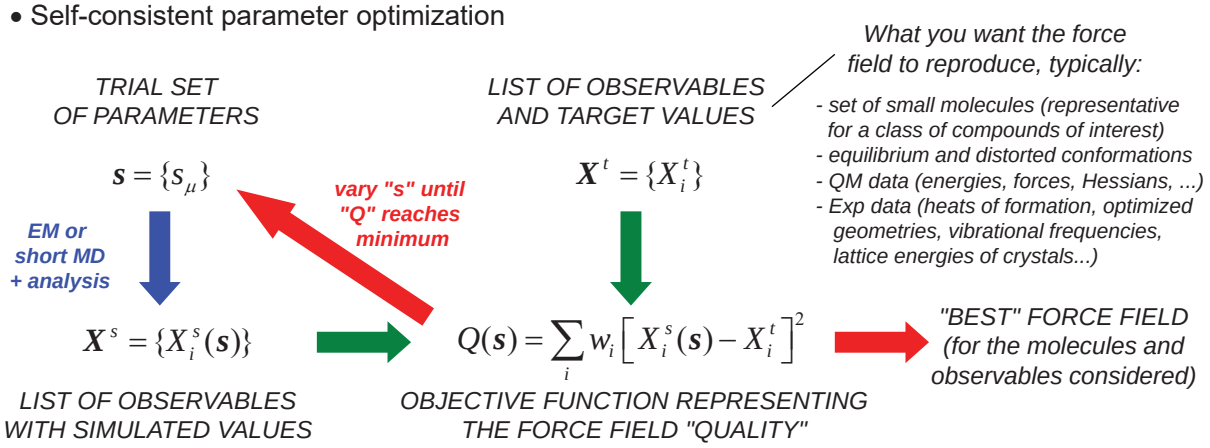
Source of data: quantum mechanics

- **Quantum mechanics** is a tempting source for *primary data*

  - Lots of information, but (1) accuracy limited by that of the QM method employed, (2) little information on intermolecular interactions and "bulk" solvation (*e.g.* limited to solute-solvent clusters)

Parametrization strategies

- Two main approaches
  - → **Self-consistent parameter optimization:** typical for class-II ("spectroscopic") force fields
  - → **Parameter optimization by trial & error:** typical for class-I ("biomolecular") force fields
- Self-consistent parameter optimization



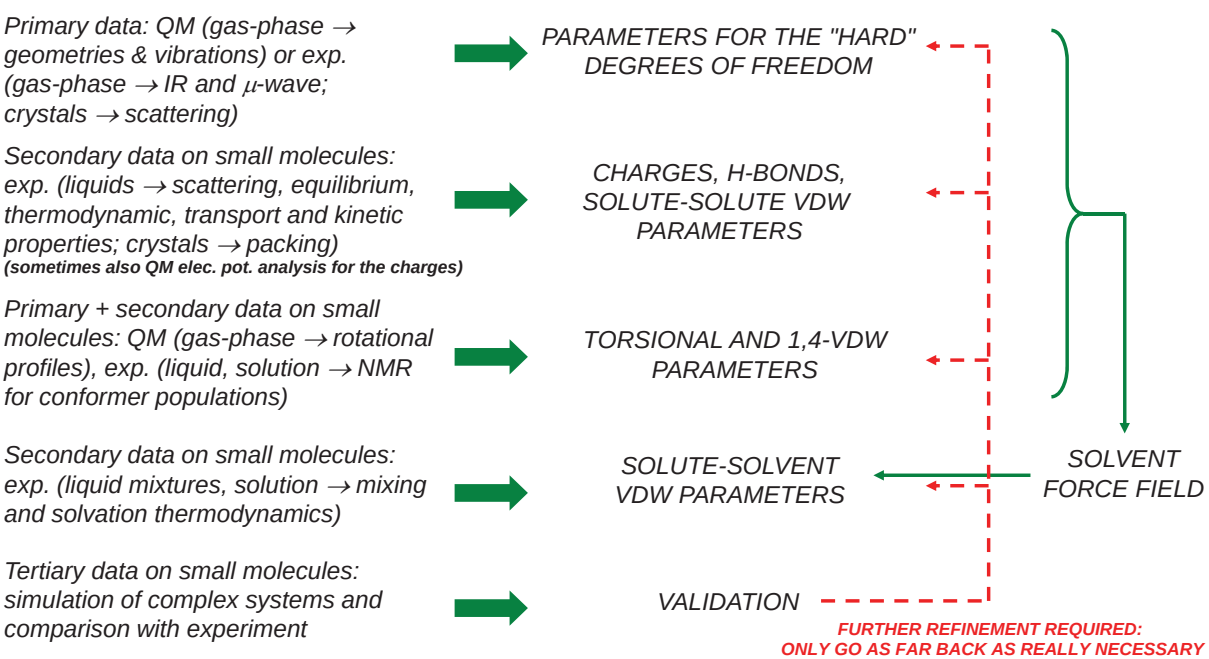
+: use of QM data (lots of data !) → many parameters can be optimized (*e.g.* anharmonic and cross terms)
+: elegant, automatic (less guesswork and human time)
+: high accuracy for the molecules and observables of interest (still limited by accuracy of QM method !)
-: QM data is approximate (*e.g.* dispersion requires 2nd order perturbation) and contains little information on intermolecular interactions and solvation
-: Observables must be primary, or secondary and cheap to compute
-: Parameters are less intuitive and transferable, and inclusion of new functional groups often requires a full reparametrization

} Often: poor charges and vdW parameters !

Parametrization strategies

- Parameter optimization by trial & error

Parameter refinement in successive stages – new changes may sometimes require an update of previously refined parameters



+: more appropriate for the use of experimental data (incl. intermolecular interactions and solvation)
+: parameters are more intuitive and transferable (sometimes, one abuses even a bit of transferability !)
- : relies on a lot on guesswork and human time
- : (arguably) lower accuracy (works well for properties that are not too force-field sensitive)

Parametrization strategies

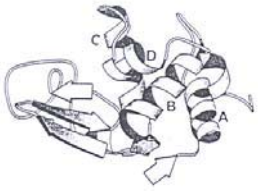
- Parameter optimization by trial & error

  → Validation & further refinement may extend over very long times (decades), triggered by the consideration of longer timescales or new compounds, or the availability of new experimental data (revealing force field deficiencies)

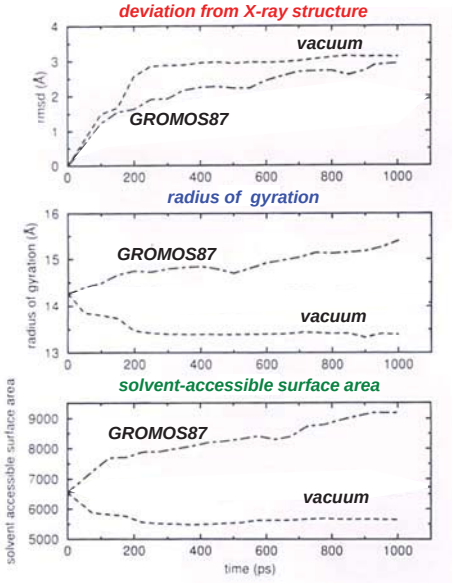
- Example: GROMOS87 → GROMOS96: the problem

Simulation timescale in 1996 ~1ns
→ now it looks bad !

hen-egg-white lysozyme (129 residues)



Simulation timescale in 1987-1995 ~1-100 ps
→ all looked fine



⇒ *large deviations from experimental structure (just as vacuum)*

⇒ *too expanded shape*

⇒ *too large solvent-exposed surface area*

→ Longer timescale simulations evidence that lysozyme is too "hydrophilic" in water !!!

Parametrization strategies

- Example: GROMOS87 → GROMOS96: the remedy

→ Simulations of pure alkanes

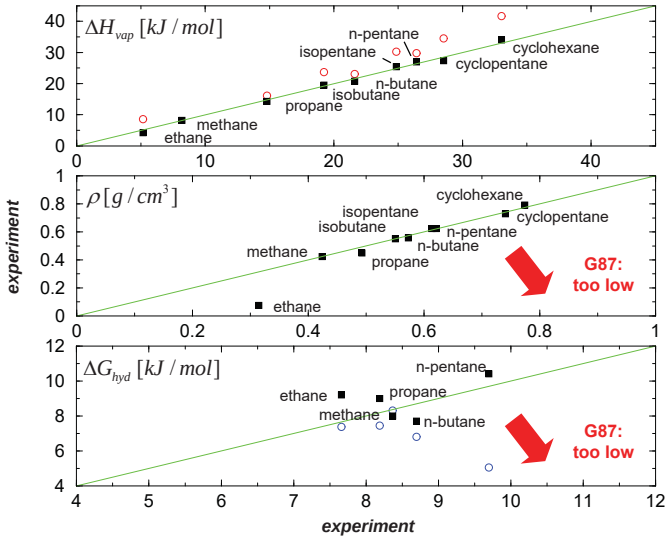
Refine: $C_6^{1/2}(CH_n;CH_n)$ and $C_{12}^{1/2}(CH_n;CH_n)$ $n=0..4$

Target: ρ and ΔH_{vap} for alkanes

→ Simulations alkanes in water

Refine: $C_6^{1/2}(OW;OW)$ and $C_{12}^{1/2}(OW;OW)$
for non H-bonding interactions (value for water-water: unaffected)

Target: ΔG_{hyd} for alkanes

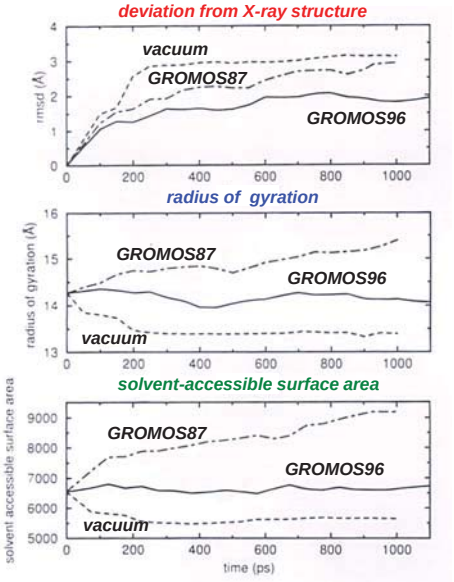


Note: pure liquid simulations at T=300K (except: methane at 111.7K) and P=P_{vap} (ethane 41 bar, propane 9 bar, butane 2 bar, isobutane 3 bar; others: 0.1-1 bar)

■ GROMOS96
○ GROMOS87
○ GROMOS87-mod

Parametrization strategies

• Example: GROMOS87 → GROMOS96: validation



→ now it looks much better !

⇒ *acceptable deviations from experimental structure (note: ideal value is not zero; there are always fluctuations !)*

⇒ *compactness stable and similar to crystal structure*

⇒ *solvent-exposed surface area stable and similar to crystal structure*

Things to keep in mind

• THERE IS NO SUCH THING AS A UNIVERSAL FORCE FIELD.
But force fields best suited for:
- a given system
- a given phase
- a given set of properties studied
- a given computer budget

• FORCE FIELD PARAMETERS ARE NOT PHYSICAL CONSTANTS
But but mere parameters which are:
- correlated among each other
- correlated with choices made in functional form & combination rules
- correlated with the choice of the considered degrees of freedom
- correlated with the force-field training set

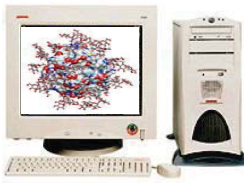
→ they are generally not transferable from a force-field to another

• THE QUALITY OF A FORCE FIELD IS LIMITED BY THE CRUDEST (NOT THE BEST !) APPROXIMATION MADE IN ITS DEFINITION

Computer Simulation in Chemistry, Biology and Physics

P.H. Hünenberger

COMPUTER SIMULATION OF MOLECULAR SYSTEMS

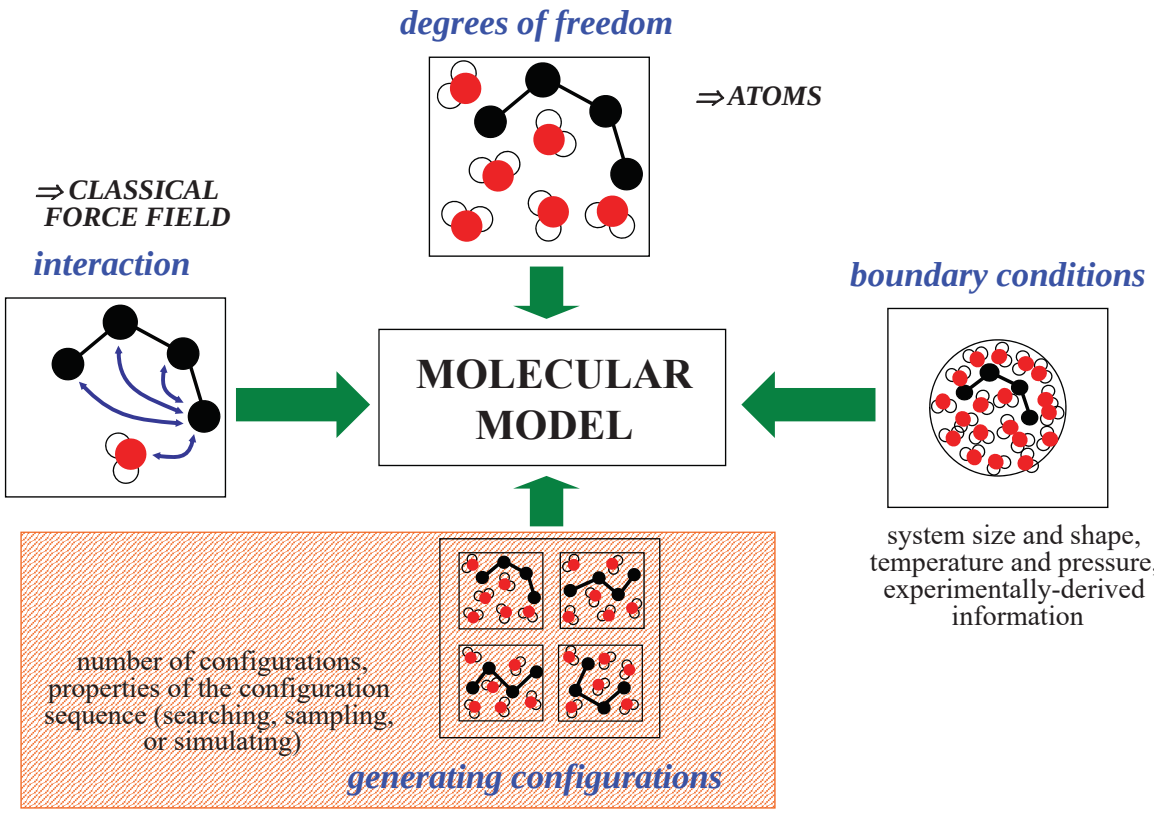


Herbstsemester 2019
Tuesday 9:45 a.m. – 11:30 p.m.
HCI D2

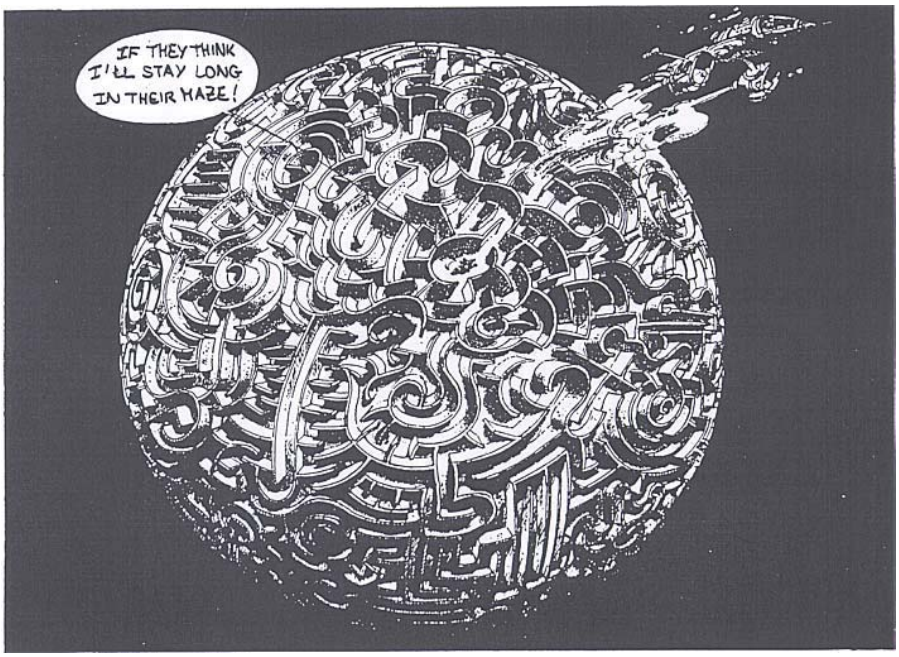
Lecture 529-0004-00
www.csms.ethz.ch/education/CSCBP

LECTURE 3 (WEEKS 3+4):
Generating configurations

Four basic choices defining a molecular model



GENERATING CONFIGURATIONS



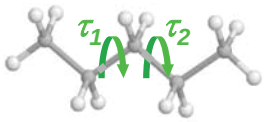
Potential energy surfaces

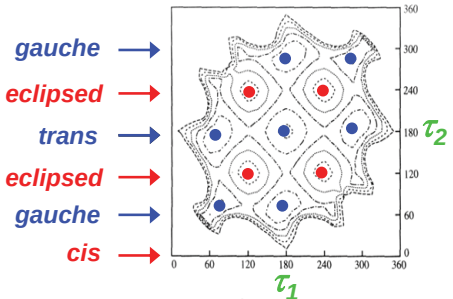
Potential energy surface (PES)

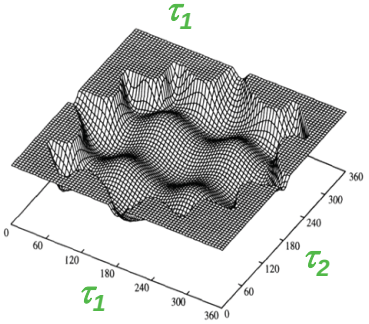
force field
or
quantum-chemical method
(within Born-Oppenheimer)

} potential energy (hyper)surface defined by $V(\boldsymbol{r})=V(\{\boldsymbol{r}_i\})$ with $3N$ (Cartesian) or $3N$-6 (internal) dimensions

Example: pentane







rigid bonds and bond-angles optimized methyl orientation (or united atom)

⇒ 2 degrees of freedom

⇒ 9 minima and 9 maxima

Potential energy surfaces

Large systems

- many dimensions

  ⇒ *no hope to visualize the PES*
  ⇒ *no hope to enumerate all configurations*

  e.g. ***100 residue protein***, [~15 atoms per residue] → ***4500 degrees of freedom***

  5000 rigid water molecules → ***30'000 degrees of freedom***

- many energy minima and barriers (the PES is "*frustrated*")

  ⇒ *no hope to enumerate all minimas*

  e.g. ***alkane with n carbons*** [~ 3 minima & barriers per torsion] → ***about 3^n minima*** (n=10: 59'049, n=20: 3'486'784'401)

Generating configurations

- needs a method that generates preferentially ***relevant configurations*** (*e.g.* low energy, Boltzmann-weighted)
- needs a good ***initial configuration*** whenever available (*e.g.* from X-ray or NMR experiments)

Properties of the method

- may generate a ***Boltzmann-weighted ensemble*** of configurations
  ⇒ ***thermodynamic properties can be calculated***
- may generate a sequence of configurations through a physically-based ***classical equation of motion***
  ⇒ ***dynamic properties can be calculated***

Generating configurations

	SEARCHING	SAMPLING	SIMULATING
Boltzmann distribution	**NO**	**YES**	**YES**
Dynamical sequence	**NO**	**NO**	**YES**

Methods

- Single configuration
  - Energy minimization (Normal-mode analysis, transition-state location)
- Multiple configurations
  - Conformational search
    - Systematic search
    - Random search
    - Stepwise build-up
    - Genetic algorithms
    - Distance-geometry
    - Homology modeling
    - …
  - Enhanced-search molecular dynamics
    - Simulated annealing
    - PEACS
    - Soft-core atoms
    - 4D molecular dynamics
    - Local elevation
    - Essential dynamics
    - Conformational flooding
    - …

followed by

- Monte Carlo
- Molecular dynamics with altered masses
- Parallel tempering
- ...

- Molecular dynamics (with thermostat)
- Stochastic dynamics
- Brownian dynamics

Search the PES as wide as possible for low-energy regions, and report all local energy minima found by EM

Energy minimization

Problem of energy minimization (EM)

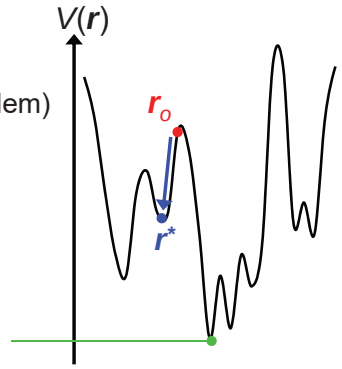
⇒ Given $V(\boldsymbol{r})$ with $\boldsymbol{r} = \{r_\alpha \mid \alpha=1\ldots 3N\}$, find a point $\boldsymbol{r}^*$ for which

stationary point $\left.\frac{\partial V(\boldsymbol{r})}{\partial r_\alpha}\right|_{\boldsymbol{r}^*} = 0 \quad \forall \alpha$ and $\left.\frac{\partial^2 V(\boldsymbol{r})}{\partial r_{\alpha\beta}^2}\right|_{\boldsymbol{r}^*} > 0 \quad \forall \alpha,\beta$ ***minimum***

Properties of energy minimization methods

- *numerical iterative* methods (due to the high complexity of the problem)
- perform successions of *downhill moves*
- find the *local minimum* closest to an *initial configuration* $\boldsymbol{r}_o$
- for frustrated PES, are *poor search methods* unless
  → applied to a large set of low-energy initial configurations $\boldsymbol{r}_o$
  → or combined with uphill moves

  … even in this case, the *global minimum* can seldom be located …



Various methods

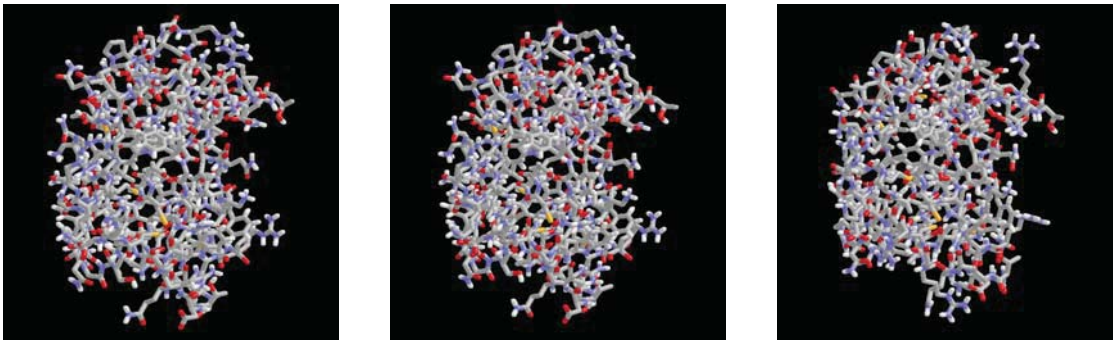
⇒ Non-derivative *(e.g.* simplex), first order *(e.g.* steepest descent) or second-order *(e.g.* Newton-Raphson)

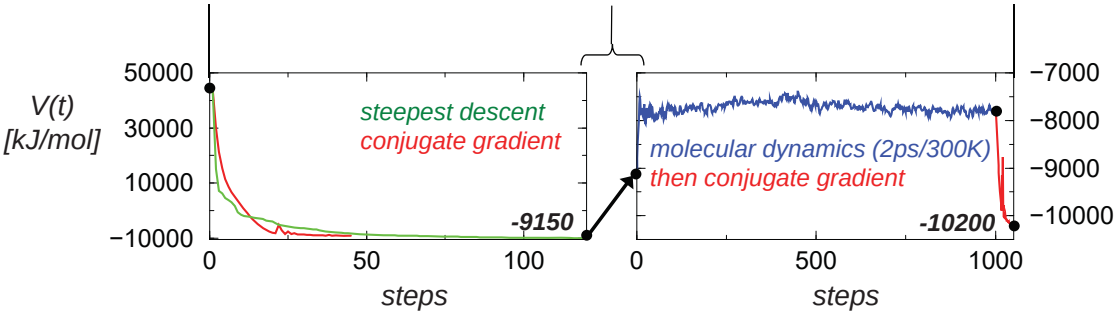
⇒ Not so important → read them in the script...

★

Energy minima and frustrated systems

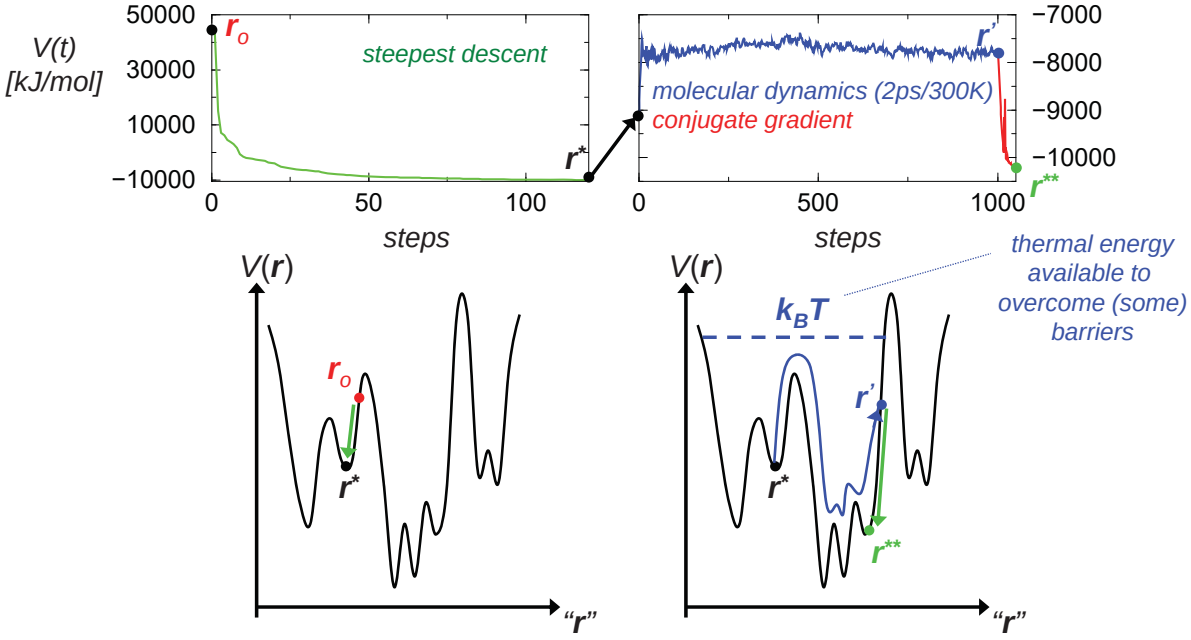
e.g. hen-egg-white lysozyme in vacuum





Energy minima and frustrated systems

Explanation



⇒ energy minimization is a *poor search method for frustrated PES* (lacks uphill moves)
⇒ it is mainly used in simulations to *relax strain*
- in configurations generated by *another search method*
- in initial configurations to be used for molecular dynamics simulations

Energy minima and populations

Population of the states of a molecular system

⇒ In the canonical (NVT) ensemble the population of states is controlled by the corresponding Helmholtz free energy, *i.e.*

$$p_i \sim \exp(-\beta A_i) \quad \text{with} \quad A_i = U_i - TS_i$$

$(k_BT)^{-1}$ — U_i: **well depth** — S_i: **well width**

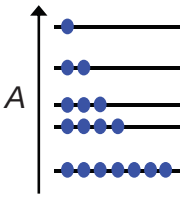
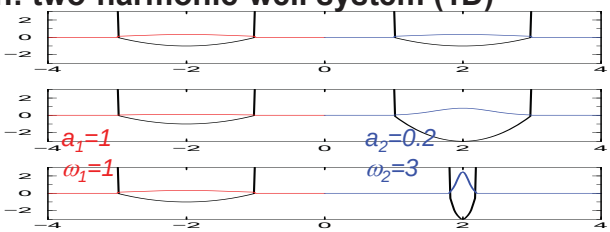


Illustration: two-harmonic-well system (1D)



f_1=0.50
f_2=0.50

cf unfolded and folded states of a protein

⇒ **a broad/shallow minimum may be more populated as a narrow/deep minimum (*i.e.* the energy is by no means the only relevant criterion) !**

Consequences

⇒ for systems with many degrees of freedom, entropic effects are very important
⇒ even *very deep minima can correspond to small populations*
⇒ proper generation of configurations to get *relevant statistical information* should be performed *according to free energy* (including entropic effects) rather than energy

→ ***sampling or simulating***

Energy minimization

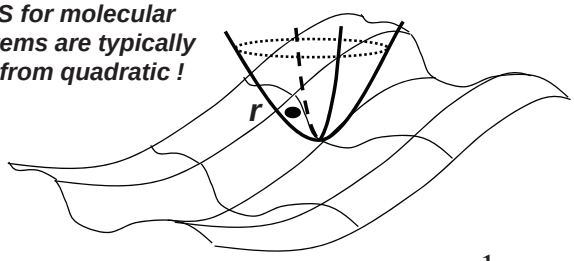
Problem of energy minimization (EM)

⇒ Given $V(\boldsymbol{r})$ with $\boldsymbol{r} = \{r_\alpha \mid \alpha=1\ldots 3N\}$, find a point $\boldsymbol{r}^*$ for which

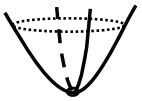
stationary point $\left.\frac{\partial V(\boldsymbol{r})}{\partial r_\alpha}\right|_{\boldsymbol{r}^*} = 0 \quad \forall \alpha$ and $\left.\frac{\partial^2 V(\boldsymbol{r})}{\partial r_{\alpha\beta}^2}\right|_{\boldsymbol{r}^*} > 0 \quad \forall \alpha,\beta$ ***minimum***

Local quadratic approximation to the PES

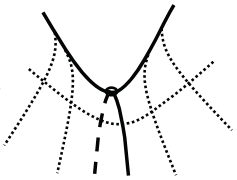
Only locally valid… PES for molecular systems are typically far from quadratic !



all eigenvalues > 0
⇒ minimum



one eigenvalue < 0
⇒ 1st order saddle point



$$V(\boldsymbol{r}+\Delta\boldsymbol{r}) \approx V(\boldsymbol{r}) + {}^t\nabla V(\boldsymbol{r})\Delta\boldsymbol{r} + \frac{1}{2}{}^t\Delta\boldsymbol{r}\,\underline{\boldsymbol{H}}(\boldsymbol{r})\Delta\boldsymbol{r}$$

Gradient vector (minus force), zero at stationary point

small displacement

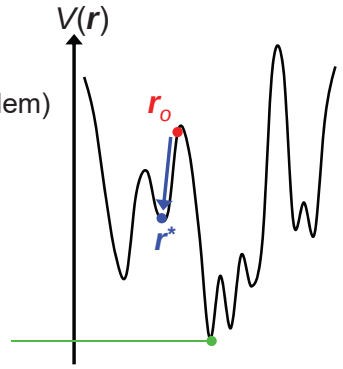
Hessian matrix, curvature information

$$H_{\alpha\beta} = \frac{\partial^2 V(\boldsymbol{r})}{\partial r_\alpha \partial r_\beta}$$

m eigenvalues < 0
→ m^th order saddle point (m=3N → maximum)

Energy minimization

Properties of energy minimization methods



- *numerical iterative* methods (due to the high complexity of the problem)
- perform successions of *downhill moves*
- find the *local minimum* closest to an *initial configuration* $\boldsymbol{r}_o$
- for frustrated PES, are *poor search methods* unless
  → applied to a large set of low-energy initial configurations $\boldsymbol{r}_o$
  → or combined with uphill moves

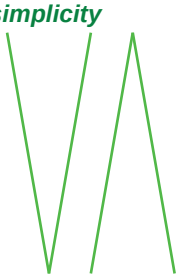
… even in this case, the *global minimum* can seldom be located …

robustness, simplicity

Classes energy minimization methods

$$V(\boldsymbol{r}+\Delta\boldsymbol{r}) \approx V(\boldsymbol{r}) + {}^t\nabla V(\boldsymbol{r})\Delta\boldsymbol{r} + \frac{1}{2}{}^t\Delta\boldsymbol{r}\,\underline{\boldsymbol{H}}(\boldsymbol{r})\Delta\boldsymbol{r}$$

non-derivative methods | *first-order methods* | *second-order methods*



convergence rate, computer time per step, memory

- non-derivative methods maybe used for *complex problems* where *derivatives are unavailable*
- first order typical for *classical systems* and quantum-mechanical calculations where *second derivatives are unavailable*
- second order typical for *quantum-mechanical calculations*, as a final refinement step
- derivatives may be computed analytically or numerically (finite-difference, expensive !)

The simplex method

a non-derivative method

Simplex

⇒ Geometrical figure with $M+1$ vertices in M (=$3N$) dimensions (*e.g.* triangle, tetrahedron, …)

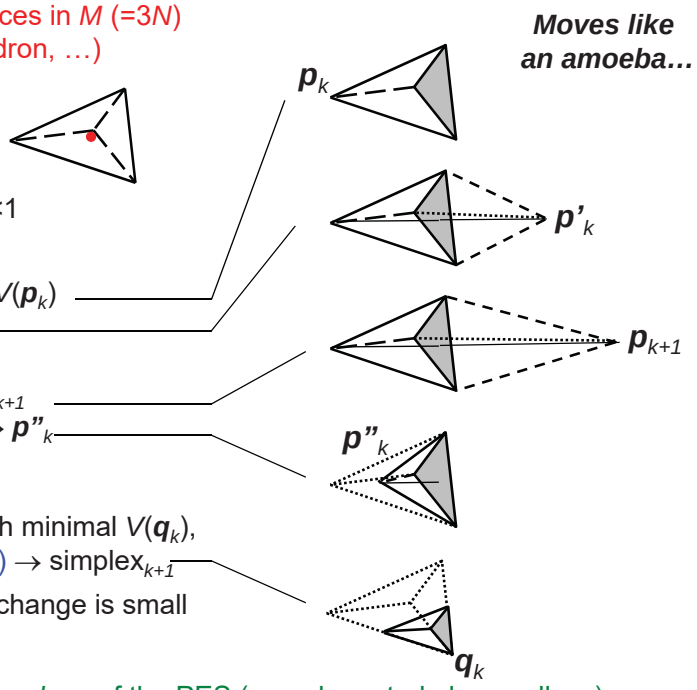
Algorithm

⇒ Start at configuration $\boldsymbol{r}_o$
Define simplex around point
Define scaling factors $\alpha>1$ and $\beta<1$

⇒ Iterate for k

- find the vertex $\boldsymbol{p}_k$ with maximal $V(\boldsymbol{p}_k)$
- reflect($\boldsymbol{p}_k$) → $\boldsymbol{p'}_k$
- if $V(\boldsymbol{p'}_k)< V(\boldsymbol{p}_k)$
  then reflect-expand($\boldsymbol{p}_k,\alpha$) → $\boldsymbol{p}_{k+1}$
  else highest-contract($\boldsymbol{p}_k,\alpha^{-1}$) → $\boldsymbol{p''}_k$
  if $V(\boldsymbol{p''}_k)<V(\boldsymbol{p}_k)$
  then $\boldsymbol{p}_{k+1} = \boldsymbol{p''}_k$
  else find the vertex $\boldsymbol{q}_k$ with minimal $V(\boldsymbol{q}_k)$,
  lowest-contract($\boldsymbol{q}_k,\beta$) → simplex$_{k+1}$

⇒ Terminate when move or energy change is small



Properties

⇒ Simplex *shape varies with local topology* of the PES (*e.g.* elongated along valleys)
⇒ Very *robust* (always converges to a minimum, normally the closest)
⇒ Very *slow convergence* (many evaluations of the potential energy)
⇒ Useful for *initial refinement* or when *derivatives are unavailable or expensive*
⇒ Alternative: sequential univariate minimization

The steepest-descent method

a first-order method

Move

⇒ in the direction of the negative gradient, *i.e.* downhill along $\boldsymbol{s}_k = \dfrac{\boldsymbol{g}_k}{g_k}$ with $\boldsymbol{g}_k = -\nabla V(\boldsymbol{r}_k)$

Algorithm (with arbitrary step)

⇒ Start at configuration $\boldsymbol{r}_o$
Define intial step size Δr_o
Define scaling factors $\alpha>1$ and $\beta<1$

⇒ Iterate for k

- $\boldsymbol{r}_{k+1} = \boldsymbol{r}_k + \Delta r_k \boldsymbol{s}_k$
- if $V(\boldsymbol{r}_{k+1})< V(\boldsymbol{r}_k)$
  then $\Delta r_{k+1} = \alpha\, \Delta r_k$
  else $\Delta r_{k+1} = \beta\, \Delta r_k$

⇒ Terminate when Δr_k or $|V(\boldsymbol{r}_{k+1})-V(\boldsymbol{r}_k)|$ small

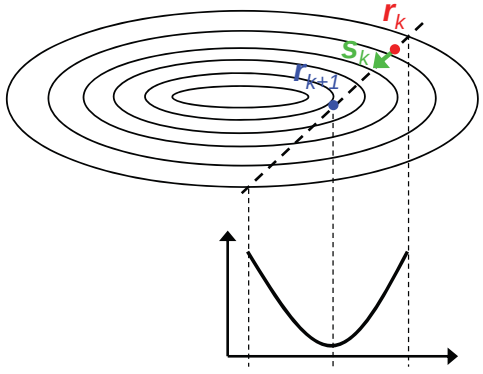
Algorithm (with line search)

⇒ Start at configuration $\boldsymbol{r}_o$

⇒ Iterate for k

- line-minimize($\boldsymbol{r}_k,\boldsymbol{s}_k$) → $\boldsymbol{r}_{k+1}$

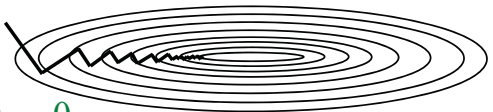
⇒ Terminate when $|V(\boldsymbol{r}_{k+1})- V(\boldsymbol{r}_k)|$ small



locating the minimum requires more than one potential evaluation !

Properties

⇒ *Robust* (always converges to a minimum)
⇒ Rather *slow convergence* (especially in long, narrow valleys). For the line search, because:



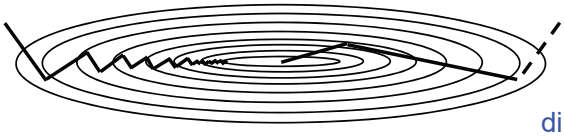
$\boldsymbol{s}_{k+1} \cdot \boldsymbol{s}_k = 0$

The conjugate-gradient method

another first-order method

For a quadratic surface

steepest descent successive line minimizations along the negative gradient $\boldsymbol{g}_k$



conjugate gradient successive line minimizations along mutually-conjugate directions $\boldsymbol{h}_k$ (*i.e.* $\boldsymbol{h}_k \underline{\boldsymbol{H}} \boldsymbol{h}_l = 0$)

$\Rightarrow \boldsymbol{g}_{k+1} \cdot \boldsymbol{g}_k = 0$
and minimization $k+1$ spoils the effect of minimization k (and all previous ones)

$\Rightarrow \boldsymbol{g}_k \cdot \boldsymbol{h}_l \ , \ \forall k,l$
on a quadratic surface, gradient is orthogonal to all previous search directions, *i.e.* the effect of line-minimizations is cumulative !

Algorithm

$\Rightarrow$ Start at configuration $\boldsymbol{r}_o$
Set $\boldsymbol{h}_o = \boldsymbol{g}_o = -\nabla V(\boldsymbol{r}_o)$ *first step along gradient*

$\Rightarrow$ Iterate for k

- line-minimize$(\boldsymbol{r}_k, \boldsymbol{h}_k) \rightarrow \boldsymbol{r}_{k+1}$
- $\boldsymbol{g}_{k+1} = -\nabla V(\boldsymbol{r}_{k+1})$
- $\boldsymbol{h}_{k+1} = \boldsymbol{g}_{k+1} + \gamma_k \boldsymbol{h}_k$

new search direction ≠ along gradient

$$\gamma_k = \frac{\boldsymbol{g}_{k+1} \cdot \boldsymbol{g}_{k+1}}{\boldsymbol{g}_k \cdot \boldsymbol{g}_k} \quad \text{or} \quad \frac{(\boldsymbol{g}_{k+1} - \boldsymbol{g}_k) \cdot \boldsymbol{g}_{k+1}}{\boldsymbol{g}_k \cdot \boldsymbol{g}_k}$$

Fletcher-Reeves *Polak-Ribiere*

equivalent for a quadratic surface

$\Rightarrow$ Terminate when $|V(\boldsymbol{r}_{k+1}) - V(\boldsymbol{r}_k)|$ small

Properties

$\Rightarrow$ Implicit use of second-derivative information
$\Rightarrow$ Converges in $M+1$ steps ($M=3N$) on a quadratic surface
$\Rightarrow$ *Converges generally faster than steepest descent* on non-quadratic surface
$\Rightarrow$ *Less robust* than steepest descent (may be trapped in a subspace of lower dimension)
$\Rightarrow$ Sometimes also used as arbitrary-step method

→ reset search direction from time to time

The Newton-Raphson method

a second-order method

Move

$\Rightarrow$ towards the closest stationary point of the local quadratic approximation to the PES

$$V(\boldsymbol{r}_k + \Delta\boldsymbol{r}_k) \approx V(\boldsymbol{r}_k) + {}^t\nabla V(\boldsymbol{r}_k)\Delta\boldsymbol{r}_k + \frac{1}{2}\,{}^t\Delta\boldsymbol{r}_k\, \underline{\boldsymbol{H}}(\boldsymbol{r}_k)\Delta\boldsymbol{r}_k$$

$$\Rightarrow \nabla V(\boldsymbol{r}_k + \Delta\boldsymbol{r}_k) = \nabla V(\boldsymbol{r}_k) + \underline{\boldsymbol{H}}(\boldsymbol{r}_k)\Delta\boldsymbol{r}_k = \boldsymbol{0}$$

stationary point (min., max., saddle)

solve for $\Delta\boldsymbol{r}_k$

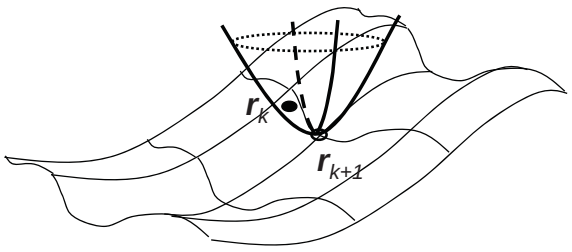
$$\Rightarrow \Delta\boldsymbol{r}_k = -\underline{\boldsymbol{H}}^{-1}(\boldsymbol{r}_k)\nabla V(\boldsymbol{r}_k)$$

Algorithm

$\Rightarrow$ Start at configuration $\boldsymbol{r}_o$
$\Rightarrow$ Iterate for k

- $\boldsymbol{r}_{k+1} = \boldsymbol{r}_k - \underline{\boldsymbol{H}}^{-1}(\boldsymbol{r}_k)\,\nabla V(\boldsymbol{r}_k)$

$\Rightarrow$ Terminate when $|V(\boldsymbol{r}_{k+1}) - V(\boldsymbol{r}_k)|$ small

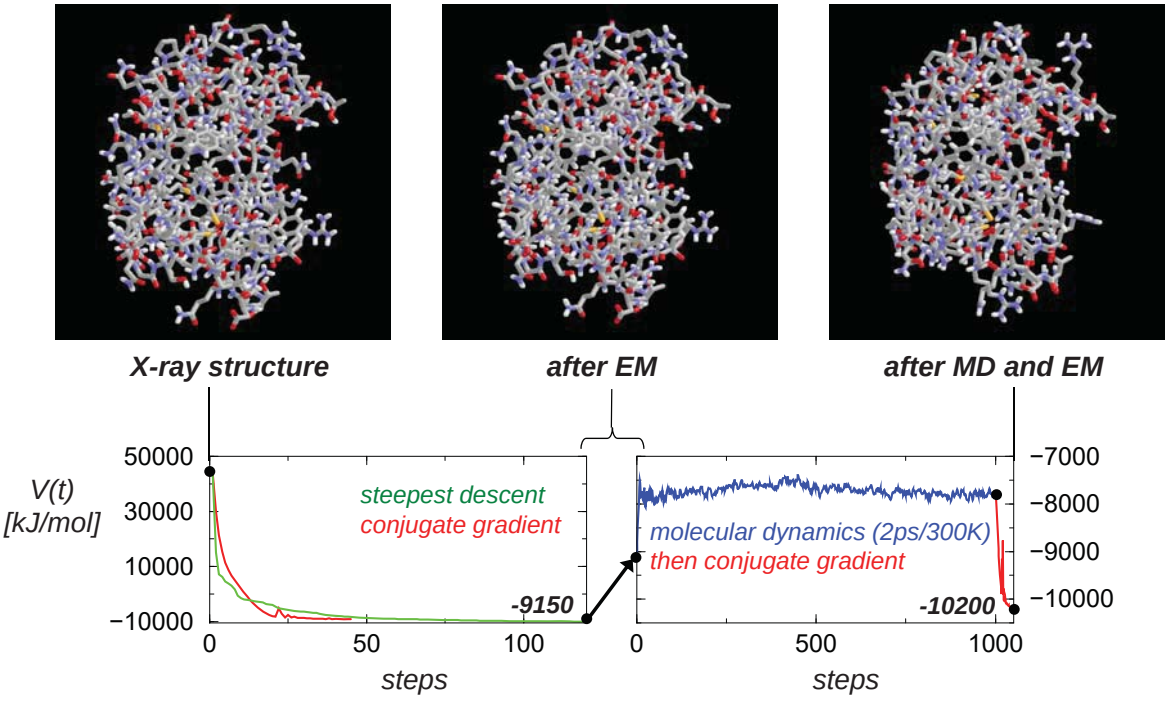


Properties

$\Rightarrow$ Converges in 1 step on a quadratic surface
$\Rightarrow$ *Converges faster than steepest descent or conjugate gradient* on non-quadratic surface
$\Rightarrow$ *Not robust* (converges to the closest stationary point – not necessary a minimum)
$\Rightarrow$ Expensive in terms of time-per-step and memory (Hessian calculation, storage and inversion)
$\Rightarrow$ Used for *final refinement in small systems* (often quantum chemistry calculations)
$\Rightarrow$ Alternative, quasi-Newton methods: gradually construct an approximation $\underline{\boldsymbol{A}}_k$ to $\underline{\boldsymbol{H}}^{-1}$ so that $lim_{k\rightarrow\infty}\, \underline{\boldsymbol{A}}_k = \underline{\boldsymbol{H}}^{-1}$ (converges in M steps on quadratic surface, first order again; memory cost remains, but no need for the expensive matrix inversion)

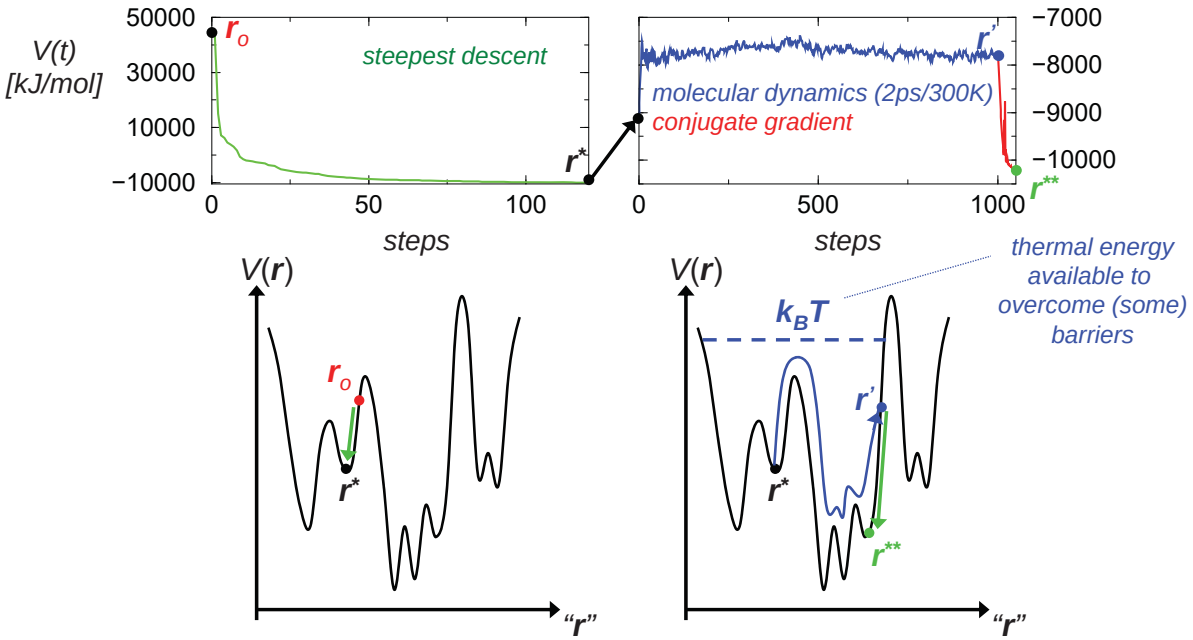
Energy minima and frustrated systems

e.g. hen-egg-white lysozyme in vacuum



Energy minima and frustrated systems

Explanation



⇒ energy minimization is a *poor search method for frustrated PES* (lacks uphill moves)
⇒ it is mainly used in simulations to *relax strain*
- in configurations generated by *another search method*
- in initial configurations to be used for molecular dynamics simulations

Energy minima and populations

Population of the states of a molecular system

⇒ In the canonical (NVT) ensemble the population of states is controlled by the corresponding Helmholtz free energy, *i.e.*

$$p_i \sim \exp(-\beta A_i) \quad \text{with} \quad A_i = U_i - TS_i$$

$(k_B T)^{-1}$ — well depth (U_i) — well width (TS_i)

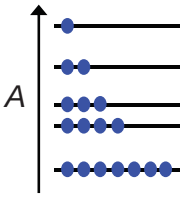


Illustration: two-harmonic-well system (1D)

- potential energy (one well)

$$V(x) = \begin{cases} \beta^{-1}\omega(x^2/a^2 - 1) & \text{for } |x| < a \\ \infty & \text{otherwise} \end{cases}$$

- partition function (one well)

$$Z = \int_{-a}^{a} dx \exp[-\beta V(x)] = \frac{\pi^{1/2} a \exp(\omega)\,\mathrm{erf}(\omega^{1/2})}{\omega^{1/2}}$$

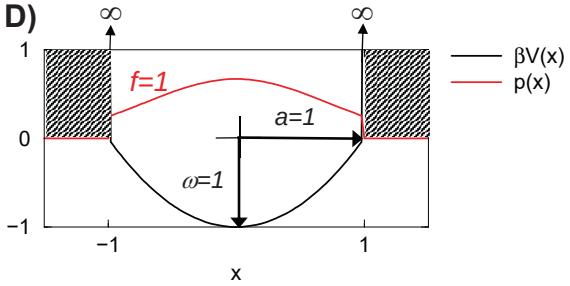
- probability distribution (two wells)

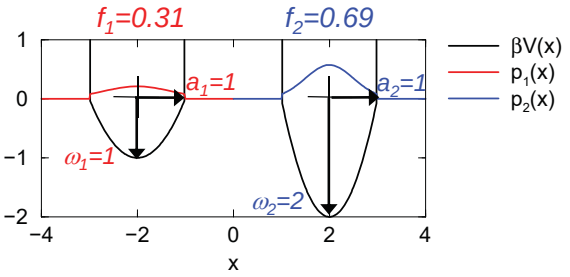
$$p_1(x) = (Z_1 + Z_2)^{-1} \exp[-\beta V_1(x)]$$

$$p_2(x) = (Z_1 + Z_2)^{-1} \exp[-\beta V_2(x)]$$

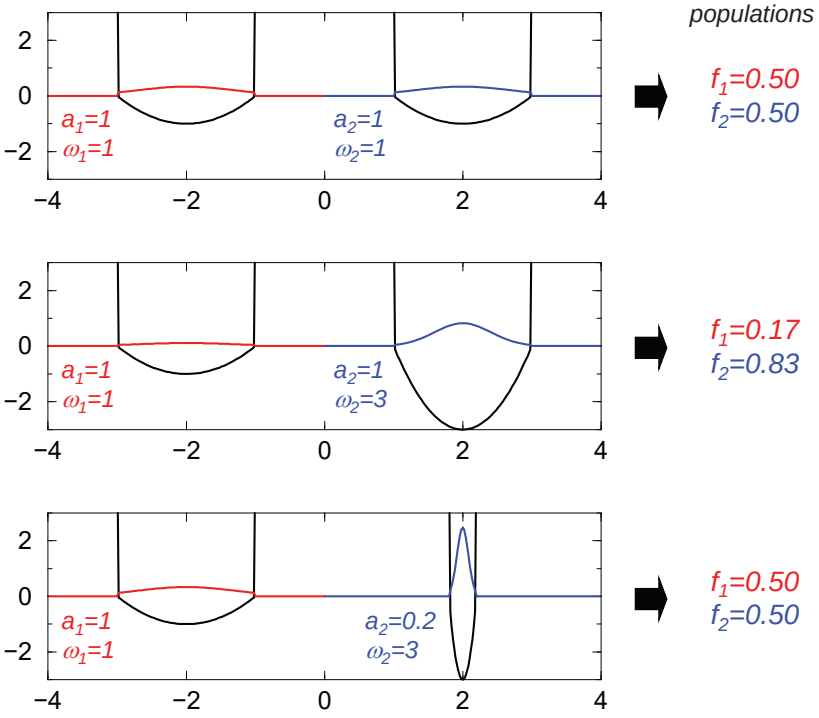
- relative populations

$$f_1 = (Z_1 + Z_2)^{-1} Z_1 \qquad f_2 = (Z_1 + Z_2)^{-1} Z_2$$





Energy minima and populations



cf unfolded and folded states of a protein

⇒ a broad/shallow minimum may be more populated as a narrow/deep minimum (*i.e.* the energy is by no means the only relevant criterion) !

Energy minima and populations

In many dimensions

$\omega_1 = \omega_2 = 1$, $a_1 = 1$, $a_2 = 1.01$ (second minimum is 1% broader in all M dimensions)

M	f_1	f_2
1	0.50	0.50
10	0.48	0.52
100	0.27	0.73
1000	0.00	1.00

balancing the populations would require $\omega_1 \cong 500'000$

Consequences

⇒ for systems with many degrees of freedom, entropic effects are very important

⇒ even *very deep minima can correspond to small populations*

⇒ proper generation of configurations to get *relevant statistical information* should be performed *according to free energy* (including entropic effects) rather than energy

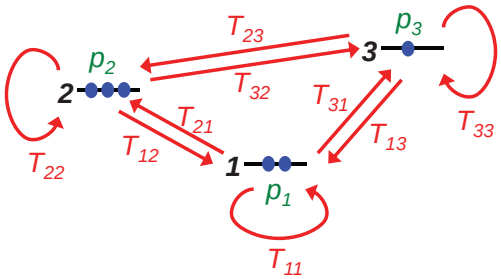
→ sampling or simulating

The Monte Carlo method

more properly named Metropolis Monte Carlo or importance sampling

Markov model for a collection of systems

- the system can be in a finite number N of states
- transition probabilities depend only on the initial and final states (no history)
- consider a large (infinite) collection of systems (●)
- relative populations of the states $\boldsymbol{p} = \{p_n \mid n=1\ldots N\}$
- transition probabilities $\underline{\boldsymbol{T}} = \{T_{mn} \mid m,n=1\ldots N\}$, where T_{mn} is the probability of $n \to m$ transition
- normalization



$$\sum_n^N p_n = 1 \quad \text{and} \quad \sum_m^N T_{mn} = 1\,,\ \forall n$$

Evolution of the relative populations

step 0	$\boldsymbol{p}(0)$ [arbitrary]
step 1	$\boldsymbol{p}(1) = \underline{\boldsymbol{T}}\,\boldsymbol{p}(0)$
step 2	$\boldsymbol{p}(2) = \underline{\boldsymbol{T}}\,\boldsymbol{p}(1) = \underline{\boldsymbol{T}}^2\,\boldsymbol{p}(0)$
⋮	⋮
step k	$\boldsymbol{p}(k) = \underline{\boldsymbol{T}}^k\,\boldsymbol{p}(0)$

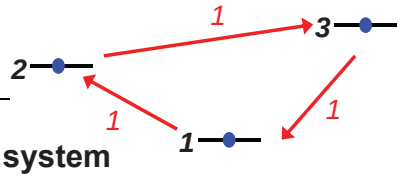
It can be shown that (in most cases)

- $\boldsymbol{p}(k)$ converges towards a value independent of $\boldsymbol{p}(0)$
- this *equilibrium value* $\boldsymbol{p}=\boldsymbol{p}(\infty)$ satisfies $\underline{\boldsymbol{T}}\,\boldsymbol{p} = \boldsymbol{p}$ (convergence)

The Monte Carlo method

Microscopic reversibility (detailed balance)

- at equilibrium, transitions between any two states occur at the same rate, *i.e.* $T_{mn}\, p_n = T_{nm}\, p_m$
- this excludes processes of the type



Generating a Markov chain of states for a single system

- from current state n select a possible new state m at random, according to a probability given by a *stochastic (or underlying) matrix* $\underline{\boldsymbol{\alpha}} = \{\alpha_{mn} \mid m,n=1\ldots N\}$, where α_{mn} $(m\neq n)$ is the probability of selecting state m for a $n\to m$ transition attempt from n (and $\alpha_{nn} = 0$)
- accept the transition attempt according to a probability given by an *acceptance matrix* $\underline{\mathbf{a}} = \{a_{mn} \mid m,n=1\ldots N\}$
- a rejected transition is considered a $n\to n$ transition
- this generates a Markov chain with

$$\begin{cases} T_{mn} = \alpha_{mn} a_{mn} & \text{if} \quad m \neq n \\ T_{nn} = 1 - \sum_{m\neq n}^{N} T_{mn} \end{cases}$$

Markov chain of states with Boltzmann weighting (Metropolis)

- use a symmetric stochastic matrix $\alpha_{mn} = \alpha_{nm}$
- use an acceptance matrix with

$$a_{mn} = \begin{cases} 1 & \text{if} \quad V_m \leq V_n \\ \exp\{-\beta(V_m - V_n)\} & \text{otherwise} \end{cases}$$

- proof using microscopic reversibility

for $n\neq m$:

$$\frac{p_m}{p_n} = \frac{T_{mn}}{T_{nm}} = \frac{\alpha_{mn}}{\alpha_{nm}} \frac{a_{mn}}{a_{nm}} = \frac{a_{mn}}{a_{nm}} = \exp\{-\beta(V_m - V_n)\}$$

- this result is assumed to remain valid for systems with an infinite number N of states

The Monte Carlo method

Application to atomic liquids

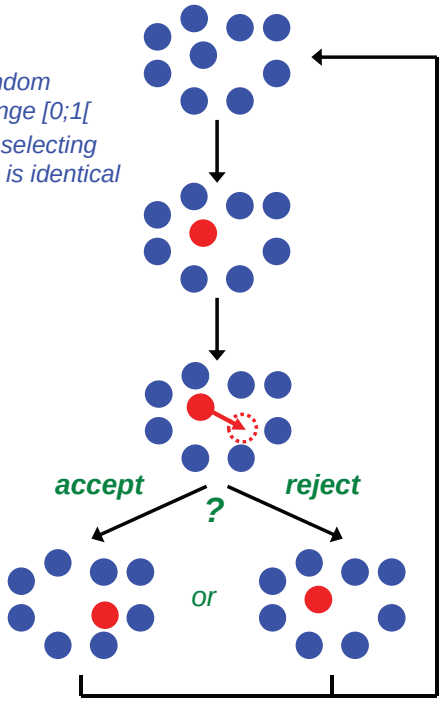
⇒ Start at configuration $\boldsymbol{r}(0)$

→ *rand(0,1): random number in range [0;1[*

⇒ Iterate for k

→ *probability of selecting reverse move is identical*

- select an atom i at random
  e.g. $i = int\,[N\ rand(0,1)+1]$
  or do all atoms in sequence
- attempt a move for this atom $\boldsymbol{r}_i(k)\to\boldsymbol{r}_i'(k)$
  e.g. $x_i' = x_i + [\ 2\ rand(0,1)-1\ ]\ \Delta r$
  $y_i' = y_i + [\ 2\ rand(0,1)-1\ ]\ \Delta r$
  $z_i' = z_i + [\ 2\ rand(0,1)-1\ ]\ \Delta r$
- if $V'=V[\boldsymbol{r}'(k)] < V=V[\boldsymbol{r}(k)]$
  then $\boldsymbol{r}(k+1) = \boldsymbol{r}'(k)$ → accept
  else if $rand(0,1) \leq exp\{-\beta(V'-V)\}$
  then $\boldsymbol{r}(k+1) = \boldsymbol{r}'(k)$ → accept
  else $\boldsymbol{r}(k+1) = \boldsymbol{r}(k)$ → reject



⇒ Terminate when $k=N_{steps}$ steps

Step size Δr

- too large → small acceptance ratio
- too small → slow sampling

tune for an acceptance ratio of ~50% (can be done on the flight)

The Monte Carlo method

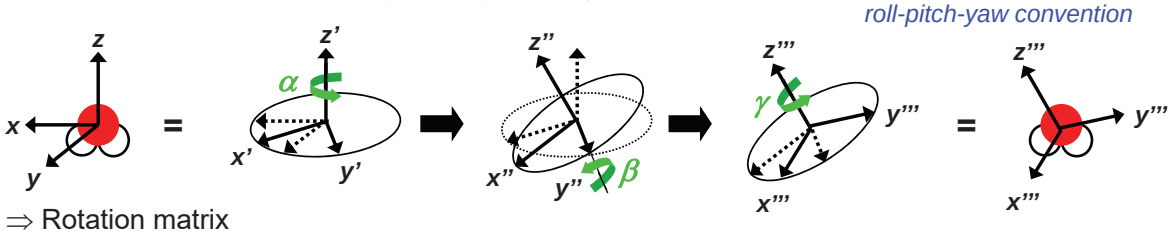
Moves for rigid molecules (e.g. molecular liquids)

⇒ Random translation of the center of mass

⇒ Random rotations about the center of mass (random changes in the Euler angles)

⇒ Rotation of the coordinate system by the angles α, β, and γ

roll-pitch-yaw convention



⇒ Rotation matrix

$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix} = \begin{pmatrix} \cos\gamma\cos\beta\cos\alpha - \sin\gamma\sin\alpha & \cos\gamma\cos\beta\sin\alpha + \sin\gamma\cos\alpha & -\cos\gamma\sin\beta \\ -\sin\gamma\cos\beta\cos\alpha - \cos\gamma\sin\alpha & -\sin\gamma\cos\beta\sin\alpha + \cos\gamma\cos\alpha & \sin\gamma\sin\beta \\ \sin\beta\cos\alpha & \sin\beta\sin\alpha & \cos\beta \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

coordinates in new system — *coordinates in old system*

⇒ Random step

α = [2 *rand(0,1)-1*] $\Delta\alpha$

cos β = [2 *rand(0,1)-1*] Δ(cos β)

γ = [2 *rand(0,1)-1*] $\Delta\gamma$

⇒ 4 maximal displacement parameters Δr, $\Delta\alpha$, Δ(cos β), and $\Delta\gamma$ to be optimized

⇒ Alternative: quaternions (q_1,q_2,q_3,q_4) with Σq_i^2=1

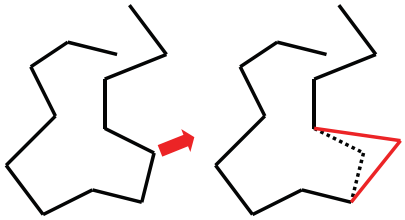
The Monte Carlo method

Moves for flexible molecules

⇒ Random translation of the center of mass

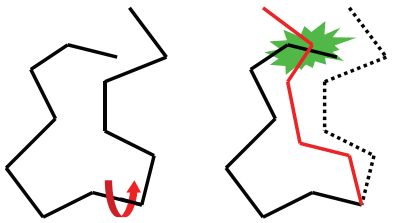
⇒ Random rotations about the center of mass

⇒ Other moves (may be difficult for macromolecules)



random atomic displacement

⇒ hard degrees of freedom (bond, angles) lead to high energies



torsional move (frozen bonds/angles)

⇒ high probability of steric clash

very small steps or more clever moves!

better idea: concerted torsions by the same angle

The Monte Carlo method

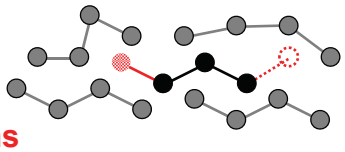
Advantages

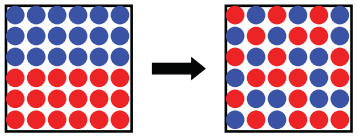
- Generates a Boltzmann-weighted ensemble of configurations (canonical or NVT ensemble)
  ⇒ *thermodynamic observables can be computed as*

$$X =< \mathrm{x}(\boldsymbol{r}) >= \frac{1}{N_{steps}} \sum_{k=1}^{N_{steps}} \mathrm{x}(\boldsymbol{r}(k))$$

X : *macroscopic observable*
x : *corresponding microscopic observable*

- Easily extended to other isothermal ensembles (NPT, μVT, ...)
- No need for derivatives of the potential energy function (forces)
  ⇒ *may be computationally less expensive*
  ⇒ *applicable to discontinuous potential energy functions*
- Efficient unphysical (but reversible) moves can be designed for improved sampling
  ⇒ *e.g.* *reptation in a dense polymer* *exchange of particles in a liquid mixture*





Limitations

- Non-deterministic (no constant of motion)
- No dynamic information (not if you are rigorous...)
- The design of random moves with a reasonable acceptance ratio may be difficult for some systems (*e.g.* macromolecules)

+ biased MC: configuration-biased or force-biased

Classical mechanics (Newton)

Newtonian formulation

- Valid in *Cartesian coordinates systems* only
- Newton's *second law for a particle* (point mass) *i*

$$\ddot{\boldsymbol{r}}_i(t) = m_i^{-1}\boldsymbol{F}_i(t)$$

equation of motion, 2nd order, no 1st order term

- *Conservation of linear (angular) momentum* in the absence of external force (torque)

$$\boldsymbol{F}_i = \boldsymbol{0} \;\Rightarrow\; \frac{d}{dt}(m_i\dot{\boldsymbol{r}}_i) = \boldsymbol{0} \qquad \boldsymbol{r}_i \times \boldsymbol{F}_i = \boldsymbol{0} \;\Rightarrow\; \frac{d}{dt}(m_i\boldsymbol{r}_i \times \dot{\boldsymbol{r}}_i) = \boldsymbol{0}$$

(with respect to any origin)

- *Potential energy* in conservative force fields (*e.g.* isolated molecular system)

$$\boldsymbol{F}(t) = \boldsymbol{F}(\boldsymbol{r}(t)) \;\text{ and }\; \oint d\boldsymbol{r}\cdot\boldsymbol{F}(\boldsymbol{r}) = 0 \;\Rightarrow\; \boldsymbol{F}(\boldsymbol{r}) = -\nabla V(\boldsymbol{r})$$

- *Energy conservation* in conservative systems

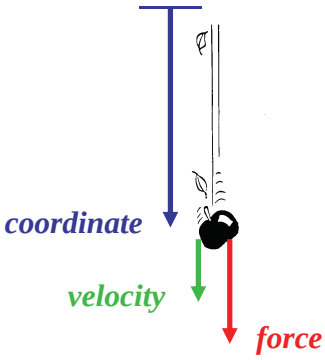
$$0 = \frac{d}{dt}\sum_i^N \int_{\boldsymbol{r}_i(t_o)}^{\boldsymbol{r}_i(t)} d\boldsymbol{r}_i \cdot [\nabla_i V(\boldsymbol{r}) + m_i\ddot{\boldsymbol{r}}_i] = \frac{d}{dt}[V(\boldsymbol{r}(t)) - V(\boldsymbol{r}(t_o)) + \sum_i^N \int_{t_o}^{t} d\tau\, \dot{\boldsymbol{r}}_i(\tau) m_i \ddot{\boldsymbol{r}}_i(\tau)]$$

$$= \frac{d}{dt}[V(\boldsymbol{r}(t)) + \sum_i^N \frac{1}{2} m_i \dot{\boldsymbol{r}}_i^2(t)] = \frac{d}{dt}[\underbrace{V(t)}_{\text{potential energy}} + \underbrace{K(t)}_{\text{kinetic energy}}]$$

⇒ ***classical dynamics for isolated systems generates a microcanonical (NVE) ensemble !***

- *Deterministic*, *i.e.* $\boldsymbol{r}(t_o), \dot{\boldsymbol{r}}(t_o),$ and $V(\boldsymbol{r})$ fully determine the future of the system
- *Time reversible*, *i.e.* the change $\dot{\boldsymbol{r}}(t) \rightarrow -\dot{\boldsymbol{r}}(t)$ implies $\boldsymbol{r}(t+\tau) = \boldsymbol{r}(t-\tau)$ for any τ

Classical mechanics (determinism)



The inventor of classical mechanics demonstrates the deterministic nature of his theory

Sir Isaac Newton
1642 -1727

Classical mechanics (N-body problem and determinism)

We ought to regard the present state of the universe as the effect of its antecedent state and as the cause of the state that is to follow. An intelligence knowing all the forces acting in nature at a given instant, as well as the momentary positions [and velocities] of all things in the universe, would be able to comprehend in one single formula the motions of the largest bodies as well as the lightest atoms in the world, provided that its intellect were sufficiently powerful to subject all data to analysis; to it nothing would be uncertain, the future as well as the past would be present to its eyes.

Pierre Laplace (1749-1827), “Philosophical essay on probabilities”

sounds like an early definition of classical MD ...

just that the powerful intelligence turned out to be a dumb computer !

Classical mechanics (Lagrange)

Lagrangian formulation

- A generalization of Newton's formulation for *generalized coordinates systems*
- Generalized coordinates:
  Any set of 3*N* scalars (*e.g.* distances, angles, dihedrals, …) *sufficient to specify the coordinates of all particles* in the system
  → 3*N*-dimensional *generalized coordinate vector* $\boldsymbol{q}$ so that $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{q})$
  → 3*N*-dimensional *generalized velocity vector* $\dot{\boldsymbol{q}}$ so that $\dot{\boldsymbol{q}} = d\boldsymbol{q}/dt$
- *Lagrangian function* of a system (no explicit time-dependence for an isolated system)

$$L(\boldsymbol{q},\dot{\boldsymbol{q}}) = K(\boldsymbol{q},\dot{\boldsymbol{q}}) - V(\boldsymbol{q})$$

potential energy / ***kinetic energy***

- *Lagrange equation of motion* for the generalized coordinates and velocities

$$\frac{d}{dt}\left(\frac{\partial L(\boldsymbol{q},\dot{\boldsymbol{q}})}{\partial \dot{\boldsymbol{q}}}\right) = \frac{\partial L(\boldsymbol{q},\dot{\boldsymbol{q}})}{\partial \boldsymbol{q}}$$

2nd order, no 1st order term

- Special case of a *Cartesian coordinate system*

$$L(\boldsymbol{r},\dot{\boldsymbol{r}}) = K(\dot{\boldsymbol{r}}) - V(\boldsymbol{r}) = \frac{1}{2}\dot{\boldsymbol{r}}^T \underline{\boldsymbol{M}}\dot{\boldsymbol{r}} - V(\boldsymbol{r})$$

mass matrix (diagonal, atomic masses as elements)

$$\Rightarrow \quad \frac{d}{dt}\left(\frac{\partial L(\boldsymbol{r},\dot{\boldsymbol{r}})}{\partial \dot{\boldsymbol{r}}}\right) = \underline{\boldsymbol{M}}\,\ddot{\boldsymbol{r}} = \boldsymbol{F} = -\frac{\partial V(\boldsymbol{r})}{\partial \boldsymbol{r}} = \frac{\partial L(\boldsymbol{r},\dot{\boldsymbol{r}})}{\partial \boldsymbol{r}}$$

Newton equation of motion

- Useful for (*i*) *enforcing constraints* or (*ii*) including *artificial dynamical degrees of freedom*

Classical mechanics (Hamilton)

Hamiltonian formulation

- An alternative to the Lagrange formulation for generalized coordinates systems
- Generalized coordinates $\boldsymbol{q}$
- Generalized momenta $\boldsymbol{p}$ defined as $\boldsymbol{p} = \dfrac{\partial L(\boldsymbol{q},\dot{\boldsymbol{q}})}{\partial \dot{\boldsymbol{q}}}$
- *Hamiltonian function* of a system (time independent for an isolated system)

$$H(\boldsymbol{q}, \boldsymbol{p}[,\dot{\boldsymbol{q}}]) = \boldsymbol{p}\cdot\dot{\boldsymbol{q}} - L(\boldsymbol{q},\dot{\boldsymbol{q}}) = \boldsymbol{p}\cdot\dot{\boldsymbol{q}} - K(\boldsymbol{q},\dot{\boldsymbol{q}}) + V(\boldsymbol{q})$$

- Hamiltonian variation

$$dH = \boldsymbol{p}\cdot d\dot{\boldsymbol{q}} + \dot{\boldsymbol{q}}\cdot d\boldsymbol{p} - \underbrace{\frac{\partial L(\boldsymbol{q},\dot{\boldsymbol{q}})}{\partial \boldsymbol{q}}}_{=\,\dot{\boldsymbol{p}}\ \textit{(Lagrange)}} d\boldsymbol{q} - \underbrace{\frac{\partial L(\boldsymbol{q},\dot{\boldsymbol{q}})}{\partial \dot{\boldsymbol{q}}}}_{=\,\boldsymbol{p}\ \textit{(definition)}} d\dot{\boldsymbol{q}} = \dot{\boldsymbol{q}}\cdot d\boldsymbol{p} - \dot{\boldsymbol{p}}\cdot d\boldsymbol{q}$$

$$\Rightarrow H = H(\boldsymbol{q}, \boldsymbol{p})$$

- *Hamilton equation of motion* for the generalized coordinates and momenta

$$\frac{\partial H(\boldsymbol{q},\boldsymbol{p})}{\partial \boldsymbol{p}} = \dot{\boldsymbol{q}} \quad \textit{and} \quad \frac{\partial H(\boldsymbol{q},\boldsymbol{p})}{\partial \boldsymbol{q}} = -\dot{\boldsymbol{p}}$$

two 1st order equations

- The Hamiltonian function for a system with kinetic energy depending quadratically on the generalized velocities (most cases) represents the *total energy of the system*

$$\textit{if} \quad K(\boldsymbol{q},\dot{\boldsymbol{q}}) = \sum_i c_i(\boldsymbol{q})\,\dot{q}_i^2 \quad \textit{then} \quad p_i = 2c_i(\boldsymbol{q})\,\dot{q}_i \quad \textit{and} \quad \boldsymbol{p}\cdot\dot{\boldsymbol{q}} = 2K(\boldsymbol{q},\dot{\boldsymbol{q}})$$

$$\Rightarrow H(\boldsymbol{q},\boldsymbol{p}) = \boldsymbol{p}\cdot\dot{\boldsymbol{q}} - K(\boldsymbol{q},\dot{\boldsymbol{q}}) + V(\boldsymbol{q}) = K(\boldsymbol{q},\boldsymbol{p}) + V(\boldsymbol{q})$$

potential energy / ***kinetic energy*** → ***total energy***

Hamiltonian formulation (continued)

- Special case of *a Cartesian coordinate system*

$$H(\boldsymbol{r},\boldsymbol{p}_r) = K(\boldsymbol{p}_r) + V(\boldsymbol{r}) = \frac{1}{2}\boldsymbol{p}_r^T\,\underline{\boldsymbol{M}}^{-1}\boldsymbol{p}_r + V(\boldsymbol{r})$$

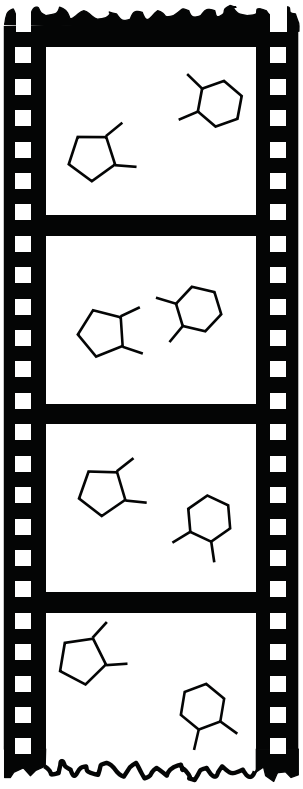
$$\Rightarrow \quad \frac{\partial H(\boldsymbol{r},\boldsymbol{p}_r)}{\partial \boldsymbol{p}_r} = \underbrace{\underline{\boldsymbol{M}}^{-1}\boldsymbol{p}_r = \dot{\boldsymbol{r}}}_{\substack{\textit{definition of}\\ \textit{the momentum}}} \quad \textit{and} \quad \frac{\partial H(\boldsymbol{r},\boldsymbol{p}_r)}{\partial \boldsymbol{r}} = \frac{\partial V(\boldsymbol{r})}{\partial \boldsymbol{r}} = \underbrace{-\boldsymbol{F} = -\dot{\boldsymbol{p}}_r}_{\substack{\textit{Newton}\\ \textit{equation of motion}}}$$

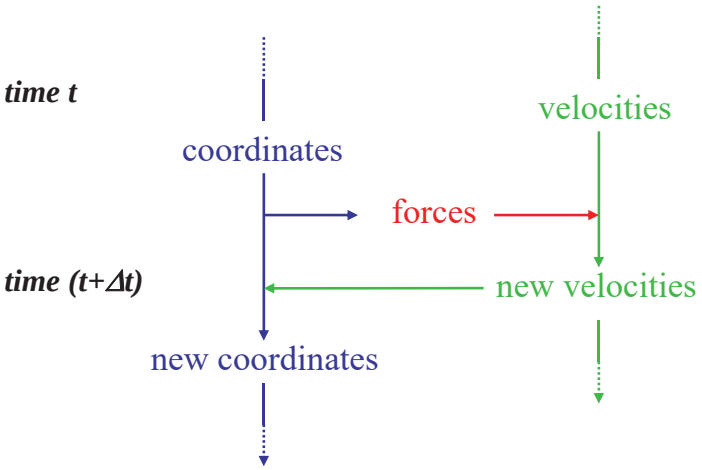
- Powerful formalism for
  - (*i*) *connecting quantum mechanics and classical mechanics* (Hamiltonian operator versus Hamiltonian function)
  - (*ii*) *derivations in statistical mechanics* (total energy is a central quantity)

★

Molecular dynamics (principle)

⇒ Making a „movie“ of a molecular system by integrating the classical equations of motion (usually Cartesian coordinate system → Newton formulation)





Molecular dynamics (integrator)

Integrator

⇒ algorithm to integrate the classical equations of motion based on a finite timestep Δt
e.g. Euler, Verlet, leap-frog, velocity-Verlet, Beeman, Gear,…

⇒ *Properties:*	*For molecular dynamics:*
Computational costs, memory requirement	*unimportant*
Number of force evaluations per step	*no more than one*
Accuracy for long timesteps	should be *good*
Momentum and energy conservation	should be *satisfied*
Time reversibility	should be *satisfied*
Compatible with thermostatting	should be *satisfied*

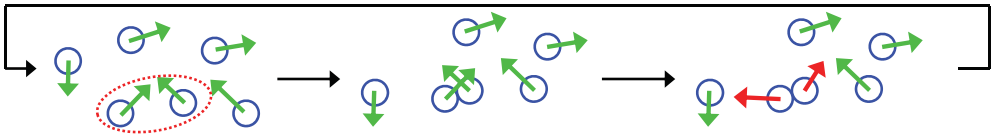
+ uniform in time (constant timestep)

A primitive integrator

⇒ Alder & Wainwright 1957 – first MD study of hard spheres in the condensed phase

⇒ Integrator

(sampling is not uniform in time; does not work for soft potentials)



locate next colliding pair and determine collision time

compute coordinates at collision time

update velocities of colliding particles (elastic choc)

★

Molecular dynamics (leap-frog)

The leap-frog integrator

⇒ Taylor expansion of coordinates and velocities at t+Δ (timestep Δ)

$$\boldsymbol{r}(t+\Delta)=\boldsymbol{r}(t)+\boldsymbol{v}(t)\Delta+\frac{1}{2}\boldsymbol{a}(t)\Delta^2+O[\Delta^3]$$

$$\boldsymbol{v}(t+\Delta)=\boldsymbol{v}(t)+\boldsymbol{a}(t)\Delta+\frac{1}{2}\boldsymbol{b}(t)\Delta^2+O[\Delta^3]$$

$\boldsymbol{b}(t)$ *acceleration derivative (no official name)*

⇒ Velocity propagation

$$\boldsymbol{v}(t+\tfrac{1}{2}\Delta)=\boldsymbol{v}(t)+\frac{1}{2}\boldsymbol{a}(t)\Delta+\frac{1}{8}\boldsymbol{b}(t)\Delta^2+O[\Delta^3] \quad +$$

$$\boldsymbol{v}(t-\tfrac{1}{2}\Delta)=\boldsymbol{v}(t)-\frac{1}{2}\boldsymbol{a}(t)\Delta+\frac{1}{8}\boldsymbol{b}(t)\Delta^2-O[\Delta^3] \quad -$$

$$\boldsymbol{v}(t+\tfrac{1}{2}\Delta)=\boldsymbol{v}(t-\tfrac{1}{2}\Delta)+\boldsymbol{a}(t)\Delta+O[\Delta^3]$$

⇒ Coordinate propagation

$$\boldsymbol{r}(t+\tfrac{1}{2}\Delta)=\boldsymbol{r}(t)+\frac{1}{2}\boldsymbol{v}(t)\Delta+\frac{1}{8}\boldsymbol{a}(t)\Delta^2+O[\Delta^3] \quad +$$

$$\boldsymbol{r}(t-\tfrac{1}{2}\Delta)=\boldsymbol{r}(t)-\frac{1}{2}\boldsymbol{v}(t)\Delta+\frac{1}{8}\boldsymbol{a}(t)\Delta^2-O[\Delta^3] \quad -$$

time shift by ½Δ

$$\boldsymbol{r}(t+\tfrac{1}{2}\Delta)=\boldsymbol{r}(t-\tfrac{1}{2}\Delta)+\boldsymbol{v}(t)\Delta+O[\Delta^3]$$

$$\boldsymbol{r}(t+\Delta)=\boldsymbol{r}(t)+\boldsymbol{v}(t+\tfrac{1}{2}\Delta)\Delta+O[\Delta^3]$$

leap-frog equations *(third-order accurate)*

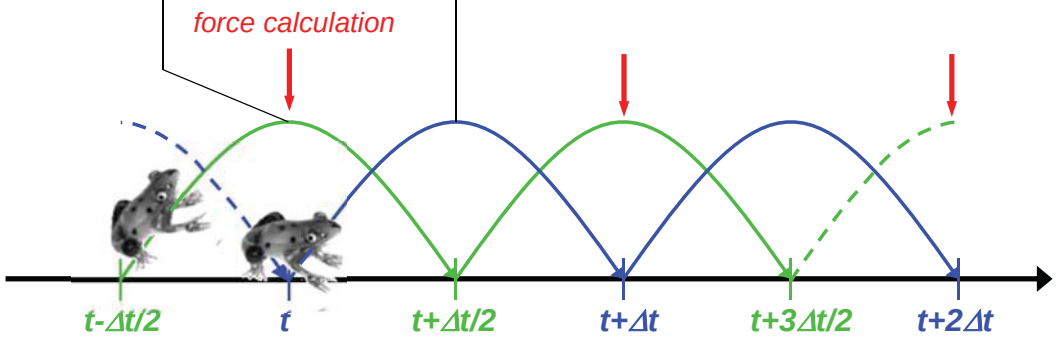
★★

Molecular dynamics (leap-frog)

The leap-frog integrator in practice

$$v(t+\frac{1}{2}\Delta t) = v(t-\frac{1}{2}\Delta t) + a(t)\Delta t$$

$$r(t+\Delta t) = r(t) + v(t+\frac{1}{2}\Delta t)\Delta t$$



Coordinates and velocities are not available simultaneously

⇒ to get the kinetic energy at time t, the corresponding velocities must be back-calculated

$$v(t) = \frac{1}{2}\left[v(t-\frac{1}{2}\Delta t) + v(t+\frac{1}{2}\Delta t)\right] + \frac{\Delta t}{16}\left[a(t-\Delta t) - a(t+\Delta t)\right] + O[(\Delta t)^4]$$

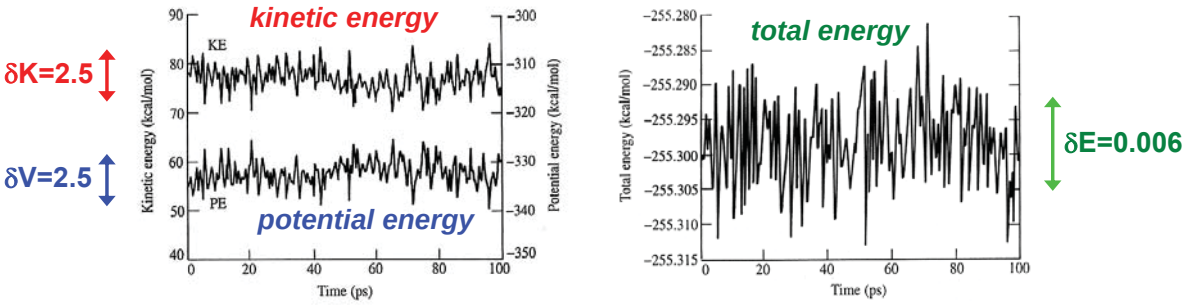
in practice: average the kinetic energy

★

Molecular dynamics (energy conservation)

Energy conservation

⇒ liquid argon (256 atoms), ρ=1.396 g/cm^3, T_{init}=100K, Velocity-Verlet, Δt=10 fs



⇒ the kinetic and potential energies fluctuate significantly

⇒ the total energy is essentially conserved

- in the absence of non-conservative forces (not always true in practice *e.g.* cutoff noise, constraints applied with finite tolerance !), it should be exactly conserved in the limit Δt→0
- at finite Δt, energy conservation (fluctuations, drift) is limited by the integration accuracy
  → intrinsic accuracy of the integrator (neglect of terms of O[(Δt)^n])
  → timestep size *e.g.* δE = 0.006 (Δt=10fs), δE = 0.002 (Δt=5fs), and δE = 0.040 (Δt=25fs)

Molecular dynamics (initial conditions)

Initial configuration

⇒ The choice of an initial configuration ***r***(0) is unimportant if the *equilibration time preceding the simulation* is longer than the *configurational relaxation time* for the system (*i.e.* the time required to loose its configurational "memory")

System type	*Relaxation time (indicative)*	
gas	~ 1 ps	***acceptable equilibration times***
pure liquid	~ 10 – 100 ps	
small organic molecule in solution	~ 10 ps – 1 ns	
short peptide in solution	~ 10 – 100 ns	
lipid aggregation in solution	~ 10 – 100 ns	
protein in solvent	~ 1 ms – 1 s	

⇒ When this is not the case, the initial (solute) configuration must be chosen carefully (*e.g.* experimentally available from X-ray crystallography or NMR structure determination)

Initial velocities

⇒ The exact choice of initial velocities ***v***(0) is generally unimportant (except for determining the initial system energy) – because the *velocity relaxation time* for most systems is short (~ps).

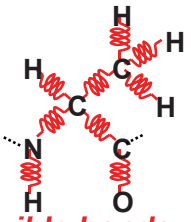
⇒ Initial velocities are generally assigned randomly from a Maxwell-Boltzmann distribution at a given temperature (statistical mechanics of NVT ensemble, system without constraints)

$$p(v_{ix}) = \left(\frac{m_i}{2\pi k_B T}\right)^{1/2} \exp\left(-\frac{m_i v_{ix}^2}{2k_B T}\right)$$

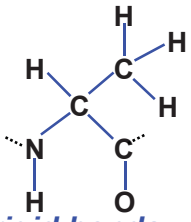
initial temperature

★

Molecular dynamics (bond constraints)



flexible bonds



rigid bonds

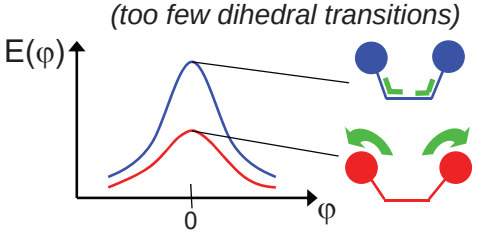
Note:
United-atom models also eliminate the aliphatic CH bonds!

→ Maximal frequency imposes Δt = 0.5-1 fs
→ Bond vibrations are weakly coupled to the system (poor exchange of kinetic energy)
→ Bond stretching frequencies (~1000-3000 cm^{-1}) are above k_BT / h (200 cm^{-1}) and are thus in the quantum-mechanical ground state (but: classical oscillators get k_BT equipartition energy!)

→ Maximal frequency now only imposes Δt = 2 fs
→ Narrower spectrum of frequencies
→ Constraints are associated with no kinetic energy / heat capacity / entropy (better representation of the quantum-mechanical state!)

Angle constraints? ⇒ not recommended (except in fully rigid molecules)

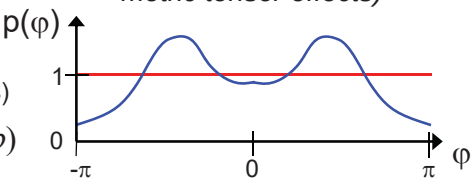
dynamic artifacts
(too few dihedral transitions)



thermodynamic artifacts
(wrong dihedral angle distribution - metric tensor effects)

(no interactions)

But: $K = K(\dot{\varphi}, \varphi)$

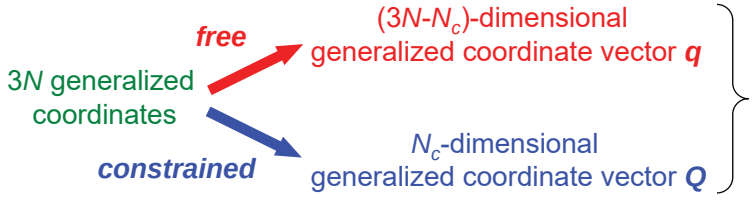


Molecular dynamics (constraints in generalized coordinates)

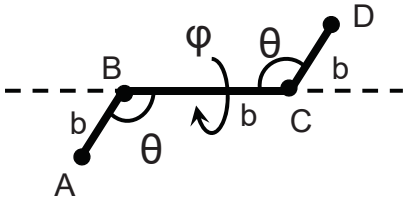
Application of constraints by Lagrangian dynamics

⇒ elegant but impractical for all but the smallest systems

⇒ principle



⇒ very simple example: united-atom butane with one torsional degree of freedom



- *points A,B, and C are fixed*
- *bond lengths and angles are constrained*

$$K(\dot\varphi;b,\theta)=\frac{1}{2}m(\dot\varphi\, b\sin\theta)^2$$

$$\frac{d}{dt}\left(\frac{\partial L(\varphi,\dot\varphi;b,\theta)}{\partial\dot\varphi}\right)=\frac{\partial L(\varphi,\dot\varphi;b,\theta)}{\partial\varphi}$$

$$\Rightarrow\quad \ddot\varphi=-(mb^2\sin^2\theta)^{-1}\frac{dV(\varphi)}{d\varphi}$$

can be integrated in time given $V(\varphi)$, $\varphi(0)$, and $\dot\varphi(0)$

⇒ Lagrangian and derivatives become quickly very complex for large systems !

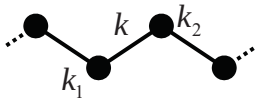
⇒ Lagrangian dynamics is difficult to apply for all but the smallest molecular systems…

Molecular dynamics (constraints in Cartesian coordinates)

Leap-frog + effect of constraints

⇒ constraint equations

$$\sigma_k(\{\boldsymbol{r}_i\})=r_{k_1k_2}^2-d_{k_1k_2}^2=0\quad,\quad k=1\ldots N_c$$



⇒ Lagrange's method of undetermined multipliers

$$m_i\frac{d^2\boldsymbol{r}_i(t)}{dt^2}=-\frac{\partial}{\partial\boldsymbol{r}_i}\left[V(\{\boldsymbol{r}_i\})+\sum_{k=1}^{N_c}l_k(t)\sigma_k(\{\boldsymbol{r}_i\})\right]$$

We add a "zero-term" to the potential energy

⇒ Constraint forces

$$\boldsymbol{F}_i^{\,c}(t)=-\sum_{k=1}^{N_c}l_k(t)\frac{\partial\sigma_k(\{\boldsymbol{r}_i(t)\})}{\partial\boldsymbol{r}_i(t)}=-2\sum_{k=1}^{N_c}l_k(t)(\delta_{ik_1}-\delta_{ik_2})\boldsymbol{r}_{k_1k_2}(t)$$

For each constraint k, a constraint force acts on the connected atoms k1 and k2, and is along their connecting vector

⇒ Unconstrained step (leap frog)

$$\boldsymbol{r}_i^{\,uc}(t+\Delta t)=\boldsymbol{r}_i(t)+\boldsymbol{v}_i(t-\Delta t/2)\Delta t+(\Delta t)^2\,m_i^{-1}\boldsymbol{F}_i^{\,uc}(t)$$

⇒ Effect of the constraints

$$\boldsymbol{r}_i^{\,c}(t+\Delta t)=\boldsymbol{r}_i^{\,uc}(t+\Delta t)+(\Delta t)^2\,m_i^{-1}\boldsymbol{F}_i^{\,c}(t)$$

⇒ Coordinates after constraining should satisfy constraints

Constraint equations only fulfilled at full timesteps!

$$\left[\boldsymbol{r}_{k_1k_2}^{uc}(t+\Delta t)+(\Delta t)^2\,m_{k_1}^{-1}\boldsymbol{F}_{k_1}^{\,c}(t)-(\Delta t)^2\,m_{k_2}^{-1}\boldsymbol{F}_{k_2}^{\,c}(t)\right]^2-d_{k_1k_2}^2=0\quad,\quad k=1\ldots N_c$$

⇒ That is...

$$\left\{\boldsymbol{r}_{k_1k_2}^{uc}(t+\Delta t)-2(\Delta t)^2\sum_{k'=1}^{N_c}l_{k'}(t)\,\boldsymbol{r}_{k_1'k_2'}(t)\left[m_{k_1}^{-1}(\delta_{k_1k_1'}-\delta_{k_1k_2'})+m_{k_2}^{-1}(\delta_{k_2k_2'}-\delta_{k_2k_1'})\right]\right\}^2-d_{k_1k_2}^2=0\ ,\quad k=1\ldots N_c$$

Molecular dynamics (SHAKE)

Solving the system of equations: the SHAKE procedure

$\Rightarrow$ set of N_c coupled non-linear equations must be solved (tricky !)

$$\left\{ \boldsymbol{r}^{uc}_{k_1k_2}(t+\Delta t) - 2(\Delta t)^2 \sum_{k'=1}^{N_c} l_{k'}(t)\, \boldsymbol{r}_{k_1'k_2'}(t) \left[m_{k_1}^{-1}(\delta_{k_1k_1'} - \delta_{k_1k_2'}) + m_{k_2}^{-1}(\delta_{k_2k_2'} - \delta_{k_2k_1'}) \right] \right\}^2 - d^2_{k_1k_2} = 0 \ , \quad k = 1 \ldots N_c$$

$\Rightarrow$ *Approximation 1* : assume that the constraints are uncoupled (only retain *k*=*k'*)

$$\left\{ \boldsymbol{r}^{uc}_{k_1k_2}(t+\Delta t) - 2(\Delta t)^2\, l_k(t)\, \boldsymbol{r}_{k_1k_2}(t) \left[m_{k_1}^{-1} + m_{k_2}^{-1} \right] \right\}^2 - d^2_{k_1k_2} = 0 \ , \ k = 1 \ldots N_c$$

$\Rightarrow$ *Approximation 2*: linearize the equation (neglect terms in l^2)

$$l_k(t) = \frac{\left[\boldsymbol{r}^{uc}_{k_1k_2}(t+\Delta t) \right]^2 - d^2_{k_1k_2}}{4(\Delta t)^2 \left[m_{k_1}^{-1} + m_{k_2}^{-1} \right] \boldsymbol{r}_{k_1k_2}(t) \cdot \boldsymbol{r}^{uc}_{k_1k_2}(t+\Delta t)} \ , \ k = 1 \ldots N_c$$

$\Rightarrow$ Coordinate resetting

$$\boldsymbol{r}^{c}_{k_1}(t+\Delta t) = \boldsymbol{r}^{uc}_{k_1}(t+\Delta t) - 2(\Delta t)^2\, m_{k_1}^{-1} l_k(t) \boldsymbol{r}_{k_1k_2}(t)$$

$$\boldsymbol{r}^{c}_{k_2}(t+\Delta t) = \boldsymbol{r}^{uc}_{k_2}(t+\Delta t) + 2(\Delta t)^2\, m_{k_2}^{-1} l_k(t) \boldsymbol{r}_{k_1k_2}(t)$$

The atoms are moved in the direction of the original bonds

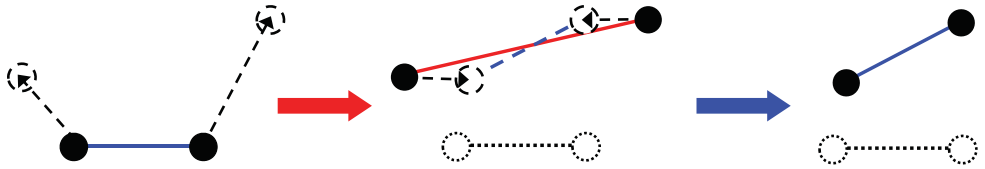
★★

Molecular dynamics (SHAKE)

The SHAKE procedure in practice

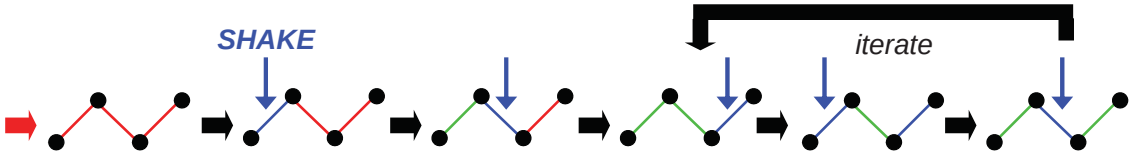
The lightest atom moves most; satisfy Newton's third law

$\Rightarrow$ for a single bond



free-flight step (unconstrained) — *coordinate resetting (SHAKE)*

$\Rightarrow$ for multiple bonds: iterate until all constraints are fulfilled within some relative tolerance



after free-flight step

relative geometric tolerance e.g. 10^{-4}

$\Rightarrow$ velocities must be corrected to remove the component along the constraints

$$\boldsymbol{v}(t + \frac{1}{2}\Delta t) = [\boldsymbol{r}(t+\Delta t) - \boldsymbol{r}(t)] / \Delta t$$

(conserves total linear and angular momenta)

$\Rightarrow$ forces must be corrected to include the constraint forces

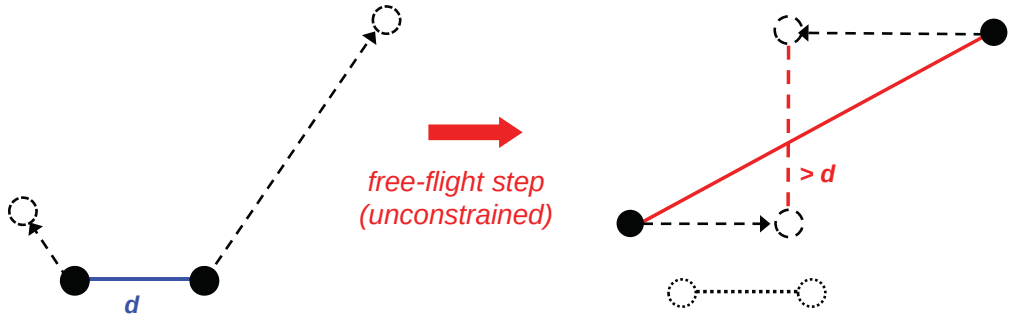
$$\boldsymbol{F}(t) = \boldsymbol{F}^{uc}(t) + \underline{\boldsymbol{M}}[\boldsymbol{r}(t+\Delta t) - \boldsymbol{r}^{uc}(t+\Delta t)] / (\Delta t)^2$$

(closely related to the Lagrangian multipliers)

Molecular dynamics (SHAKE)

SHAKE failures

⇒ If (one of) the atoms move too much in a single timestep, then SHAKE does not converge!



⇒ This is usually an (early) indication of a **bad simulation setup** (rather than a problem with SHAKE)

⇒ Shake failures frequently occur after starting a simulation from a high-energy configuration, e.g. by skipping energy minimization of after changing the force field

Molecular dynamics (constraints)

Constraint versus *restraint*

⇒ two very different concepts !

restraint:
- potential+kinetic energy
- k_BT equipartition energy
- frequency included in dynamics (→timestep!)

constraint:
- no energy contribution (neither potential nor kinetic)
- metric tensor effects
- no vibration

Other type of constraints

⇒ angle constraints should be avoided in flexible molecules with bond constraints

⇒ in fully rigid molecules (N atoms) the geometry is generally enforced by $3N$-6 distance constraints

⇒ SHAKE can be generalized to other type of internal coordinates (*e.g.* angle, dihedral angle, radius of gyration, box dipole moment, RMSD, …)

Alternative to SHAKE

⇒ M-SHAKE (matrix inversion)

⇒ SETTLE (analytical non-iterative version for triatomic molecules)

⇒ RATTLE (SHAKE analog for the velocity-Verlet integrator)

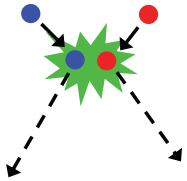
⇒ LINCS (variation of SHAKE)

Molecular dynamics (timestep)

Choice of a timestep

reasonable

too short
→ poor sampling



too large
→ poor energy conservation
→ program overflow or failure

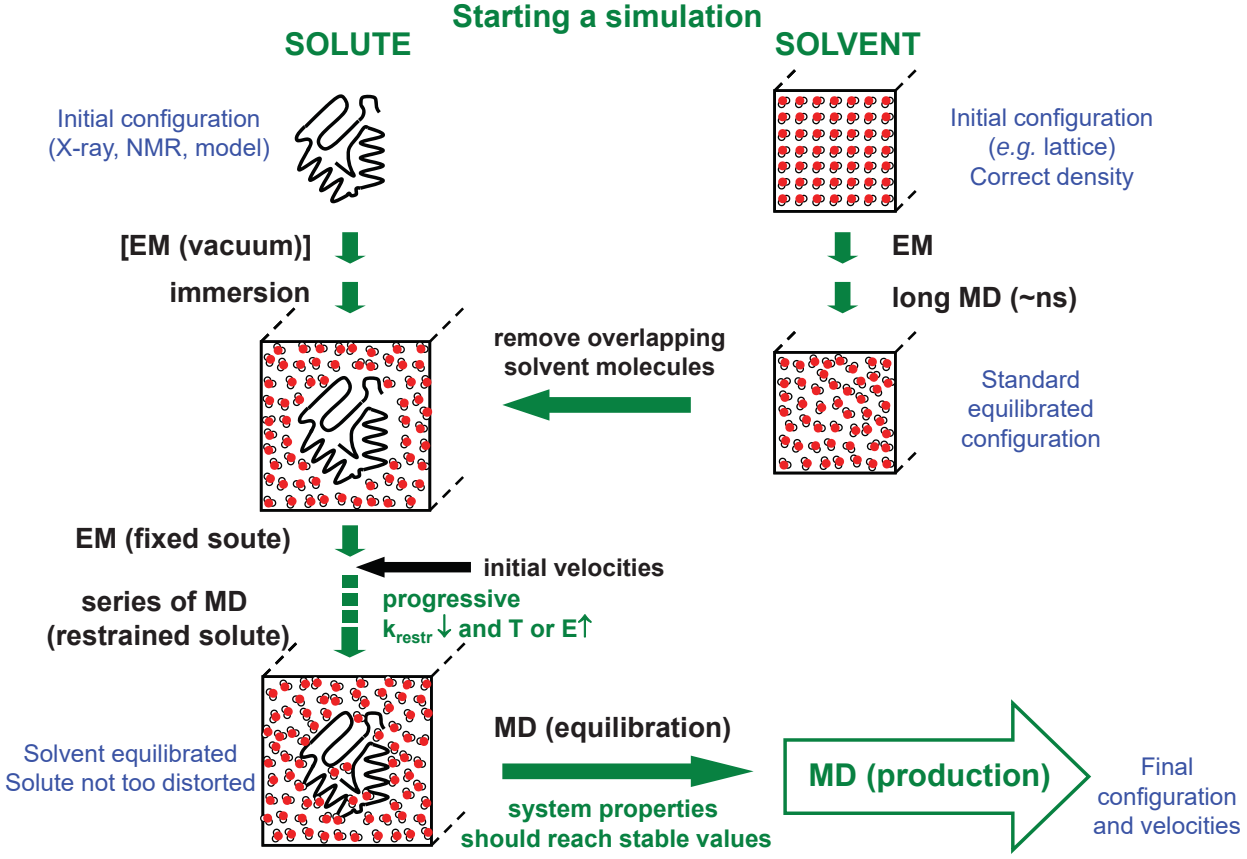
⇒ ***Rule of thumb:*** $\Delta t = \tau / 10$

where τ is the period of the fastest motion in the system

System	*Motions*	*Timescale*	*Timestep*
Atomic liquid (*e.g.* argon)	translation, vibration at contact	≤ 1000 fs (from LJ curve)	10 fs
Molecular liquid (rigid molecules)	*idem* + rotation (libration)	≤ 500 cm^{-1} (μ-wave)	5 fs
Flexible molecules (rigid bonds)	idem + torsion + bond-angle vibration	≤ 2000 cm^{-1} (IR)	2 fs
Flexible molecules	*idem* + bond-stretching vibration	≤ 3000 cm^{-1} (IR; for C-H)	0.5-1 fs

Molecular dynamics (setup)

Starting a simulation



Stochastic dynamics

Application

⇒ MD simulations with an **implicit** representation of the solvent (computationally cheaper)

Explicit solvent (MD)

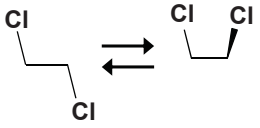
~10 water molecules per solute atom

Implicit solvent (SD)

solvent is only interesting for its mean effect on the solute

Solvent effect on the solute

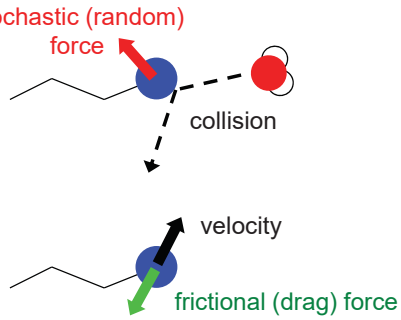
e.g.



77% 23% (gas phase)
44% 56% (liquid)

SOLVENT

- **mean effect (electrostatic+hydrophobic)**
- **stochastic collisions (often assumed uncorrelated in space and time)**
- **frictional drag (velocity- and possibly coordinate-dependent)**



Stochastic dynamics

Langevin equation of motion

equation of motion, 2nd order with 1st order term

$$m_i\ddot{\boldsymbol{r}}_i = \boldsymbol{F}_i^{mean}\big(r(t)\big) + \boldsymbol{R}_i(t) - m_i\gamma_i\dot{\boldsymbol{r}}_i(t)$$

mean force | **stochastic force** | **frictional force**

- incorporates the mean effect of the omitted solvent
- defines a potential of mean force (PMF)

$$\boldsymbol{F}_i^{mean}(\boldsymbol{r}) = -\frac{\partial V^{mean}(\boldsymbol{r})}{\partial \boldsymbol{r}_i}$$

V^{mean}: is actually a free energy

- accurate solvent-averaged PMFs are very difficult to determine (models are always more or less empirical...)

- Gaussian probability distribution

$$p(R_\alpha) = (2\pi\sigma_\alpha)^{-1/2}\exp(\frac{-R_\alpha^2}{2\sigma_\alpha})$$

- common approximations (for components R_α and R_β of the *3N*-dimensional vector **R**)

  - zero average $< R_\alpha(t) >= 0$

  - no correlation in space and time

  $$< R_\alpha(0)R_\beta(t) >= \sigma_\alpha\delta_{\alpha\beta}\delta(t)$$

  - no correlation with previous atomic velocities

  $$< \dot{r}_\alpha(0)R_\beta(t) >= 0 \quad for \quad t \geq 0$$

- the collision frequencies γ_i (unit: inverse time) may be assumed constant (γ) or depend on **r** (*e.g.* proportional to solvent accessibility of atom i)

- the value of γ may be estimated from Einstein's or Stokes' (~spherical solvent) laws for the pure solvent

$$m_{slv}\gamma = \frac{k_B T}{D_{slv}} \text{ or } m_{slv}\gamma = 6\pi\eta_{slv}a_{slv}$$

D_{slv} : diffusion constant
η_{slv} : viscosity
a_{slv} : radius
m_{slv} : mass

Stochastic dynamics

Steady-state balance

fluctuation-dissipation theorem

⇒ The stochastic forces introduce energy
⇒ The frictional forces remove energy
} balanced when $\sigma_\alpha = 2m_\alpha k_B T_o \gamma_\alpha$

⇒ although **plain MD** (no thermostat) samples a **microcanonical** (NVE) ensemble, **SD** samples a **canonical** (NVT) ensemble at a temperature T_o determined by the choice of the γ_α and σ_α

⇒ stochastic and frictional forces influence the **dynamics** of the system, but **not its thermodynamics** (compared *e.g.* to thermostatized MD; as long as the stochastic and frictional forces are non-zero and finite...)

Brownian dynamics

⇒ Limiting case for high viscosity

⇒ Neglect the inertial term in the Langevin equation

$$m_i \ddot{\boldsymbol{r}}_i = \boldsymbol{F}_i^{mean}(\boldsymbol{r}(t)) + \boldsymbol{R}_i(t) - m_i \gamma_i \dot{\boldsymbol{r}}_i(t) = \boldsymbol{0}$$

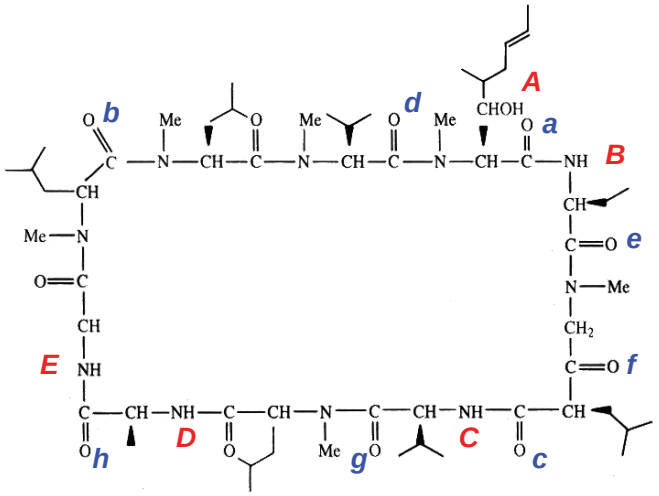
$$\Rightarrow \quad m_i \dot{\boldsymbol{r}}_i = \gamma_i^{-1} [\, \boldsymbol{F}_i^{mean}(\boldsymbol{r}(t)) + \boldsymbol{R}_i(t) \,]$$

equation of motion, 1st order

Stochastic dynamics

Example:

- Simulations of cyclosporin A (CPA) - cyclic undecapeptide with immunosuppressive propeties
- MD: CPA in vacuum / CPA + 632 H_2O / CPA + 591 CCl_4
- SD: CPA in vacuum / CPA + γ=91 ps^{-1} (H_2O) / γ=24 ps^{-1} (CCl_4) [about 10 times less expensive]
- At 300K, 1atm, GROMOS87 force field, 40 ps



MeBmt-Abu-Sar-MeLeu-Cal-MeLeu-Ala-D-Ala-MeLeu-MeLeu-MeVal

Stochastic dynamics

H-bonds: [% occurrence along the simulations]

Donor-acceptor		MD vacuum	MD CCl_4	SD CCl_4	SD H_2O	MD H_2O
A	a	84	87	73	11	0
A	b	0	0	0	8	0
B	c	61	30	31	68	23
B	d	31	58	56	40	28
C	e	88	65	72	96	72
C	f	0	0	6	1	0
C	c	0	0	5	3	0
D	c	36	49	54	20	0
D	d	89	49	44	67	54
E	g	69	60	67	78	11
E	h	2	0	0	0	0
Mean		42	36	37	36	17

GOOD ***BAD***

Stochastic dynamics

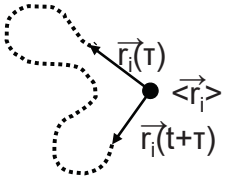
Explanation:

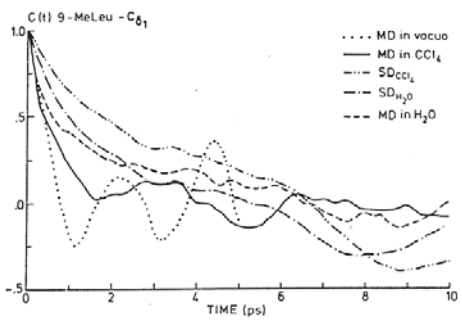
The simulations use the same force field, calibrated for use with explicit-solvent simulations
⇒ In the SD simulations, the mean effect of the solvent is not included

CCl4: mean effect of solvent on H-bonds is negligible
⇒ MD and SD give comparable results

H2O: solvent competes efficiently with intramolecular H-bonds
⇒ SD and MD/vacuum give too strong intramolecular H-bonds

Dynamical properties:





position autocorrelation function

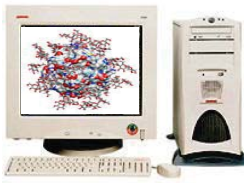
MD/vacuum: oscillatory (harmonic vibrations)

All others: damped random fluctuations

Computer Simulation in Chemistry, Biology and Physics

P.H. Hünenberger

COMPUTER SIMULATION OF MOLECULAR SYSTEMS

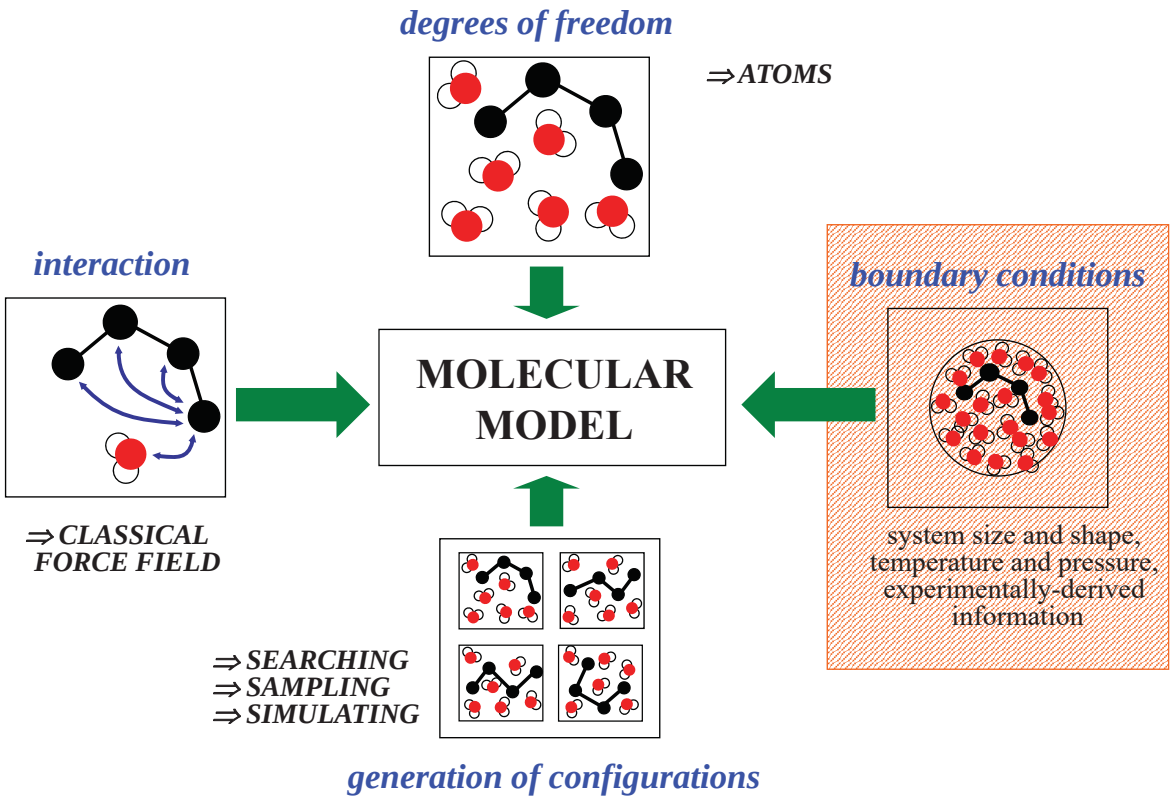


Herbstsemester 2019
Tuesday 9:45-11:30 a.m.
HCI D2

Lecture 529-0004-00
www.csms.ethz.ch/education/CSCBP

LECTURE 4 (WEEK 5):
Boundary conditions

Four basic choices defining a molecular model



Types of boundary conditions

The term *boundary condition* applies to any *global constraint* (*i.e.* concerning the whole system) enforced during a simulation

A *hard* boundary condition is a constraint imposed on an *instantaneous observable* (*i.e.* enforced at all times during the simulation)

A *soft* boundary condition is a constraint imposed on an *average observable* (*i.e.* enforced on its average value over a given timescale τ).

BOUNDARY CONDITIONS

SPATIAL	***THERMODYNAMIC***	***GEOMETRIC***
(→ simulated sample)	(→ simulated ensemble)	(→ internal coordinates)
• size (volume)	• constant E (H,L,R) or constant T	• bond-length constraints
• shape	• constant V or constant P	• structure factors (X-ray)
• surroundings	• constant N or constant μ	• inter-proton distances (NMR)
• boundary	• constant linear momentum	• J-coupling constants (NMR)
• buffer region	• constant angular momentum	• …

■ ***if applied, should be hard***
■ ***if applied, should be soft***

Spatial boundary conditions

⇒ size, shape, surroundings, boundary and buffer region of the simulated sample

The *sample size* is often *prescribed by the system of interest* (*e.g.* size of a biomolecule, need for adequate solvation, size of a crystallographic unit cell), in balance with the *available computational resources* (limit at ~10^4-10^5 atoms)

⇒ when simulating *condensed-phase systems*, one generally wants to model *macroscopic (bulk) properties* by simulation of a *truly microscopic system*

Problem:

MICROSCOPIC SAMPLE

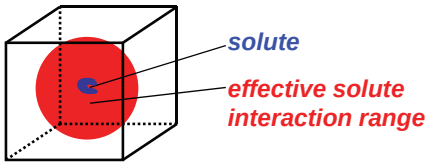
one of the major plagues in molecular simulations…

FINITE-SIZE EFFECTS **SURFACE EFFECTS**

→ ***Simulated properties may heavily depend on specific choices made in spatial boundary conditions***

Finite-size effects in microscopic samples

Macroscopic sample | ***Microscopic (simulated) sample***



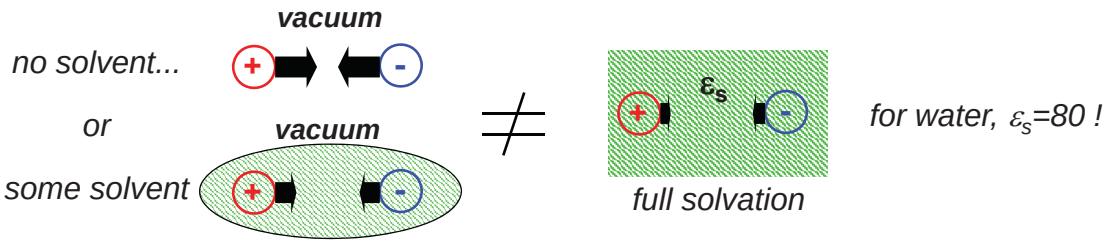
1 l water
→ interaction range is entirely within sample ➡ *bulk properties do not depend on sample size and shape...*

8×8×8 nm³ water
→ interaction range goes beyond sample ➡ *finite-size effects...*

finite-size effects are dominated by electrostatics (longest range)

e.g. to get the hydration free energy of a Na^+ ion within k_BT (2.5 kJ·mol^{-1}) requires a droplet of ~28 nm radius (~3'000'000 molecules) !!!

⇒ Surroundings as vacuum → lack of dielectric screening

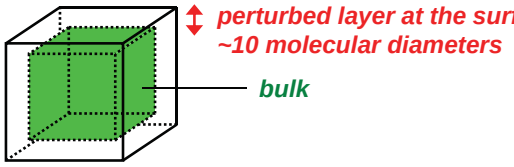


for water, ε_s=80 !

→ electrostatic interactions (ion pairing, hydrogen bonds, ...) are overweighted in the sample

Surface effects in microscopic samples

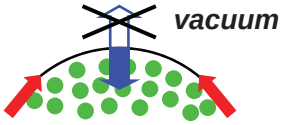
Macroscopic sample | ***Microscopic (simulated) sample***



1 l water, ~3.3·10²⁵ molecules
→ ~2.8·10¹⁸ perturbed molecules
***i.e.* ~1 out of 10'000'000 (negligible)**

8×8×8 nm³ water, ~17'000 molecules
→ ~12'400 perturbed molecules,
***i.e.* ~75% (majority !)**

⇒ Interface to vacuum → surface tension effects

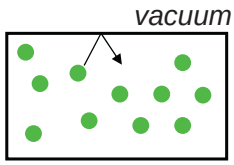


→ tendency to minimize the surface area (sphere)
→ increased pressure in the sample (compactness)
→ effects become ***very large*** for microscopic samples

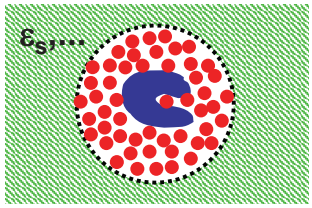
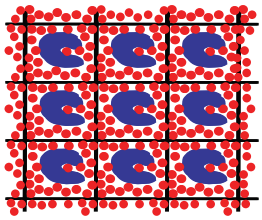
⇒ Interface explicit-solvent to vacuum

→ solvent evaporation
→ inhomogeneous solvent distribution at surface (≠ bulk)
→ preferential orientation of solvent dipoles (≠ bulk)

Spatial boundary conditions in the absence of explicit solvation

	Container with hard (reflecting) walls	Vacuum boundary conditions	Implicit-solvent model (polar & non-polar)
			
finite-size effects:	large	large	no (in principle)
surface effects:	large	large	reduced
sampling:	MD	SD [collisions, drag, ***no*** mean solvation]	SD [collisions, drag mean solvation]
comment:	OK for gases close to ideality	→ spherical shape → compactness → too strong electrostatics ***NOT RECOMMENDED !***	difficult to design and parameterize, parameter-sensitive, cheap ***BETTER...***

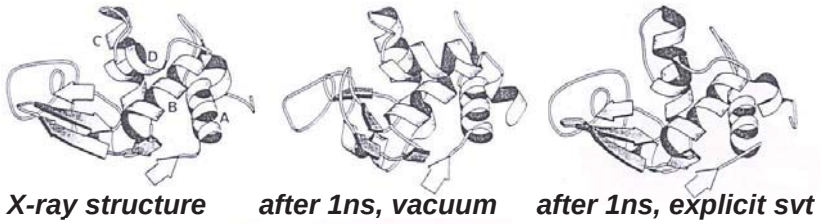
Spatial boundary conditions in the presence of explicit solvation

	Solvent droplet (or layers) in vacuum	Droplet with implicit-solvent model	Periodic boundary conditions (PBC)
			
finite-size effects:	reduced	no (in principle)	reduced
surface effects:	yes (solvent-vacuum)	reduced	eliminated !
sampling:	MD (evtl. stochastic boundary)	MD (evtl. stochastic boundary)	MD
comment:	→ solvent evaporation → inhomogeneous svt distribution at surface → preferential svt dipole orientation at surface → finite-size effects ***RATHER CRUDE !***	difficult to design and parameterize, parameter-sensitive	→ anisotropy → artificial periodicity → lots of solvent → finite-size effects

MOST POPULAR BOUNDARY CONDITIONS

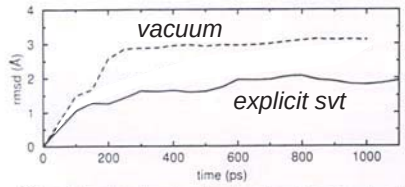
Distortive effects of vacuum boundary conditions

e.g. hen-egg-white lysozyme (protein, 129 residues)



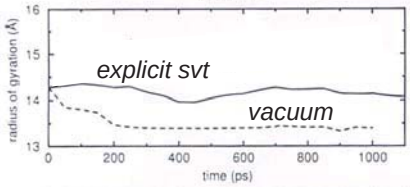
X-ray structure | *after 1ns, vacuum* | *after 1ns, explicit svt (PBC)*

deviation from X-ray structure



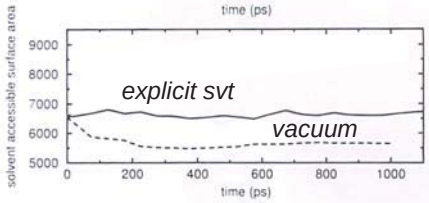
⇒ larger deviations from experiment

radius of gyration



⇒ too compact shape

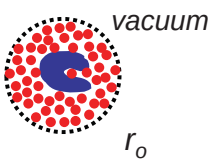
solvent-accessible surface area



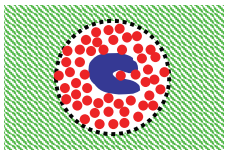
⇒ more spherical shape, reduced surface area

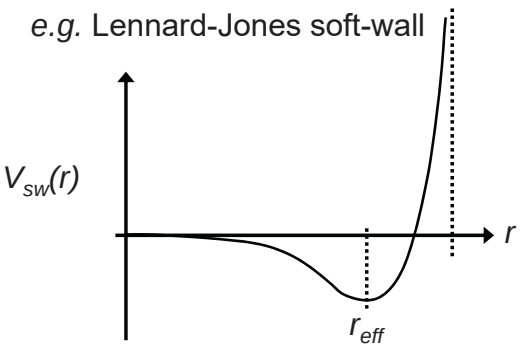
⇒ ***+ too stable salt bridges***
+ too stable H-bonds
+ slower dynamics

Soft-wall potential



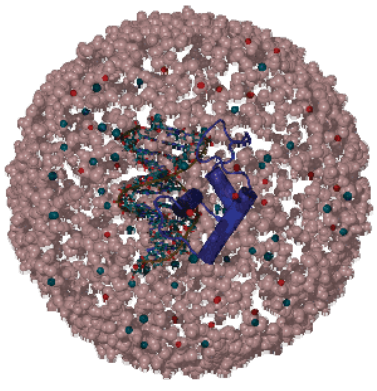
or





$$V_{sw}(r;r_o) = C_{12}(r_o - r)^{-12} - C_6(r_o - r)^{-6}$$

LJ parameters for water-water interaction

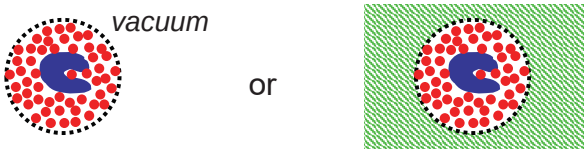


e.g. Lac headpiece-DNA complex

⇒ Confines sample and prevents solvent evaporation
⇒ Mimics dispersion interactions with omitted solvent

⇒ Does not prevent inhomogeneous solvent distribution at surface
⇒ Does not prevent preferential solvent dipole orientation at surface

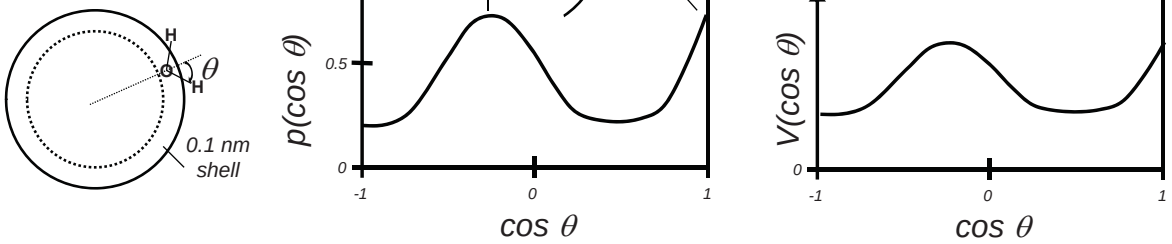
Surface dipole orientational potential



e.g. water droplet
+ soft-wall
+ implicit svt outside

⇒ *undesired effects*

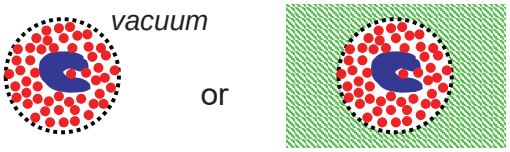
⇒ *correction potential* (applied to molecules in surface shell)

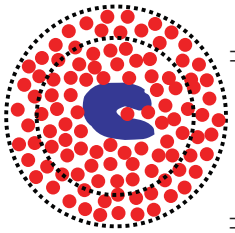


⇒ Prevents preferential solvent dipole orientation at surface
⇒ Partially prevents inhomogeneous solvent distribution at surface

⇒ Rather *ad hoc*
⇒ Poorly transferable (takes time to calibrate; specific to one continuum model; has to be redone for different droplet sizes or solutes in the droplet [especially if charged])

Extended wall and stochastic boundary conditions

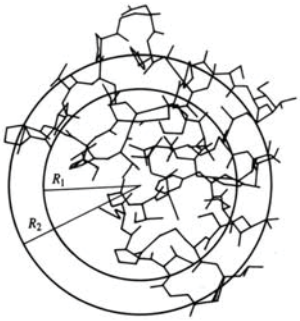




⇒ *Extended-wall (buffer) region*
→ atoms/molecules are positionally restrained

⇒ Possibly *stochastic boundary conditions* in the extended-wall region
→ apply Langevin equation (SD) rather than Newton equation (MD)

⇒ Possibly apply *surface dipole-orientational potential* in the extended-wall region



e.g. simulation of an enzyme active site (surrounded by vacuum or fixed protein atoms and possibly solvent)

⇒ Confines sample and prevents solvent evaporation
⇒ Partly remedies inhomogeneous solvent distribution at surface
⇒ Partly remedies preferential solvent-dipole orientation at surface (with dipole-orientational potential)
⇒ Introduces thermostating through the surface (with stochastic boundary conditions; microcanonical→canonical)

⇒ Difficult to calibrate

Implicit-solvent models

For solutes of arbitrary shapes:

⇒ Electrostatic component
- → distance-dependent dielectric "constant"
- → numerical solution of the Poisson(-Boltzmann) equation, generalized Born (GB) model, mean-spherical approximation (MSA), or integral equations (RISM)
- → Langevin dipoles

⇒ Non-polar component (van der Waals + hydrophobic)
- → often related to the total solvent-exposed surface area (SASA) of the solute
- → sometimes empirically related to the solvent-exposed surface area sorted by the type of contributing atoms

⇒ To be used together with SD
- → stochastic collision
- → drag force

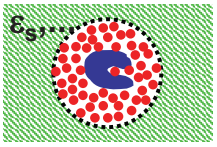
sometimes made proportional to the solvent-exposed surface area of each atom

- → same equilibrium properties, improved dynamic properties

KEY ISSUES: validity of macroscopic continuum electrostatics at the microscopic level, exact location of solute-solvent boundary, atomic parameters (charges, radii, surface-area parameters), solute dielectric constant, surface-area coefficient ...

Implicit-solvent models

For a spherical droplet:



[hide the boundary problem from the solute – but: problem remains at the solvent-vacuum interface]

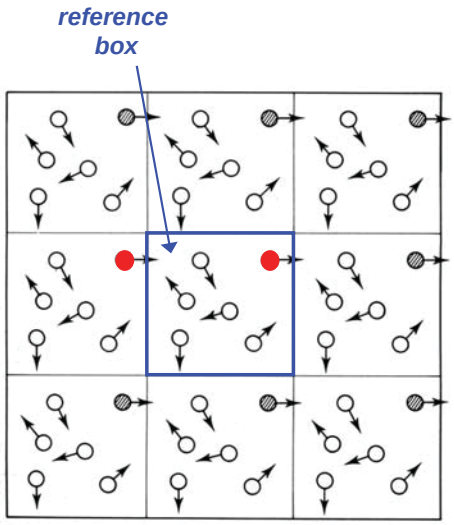
⇒ Electrostatic component
- → Kirkwood reaction field (spherical droplet)
- → Friedman image-charge method

⇒ Non-polar component
- → soft-wall potential
- → extended-wall region
- → boundary dipole-orientation potential

⇒ To be used together with MD
- → possibly: stochastic boundary conditions

KEY ISSUES: same as before (+singularity at boundary)

Periodic boundary conditions



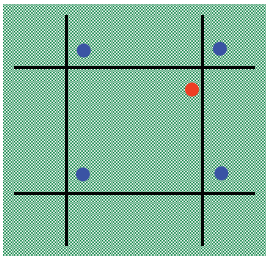
- The simulated system (solute + solvent) consists of particles within a reference computational box of space-filling shape (*e.g.* cube)
- At each simulation step, particles exiting the box through one face are translated so that they reenter the box through the opposing face
- This procedure mimics a system consisting of an infinite lattice of periodic copies of the reference box (→no interface to vacuum !)
- Only the coordinates of particles in the central box are actually stored in the computer

Achille kindly illustrates the concept of periodic boundary conditions

Periodic boundary conditions

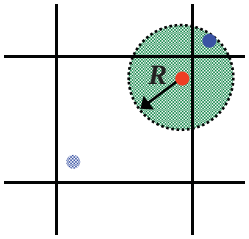
In principle, each particle in the reference box has non-bonded interactions with all other particles in the reference box and all their replicas in the periodic system

→ interaction may be evaluated using *lattice-sum methods* (Fourier series)



In practice, the non-bonded interaction is often truncated at a certain *cutoff distance R*

→ interaction are evaluated using a *double sum* over particles with a distance smaller than *R* (in general smaller than the half box edge; selects a subset of the minimum-image pairs)



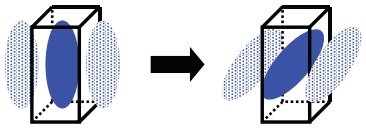
Covalent interactions are short-ranged and only act between minimum images

(minimum-image pair: atom and the closest periodic replica of another one)

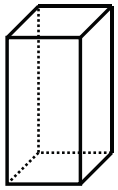
Periodic boundary conditions

Common box shapes (space-filling):

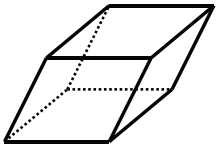
- Rectangle: for elongated macromolecules, but watch out if the molecule rotates



Fix: roto-translational constraints

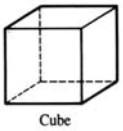


Rectangular



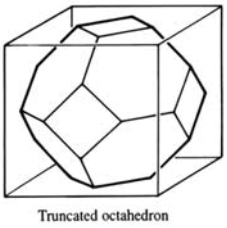
Triclinic

- Hexagonal prism: idem (may be used for DNA)
- Cube: isotropic, but requires a lot of solvent



Cube

- Truncated octahedron: almost isotropic, and requires less solvent for spherical molecules



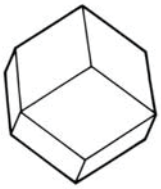
Truncated octahedron

$V=L^3$
$R_i=L/2$
$\Rightarrow V / V_i \approx 1.9$

$V=L^3/2$
$R_i=\sqrt{3}L/4$
$\Rightarrow V / V_i \approx 1.5$

needs 20% less solvent

Hexagonal prism



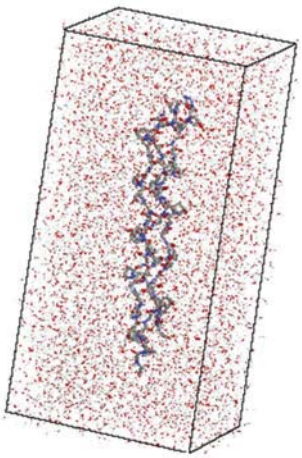
Rhombic dodecahedron

- Triclinic: for crystal simulations and, more recently, for implementing any (optimal) box shape into a single simulation code

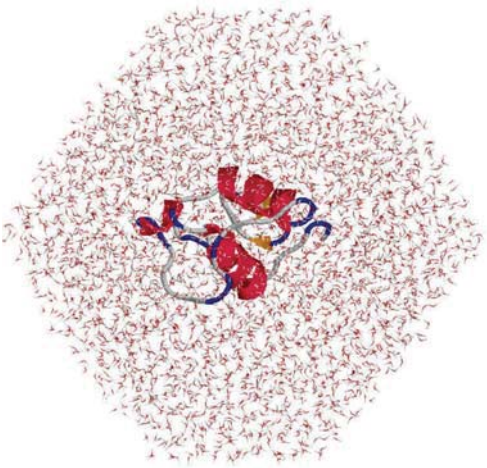
[as shown by Bekker, JCC ***18*** 1930 (1997), a simulation in any box shape can equivalently be carried our in a triclinic box]

Periodic boundary conditions

Examples:



collagen peptide in a rectangular box [possibly not so clever]



prion protein in a truncated octahedron

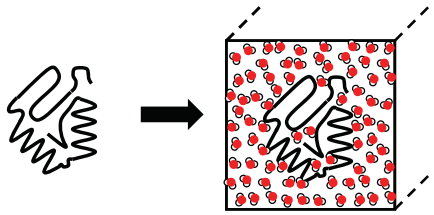
Periodic boundary conditions: practical considerations

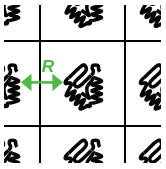
Solvating a macromolecule

Minimum distance solute-to-wall should be chosen large enough

⇒ when applying a cutoff R:

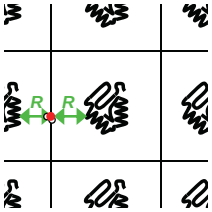
→ at least R/2





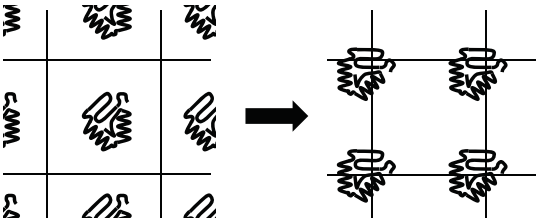
no solute atom interacts with solute atom in periodic copy

→ better at least R



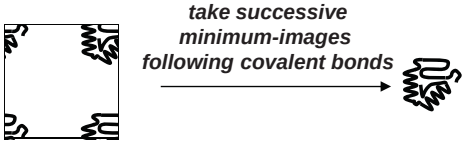
no solvent molecule interacts with solute atom in two solute periodic copies

Restoring the covalent connectivity



Macromolecules may drift in the periodic system

⇒ need to restore the covalent connectivity before applying analyses



Periodic boundary conditions: practical considerations

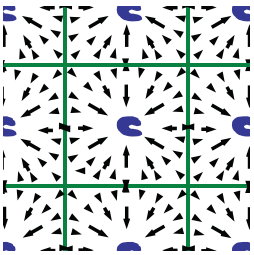
- **Periodic boundary conditions is probably the best remedy for *surface effects***

  ⇒ There is in effect no system surface (pushed to infinity)

  Note that in some cases, surface effects may be physical.
  E.g. ionic solvation free energies are affected by the water-vacuum interface potential, which is not accounted for under periodic boundary conditions

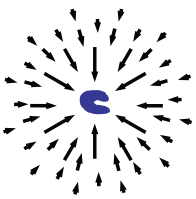
- **But it is only a partial remedy for *finite-size effects* in simulations of solutions (these effects are dominated by long-range electrostatic interactions)**

simulated sample under PBC



→ : solvent polarization

≠



ideal macroscopic system at infinite dilution

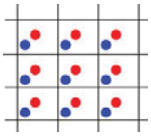
≈

simulated droplet

Periodic boundary conditions: a word of warning

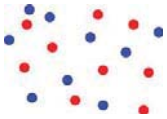
⇒ **Periodic system implies finite concentration *e.g.* 1 solute + 1000 water → 0.05 molal solution (charge 10 *e* → 5 molal ionic strength !)**

⇒ **Yet, this is an approximate model for a real solution at this concentration**



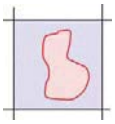
Anisotropy
No interaction at L/2

≠

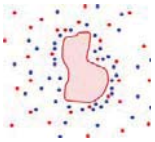


⇒ **This is an approximate model for the counter-ion distribution around a macromolecule**

non-neutral solute + background



≠



non-neutral solute + counter-ions

⇒ **Artificial anisotropy and periodic long-range correlations (range of interaction longer than box size)**

⇒ **Not possible to reproduce correct (non-periodic) correlations on lengthscales larger than the box size (e.g. undulations in a membrane)**

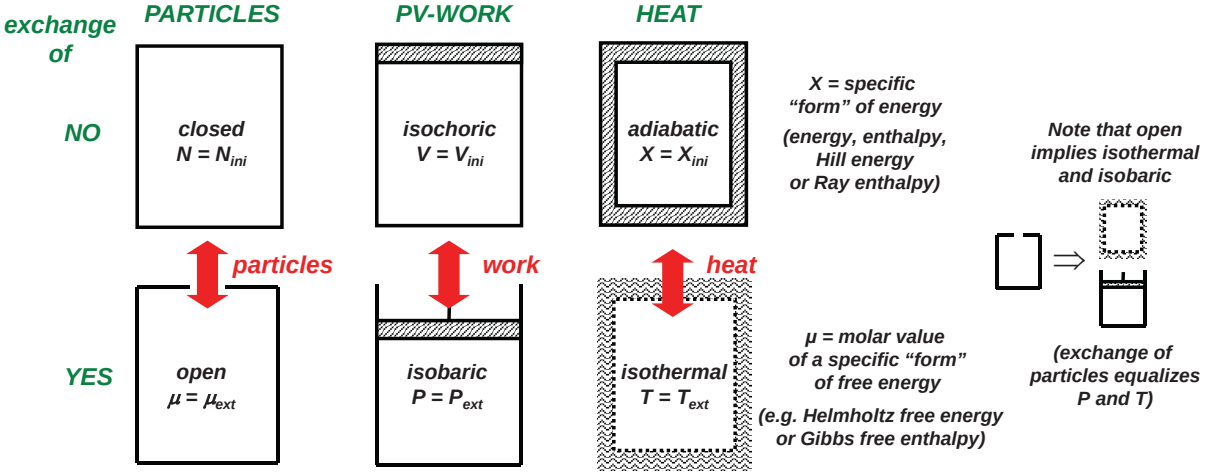
- **Almost OK for the simulation of crystals**
  ⇒ only neglects static & dynamic disorder
- **Sometimes needs too much solvent (very large systems)**
- **Probably still the best nowadays**
  but: → keep restrictions in mind, use large enough boxes

Thermodynamic boundary conditions

- **Thermodynamic systems** may be classified according to what they are allowed or not to **exchange** with their **surroundings**
- If the properties of the **surroundings** are **homogeneous** and **time-independent**, the system will ultimately reach a state **equilibrium**, *i.e.* characterized by **system** properties that are also **homogeneous** and **time-independent**

  → it makes sense to characterize this equilibrium state using a **specific choice** of **independent variables**, selected either because they are strictly **conserved** (constant) or because they **match** (on average) the corresponding (given) value in the **surroundings**

  → for a **one-component one-phase** system with work exchange exclusively involving **isotropic volume variations**, we need **three independent variables**



Thermodynamic boundary conditions

- Things get **more complex** when one considers the following **generalizations**

→ the system is **not confined**

Then P=0 and V is undefined — *E.g. the solar system!* — *Note: I am not sure this makes sense in thermodynamics, as non-confined systems should always "evaporate"*

→ the volume variations are **anisotropic**

Scalar V and P related quantities become tensors

→ the system exchanges work **other than PV-work** with its surroundings

E.g. electrical work (electrochemistry) or light (photochemistry)

→ the system has **multiple components** (or phases)

The composition N becomes a vector — *The system may be semi-open (i.e. open to a specific type of particles only)* — *Note: semi-open systems are still automatically isothermal, but not necessarily isobaric (osmotic pressure)*

→ **other boundary conditions** (or constraints) are introduced

E.g. wall within the system, external (electrical or gravitational) field value, stoechiometry constraint along a reaction, fixed value of the extent of reaction, ...

→ **non-equilibrium** situations are considered

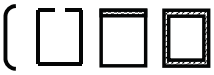
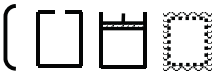
If the long-time behavior of the system is not yet reached (relaxation after a perturbation) or if the boundary conditions are inhomogeneous or/and time-dependent

- But let's **forget about all this** for the moment...

Thermodynamic systems

** = unofficial names*

- This leads to eight types of equilibrium **thermodynamic systems** *(only 5 are actually relevant)*

boundary conditions	*name of the system*	*natural choice of independent variables*	*"generalized" energy*	
	microcanonical	NVE	*internal energy*	E
	canonical	NVT		
	isoenthalpic-isobaric	NPH	*enthalpy*	$H = E + PV$
	isothermal-isobaric (Gibbs)	NPT		
	grand-microcanonical*	μVL	*Hill energy*	$L = E - \mu N$
	grand-canonical	μVT		
	grand-isoenthalpic-isobaric*	μPR	*Ray* enthalpy*	$R = E + PV - \mu N$
	generalized	μPT	*size undefined, i.e. not really a valid choice*	

→ In **green** are **intensive** variables (*local* observables): value matches the corresponding (constant) value in the surroundings

→ In **red** are **extensive** variables (*subsystem-additive* observables): value is strictly conserved (constant)

as $T=T_{ext}$, better use grand-canonical

Sampling/simulation methods and generated ensembles

- Ensembles generated

boundary conditions	*name of the system*	*natural choice of independent variables*	*sampling/simulation method*
	microcanonical	NVE	***plain MD***
	canonical	NVT	***MD + thermostat, MC, SD***
	isoenthalpic-isobaric	NPH	***MD + barostat***
	isothermal-isobaric (Gibbs)	NPT	***MD + thermostat + barostat***
	grand-canonical	μVT	***grand-canonical MD***

Thermodynamics

- In **thermodynamics**, we consider

  → **macroscopic systems**, *i.e.* the so-called *thermodynamic limit*

  → **processes** resulting from a change in the boundary conditions (*i.e.* in their type or/and in the value of the independent variables) and connecting **two equilibrium situations**

The goals are to quantify

- ***The associated change of the dependent variables***
- ***The direction of spontaneity***
- ***The exchanges (heat, work, particles) with the surroundings***

Note: the latter usually depend on the path of the process (boundary conditions, reversibility); the two former do not

  → for the **equilibrium situation itself**, the boundary conditions are **largely irrelevant** in the thermodynamic limit

E.g. whether you specify a state by N,V,E or N,P,H or N,V,T or N,P,T is up to you

  → this is no longer the case in **statistical mechanics...**

Statistical mechanics

- In **statistical mechanics**, we consider

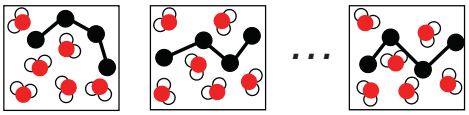
  → **all system sizes**, *i.e.* not necessarily the *thermodynamic limit* of macroscopic systems

  → the "composition" of equilibrium situations in terms of **microscopic configurations**

coordinates and momenta of all particles in the system $\boldsymbol{x} = (\boldsymbol{r}, \boldsymbol{p})$

- The key quantities are

  → the **probability distribution** of the microscopic configurations in the ensemble



Thermodynamic (equilibrium) state with boundary conditions XYZ

$\rho_{XYZ}(\boldsymbol{x})$ *probability of each configuration in the ensemble (normalized!)*

Hamiltonian (total energy) [normally written H]

e.g. microcanonical NVE $\rho_{NVE}(\boldsymbol{x}) \sim \delta(\mathcal{E}(\boldsymbol{x}) - E)$

e.g. canonical NVT $\rho_{NVT}(\boldsymbol{x}) \sim \exp\left(-\dfrac{\mathcal{E}(\boldsymbol{x})}{k_B T}\right)$

  → **microscopic observables** describing a given property in each microscopic configuration

e.g. microscopic (instantaneous) temperature $\mathcal{T}(\boldsymbol{x})$ *e.g. microscopic (instantaneous) pressure* $\mathcal{P}(\boldsymbol{x})$

- Most macroscopic observables are then obtained by **ensemble averaging**

Exceptions: free energy, entropy, chemical potential (discussed in later lectures)

$$A = A(X,Y,Z) = \int d\boldsymbol{x}\, \rho_{XYZ}(\boldsymbol{x})\, \mathcal{A}(\boldsymbol{x}) = \left\langle \mathcal{A}(\boldsymbol{x}) \right\rangle_{XYZ}$$

e.g. pressure in the canonical NVT ensemble $P = P(N,V,T) = \left\langle \mathcal{P}(\boldsymbol{x}) \right\rangle_{NVT}$

Statistical mechanics

- We can also look at the corresponding **fluctuations** and even **distributions**

$$\Delta A = \left\langle \left[\mathcal{A}(\boldsymbol{x}) - A \right]^2 \right\rangle_{XYZ}^{1/2} = \left[\left\langle \mathcal{A}^2(\boldsymbol{x}) \right\rangle_{XYZ} - A^2 \right]^{1/2}$$

$$p(\mathcal{A}') = \left\langle \delta\left[\mathcal{A}(\boldsymbol{x}) - \mathcal{A}' \right] \right\rangle_{XYZ}$$

  → **independent extensive** quantities are **rigorously conserved** (no fluctuations)

e.g. microcanonical NVE	*e.g. isoenthalpic-isobaric NPH*	*e.g. canonical NVT*
$\mathcal{N} = N,\ \Delta N = 0$	$\mathcal{N} = N,\ \Delta N = 0$	$\mathcal{N} = N,\ \Delta N = 0$
$\mathcal{V} = V,\ \Delta V = 0$	$\mathcal{P}$ $\quad \Delta P \neq 0$	$\mathcal{V} = V,\ \Delta V = 0$
(total) energy = Hamiltonian → $\mathcal{E} = E,\ \Delta E = 0$	*enthalpy* $\mathcal{H} = \mathcal{E} + P\mathcal{V}$ → $\mathcal{H} = H,\ \Delta H = 0$	$\mathcal{T}$ $\quad \Delta T \neq 0$

  → **dependent extensive** quantities **fluctuate** around **their averages**, and these fluctuations **do not vanish** in the **thermodynamic limit**

e.g. canonical NVT

$$E = \left\langle \mathcal{E} \right\rangle \qquad \Delta E = (k_B C_V)^{1/2} T \quad \Rightarrow \quad \Delta E \sim N^{1/2}$$

The fluctuations are related to thermodynamic derivative quantities

  → **intensive** quantities **fluctuate** around their **averages**, and these fluctuations **vanish** in the **thermodynamic limit**

independent: imposed values
dependent: dependent values

so: not for the microscopic simulated systems!

e.g. canonical, if we define $\mathcal{T} = (3k_B N/2)^{-1} \mathcal{K}$ *(total) kinetic energy*

$$T = \left\langle \mathcal{T} \right\rangle \qquad \Delta T = [(3/2)\, N]^{-1/2} T \quad \Rightarrow \quad \Delta T \to 0 \text{ for } N \to \infty$$

Statistical mechanics

- Example of instantaneous **observables**

Note: the definitions are not unique, i.e. a given thermodynamic quantity may be obtained by averaging different instantaneous observables (the fluctuations may then not be the same for finite systems !)

Number of particles	$\mathcal{N}$
Volume	$\mathcal{V}$
Energy (Hamiltonian)	$\mathcal{E} = \mathcal{U} + \mathcal{K}$
Temperature	$\mathcal{T} = \dfrac{2\mathcal{K}}{3k_B\mathcal{N}}$
Pressure	$\mathcal{P} = \dfrac{2\left(\mathcal{K} - \mathcal{W}\right)}{3\mathcal{V}}$
Enthalpy	$\mathcal{H} = \mathcal{E} + P\mathcal{V}$
Hill energy	$\mathcal{L} = \mathcal{E} - \mu\mathcal{N}$
Ray enthalpy	$\mathcal{R} = \mathcal{E} + P\mathcal{V} - \mu\mathcal{N}$

Hamiltonian = total potential energy + total kinetic energy

see next slides

Instantaneous temperature

- From the **equipartition theorem**, we know that the average kinetic energy per degree of freedom (dof) is related to the **macroscopic temperature** as

(for a system at equilibrium at temperature T)

for one dof α (e.g. x, y or z Cartesian dof of one atom)

$$\left\langle \mathcal{K}_\alpha \right\rangle = \left\langle \frac{1}{2} m_\alpha v_\alpha^2 \right\rangle = \frac{k_B T}{2}$$

for a set of dof's ⇒

$$\left\langle \mathcal{K} \right\rangle = \frac{\mathcal{N}_D k_B T}{2}$$

$\mathcal{N}_D$ *number of degrees of freedom (dof)*

→ an acceptable definition for the **instantaneous temperature** is then

$$\mathcal{T} = \frac{2}{\mathcal{N}_D k_B}\mathcal{K}$$

this ensures at equilibrium ⇒

$$\left\langle \mathcal{T} \right\rangle = T$$

discussed later

→ in **three dimensions**, in the absence of **constraints** and **uncoupled dof's**, and considering the **entire system**

separate temperatures of subsystems or/and dof types may also be of interest

$$\mathcal{N}_D = 3N$$

N *number of particles in the system*

Instantaneous pressure

- From the **virial theorem**, we know that each dof contributes to the **isotropic macroscopic pressure** of the **entire system** as

$$\left\langle \frac{2(\mathcal{K}_\alpha - \mathcal{W}_\alpha)}{3\mathcal{V}} \right\rangle = \left\langle \frac{m_\alpha v_\alpha^2 + r_\alpha F_\alpha}{3\mathcal{V}} \right\rangle \quad \text{for the entire system} \Rightarrow \quad P = \left\langle \frac{2(\mathcal{K} - \mathcal{W})}{3\mathcal{V}} \right\rangle$$

(for a system at equilibrium at pressure P)

anisotropic discussed later

→ the **instantaneous isotropic virial** of the system is defined as

$$\mathcal{W} = -\frac{1}{2}\sum_{i=1}^{N} \boldsymbol{r}_i \cdot \boldsymbol{F}_i$$

Watch out: this is the GROMOS definitions (various books use +/- or/and 2x this definition)

assumed again: $\mathcal{N}_D = 3N$

three-dimensions, no constraints, no uncoupled dof's (discussed later)

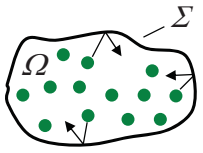
→ an acceptable definition for the **instantaneous pressure** is then

$$\mathcal{P} = \frac{2(\mathcal{K} - \mathcal{W})}{3\mathcal{V}} = \frac{Nk_B\mathcal{T}}{\mathcal{V}} - \frac{\mathcal{W}}{3\mathcal{V}} \quad \text{this ensures at equilibrium} \Rightarrow \quad \langle \mathcal{P} \rangle = P$$

Instantaneous pressure

- Reminder: the **virial theorem**

→ consider a **finite system at equilibrium** that is bounded by a **confinement wall**



wall exerts an external isotropic pressure P

→ **dynamical virial** definition

$$\mathcal{Q} = \sum_{i=1}^{N} m_i \boldsymbol{r}_i \cdot \dot{\boldsymbol{r}}_i \Rightarrow \dot{\mathcal{Q}} = 2(\mathcal{K} - \mathcal{W}^{tot})$$

time derivative

because

$$\mathcal{K} = \frac{1}{2}\sum_{i=1}^{N} m_i \dot{\boldsymbol{r}}_i^2$$

$$\mathcal{W}^{tot} = -\frac{1}{2}\sum_{i=1}^{N} \boldsymbol{r}_i \cdot \boldsymbol{F}_i^{tot} = -\frac{1}{2}\sum_{i=1}^{N} m_i \boldsymbol{r}_i \cdot \ddot{\boldsymbol{r}}_i$$

between particles + from the wall

Newton

→ **virial theorem**

$$\langle \dot{\mathcal{Q}} \rangle = \lim_{t\to\infty} \frac{1}{t}\int_0^t dt\, \dot{\mathcal{Q}}(t) = \lim_{t\to\infty} \frac{\mathcal{Q}(t) - \mathcal{Q}(0)}{t} = 0$$

because $\mathcal{Q}$ *is always finite*

→ *coordinates are bounded to system volume*

→ *velocities are bounded by temperature*

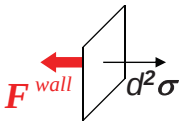
use long-time average to calculate ensemble average (ergodicity)

$$\Rightarrow \quad \langle \mathcal{W}^{tot} \rangle = \langle \mathcal{K} \rangle$$

→ virial of the **wall forces**

between particles / *from the wall*

$$\mathcal{W}^{tot} = \mathcal{W} + \mathcal{W}^{wall}$$



d³ω: volume element
d²σ: surface element

$$\langle \mathcal{W}^{wall} \rangle = -\frac{1}{2}\sum_{n=1}^{N} \boldsymbol{r}_i \cdot \boldsymbol{F}_i^{wall} = \frac{1}{2} P \int_\Sigma d^2\boldsymbol{\sigma}(\boldsymbol{r}) \cdot \boldsymbol{r}$$

$$= \frac{1}{2} P \int_\Omega d^3\omega(\boldsymbol{r}) \nabla \cdot \boldsymbol{r} = \frac{3}{2} P \langle \mathcal{V} \rangle$$

divergence (Gauss) theorem

$$\Rightarrow \quad \langle \mathcal{W} \rangle = \left\langle \mathcal{K} - \frac{3}{2} P\mathcal{V} \right\rangle$$

Instantaneous pressure

- So, we have indeed at equilibrium

$$P = \left\langle \frac{2(\mathcal{K} - \mathcal{W})}{3\mathcal{V}} \right\rangle \qquad \text{where} \qquad \mathcal{W} = -\frac{1}{2}\sum_{i=1}^{N} \boldsymbol{r}_i \cdot \boldsymbol{F}_i$$

→ **interpretation**

Ideal-gas contribution (due to the kinetic energy, i.e. the particles hitting the walls)

$$P = \frac{N k_B T}{V} - \left\langle \frac{\mathcal{W}}{3\mathcal{V}} \right\rangle$$

contribution from intermolecular forces (cf real-gas equations of state)

attractive → "pull inwards" → W > 0 → P ↓

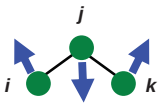
repulsive → "pull outwards" → W < 0 → P ↑

→ **virial calculation**: simultaneously with the potential energy and forces

- The same equation can be used under **periodic boundary conditions**, but one has to pay a bit attention to select the right periodic images for the coordinates $\boldsymbol{r}_i$ *ambiguous*

→ covalent terms: **gather** the atoms by applying periodic shifts

e.g. bond-angle bending term



$$W_{ijk} = -\frac{1}{2}\left(\boldsymbol{r}_i \cdot \boldsymbol{F}_i + \boldsymbol{r}_j \cdot \boldsymbol{F}_j + \boldsymbol{r}_k \cdot \boldsymbol{F}_k\right)$$

... but this one will be zero (see later)

→ non-bonded minimum-image pairwise terms: rewrite **as double sum**

$$\boldsymbol{F}_{ij} \quad \boldsymbol{r}_{ij} = \boldsymbol{r}_j - \boldsymbol{r}_i \quad \boldsymbol{F}_{ji} = -\boldsymbol{F}_{ij}$$

Newton's 3rd law

use minimum-image vector for this one

$$W_{ij} = -\frac{1}{2}\sum_i^N \sum_{j\neq i}^N \boldsymbol{r}_i \cdot \boldsymbol{F}_{ij} = -\frac{1}{4}\sum_i^N \sum_{j\neq i}^N \left[\boldsymbol{r}_i \cdot \boldsymbol{F}_{ij} + \boldsymbol{r}_j \cdot \boldsymbol{F}_{ji}\right] = \frac{1}{4}\sum_i^N \sum_{j\neq i}^N \boldsymbol{r}_{ij} \cdot \boldsymbol{F}_{ij} = \frac{1}{2}\sum_i^N \sum_{j>i}^N \boldsymbol{r}_{ij} \cdot \boldsymbol{F}_{ij}$$

e.g. repulsive (as drawn) → W < 0 → P ↑

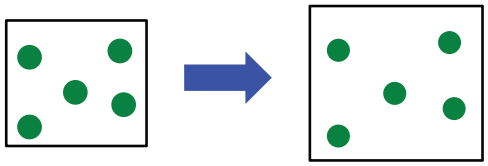
→ note that **constraint forces** also have a virial contribution

calculated from the SHAKE Lagrange multipliers

Instantaneous pressure

- Another way of looking at the virial

→ imagine we scale a system isotropically by a factor *1+α* along each Cartesian direction



$$\boldsymbol{r}_i \to (1+\alpha)\boldsymbol{r}_i$$
$$\mathcal{V} \to (1+\alpha)^3 \mathcal{V}$$

→ close to α = 0, the changes in potential energy and volume are

$$d\mathcal{U} = \sum_{i=1}^{N} \frac{\partial \mathcal{U}}{\partial \boldsymbol{r}_i} \cdot \frac{d\boldsymbol{r}_i}{d\alpha} d\alpha = 2\mathcal{W} d\alpha$$
$$d\mathcal{V} = 3\mathcal{V} d\alpha$$

$$\mathcal{W} = \frac{3\mathcal{V}}{2}\frac{\partial \mathcal{U}}{\partial \mathcal{V}}$$

in line with previous equations:

attractive → "pull inwards" → W > 0 → P ↓

repulsive → "pull outwards" → W < 0 → P ↑

→ the isotropic virial of all angle-dependent covalent terms is zero

$$\frac{\partial \mathcal{U}}{\partial \mathcal{V}} = 0$$

because an angle does not change upon isotropic coordinate scaling!

this is true for bond-angle and proper/improper dihedral-angle terms

→ the isotropic virial of a pairwise homogeneous term is related to its potential energy

$$\mathcal{U}_n = \sum_{i,j>i} C_{ij} r_{ij}^n \quad \Rightarrow \quad d\mathcal{U}_n = d\sum_{i,j>i} C_{ij}\left[(1+\alpha) r_{ij}\right]^n = n\mathcal{U}_n d\alpha \quad \Rightarrow \quad \mathcal{W}_n = \frac{n}{2}\mathcal{U}_n$$

→ things get more complicated when you consider anisotropic volume variations

then the virial becomes a tensor and the two above simplifications no longer hold (see later)

e.g. $\mathcal{W} = -\frac{1}{2}\mathcal{U}_{Cb} - 3\mathcal{U}_{C6} - 6\mathcal{U}_{C12}$

Thermostating / barostating: need

- The reasons why we need **temperature / pressure control** are
  - → simulations in NVT and NPT (evtl. NPH, μVT) ensembles *to match experimentally common conditions*
  - → study T- or/and P-dependent processes *e.g. phase or conformational transitions*
- Other reasons why we need **temperature control** are
  - → stabilize the energy *prevent energy increase in simulations caused by numerical errors, e.g. cutoff truncation, finite timestep, ...* *(microcanonical: we could also use an "ergostat")*
  - → study non-equilibrium (steady-state) processes *e.g. where heat must be evacuated from the system*
  - → to enhance conformational searches *e.g. MD at high T, simulated annealing, parallel/serial tempering*
- Boundary condition should in principle be **soft** (P and T, as intensive quantities, have non-negligible fluctuations in microscopic systems)
- Common approaches
  - → constraining (fix to target value, no fluctuations)
  - → weak coupling (first-order relaxation)
  - → extended-system coupling (second-order relaxation)
  - → stochastic coupling *e.g. SD, stochastic volume variations*

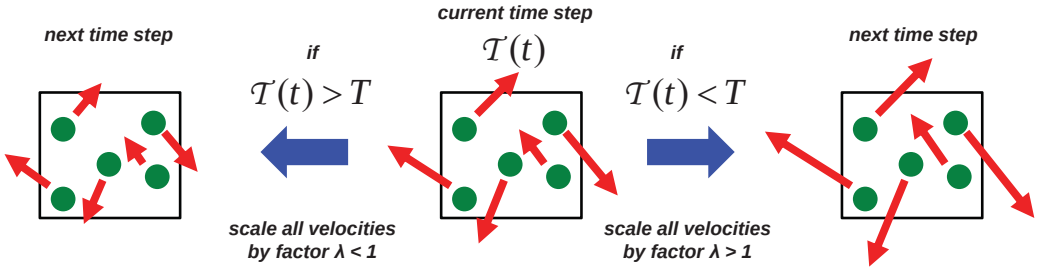
As sean earlier, the Langevin equation of motion leads to a constant temperature (balance between stochastic collisions and frictional drag) and thus implicitly involves a thermostat (sometimes also used in explicit-solvent MD) !

Thermostating / barostating: working principle

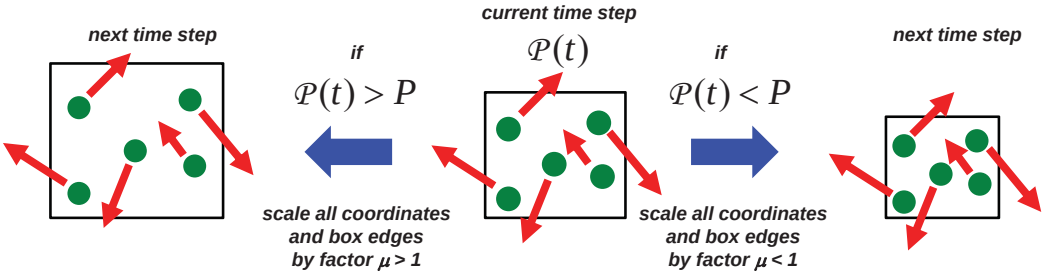
- We have **instantaneous** definitions for the **temperature** and the **pressure**

$$\mathcal{T} = \frac{2}{\mathcal{N}_D k_B}\mathcal{K} \qquad \textit{and} \qquad \mathcal{P} = \frac{2(\mathcal{K}-\mathcal{W})}{3\mathcal{V}}$$

- Working principle of **thermostating** by velocity scaling (most common)



- Working principle of **barostating** by isotropic coordinate/box scaling



Thermostating *via* temperature constraining

Woodcock WO71.1
Hoover HO82.1
Evans EV83.3

- **Equations of motion** → *isokinetic, closed, isochoric*

$$\dot{\mathcal{N}} = 0 \qquad \dot{\boldsymbol{r}} = \nabla_{\boldsymbol{p}}\mathcal{K} \qquad \dot{\boldsymbol{p}} = -\nabla_{\boldsymbol{r}}\mathcal{U} - \zeta_T\,\boldsymbol{p} \qquad \dot{\mathcal{V}} = 0$$

with $\zeta_T = \dfrac{-\nabla_{\boldsymbol{r}}\mathcal{U}\cdot\nabla_{\boldsymbol{p}}\mathcal{K}}{2K}$ → *effective "friction coefficient"*

$-\nabla_{\boldsymbol{r}}\mathcal{U}\cdot\nabla_{\boldsymbol{p}}\mathcal{K} = \boldsymbol{F}\cdot\dot{\boldsymbol{r}}$ → *"tendency" of the forces to increase the kinetic energy*

→ **brake** particles if ***F*** would increase $\mathcal{K}$, **accelerate** them if ***F*** would decrease $\mathcal{K}$

→ **kinetic energy** stays **constant** at *K* if we start with $\mathcal{K}(0) = K$

proof $$\dot{\mathcal{K}} = \frac{d}{dt}\left(\frac{1}{2}\boldsymbol{p}\cdot\underline{\boldsymbol{M}}^{-1}\boldsymbol{p}\right) = \dot{\boldsymbol{p}}\cdot\underline{\boldsymbol{M}}^{-1}\boldsymbol{p} = -\nabla_{\boldsymbol{r}}\mathcal{U}\cdot\nabla_{\boldsymbol{p}}\mathcal{K} - \zeta_T\underbrace{\boldsymbol{p}\cdot\nabla_{\boldsymbol{p}}\mathcal{K}}_{2\mathcal{K}} = -\nabla_{\boldsymbol{r}}\mathcal{U}\cdot\nabla_{\boldsymbol{p}}\mathcal{K}\left(1-\frac{\mathcal{K}}{K}\right) = 0$$

- **Conserved quantities**

$$\dot{\mathcal{N}}_D = 0 \qquad \dot{\mathcal{V}} = 0 \qquad \dot{\mathcal{K}} = 0 \;\; (\dot{\mathcal{T}} = 0)$$

- **Phase-space probability distribution**

isokinetic for the momenta / *canonical for the coordinates*

$$\rho = C\,\underbrace{\delta(\mathcal{N}_D - N_D)}_{\text{closed}}\,\underbrace{\delta(\mathcal{V}-V)}_{\text{isochoric}}\,\delta(\mathcal{K}-K)\exp[-\beta\mathcal{U}] \qquad \textit{with}\quad \beta = \frac{N_D - 1}{2K}$$

→ the kinetic-energy constraint **removes one dof**

so, if we want $\beta = \dfrac{1}{k_B T}$ *we need to adjust the temperature definitions to* $\mathcal{T} = \dfrac{2\mathcal{K}}{(\mathcal{N}_D - 1)k_B}$ *and* $T = \dfrac{2K}{(N_D - 1)k_B}$

Thermostating *via* temperature constraining

- Practical **implementation** (leap-frog, velocity scaling)

say that $\mathcal{K}(t-\Delta/2)$ *is* K

leap $\boldsymbol{v}(t+\Delta/2) = \boldsymbol{v}(t-\Delta/2) + \underline{\boldsymbol{M}}^{-1}\boldsymbol{F}(t)$

calculate $\mathcal{K}(t+\Delta/2)$

scale $\boldsymbol{v}(t+\Delta/2)$ *by* $\lambda = \left[\mathcal{K}(t-\Delta/2)/\mathcal{K}(t+\Delta/2)\right]^{1/2}$ *or* $\lambda = \left[K/\mathcal{K}(t+\Delta/2)\right]^{1/2}$

after scaling, the kinetic energy is again K

Hoover-Evans, imposes $\dot{\mathcal{K}} = 0$ — *unwise: numerical noise may lead to drifts*

Woodcock, imposes $\mathcal{K} = K$ — *wise: the target value is explicitly maintained*

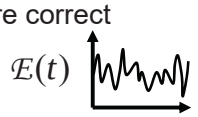
referred to as "velocity scaling"

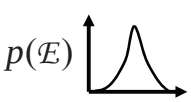
- **Properties**

→ +: we sample the **canonical** (NVT) ensemble **in terms of coordinates** *provided that we remove one dof in the temperature definition!*

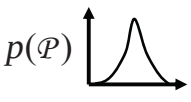
→ -: but we sample an **isokinetic** (≠canonical) ensemble **in terms of momenta** (zero temperature fluctuations – incorrect for a microscopic system!) *still, if we only care about configuration-dependent properties, it is not a bad choice !*

→ +: the total energy and the pressure **fluctuate** in time; the fluctuations are correct





fluctuations depend on C_V *and scale as* $\mathcal{N}_D^{-1/2}$



fluctuations depend on C_V *and* K_T *and scale as* $\mathcal{N}_D^{-1/2}$

→ +: with the Woodcock variant, the energy may **no longer drift** (even when there is noise) because a target temperature is specified explicitly

Thermostating *via* weak coupling

- **Equations of motion** → *Berendsen, closed, isochoric*

$$\dot{\mathcal{N}} = 0 \qquad \dot{\boldsymbol{r}} = \nabla_p \mathcal{K} \qquad \dot{\boldsymbol{p}} = -\nabla_r \mathcal{U} - \zeta_T \, \boldsymbol{p} \qquad \dot{\mathcal{V}} = 0$$

with $\zeta_T = \frac{1}{2\tau_T}\frac{\mathcal{K}-K}{\mathcal{K}}$ → *effective "friction coefficient"* τ_T *thermostat coupling time*

→ **brake** particles if $\mathcal{K}>K$, **accelerate** them if $\mathcal{K}<K$, in proportion to the **relative difference**

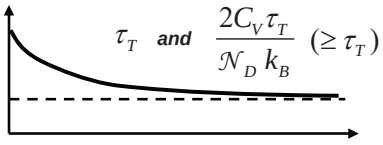
→ in an **ideal-gas situation**, the temperature **relaxes exponentially** towards its target

proof $\dot{\mathcal{K}} = \frac{d}{dt}\left(\frac{1}{2}\boldsymbol{p}\cdot\underline{\boldsymbol{M}}^{-1}\boldsymbol{p}\right) = \dot{\boldsymbol{p}}\cdot\underline{\boldsymbol{M}}^{-1}\boldsymbol{p} = -\zeta_T \, \boldsymbol{p}\cdot\nabla_p\mathcal{K} = -2\mathcal{K}\zeta_T = -\frac{1}{\tau_T}(\mathcal{K}-K)$ *since* $\mathcal{U}=0$

→ in a **realistic situation** and for **long coupling times**, the relaxation may be **slower** as the energy added also redistributes to potential energy

$\dot{\mathcal{T}} \approx -\frac{1}{C_V\tau_T}(\mathcal{K}-K)$ *with* $C_V = \left(\frac{\partial E}{\partial T}\right)_V$ ➡ $\dot{\mathcal{K}} \approx -\frac{\mathcal{N}_D k_B}{2C_V\tau_T}(\mathcal{K}-K)$

relaxation time between τ_T *and* $\frac{2C_V\tau_T}{\mathcal{N}_D k_B}$ $(\geq \tau_T)$



- **Conserved quantities**

$$\dot{\mathcal{N}}_D = 0 \qquad \dot{\mathcal{V}} = 0$$

third one unknown (to my knowledge), will depend on C_V and τ_T

averages are correct, e.g. $\langle\mathcal{T}\rangle = T$ *but fluctuations are not*

- **Phase-space probability distribution**

$$\rho = C\,\delta(\mathcal{N}_D - N_D)\,\delta(\mathcal{V}-V)\,?(\mathcal{K},\mathcal{U})$$

not canonical, not analytical (to my knowledge), will depend on C_V and τ_T

Thermostating *via* weak coupling

- Practical **implementation** (leap-frog, velocity scaling) τ_T *typically set to 0.1 ps in GROMOS*

scale $\boldsymbol{v}(t+\Delta/2)$ *by*

$$\lambda = \left[1 - \frac{\Delta}{\tau_T}\frac{\mathcal{K}(t+\Delta/2)-K}{\mathcal{K}(t+\Delta/2)}\right]^{1/2}$$

➡

after scaling, the kinetic energy is increased by

$$-\frac{\Delta}{\tau_T}\left[\mathcal{K}(t+\Delta/2)-K\right]$$

→ **limiting cases**

$\tau_T = \Delta$ ⟶ *we recover temperature-constraining*

$\tau_T \to \infty$ ⟶ *we recover a microcanonical situation*

- **Properties**

→ +: the energy may **no longer drift** (even when there is noise) because a target temperature is specified explicitly *≠microcanonical, Hoover-Evans*

→ +: the **temperature fluctuates,** which is **physical** *fluctuations scale as* $\mathcal{N}_D^{-1/2}$ *≠temperature-constraining*

→ +: the **temperature relaxation** behavior is **exponential**, which is **physical** *≠Nosé-Hoover*

→ -: we sample a somewhat **ill-defined ensemble** (Berendsen ensemble); the **averages** of the total energy, temperature and pressure are correct, but the **fluctuations/distributions** depend on C_V and τ_T and are **not rigorously canonical**

For most purposes, it does not matter ➡

Possible exceptions:
- when coupling very small (sub)systems (i.e. with few dof's)
- when calculating properties using fluctuation formulae
- when using serial/parallel tempering schemes (maybe)

Thermostating *via* extended-system coupling

Nosé NO84.2
Hoover HO85.1

- **Equations of motion** → *canonical*

$$\dot{\mathcal{N}} = 0 \qquad \dot{\boldsymbol{r}} = \nabla_{\boldsymbol{p}} \mathcal{K} \qquad \dot{\boldsymbol{p}} = -\nabla_{\boldsymbol{r}} \mathcal{U} - \zeta_T \, \boldsymbol{p} \qquad \dot{\zeta}_T = \frac{1}{\tau_T^2}\left(\frac{\mathcal{K}}{K} - 1\right) \qquad \dot{\mathcal{V}} = 0$$

ζ_T *thermostat variable (units of time^{-1})*

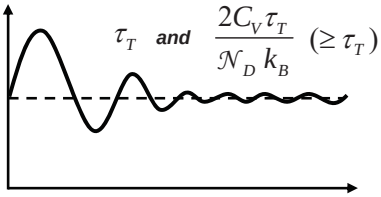
τ_T *thermostat coupling time*

now, the "friction coefficient" has become an extra variable in the dynamics, with its own equation of motion

→ **increase friction** if $\mathcal{K}$>K, **decrease it** if $\mathcal{K}$<K, in proportion of the **relative difference**

→ in an **ideal-gas situation**, the temperature **relaxes with damped-oscillations** towards its target

oscillation time between τ_T *and* $\frac{2C_V \tau_T}{\mathcal{N}_D k_B}$ $(\geq \tau_T)$



- **Conserved quantities**

integrate extra equation $\dot{\eta} = 2K\zeta_T$

$$\mathcal{A} = \mathcal{H} + \eta + K\tau_T^2\zeta_T^2$$

$$\dot{\mathcal{N}}_D = 0 \qquad \dot{\mathcal{V}} = 0 \qquad \dot{\mathcal{A}} = 0$$

- **Phase-space probability distribution**

$$\rho = C\,\delta(\mathcal{N}_D - N_D)\delta(\mathcal{V} - V)\exp[-\beta\mathcal{H}]\exp[-\beta K\tau_T^2\zeta_T^2] \qquad \beta = \frac{1}{k_B T} = \frac{N_D}{2K}$$

closed *isochoric* *canonical!* *gaussian*

Thermostating *via* extended-system coupling

- Practical **implementation** (leap-frog, velocity scaling)

τ_T *typically set to 0.1 ps in GROMOS*

propagate ζ_T $\quad \zeta_T(t+\Delta) = \zeta_T(t) + \frac{1}{\tau_T^2}\left(\frac{\mathcal{K}(t+\Delta/2)}{K} - 1\right)$

interpolate ζ_T $\quad \zeta_T(t+\Delta/2) = \left[\zeta_T(t) + \zeta_T(t+\Delta)\right]/2$

scale $\boldsymbol{v}(t+\Delta/2)$ *by* $\lambda = 1 - \Delta\,\zeta_T(t+\Delta/2)$

- **Properties**

→ +: the energy may **no longer drift** (even when there is noise) because a target temperature is specified explicitly

≠microcanonical, Hoover-Evans

→ +: the **temperature fluctuates,** which is **physical**

fluctuations scale as $\mathcal{N}_D^{-1/2}$ *≠temperature-constraining*

→ +: we sample rigorously the **canonical ensemble**, *i.e.* the **averages** of the total energy, temperature and pressure are correct, as well as their **fluctuations/distributions**

→ -: the **temperature relaxation** behavior is **damped-oscillatory**, and **remains oscillatory** at equilibrium which is **unphysical**

≠weak-coupling

Possible remedy: Nosé-Hoover chain!

→ -: for systems with few dof's, there may be **ergodicity violations**, *i.e.* the system gets trapped in periodic oscillations with the thermostat and does not sample its entire accessible phase space

For most purposes, it does not matter ➡ *Possible exceptions:*
- when coupling very small (sub)systems (i.e. with few dof's)
- when the detailed dynamics of the system is of importance

Barostating algorithms

- Similar to thermostating, there also exists
  - → barostating *via* pressure constraining *Evans 1982*
  - → barostating *via* weak coupling *Berendsen 1984*
  - → barostating *via* extended-system coupling *Andersen 1981*

→ many features are similar to the thermostating case (not described in details)

- working principle of the weak-coupling approach (leap-frog, GROMOS)

after thermostating (if applied)

interpolate kinetic energy $\mathcal{K}(t) = \left[\mathcal{K}(t-\Delta/2) + \mathcal{K}(t+\Delta/2)\right]/2$

calculate virial $\mathcal{W}(t)$

calculate pressure $\mathcal{P}(t) = \dfrac{2\left[\mathcal{K}(t) - \mathcal{W}(t)\right]}{3\mathcal{V}(t)}$

scale coordinates and box edges by $\mu = \left\{1 + \dfrac{\kappa_T \Delta}{\tau_P}\left[\mathcal{P}(t) - P\right]\right\}^{1/3}$ *with* $\kappa_T = -\dfrac{1}{V}\left(\dfrac{\partial V}{\partial P}\right)_T$

τ_P *typically set to 0.5 ps in GROMOS*

experimental compressibility (provided on input)

approximate because we use the experimental average compressibility instead of the instantaneous one

after scaling, the volume is increased by $\Delta\mathcal{V} = \mathcal{V}\dfrac{\kappa_T \Delta}{\tau_P}\left[\mathcal{P}(t) - P\right]$

and the pressure increased by $\Delta\mathcal{P} \approx \left(\dfrac{\partial P}{\partial V}\right)_T \Delta\mathcal{V} = -\dfrac{\Delta}{\tau_P}\left[\mathcal{P}(t) - P\right]$

Additional issues / extensions

[only briefly mentioned]

- Extension to **open systems** (grand-canonical simulations)
  - → idea: simulate two systems 1 and 2 with total N_{tot} particles
    system 1: $\mathcal{N}$ *real system, open*
    system 2: $N_{tot} - \mathcal{N}$ *e.g. ideal gas at given pressure*
    } $\mathcal{N}$ *is made to vary dynamically*
- **Artificial ensembles** (non-physical phase-space probability distributions)
  - → *e.g.* Tsallis ensemble, stretched probability extended-ensemble dynamics (SPEED)
- **Uncoupled** degrees of freedom *flying ice cube*
- **Constrained** degrees of freedom

} *discussed below*

- **Anisotropic** volume variations *Parrinello & Rahman 1980*
  - → the virial and pressure become 3x3 tensors; the reference pressure may also
  - → (semi-)anisotropic coupling (*e.g.* lipid membranes), unit-cell deformations (crystals)
- Alternative **temperature and pressure definitions**
  - → configurational temperatures and pressures
  - → molecule-based kinetic energy, virial and pressure definitions
  - → separate temperature of subsystems (*e.g.* solute vs solvent) or/and dof types (*e.g.* molecule translation, rotation and vibration)
- **Proofs** of constants of motion / phase-space probability distributions *generalized Liouville equation*
- **Ergodicity** problems *hot solvent / cold solute; ergodicity violations in NH for small systems*
- Choice of **coupling scheme and constants** *separate baths; tauT < tauP*

} *discussed below*

Uncoupled degrees of freedom

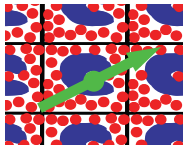
- Ensembles generated by simulation may possess **additional independent quantities** such as the system total **linear** or/and **angular momentum**

  → These are **constants of the** motion, *i.e.* **uncoupled degrees of freedom** (do not exchange kinetic energy with the others)

MD system in vacuum

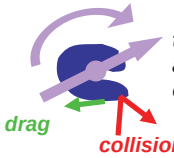
translation and rotation uncoupled

MD system under PBC



translation uncoupled, rotation coupled

SD system



translation and rotation coupled

- The conserved value of the **linear momentum** (MD simulation) *does not affect* the physical properties of the system (it is thus normally set to zero)

  → *e.g.* a glass of water that is moved around is not warmer than one sitting on the table…

- The conserved value of the **angular momentum** (MD simulation in vacuum) affects the physics of the system (through centrifugal forces)

  → value *should be specified* along with the other independent thermodynamic variables characterizing the ensemble (not given generally means set to zero)

- Uncoupled degrees of freedom can take arbitrary values and should not enter into the calculation of the instantaneous temperatures and pressures!

Uncoupled/constrained degrees of freedom

- When calculating/stabilizing the temperature and pressure

  → the velocities along uncoupled degrees of freedom should be omitted

$$\tilde{\mathcal{K}} = \frac{1}{2}\sum_{i=1}^{\mathcal{N}} m_i \tilde{\dot{\boldsymbol{r}}}_i^2$$

peculiar velocities i.e. after subtraction of possible contributions along uncoupled degrees of freedom

note: the velocities along constraints are zero, and so is the corresponding kinetic energy contribution

$$\mathcal{P} = \frac{2\left(\tilde{\mathcal{K}} - \mathcal{W}\right)}{3\mathcal{V}}$$

note: the forces along constraints and uncoupled dof's are zero, and so is the corresponding virial contribution

  → the number of degrees of freedom should be adjusted to account for uncoupled and constrained degrees of freedom

see NDFMIN in GROMOS

$$\mathcal{T} = \frac{2}{\mathcal{N}_D k_B}\tilde{\mathcal{K}} \quad with \quad \mathcal{N}_D = 3\mathcal{N} - \mathcal{N}_c - \mathcal{N}_r - \mathcal{N}_e$$

$\mathcal{N}_c =$ ***number of constraints***

$\mathcal{N}_r =$ ***6 (MD in vacuum)*** / ***3 (MD under PBC)*** / ***0 (SD)***

$\mathcal{N}_e =$ ***constrained dof from thermostat or barostat (zero unless specified)***

e.g. use "1" here for the temperature-constraining thermostat

The flying ice-cube effect

- If **thermostating** is *incorrectly* applied to **overall velocities**, the kinetic energy of uncoupled dof will fluctuate, which is *unphysical*

  → *in principle* not a problem if this kinetic energy is zero...

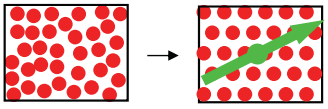
  → *in practice* (for complicated reasons) kinetic energy will **accumulate** into the uncoupled dof

⇒ spinning and unfolding !

- Even worse if the **temperature** is incorrectly calculated from *overall velocities*

  → accumulation of kinetic energy in the uncoupled dof implies a loss of kinetic energy for the internal motions, called the **flying-ice cube** effect

  → slowly during nanoseconds and then **all of a sudden**...



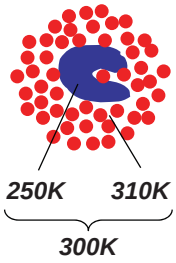
⇒ frozen !

The hot-solvent/cold-solute effect

- The solvent is typically subject to more heating (cutoff noise) than the solute, but the exchange of kinetic energy is slow

  → if these are coupled to a common thermostat, the average system temperature may be correct, but the solute is actually colder than the solvent

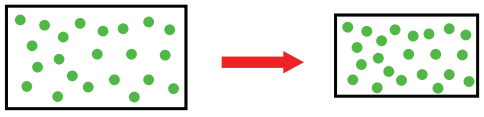
  → fix: couple solute and solvent degrees of freedom to separate thermostats



Thermostating/barostating coupling times

- Simultaneous **barostatting and thermostatting**

  → barostatting through the scaling of atomic coordinates and box volume provokes large changes in the potential energy



e.g. large increase in van der Waals energy (condensed phase)

  → partial conversion of this potential energy to kinetic energy occurs on a very short timescale

  → the temperature control (thermostatting) should occur on a shorter timescale than the barostatting to prevent anomalously-high temperature fluctuations

  → Thus, *e.g.* for the Berendsen thermostat, one **must have** $\tau_T < \tau_P$

Note:
Pressure fluctuations are typically very large !!!

typical values for weak coupling

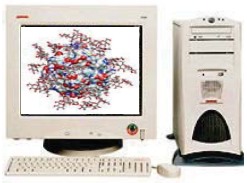
$\tau_T \approx 0.1\,ps$

$\tau_P \approx 0.5\,ps$

Computer Simulation in Chemistry, Biology and Physics

P.H. Hünenberger

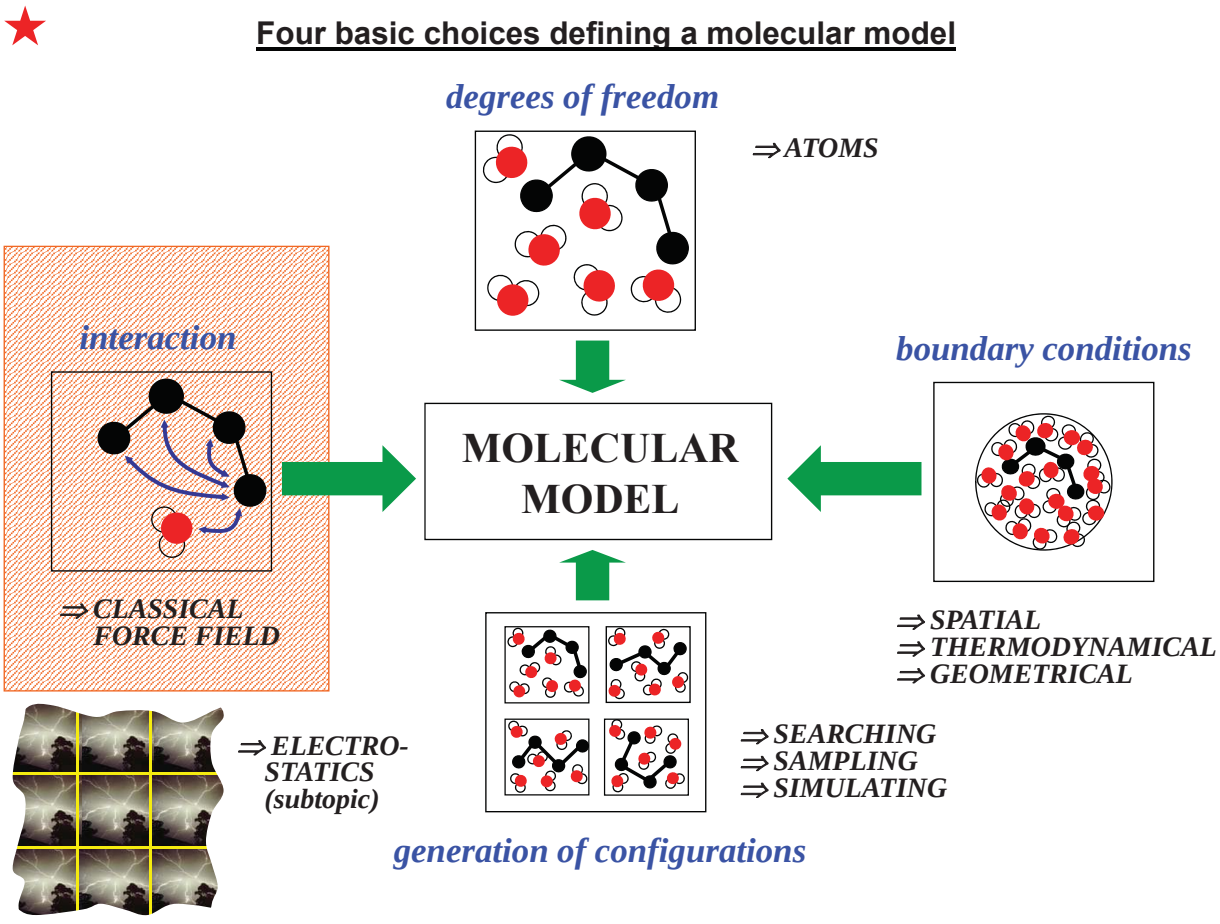
COMPUTER SIMULATION OF MOLECULAR SYSTEMS



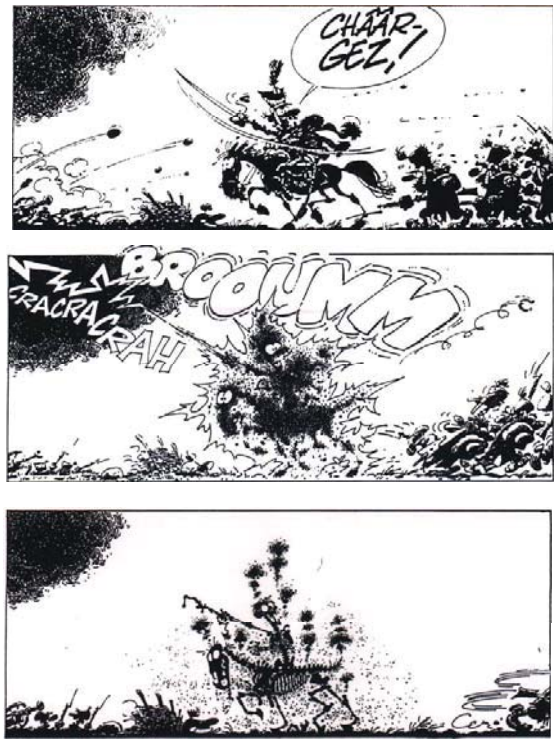
Herbstsemester 2019
Tuesday 9:45-11:30 a.m.
HCI D2

Lecture 529-0004-00
www.csms.ethz.ch/education/CSCBP

LECTURE 5 (WEEK 6):
Electrostatic interactions



Electrostatic interactions



(the fate of those who underestimate the problem of electrostatic interactions)

★

Electrostatic interactions

At the quantum-mechanical level:

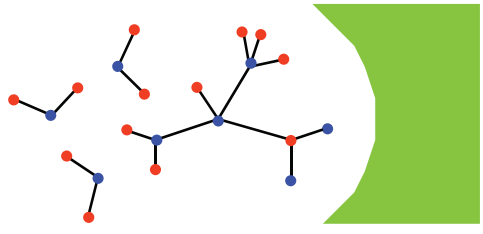
Chemistry ***is*** the science of electrostatic interactions !

At the classical (force-field) level:

Electrostatic interactions are the ***long-range (r^{-1}) residual*** of quantum-mechanical electrostatic interactions, after removal of all short-range contributions (covalent, exchange/repulsion and dispersion).

Usual description:

The monopole (partial charge) approximation + Coulomb's law



explicit solvent ⟷ ***solute*** ⟷ ***implicit solvent***

- ***positive partial charge $\delta+$***
- ***negative partial charge $\delta-$***

$$V_{ij} = \frac{q_i q_j}{4\pi\varepsilon_o r_{ij}}$$

(+ evtl. implicit solvation term)

ε_o **permittivity of vacuum (a physical constant)**

Explicit- versus implicit-solvent simulations

Typically ~10 solvent molecules per solute atom, thus computationally very expensive (biomolecules → 0.1-10 μs) !

Explicit solvent (with MD)

- Polar and "non-polar" solvation simultaneously (incl. solute-solvent van der Waals interactions, dielectric solvation/screening, H-bonding, solvation entropy and hydrophobic effect)
- Discrete solvent molecules

  → **good short-range solvation !**
- Microscopic system size (+ evtl. artificial periodicity)

  → **poor long-range solvation !**

in general, the solvent is only interesting for its mean effect on the solute...

Computationally much cheaper, but accuracy limited by that of the chosen implicit-solvent model

Implicit solvent (with SD)

- Polar solvation (*e.g.* CE, GB, PDLD) + "non-polar" solvation (*e.g.* SASA), contributions assumed separable
- Typically neglected (to some degree): structure of solvation shells, correlations among solvent molecules, specific solute-solvent interactions (H-bonding), saturation, electrostriction, ...

  → **poor short-range solvation !**
- Coulombic potential and accurate boundary conditions

  → **good long-range solvation (in favorable cases) !**

ELECTROSTATIC INTERACTIONS IN IMPLICIT-SOLVENT SIMULATIONS

Vacuum boundary conditions

→ A (bio)molecule is simulated in the total absence of solvent molecules

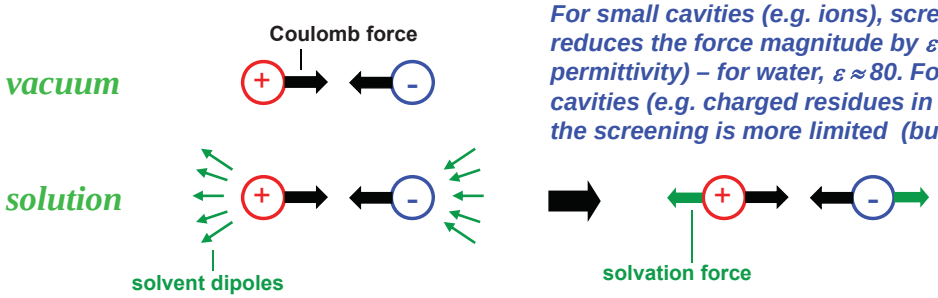
vacuum

- Surface effects (surface tension)
- No solvent dielectric screening

⇒ ***IN GENERAL A BAD IDEA*** !!!

Dielectric screening:

Reduction of the magnitude of electrostatic interactions due to the polarization of the surrounding solvent

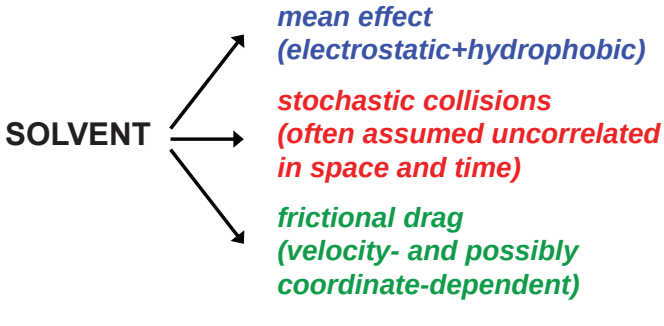


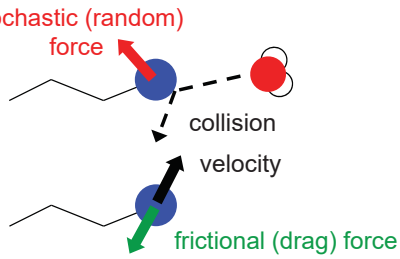
For small cavities (e.g. ions), screening reduces the force magnitude by ε (solvent permittivity) – for water, $\varepsilon \approx 80$. For large cavities (e.g. charged residues in proteins), the screening is more limited (but still large).

⇒ Charge-charge (*e.g.* ion-pairs), charge-dipole, and dipole-dipole (*e.g.* H-bonds) interactions in vacuum simulations are largely overweighted compared to what they would be in solution !

Implicit-solvent simulations

→ Mimic the forces exerted by water molecules without including these explicitly





→ ***Langevin equation of motion***

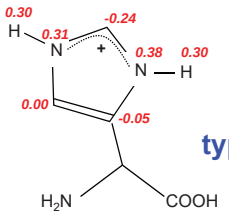
$$m_i \ddot{\boldsymbol{r}}_i = \underbrace{\boldsymbol{F}_i^{mean}(\boldsymbol{r}(t))}_{\text{mean force}} + \underbrace{\boldsymbol{R}_i(t)}_{\text{stochastic force}} - \underbrace{m_i \gamma_i \dot{\boldsymbol{r}}_i(t)}_{\text{frictional force}}$$

- The random and frictional forces are included by application of stochastic dynamics (SD; Langevin equation) rather than molecular dynamics (MD; Newton equation, appropriate for explicit-solvent simulations).
- SD generates a canonical (NVT) ensemble, while MD (without thermostat) generates a microcanonical (NVE) ensemble
- The mean force typically includes
  - → a polar (electrostatic screening) component – ***described in subsequent slides***
  - → a non-polar (van der Waals + hydrophobic) component - typically proportional to the solvent-accessible surface area (SASA) of the solute (or to fractions of the SASA contributed by different types of atoms - *e.g.* "polar" vs "non-polar" SASA)

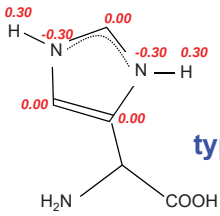
Electrostatic interactions in implicit-solvent simulations: Ad hoc fixes

- Reduction of the atomic charges (*e.g.* GROMOS96 type-B force field)

e.g. protonated His residue



type-A



type-B

charge set for explicit-solvent (+1) *charge set for vacuum (0)*

⇒ ***better than nothing - but very (very) crude !!!***

- Distance-dependent dielectric constant

$$V_{el}(r) = \frac{1}{4\pi\varepsilon_o}\frac{q\,q'}{\varepsilon_{dd}(r)\,r} \quad \text{with} \quad \varepsilon_{dd}(r) = n\,r$$

Typically used with n = 10, 40 or 80 nm^{-1} *(why do people always forget the units ?)*

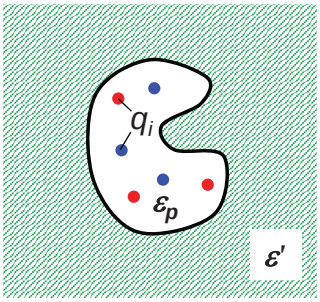
⇒ ***better than nothing - but absolutely no physical basis (turns Coulomb's law into a r^{-2} law) !!!***

- Screening functions

$$V_{el}(r) = \frac{1}{4\pi\varepsilon_o}\frac{q\,q'}{r}\frac{e^{-\kappa r}}{\varepsilon_{eff}}$$

⇒ ***better than nothing - but absolutely no physical basis (merely inspired from the Debye-Hückel model of ionic solutions) !!!***

Electrostatic interactions in implicit-solvent simulations: Continuum electrostatics (CE)



- Solve Poisson's equation (or Poisson-Boltzmann's)

$$\nabla\cdot\left[\varepsilon(\mathbf{r};\varepsilon_p,\varepsilon')\nabla\phi(\mathbf{r};\varepsilon_p,\varepsilon')\right] = -\varepsilon_o^{-1}\sum_i q_i\delta(\mathbf{r}-\mathbf{r}_i)$$

for the medium of heterogeneous permittivity

- Resolution possible analytically (sphere, ellipsoid) or numerically (*e.g.* by finite difference on a grid)

- Calculate solvation free energy as

$$\Delta G_s = (1/2)\sum_i q_i\left[\phi(\mathbf{r}_i;\varepsilon_p,\varepsilon_s) - \phi(\mathbf{r}_i;\varepsilon_p,\varepsilon_p)\right]$$

- Calculate Coulombic energy as

$$E_{Cb} = \frac{1}{4\pi\varepsilon_o\varepsilon_p}\sum_i\sum_{j>i}\frac{q_i q_j}{r_{ij}}$$

- The corresponding forces can also be computed (for MD or - better - SD)

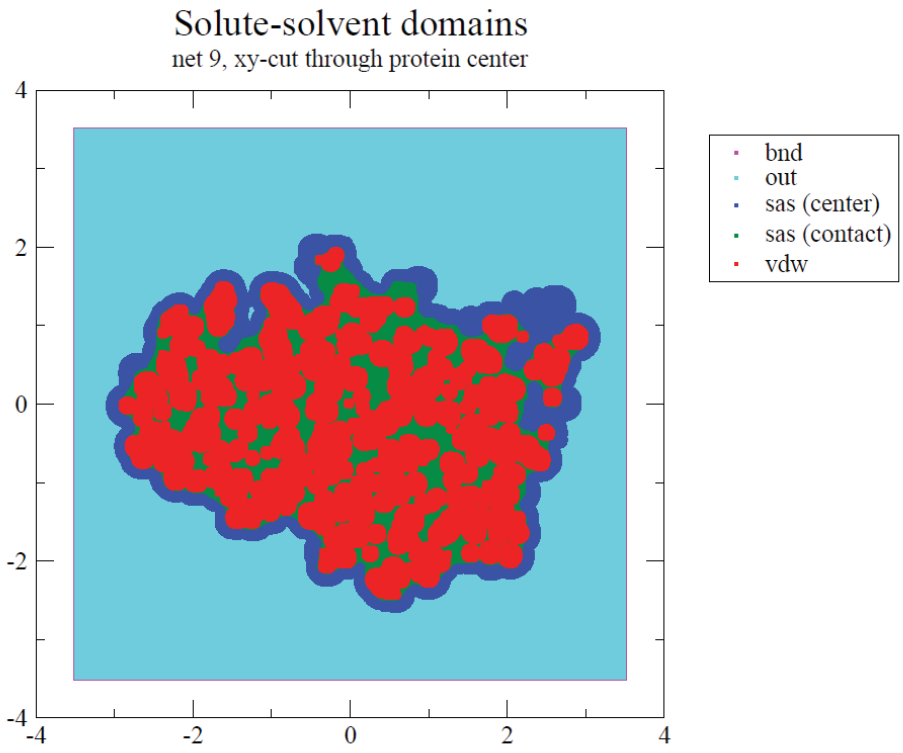
⇒ ***Problems:***

→ ***application of a macroscopic theory to microscopic systems***

→ ***value of ε_p (1, 2, 4, 20 ?) and surface definition (van der Waals or solvent-accessible; choice of atomic radii; choice of atomic charges ?) are crucial but ambiguous***

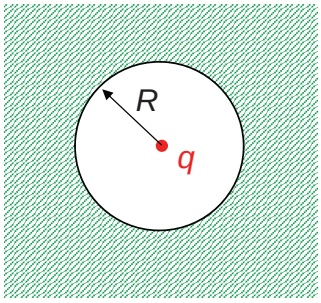
Surface definitions in continuum-electrostatics calculations

- Different **dielectric boundary** definitions for a protein



Continuum electrostatics: Two important solutions for the spherical case

- ***The case of a charge in a spherical cavity is treated by the Born model (1920)***



$$\phi_o = -\frac{q}{4\pi\varepsilon_o}\frac{\varepsilon_S - 1}{\varepsilon_S}\frac{1}{R}$$

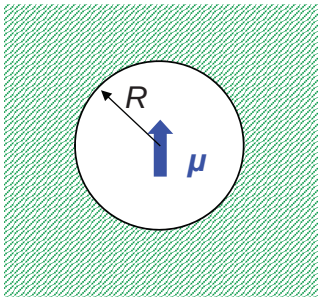
reaction potential from solvent at cavity center

$$\Delta G_s = -\frac{q^2}{8\pi\varepsilon_o}\frac{\varepsilon_S - 1}{\varepsilon_S}\frac{1}{R}$$

solvation free energy

$(\varepsilon_p = 1)$

- ***The case of a dipole in a spherical cavity is treated by the Onsager model (1936)***



$$\boldsymbol{E}_o = \frac{1}{4\pi\varepsilon_o}\frac{2(\varepsilon_S - 1)}{2\varepsilon_S + 1}\frac{\boldsymbol{\mu}}{R^3}$$

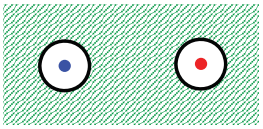
reaction field from solvent at cavity center

$$\Delta G_s = -\frac{1}{8\pi\varepsilon_o}\frac{2(\varepsilon_S - 1)}{2\varepsilon_S + 1}\frac{\boldsymbol{\mu}^2}{R^3}$$

solvation free energy

$(\varepsilon_p = 1)$

Electrostatic interactions in implicit-solvent simulations: Generalized Born (GB)

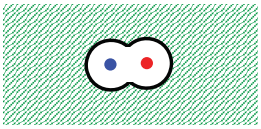


- Long-distance approximation for spherical ions

$$\Delta G_s = \frac{1}{4\pi\varepsilon_o}\left(\frac{1}{\varepsilon_s}-1\right)\left(\sum_i\sum_{j>i}\frac{q_i q_j}{r_{ij}}+\sum_i\frac{q_i^2}{2a_i}\right)$$

screened Coulomb + Born solvation - Coulomb

a_i: Born radius of the ion

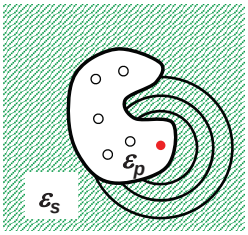


- Improved approximation (short distances + non-spherical case)

$$\Delta G_s = \frac{1}{4\pi\varepsilon_o}\left(\frac{1}{\varepsilon_s}-1\right)\left(\sum_i\sum_{j>i}\frac{q_i q_j}{f(r_{ij},a_{ij})}+\sum_i\frac{q_i^2}{2a_i}\right) \qquad f(r_{ij},a_{ij})=\left[r_{ij}^2+a_{ij}^2 e^{-r_{ij}^2/(2a_{ij}^2)}\right]^{1/2} \qquad a_{ij}=(a_i a_j)^{1/2}$$

a_i: effective Born radius

- The effective Born radii a_i are those reproducing the solvation free energy of each individual charge within the solute cavity through the Born equation



$$\Delta G_i = \frac{1}{4\pi\varepsilon_o}\left(\frac{1}{\varepsilon_s}-1\right)\frac{q_i^2}{2a_i}$$

Born equation

- obtained from continuum electrostatics (expensive)
- obtained from a shell-based estimate (approximate but fast)

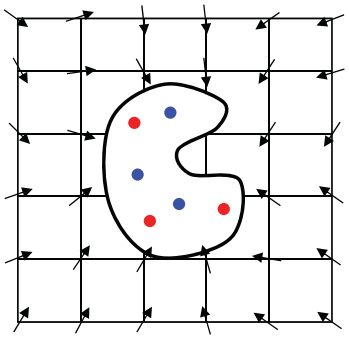
⇒ ***Problems:***

→ ***effective Born radii must be recalculated periodically***

→ ***value of ε_p and surface definition (including atomic radii) are crucial but ambiguous***

→ ***inherent approximations of the model...***

Electrostatic interactions in implicit-solvent simulations: Langevin dipoles (LD)



sometimes also called: protein-dipole Langevin-dipole (PDLD)

- Grid matching the solvent density
- Rotatable point dipoles at grid points
- Maximal dipole magnitude μ_o is that of a solvent dipole

- Dipoles obey a Langevin-type equation

$$\boldsymbol{\mu}_i = \mu_o \frac{\boldsymbol{E}_i}{E_i}\left[\frac{e^{\alpha_i}+e^{-\alpha_i}}{e^{\alpha_i}-e^{-\alpha_i}}-\frac{1}{\alpha_i}\right] \quad \text{with} \quad \alpha_i = C\frac{\mu_o E_i}{k_B T}$$

field (solute + solvent) at grid point

[iterative solution]

C: „willingness" of solvent dipoles to reorient according to the local field (related to ε_s and parameterized e.g. against explicit-solvent simulations)

$$C\to\infty \;\Rightarrow\; \alpha_i\to\infty \;\Rightarrow\; \boldsymbol{\mu}_i = \mu_o\frac{\boldsymbol{E}_i}{E}$$

(fully oriented)

⇒ ***Problems:***

→ ***finite-grid effects (especially at the solute-solvent boundary)***

→ ***dependency on the grid parameters and molecular orientation***

Short-range problem in implicit-solvent simulations

Implicit-solvent simulations:

- are computationally much less expensive than explicit-solvent simulations ($\rightarrow$ permit to tackle larger systems or/and longer timescales)
- imply instantaneous averaging over solvent configurations
- often handle appropriately the *long-range component of electrostatic interactions* ($\neq$ explicit-solvent simulations)

But...

- often involve many *ad hoc* approximations
- have difficulties accounting for the balance of solute-solute, solute-solvent and solvent-solvent interactions involved in biomolecular conformational equilibria (*e.g.* peptide or protein folding)
- do not handle accurately *short-range solute-solvent interactions* (*e.g.* solvation structure, saturation, electrostriction, hydrogen bonding, water molecules in protein cavities, ...)

$\Rightarrow$ ***Great qualitative tool - but quantitative results are often questionable...***

ELECTROSTATIC INTERACTIONS IN EXPLICIT-SOLVENT SIMULATIONS

Long-range problem in explicit-solvent simulations

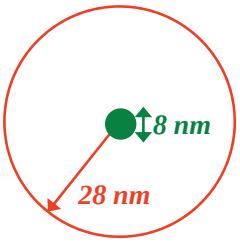
Example: ionic solvation

⇒ According to the Born model, the bulk solvation of an ion is reached within k_BT (2.5 kJ·mol^{-1} at 298K) for a droplet of radius

$$R_s = \frac{q_I^2}{8\pi\varepsilon_o k_B T}\frac{\varepsilon_S - 1}{\varepsilon_S}$$

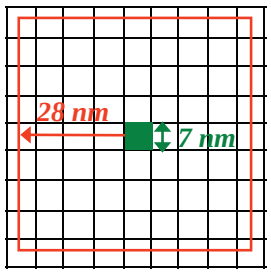
For a monovalent ion in water (ε_S=78), this evaluates to ~28 nm

⇒ Typical system size (Y2K+18): ~10'000 water molecules



droplet

⇒ ~3´000´000 molecules (maybe in ~15 years)

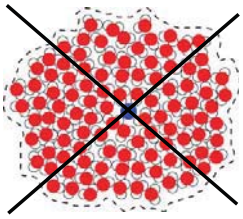


periodic box

⇒ ~6´000´000 molecules

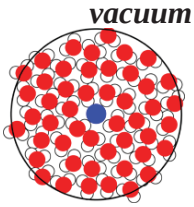
Importance:
- Many simulated observables are **very sensitive** to boundary conditions (system shape, size and surroundings) and treatment of electrostatics (examples follow...)
- The **most expensive part** of current simulations
- Quality of a model ⇒ **crudest approximation**, probably nowadays: (*i*) force-field, (*ii*) sampling, and (*iii*) electrostatics

Electrostatic interactions in explicit-solvent simulations

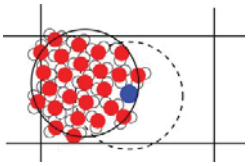


non-periodic boundary conditions (NPBC), macroscopic system [intractable]

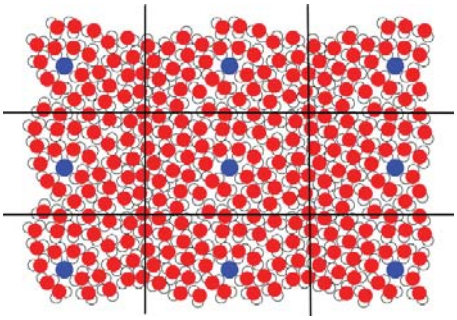
approximations (microscopic systems)



fixed boundary conditions (FBC) [+ evtl. soft wall, surface term, reaction-field term, ...]



periodic boundary conditions (PBC) with cutoff truncation (CT) [+ evtl. SH, SW, RF]

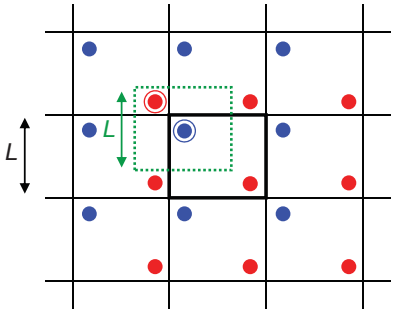


periodic boundary conditions (PBC) with lattice sum (LS) [incl. Ewald, P^{3}M, PME]

Explicit-solvent simulations with PBC/CT

⇒ Periodic boundary conditions with cutoff truncation (electrostatic interactions neglected beyond a specified cutoff range) – for reducing computational costs and artificial periodicity

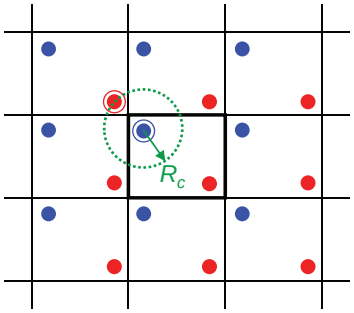
minimum-image convention



Each charge in the reference box only interacts with the closest periodic image (minimum-image) of each other charge

→ *Calculation is O[N^2], often too expensive*
→ *Enhances anisotropy in the system*
→ *Nowadays seldom used*

spherical cutoff



Each charge in the reference box only interacts with the minimum-image of another charge if the corresponding minimum-image distance is smaller than a cutoff distance R_c

→ *Calculation is O[$N \cdot R_c^3$], with R_c a free parameter*
→ *Typically $R_c \leq L/2$ (no interactions with multiple copies, no self-interactions, simpler code)*
→ *Partially suppresses anisotropy*
→ *More commonly used*

The problem with cutoff simulations

Computationally-affordable cutoff distances R_c (*e.g.* 0.8-1.4 nm) are *much shorter* than the range of electrostatic interactions in typical molecular systems !

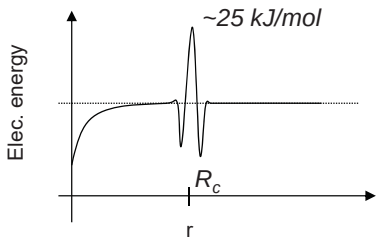
Consequences:

- Cutoff noise and heating
- Distortion of pair properties at the cutoff distance
- Error in simulated observables (liquids, ionic solvation, ionic solutions, biomolecular systems...) – which depend on the cutoff distance as well as system shape, size and boundary conditions

PBC/CT: Atom- vs molecule-based cutoff

⇒ Cutoff truncation leads to noise (heating) and distortions in pair properties

atom-based cutoff truncation



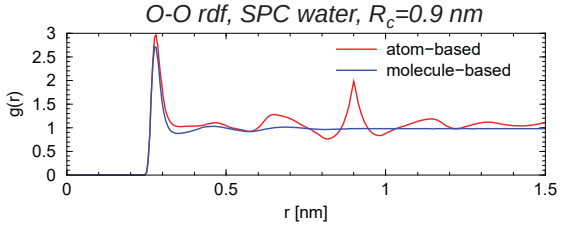
→ *Cutoff noise of the order of*

$$(4\pi\varepsilon_o)^{-1} q\, q'\, R_c^{-1}$$

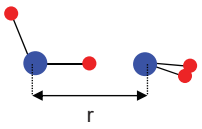
→ *Continuous but energy drift (heating), due to the inaccurate integration of the equations of motion (sharp energy variations) at finite timestep*

→ *Strong artifacts in pair properties*

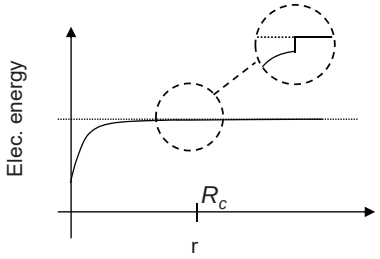
O-O rdf, SPC water, R_c=0.9 nm



e.g. water dimer



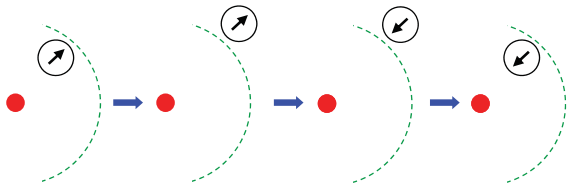
molecule-based cutoff truncation



→ *Cutoff noise of the order of*

$$(4\pi\varepsilon_o)^{-1} \mu\, \mu'\, R_c^{-3}$$

→ *Discontinuous, thus energy drift (heating) due to non-conservative processes*

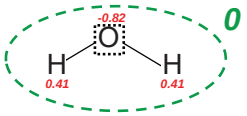


Net increase in the system energy !!!

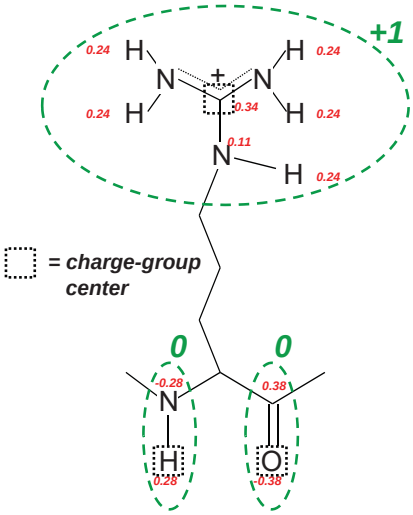
PBC/CT: Group-based cutoff

⇒ Generalization of the molecule-based cutoff

water (SPC)

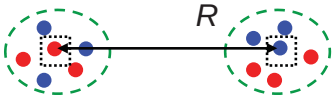


arginine (GROMOS)



→ *Relies on the definition of charge groups*

→ *Charge groups interact fully or not at all depending on the minimum-image distance between their centers*



R ≤ R_c: full interaction
R > R_c: no interaction

→ *Charge groups should be neutral as much as possible, to reduce cutoff noise*

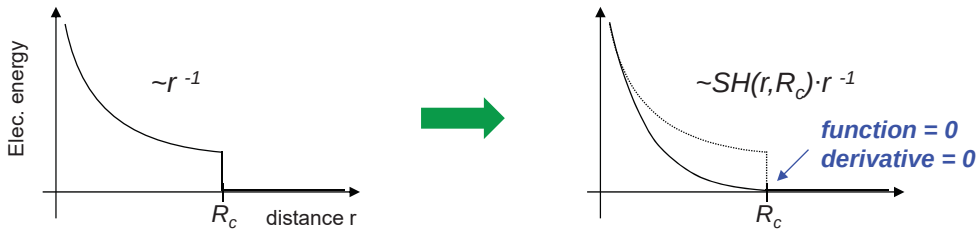
charge-charge	~	$(4\pi\varepsilon_o)^{-1} q\, q'\, R_c^{-1}$
charge-dipole	~	$(4\pi\varepsilon_o)^{-1} q\, \mu\, R_c^{-2}$
dipole-dipole	~	$(4\pi\varepsilon_o)^{-1} \mu\, \mu'\, R_c^{-3}$

→ *Charge-groups should be reasonably small, to preserve accurate short-range interactions*

→ *May require modification of the "best" atomic charge set*

PBC/CT+SH: Shifting functions

⇒ Avoid abrupt truncation of the electrostatic energy/forces through multiplication by a function $SH(r;R_c)$



→ *Shifting function is generally a polynomial*

→ *Applied with either atom-based or group-based cutoff*

→ *In both cases, entirely removes cutoff noise (heating)*

But:

→ *Interaction is unphysical (i.e. no longer Coulombic), and altered over the whole distance range*

→ *Charges must be parametrized consistently (to compensate for this change)*

examples:

CHARMm shifted dielectric:

$$SH(r,R_c)=1-\frac{2r^2}{R_c^2}+\frac{r^4}{R_c^4}$$

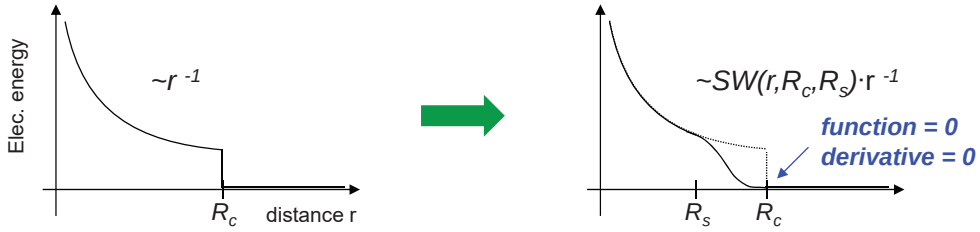
generalized force shift:

$$SH(r,R_c)=1+\frac{r^{\beta+1}}{\beta R_c^{\beta+1}}-\frac{r}{R_c}+\frac{r}{\beta R_c}$$

with β=1,2 or 3

PBC/CT+SW: Switching functions

⇒ Avoid abrupt truncation of the electrostatic energy/forces through multiplication by a function $SW(r;R_s,R_c)$ which is one below R_s



→ *Similar to shifting function, but the short-range component of the interaction remains Coulombic (more physical)*

→ *Applied with either atom-based or group-based cutoff*

→ *In both cases, entirely removes cutoff noise (heating)*

But:

→ *If R_s is too close to R_c, there may be very large forces on atom pairs close to the cutoff distance !*

→ *Charges must be parametrized consistently*

Group-based cutoff, shifting and switching functions are ways to remove the cutoff noise (heating) and – to some extent – the distortions of pair properties near the cutoff

But:

They are generally not sufficient to prevent numerous other artifacts arising from cutoff truncation !

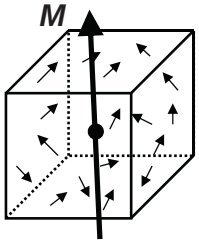
PBC/CT: Cutoff artifacts in simulations of liquids

⇒ The dielectric permittivity (related to the dipole-moment fluctuations) is largely underestimated when cutoff truncation is applied

The *dielectric permittivity* ε of a substance describes its ability to screen electrostatic interactions between charges:

in vacuum ***in medium***

It can be evaluated from a pure-liquid simulation from the fluctuations of the box dipole moment. For straight cutoff truncation:



$$\varepsilon = \frac{9\varepsilon_o V k_B T + 2[\langle \boldsymbol{M}^2 \rangle - \langle \boldsymbol{M} \rangle^2]}{9\varepsilon_o V k_B T - [\langle \boldsymbol{M}^2 \rangle - \langle \boldsymbol{M} \rangle^2]}$$

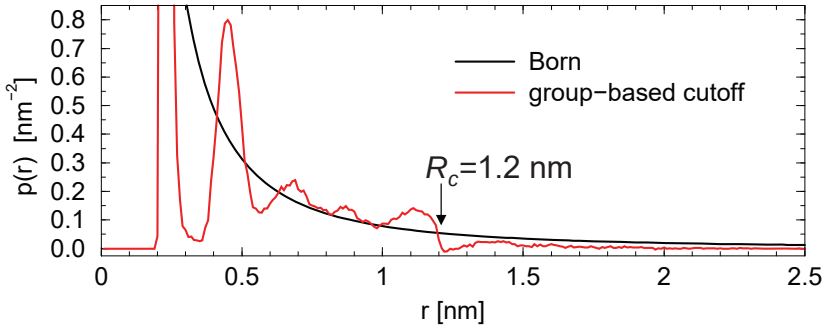
***M**: box dipole moment*
V: box volume
k_B: Boltzmann's cst
T: absolute temperature
ε_o: vacuum permittivity

→ *for cutoff simulations of SPC water with R_c=0.9 nm, one finds $\varepsilon \approx 5$ (to be compared to the experimental value of 78 !)*

→ *dipole-moment fluctuations are not stabilized because the medium outside the cutoff sphere is vacuum !*

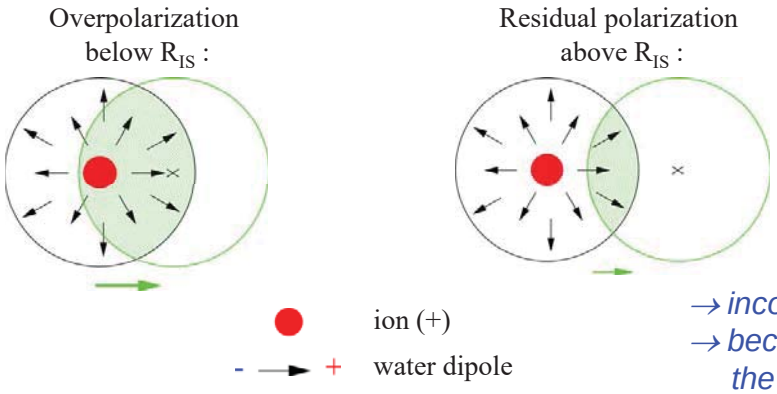
PBC/CT: Cutoff artifacts in simulations of ionic solvation

⇒ Radial polarization around a solvated sodium ion



$$p_{Born} = \frac{\varepsilon - 1}{\varepsilon} \frac{q}{4\pi r^2}$$

$$p_{sim} = \frac{1}{4\pi r^2 dr} \sum_{i \in shell(r,dr)} \frac{\boldsymbol{\mu}_i \cdot \boldsymbol{r}_i}{r_i}$$

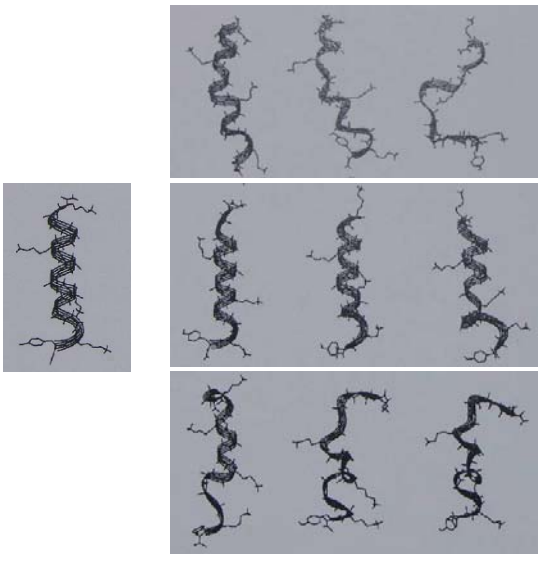


→ *incorrect solvation free energy*
→ *because the medium outside the cutoff sphere is vacuum !*

★

PBC/CT: Cutoff artifacts in biomolecular simulations

⇒ Simulations of a 17-residue peptide in explicit water
Ac – Tyr – [Lys - $(Ala)_4]_3$ – Lys – NH_2
Schreiber & Steinhauser, Biochemistry 31 5856 (1992)

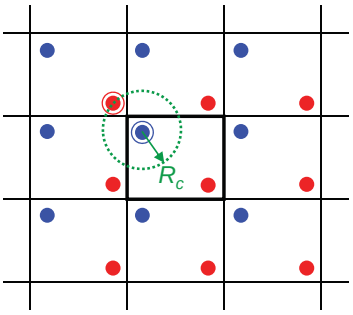


R_c	
1.4 nm	→ *unfolds*
1.0 nm	→ *stable*
0.6 nm	→ *unfolds*

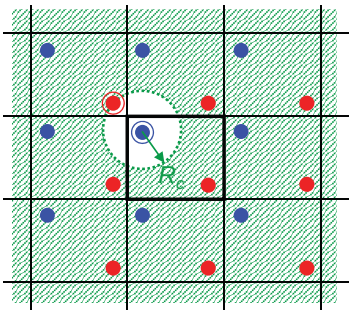
→ *System properties strongly depend on cutoff (cutoff vs Lys-Lys distance)*

⇒ Assume that the medium outside the cutoff sphere of each particle is a dielectric continuum of permittivity ε_s equal to that of the solvent [Barker & Watts, 1973]

straight cutoff

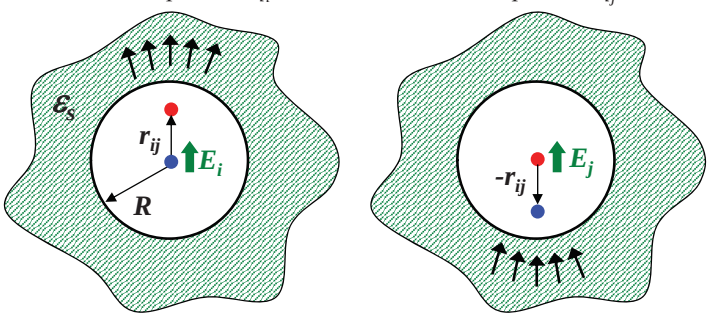


cutoff with RF correction



PBC/CT+RF: Reaction-field correction

cutoff sphere of q_i | cutoff sphere of q_j



Onsager reaction field:

$$\boldsymbol{E}_i = \frac{q_j}{4\pi\varepsilon_o}\frac{2(\varepsilon_s-1)}{2\varepsilon_s+1}\frac{\boldsymbol{r}_{ij}}{R_c^3}$$

$$\boldsymbol{E}_j = -\frac{q_i}{4\pi\varepsilon_o}\frac{2(\varepsilon_s-1)}{2\varepsilon_s+1}\frac{\boldsymbol{r}_{ij}}{R_c^3}$$

⇒ The corresponding forces $\boldsymbol{F}_i = q_i\,\boldsymbol{E}_i$ and $\boldsymbol{F}_j = q_j\,\boldsymbol{E}_j$ "appear" to derive from an effective interaction energy $V_{corr}(r_{ij}) = \frac{q_i q_j}{4\pi\varepsilon_o}\frac{\varepsilon_s-1}{2\varepsilon_s+1}\frac{r_{ij}^2}{R_c^3}$

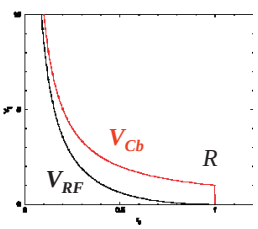
⇒ Reaction-field corrected interaction energy ($R < L/2$) $V_{RF}(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_o}\sum_{i,j>i,\bar{r}_{ij}\le R_c} q_i q_j \left[\frac{1}{\bar{r}_{ij}} + \frac{\varepsilon_s-1}{2\varepsilon_s+1}\frac{\bar{r}_{ij}^2}{R_c^3} - \frac{3\varepsilon_s}{2\varepsilon_s+1}\frac{1}{R_c}\right]$

→ formally correct for ***homogeneous dipolar*** systems only (ions ? biomolecules ?)

→ usually implemented with a group-based cutoff (why ?)

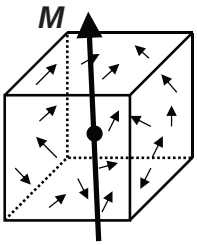
→ for large ε_s, a physically-based ***shifting function***

→ ***cutoff damps periodicity effects*** compared to lattice-sums



PBC/CT+RF: Improvement of the simulated properties

⇒ Dielectric permittivity of liquids:



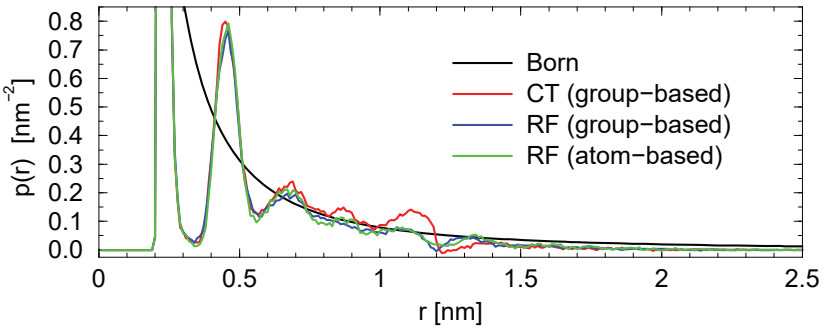
$$\varepsilon = \frac{3(2\varepsilon_s+1)\varepsilon_o V k_B T + 2\varepsilon_s[\langle \boldsymbol{M}^2\rangle - \langle \boldsymbol{M}\rangle^2]}{3(2\varepsilon_s+1)\varepsilon_o V k_B T - [\langle \boldsymbol{M}^2\rangle - \langle \boldsymbol{M}\rangle^2]}$$

***M**: box dipole moment*
V: box volume
k_B: Boltzmann's cst
T: absolute temperature
ε_o: vacuum permittivity
ε_s: reaction-field permittivity (BW)

→ *for reaction-field simulations of SPC water with R_c=0.9 nm, one finds $\varepsilon \approx 65$ (to be compared to the experimental value of 78)*

→ *dipole-moment fluctuations are stabilized because the medium outside the cutoff sphere is solvent !*

⇒ Radial polarization around a solvated sodium ion

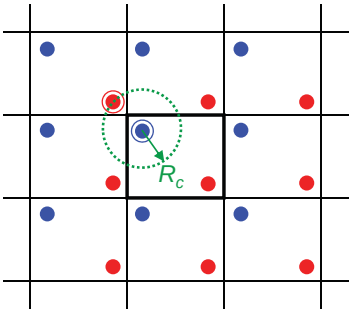


★

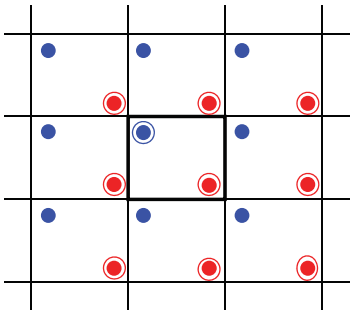
PBC/LS: Lattice-sum methods

⇒ The full periodicity of the system is taken into account

straight cutoff

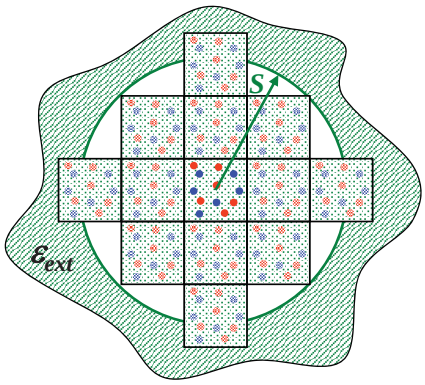


Lattice sum



PBC/LS: Ewald interactions

Considering Coulomb interactions in a strictly periodic system



Charges $\{q_i\}$ at $\{\boldsymbol{r}_i\}$ within reference box

$\Rightarrow$ In the limit $S\to\infty$

$$\phi(\boldsymbol{r}) = \phi_{ext}(\boldsymbol{r}) + \frac{1}{4\pi\varepsilon_o}\sum_i q_i\psi(\boldsymbol{r}-\boldsymbol{r}_i)$$

$\Rightarrow$ Extrinsic potential [neglected$\to$"tinfoil"]

$$\phi_{ext}(\boldsymbol{r}) = \underbrace{\left[\frac{1}{\varepsilon_o(2\varepsilon_{ext}+1)V}\sum_i q_i\boldsymbol{r}_i\right]\cdot\boldsymbol{r}}_{\text{minus extrinsic field}} - \underbrace{\frac{1}{6\varepsilon_o V}\sum_i q_i\boldsymbol{r}_i^2}_{\text{extrinsic potential (still discussed)}}$$

$\Rightarrow$ Intrinsic potential

$$\nabla^2\psi(\boldsymbol{r}) = -4\pi\left[\delta_p(\boldsymbol{r}) - V^{-1}\right]$$

compensating (background) charge density

$$\langle\psi(\boldsymbol{r})\rangle = 0 \;\; and \;\; \langle\nabla\psi(\boldsymbol{r})\rangle = \boldsymbol{0}$$

$\Rightarrow$ Potential energy (reversible charging, intrinsic)

$$V(\boldsymbol{r}) = \frac{1}{8\pi\varepsilon_o}\sum_{i,j} q_i q_j\tilde{\psi}(\boldsymbol{r}_{ij}) \quad with \quad \tilde{\psi}(\boldsymbol{r}) = \begin{cases} \psi(\boldsymbol{r}) & for\;\; r\neq 0 \\ \psi^o = \lim_{r\to 0}[\psi(\boldsymbol{r}) - r^{-1}] & for\;\; r = 0 \end{cases}$$

removes Coulomb singularity

→ formally correct for ***exactly periodic*** systems only

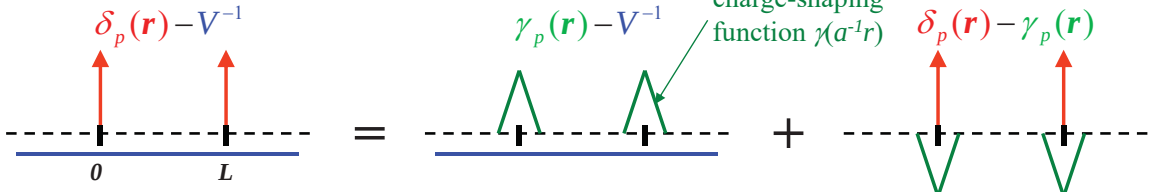
→ cutoff bias removed at the expense of enforcing ***artificial periodicity***

PBC/LS: splitting method for Ewald interactions

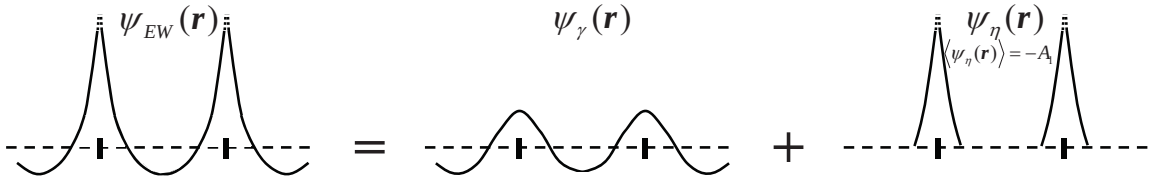
$\Rightarrow$ To solve

$$\nabla^2\psi_{EW}(\boldsymbol{r}) = -4\pi\left[\delta_p(\boldsymbol{r}) - V^{-1}\right] \quad with \quad \langle\psi_{EW}(\boldsymbol{r})\rangle = 0 \;\; and \;\; \langle\nabla\psi_{EW}(\boldsymbol{r})\rangle = \boldsymbol{0}$$

$\Rightarrow$ Splitting of the charge density



$\Rightarrow$ Splitting of the potential $\psi_{EW}(\boldsymbol{r}) = \psi_\gamma(\boldsymbol{r}) + \psi_\eta(\boldsymbol{r}) + A_1$



smooth & long-ranged → *k*-space singular & short-ranged → *r*-space

$$\psi_\gamma(\boldsymbol{r}) = \frac{4\pi}{V}\sum_{\boldsymbol{l}\in\mathbb{Z}^3,\boldsymbol{l}\neq 0}\frac{\hat{\gamma}(ak)}{k^2}\cos(\boldsymbol{k}\cdot\boldsymbol{r}) \qquad \psi_\eta(\boldsymbol{r}) = \sum_{\boldsymbol{n}\in\mathbb{Z}^3,\boldsymbol{n}\neq 0}\left\|\boldsymbol{r}_{ij}+\underline{\boldsymbol{L}}\boldsymbol{n}\right\|^{-1}\eta(a^{-1}\left\|\boldsymbol{r}_{ij}+\underline{\boldsymbol{L}}\boldsymbol{n}\right\|)$$

$\Rightarrow$ Derived (analytical) quantities

$$\hat{\gamma}(ak) = \begin{cases} 4\pi k^{-1}a^{-3}\int_0^\infty dr\, r\sin(kr)\gamma(a^{-1}r) & for\;\; k\neq 0 \\ 1 & for\;\; k = 0 \end{cases} \qquad \eta(a^{-1}r) = 4\pi a^{-3}\int_r^\infty d\rho\,(\rho - r)\gamma(a^{-1}r)$$

$$A_1 = -\frac{4\pi}{V}\int_0^\infty dr\, r\,\eta(a^{-1}r)$$

PBC/LS: splitting method for Ewald interactions

⇒ Potential energy (intrinsic)

$$V_{EW}(\boldsymbol{r}) = V_{\gamma}(\boldsymbol{r}) + V_{\eta}(\boldsymbol{r}) + V_A + V_{slf}$$

$$V_{\gamma}(\boldsymbol{r}) = \frac{1}{2\varepsilon_o V}\sum_{i,j} q_i q_j \sum_{\boldsymbol{l}\in\mathbb{Z}^3,\boldsymbol{l}\neq 0} \frac{\hat{\gamma}(ak)}{k^2}\cos(\boldsymbol{k}\cdot\boldsymbol{r}_{ij})$$ ..... *k*-space (Ewald or FFT) contains self term A_2

$$V_{\eta}(\boldsymbol{r}) = \frac{1}{4\pi\varepsilon_o}\sum_{i,j>i} q_i q_j \sum_{\boldsymbol{n}\in\mathbb{Z}^3} \left\|\boldsymbol{r}_{ij}+\underline{\boldsymbol{L}}\boldsymbol{n}\right\|^{-1}\eta(a^{-1}\left\|\boldsymbol{r}_{ij}+\underline{\boldsymbol{L}}\boldsymbol{n}\right\|)$$ ..... *r*-space contains average term A_1; $R < L/2 \rightarrow$ one $\boldsymbol{n}$-term with $\bar{\boldsymbol{r}}_{ij}$; watch out exclusions

$$V_A = \frac{1}{8\pi\varepsilon_o}[A_1\left(\sum q_i\right)^2 - (A_1+A_2)\sum q_i^2]$$ ..... removes inappropriate terms in pairwise component

$$V_{slf} = \frac{1}{8\pi\varepsilon_o}(A_1+A_2+A_3)\left(\sum q_i\right)^2$$ ..... self term (Wigner, ψ_o)

⇒ Derived quantities

$$A_2 = \frac{4\pi}{V}\sum_{\boldsymbol{l}\in\mathbb{Z}^3,\boldsymbol{l}\neq 0}\frac{\hat{\gamma}(ak)}{k^2}$$

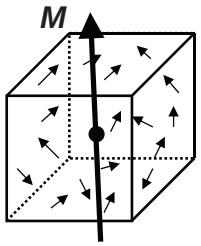
numerical
(quasi-analytical for cubic box)

$$A_3 = \lim_{r\to 0} r^{-1}[\eta(a^{-1}r)-1] + \sum_{\boldsymbol{n}\in\mathbb{Z}^3,\boldsymbol{n}\neq 0}\left\|\underline{\boldsymbol{L}}\boldsymbol{n}\right\|^{-1}\eta(a^{-1}\left\|\underline{\boldsymbol{L}}\boldsymbol{n}\right\|)$$

first term is analytical
$R < L \rightarrow$ second term is often zero

PBC/LS: improvement of the simulated properties

⇒ Dielectric permittivity of liquids:

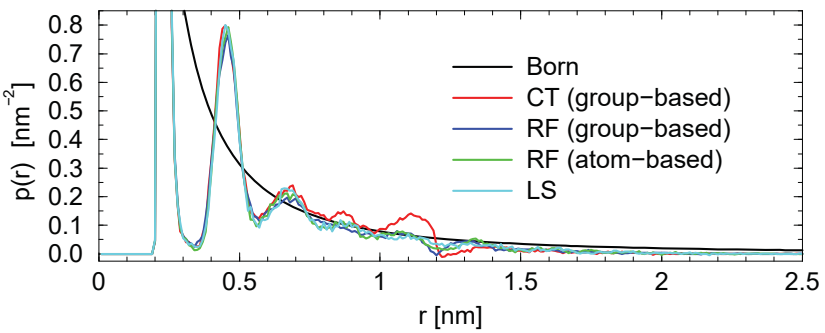


$$\varepsilon = \frac{3(2\varepsilon_s+1)\varepsilon_o V k_B T + 2\varepsilon_s[\langle \boldsymbol{M}^2\rangle - \langle \boldsymbol{M}\rangle^2]}{3(2\varepsilon_s+1)\varepsilon_o V k_B T - [\langle \boldsymbol{M}^2\rangle - \langle \boldsymbol{M}\rangle^2]}$$

$\boldsymbol{M}$: box dipole moment
V: box volume
k_B: Boltzmann's cst
T: absolute temperature
ε_o: vacuum permittivity
ε_s: external permittivity (LS)

→ for lattice-sum (tinfoil) simulations of SPC water, one finds $\varepsilon \approx 65$, similarly to RF (to be compared to the experimental value of 78)

⇒ Radial polarization around a solvated sodium ion



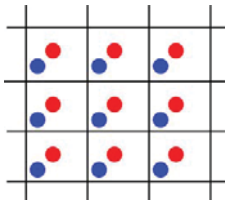
PBC/LS: periodicity-induced artifacts

LS (Ewald/P³M/PME) methods :

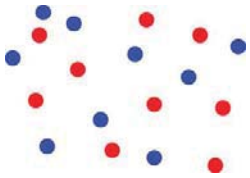
- Now routinely used for the simulation of proteins/nucleic acids ⇒ stable trajectories
- Often assumed exact ⇒ OK if we **assume periodicity to be an intrinsic property of the system** (for solutions, this is an approximation !)
- Periodic system ⇒ finite concentration

  e.g. 1 solute + 1000 water → 0.05 molal solution
  solute charge 10 *e* → 5 molal ionic strength !

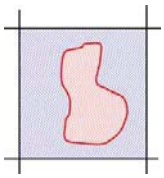
- Periodic system ≠ Realistic solution at the same concentration

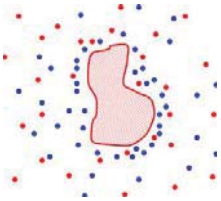


Anisotropy
No interaction at L/2



- Non-neutral solute + background ≠ Non-neutral solute + counter-ions

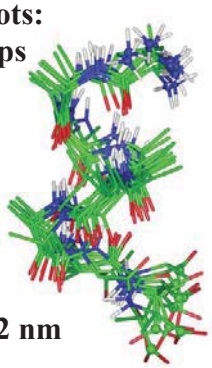




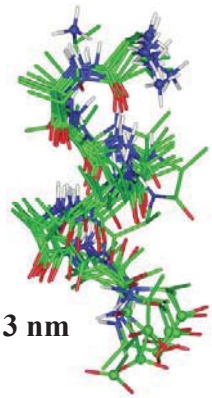
⇒ the nature and magnitude of possible artifacts should be assessed

PBC/LS: periodicity-induced artifacts in simulations of a zwitterionic polyalanine octapeptide

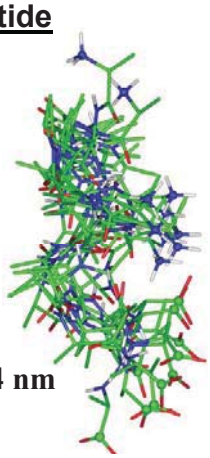
snapshots: 10x100ps



L = 2 nm

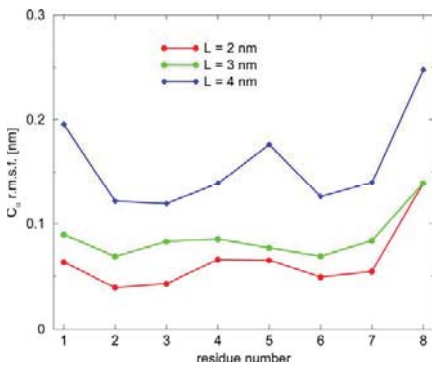


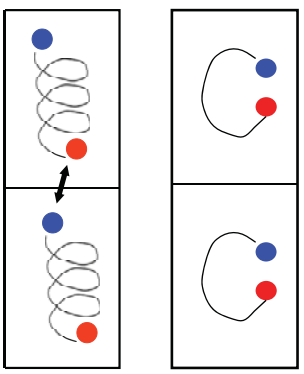
L = 3 nm



L = 4 nm

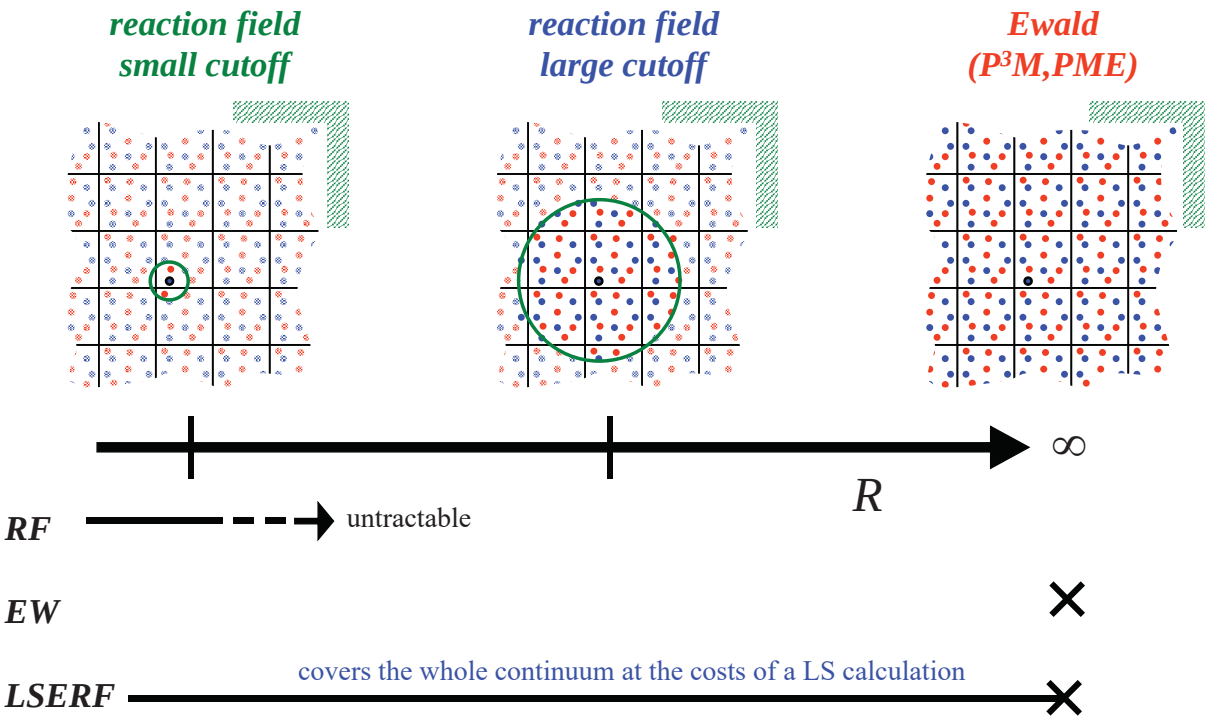
LS (P3M) electrostatics





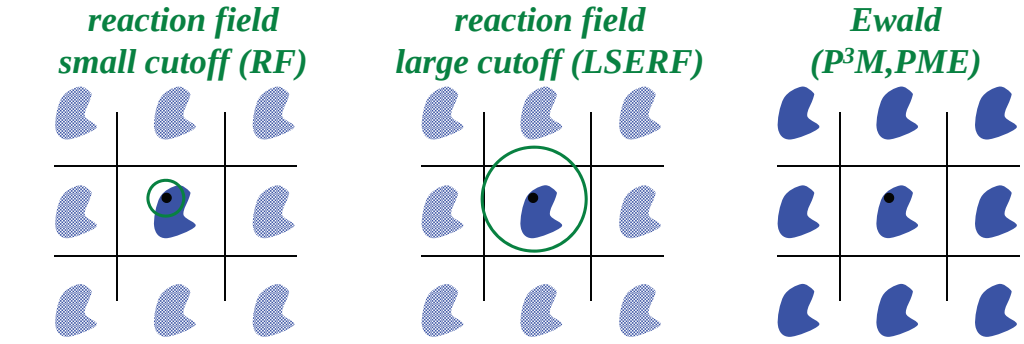
Combining the LS and RF methods: LSERF

(Heinz & Hünenberger, JCP, 2005)



Combining the LS and RF methods: LSERF

⇒ ***Applications:*** use a cutoff value in (bio-)molecular simulations reducing the inconveniences of both methods



medium beyond cutoff is not homogeneous solvent

compromise solution (R~L/2 or slightly larger)

artificial periodicity in the solution

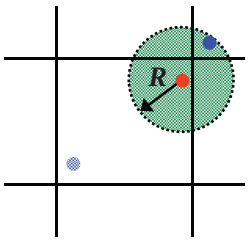
TIME-SAVING TECHNIQUES

Non-bonded interactions

⇒ **Non-bonded interactions (energy, forces, virial) are the CPU-intensive component of classical simulations.**

→ **a number of time-saving techniques can speed up the evaluation of the short-range interaction (within a cutoff distance)**

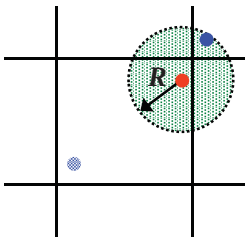
cutoff-based methods



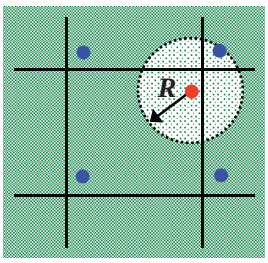
→ *minimum image only*
→ *interaction neglected beyond R*
→ *cutoff R is a physical parameter*
→ *also typical for van der Waals interactions*

lattice-sum methods

real-space



reciprocal-space



→ *all periodic images*

→ *cutoff R is a numerical parameter*
→ *also possible for van der Waals interactions*

Verlet pairlist algorithm

In both cases, pairs with a minimum-image distance smaller than R must be found

⇒ ***Standard algorithm***

```
Loop over unique atom pairs
   Calculate minimum-image vector   r̄
   Evaluate interaction if   r̄ ≤ R
End Loop
```

⇒ O[$N(N\text{-}1)/2$]~O[N^2]

⇒ ***Verlet pairlist algorithm***

pairlist generation

```
Loop over unique atom pairs
   Calculate minimum-image vector   r̄
   Pair → pairlist array if   r̄ ≤ R
End Loop
```

→ every n simulation steps (5-10)
→ use same PL between updates
⇒ O[$n^{-1}N^2$]

pairlist array → ***PL*** →

interaction evaluation

```
Loop over pairlist
   Evaluate pair interaction
End Loop
```

→ every simulation step
⇒ O[NR^3]~O[N]

⇒ ***Verlet extended-pairlist algorithm***

→ pairlist generation: use extended cutoff $R' > R$
→ interaction evaluation: test for $\bar{r} \le R$

} reduces the error of approximate pairlist between updates

Time-saving techniques

⇒ ***Twin-range method***

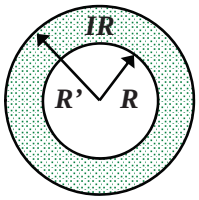
```
Loop over unique atom pairs
   Calculate minimum-image vector   r̄
   Pair → pairlist array if   r̄ ≤ R
   Evaluate pair IR interaction if   R < r̄ ≤ R'
End Loop
```

→ use same PL and IR between updates

pairlist array → ***PL*** , ***IR*** (intermediate range interactions) →

```
Loop over pairlist
   Evaluate pair interaction
End Loop
Add IR interaction
```

→ every simulation step



} cutoff effectively extended to R' (neglect of high-frequency fluctuations in IR interaction)

⇒ ***Use of atom groups***

```
Loop over unique group pairs
   Calculate minimum-image vector   r̄_g
   Pair → group pairlist array if   r̄_g ≤ R'
End Loop
```

→ groups of *e.g.* covalently-linked atoms
→ R'-R = twice the radius of the largest group
⇒ O[$m^{-2}N^2$], m = average number of atoms per group

group-pairlist array → ***PL*** →

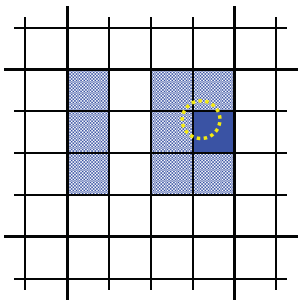
```
Loop over group pairlist
   Loop over atoms in groups
      Calculate minimum-image
         interatomic vector   r̄
      Evaluate pair interaction if   r̄ ≤ R
   End Loop
End Loop
```

→ every simulation step

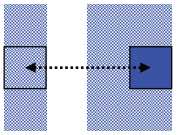
Fast pairlist generation

The standard double-loop algorithm for pairlist generation scales as $O[N^2]$
⇒ Other algorithms may speed up the process and bring the scaling to $O[N]$
⇒ They rely on the discretization of the computational box by smaller grid cells

linked-list methods



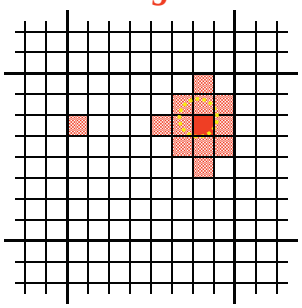
cell-pair array:



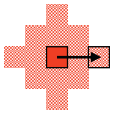
list of interacting cell pairs (periodicity)

→ cell size slightly larger than R
→ many extra pairs included
→ fast and linear only if $R<<L$

grid-cell methods



mask array:



list of relative offsets between interacting cells (no periodicity)

→ cell size comparable to atom size
→ fewer extra pairs included
→ fast and linear with carefully-optimized cell size

⇒ Methods with arbitrary or mixed cell sizes also exist
⇒ Many attempts to combine these with parallelization and vectorization

Linked-list and grid-cell algorithms

Drawbacks (both methods):

⇒ Initial pairlist contains extra pairs at distances larger than R
→ May induce anisotropy ("cube-corner" effect; especially for large cells)
→ Can be removed by filtering the initial pairlist

⇒ Pairlist is unsorted (pairs are listed in arbitrary order)
→ More difficult to handle exclusions and grouping (*e.g.* solute *vs* solvent)
→ Non-sequential array access may slow down subsequent interaction evaluation

Grid-cell methods:

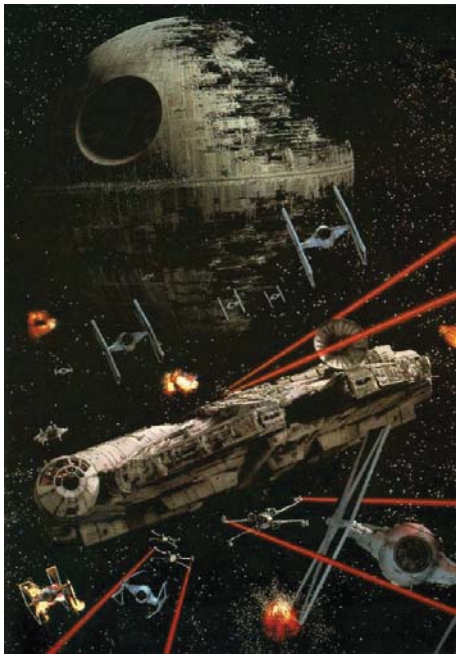
⇒ Potentially more efficient than linked-list methods
→ No restriction on grid-cell size
→ Fewer extra pairs in initial pairlist (better approximation of cutoff sphere)

⇒ Efficiency loss for small cells (requires careful tuning of cell size)
→ High memory requirements (mask array)
→ Handling of numerous empty cells
→ Handling of periodicity for all cell pairs

Can be alleviated
Heinz & Hünenberger, JCC **25** 1474 (2004)

may be alleviated by Bekker algorithm
Mol. Simul. **14** 137 (1995)
[27 pairlists with primary cell in specific neighbour box and secondary cell in reference box]

The real motivation...



"An efficient solution of the near neighbors problem would advance many important applications [including] futuristic battle area management. A one-pass engagement against many thousands of high-speed opponents requires the fast redetermination of near neighbors to ensure effective targeting in real time"

J.Boris, *J. Comput. Phys.* **66**, 1 (1986)

Laboratory for computational physics
US Naval Research Laboratory

Computer Simulation in Chemistry, Biology and Physics

P.H. Hünenberger

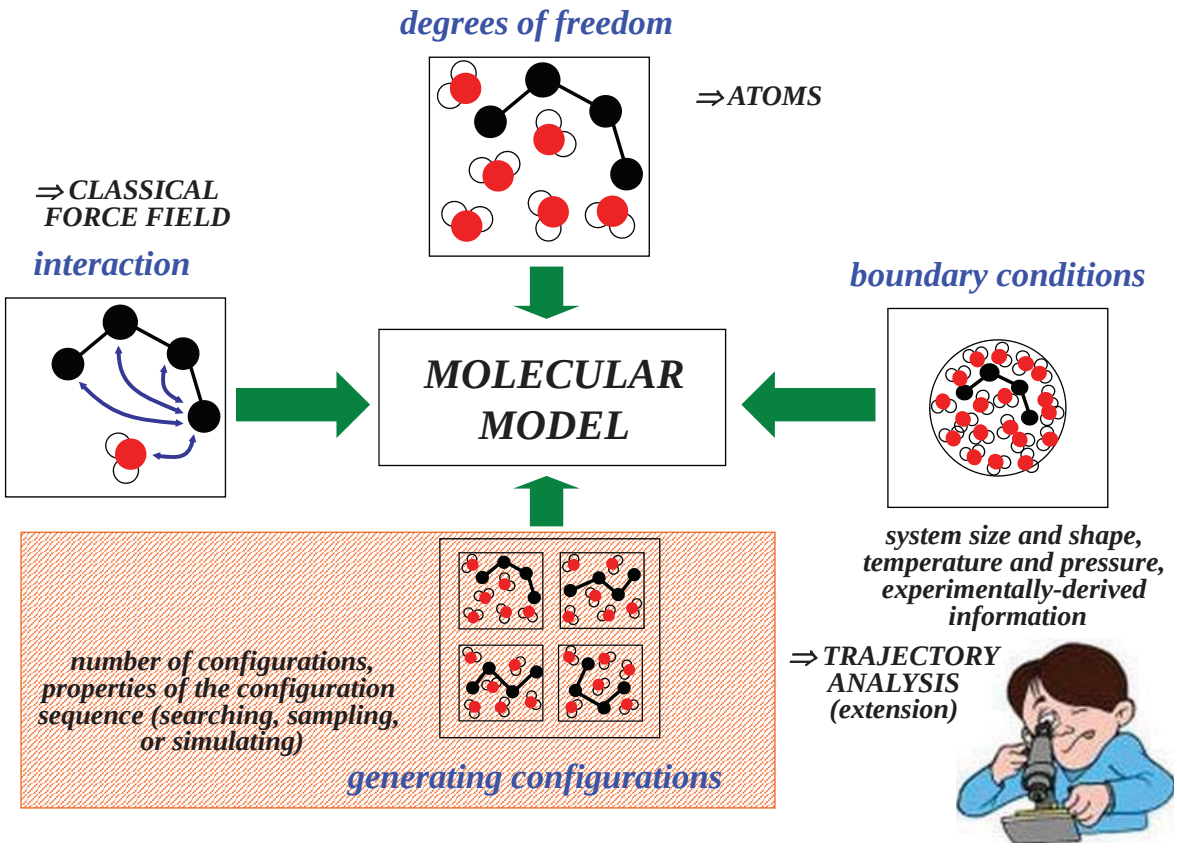
COMPUTER SIMULATION OF MOLECULAR SYSTEMS

Herbstsemester 2019
Tuesday 9:45-11:30 a.m.
HCI D2

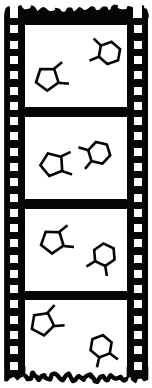
Lecture 529-0004-00
www.csms.ethz.ch/education/CSCBP

LECTURE 6 (WEEK 7): Simulation analysis

Four basic choices defining a molecular model



Analysis of trajectories: Overview



Trajectory

$\boldsymbol{r}(t)$ *Positions*

$\boldsymbol{p}(t)$ *Momenta*

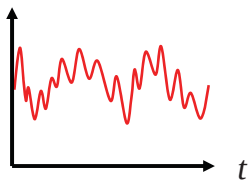
3N-dimensional

within reference box if PBC

Microscopic (instantaneous) observable

$Q(\boldsymbol{r},\boldsymbol{p})$

$Q(t)$



Time series of the instantaneous observable

(value associated to each trajectory configuration)

Statistical analysis (operator)

$\hat{O}_t$

$Q = \hat{O}_t Q(t)$

Q

Macroscopic observable

(can be compared with experimental data)

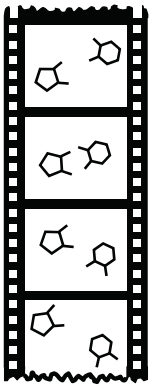
Microscopic observable

- *can be a scalar, a vector or a matrix*
- *can be bounded, unbounded or periodic*
- *can be continuous (real/integer value) or discrete (on-off)*
- *can also be a probability distribution function (histogram) along multiple occurrences of some scalar observable (e.g. all O-O distances in water)*

Statistical operator

- *acts on the time series*
- *can act on one observable or multiple observables*
- *can also involve time series from distinct simulations (with different system parameters, e.g. V, P, T, ...)*

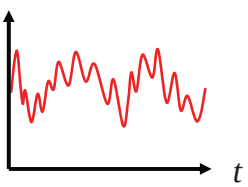
Calculating the time series of a microscopic observable



$\boldsymbol{r}(t)$ $\boldsymbol{p}(t)$

$Q(\boldsymbol{r},\boldsymbol{p})$

$Q(t)$



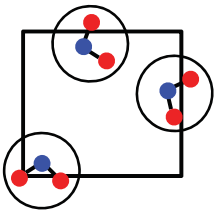
Periodic gathering

- When simulating under **periodic boundary conditions** (PBC), we monitor the coordinates of **one periodic copy** of each particle, but we actually simulate an **infinite periodic system**

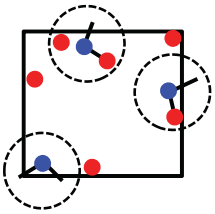
  → This makes no difference for the **momenta** (velocities)

  → For the **coordinates**, the restriction to a single copy (in the output trajectory) may be performed in different ways, *e.g.* by atoms, by groups or by molecules

e.g. water



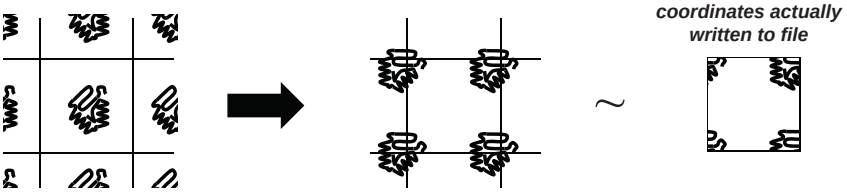
by molecule (water oxygen atoms always within reference box)



by atom (all atoms always within the reference box)

GROMOS: by charge groups (i.e. intact charge groups based on the position of the charge-group center [for water: the oxygen atom])

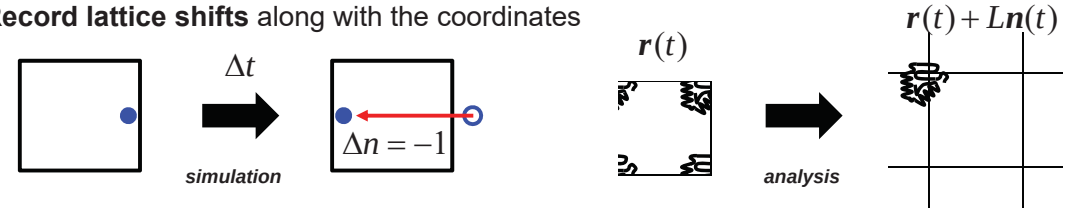
- For many analyses, we need apply **periodic gathering**, so as to follow the trajectory of a molecule that is initially «gathered» but may fail to remain so along time



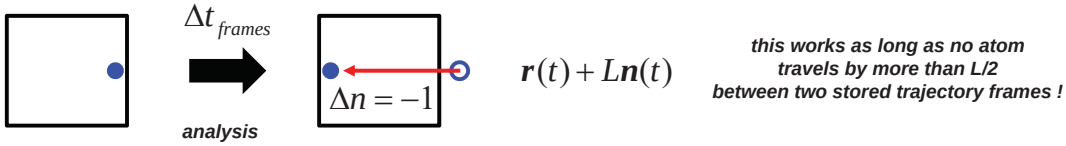
Periodic gathering

- The **periodic gathering** can be performed in different ways

  → **Record lattice shifts** along with the coordinates

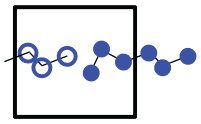


  → **Infer lattice shifts** from the atom displacements between stored trajectory frames



this works as long as no atom travels by more than L/2 between two stored trajectory frames !

  → Gather by following the **covalent connectivity** of the molecules (*i.e.* along bonds)



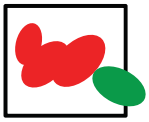
you have to pick a first (reference) atom that will always remain in the reference box

  → Gather using a **reference structure**

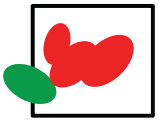
pick periodic copy of atom closest to atom position in the reference structure

- All these methods have trouble when you want to gather **different molecules**

like this



or like that



???

Periodic gathering

- For this reason, there are many possible **gathering methods** in GROMOS

  → **@pbc** flag in the GROMOS++ analysis programs

`@pbc arg1 [arg2] [arg3]`
Periodic boundary type and gathering parameters are read from `@pbc`. The first argument is the boundary type which may take the following values:

- `v` vacuum, non-periodic boundary conditions
- `r` rectangular periodic boundary conditions
- `c` triclinic periodic boundary conditions
- `t` truncated octahedral periodic boundary conditions

The second and third arguments determine the gathering method. The available gathering methods (`arg2`) are:

`nog` or `0`	do not gather
`glist` or `1`	(default) gathering, based on a list of atoms
`gtime` or `2`	gathering based on previous frame
`gref` or `3`	gathering based on a reference structure
`gltime` or `4`	gather first frame based on a list, next frames based on previous frame
`grtime` or `5`	gather first frame based on a reference structure, next frames based on previous frame
`gbond` or `6`	gathering based on bond connectivity
`cog` or `7`	gathering with respect to the centre of geometry of the all atoms of the first molecule in the system

The third argument is used for specific gathering methods. If `glist` or `gltime` is used, then `arg3` should have the form:
`list <AtomsList>`
If `gref` or `grtime` is used, then `arg3` should have the form:
`refg <ReferenceStructure>`

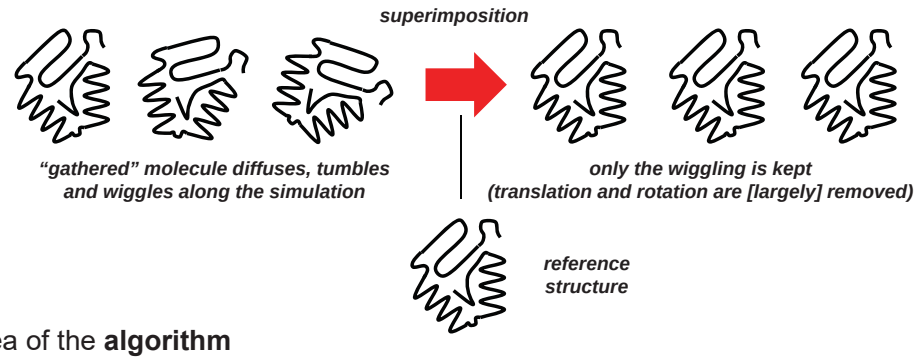
GROMOS manual Volume 5

(surprisingly, GROMOS records lattice shifts in the trajectory, but does not yet offer a method to gather based on them ;-)

Rototranslational superimposition

- For some observables related to molecular structure, gathering is not sufficient; we want to be able to **compare Cartesian coordinates**, but in such a way that the effect of the **translational and rotational diffusion is removed**

  → This can be done using **rototranslational least-squares fitting** onto a **reference structure**



  → Idea of the **algorithm**

for each new structure, minimize $X = \sum_{i=1}^{N} \left[(\boldsymbol{C} + \boldsymbol{T}\boldsymbol{r}_i) - \boldsymbol{r}_{i,ref} \right]^2$ *with respect to* $\boldsymbol{C}$ *translation vector*, $\underline{\boldsymbol{T}}$ *rotation matrix*

(boils down to a 6x6 matrix diagonalization)

  → Watch out that the result may depend on the **choice of the reference structure** and of a **subset of fitting atoms** - and that the translational and rotational motions **cannot be decoupled in an entirely strict fashion** from the wiggling !

Rototranslational superimposition

- For this reason, a **rototranslational superimposition method** must often be specified in GROMOS

  → **@ref** and **@atomsfit** flags in the GROMOS++ analysis programs

`@ref`	⟨reference coordinates (if absent, the first frame of `@traj` is reference)⟩
`@atomsfit`	⟨atom specifier: atoms to consider for fit⟩

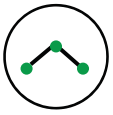
Types of instantaneous observables

- (1) **System quantities** derived **directly from the configuration** $Q(\boldsymbol{r}, \boldsymbol{p})$

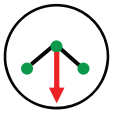
  → Depending on the case, we may have to apply periodic gathering of distinct molecules and sometimes also to follow atoms (or molecules) across periodic boundaries

  → Many such properties are *already calculated during the MD run* and stored along with the configuration trajectory (*e.g.* energy components and other thermodynamic parameters)

  → Examples of observables of this type

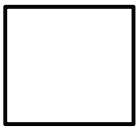


atom positions, velocities, forces



bond or dipole-moment vector

box dipole moment vector



box volume (density) V, ρ

pressure P

kinetic energy (temperature) K, T

potential energy U

total or components

r

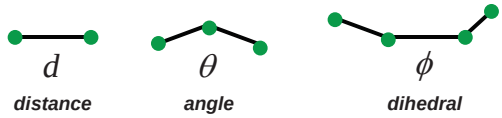
minimum-image distance between two non-bonded atoms

Types of instantaneous observables

- (2) **Molecular observables** that are **rototranslationally invariant** *i.e. so called internal coordinates of the molecule*

→ We must apply **periodic gathering**, but rototranslational superimposition is unnecessary

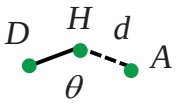
→ Examples of observables of this type



distance *angle* *dihedral*

distance, angle, dihedral between covalently linked atoms or not

Note: dihedral is a periodic coordinate !



H-bond existence

e.g. GROMOS H-bond exists if:

$$b \le 0.25\ \text{nm}$$
$$\theta \ge 135°$$

Note: H-bond existence is an on-off coordinate

Radius of gyration

$$R_{gyr} = \left[\frac{1}{M} \sum_{i=1}^{N} m_i (\boldsymbol{r}_i - \boldsymbol{R}_{cm})^2 \right]^{1/2}$$

$$\boldsymbol{R}_{cm} = \frac{1}{M} \sum_{i=1}^{N} m_i \boldsymbol{r}_i$$

center-of-mass position

$$M = \sum_{i=1}^{N} m_i$$

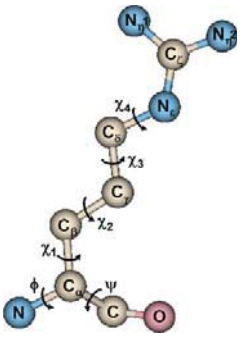
total mass

Note: definition without masses also possible

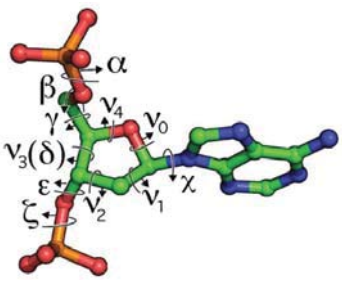
Types of instantaneous observables

→ Usual nomenclature of **dihedral angles** relevant for different **biomolecules**

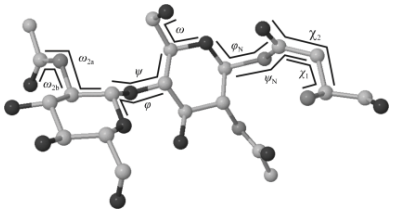
Polypeptides (proteins)

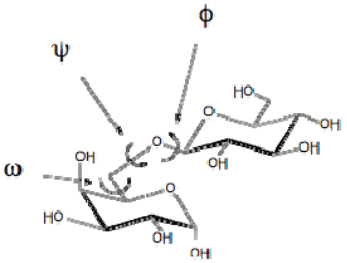


Nucleic acids



Polysaccharides

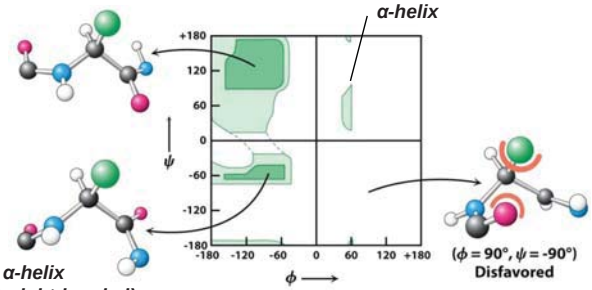




(1→6)-linked

β-sheet

left-handed α-helix



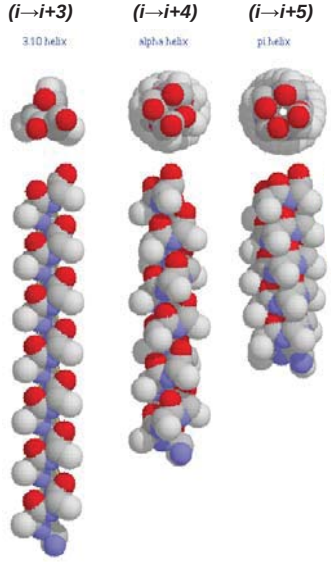
Ramachandran map (correlates with the secondary structure)

α-helix (default: right-handed)

also...

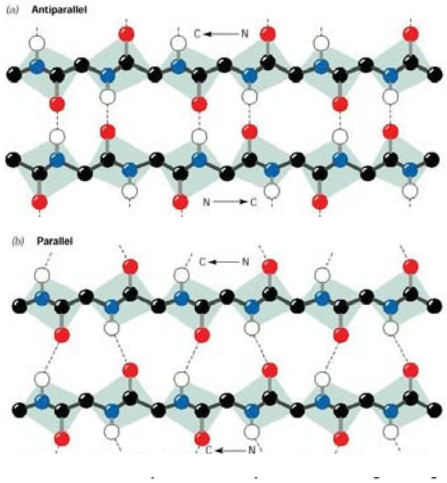
Types of instantaneous observables

→ Characterization of the (local) **secondary structure** of a protein



3_{10}-helix (i→i+3) *α-helix (i→i+4)* *π-helix (i→i+5)*

β-sheet



Assignment of local secondary structure for a residue, typically based on

- Empirically estimated H-bond energies (from H-bond distance and angle)

- Ramachandran map location

Program	Internet
DSSP	http://www.cmbi.kun.nl/gv/dssp
STRIDE	http://webclu.bio.wzw.tum.de/stride/
DSSPcont	http://cubic.bioc.columbia.edu/services/DSSPcont
VoTAP	http://www.lmcp.jussieu.fr/%7Emornon/voronoi.html
Beta-spider	http://www-lbit.iro.umontreal.ca/bSpider/
XTLsstr	http://oregonstate.edu/dept/biochem/faculty/johnson.html
KAKSI	http://migale.jouy.inra.fr/mig/mig_fr/servlog/kaksi/
PALSSE	http://prodata.swmed.edu/palsse/palsse.php
SEGNO	http://www.bioinf.man.ac.uk/~lovell/segno.shtml
SecStr	http://www.mbfys.lu.se/Services/SecStr/

Types of instantaneous observables

→ Example: thermal unfolding of lysozyme

(my first paper in MD...)
GROMOS87
500 K
180 ps

PROTEINS: Structure, Function, and Genetics 21:196–213 (1995)

Computational Approaches to Study Protein Unfolding: Hen Egg White Lysozyme as a Case Study

P.H. Hünenberger, A.E. Mark, and W.F. van Gunsteren
Laboratorium für Physikalische Chemie, ETH-Zentrum, CH-8092 Zürich, Switzerland

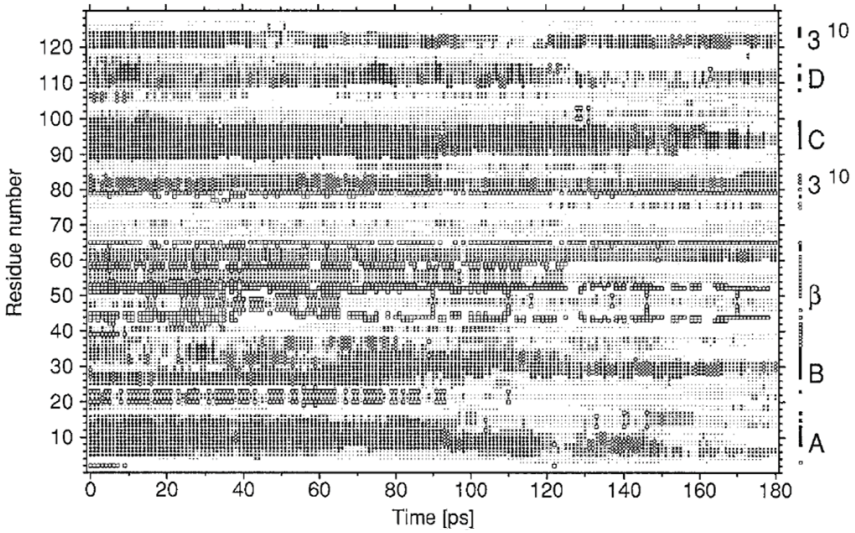
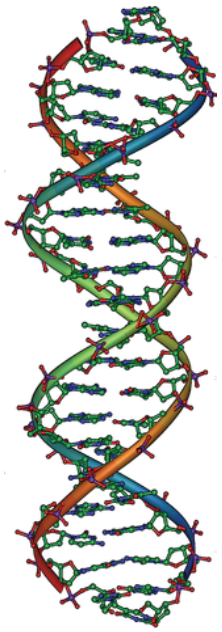
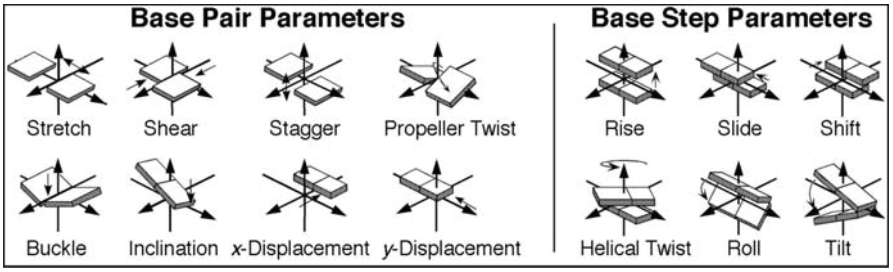


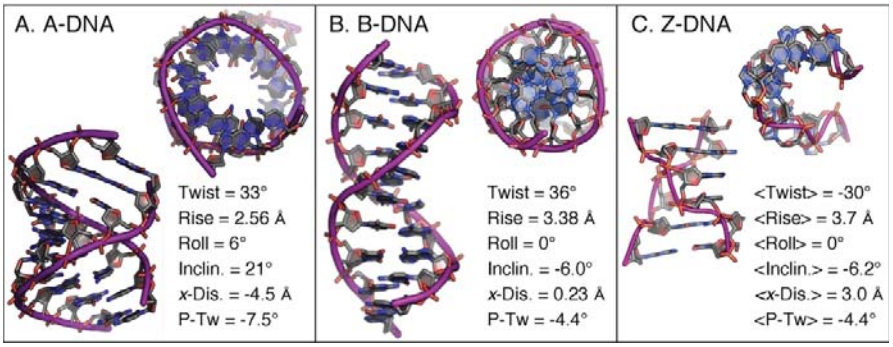
Fig. 2. Secondary structure as a function of time, for the T-run, as given by the "SUMMARY" entry of DSSP program output.[48] (●) α-helix, (□) β-bridge or β-sheet, (◇) 3^{10} helix, (♦) π helix, (×) hydrogen-bonded turn, and (·) bend. On the right hand side, a symbol indicates location of helices A, B, C, and D, the two 3^{10} helices and the β-sheet. On this side, (⊐) indicates observed but rapidly exchanging protons upon folding[34] and (■) slowly exchanging ones.

Types of instantaneous observables

→ Characterization of the (local) **base pair arrangement** of a nucleotide







Types of instantaneous observables

- (3) **Molecular observables** derived from the **Cartesian coordinates**

i.e. providing overall structural information

→ We must apply **periodic gathering** and **rototranslational superimposition**

→ Examples of observables of this type

$$\{\mathbf{r}_i, i = 1..N\}$$

e.g. movie of your favorite protein !!!

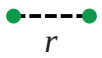
fitted atom positions – rototranslation removed (sometimes also used: velocities, forces)

Types of instantaneous observables

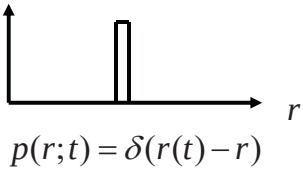
- (4) **Distribution functions** over **multiple occurrences** of any of the previous types (1-3) of coordinates

→ If a system involves multiple occurrences of a given coordinate, we can calculate the corresponding distribution for a given trajectory frame, called a **distribution function**

example:



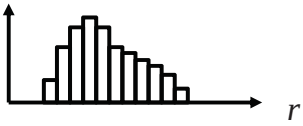
minimum-image distance between two types of atoms in two different molecules



$$p(r;t) = \delta(r(t) - r)$$

Dirac delta function → approximated in practice by a finite-width bin function

upon averaging over all N equivalent pairs of the same atoms in any pair of molecules



probability distribution → approximated in practice by a histogram

$$Q(r;t) = \frac{1}{N}\sum_{n=1}^{N} p(r_n;t) = \left\langle p(r;t) \right\rangle_{crd}$$

example: the distribution of O-O distances between molecules in a water sample (scaled by the Jacobian factor) is called the radial distribution function (RDF)

the instantaneous observable is here a function, called a distribution function

$\langle \cdots \rangle_{crd}$ *average over equivalent coordinates (≠<...> which would be over time [default of the notation]!)*

→ Observables that correspond to **averages over multiple molecules** (or, even better, molecule pairs) in a sample are typically much **easier to calculate accurately** (needs less sampling) compared to observables that have a single value in each trajectory frame !

Equilibration vs sampling

- The **ergodic hypothesis** postulates that in the limit of infinite time, any trajectory will end up visiting **all possible phase-space points** compatible with the hard macroscopic constraints

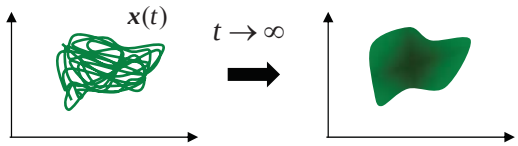
e.g. NVT ensemble

$\boldsymbol{x} = \{\boldsymbol{r}, \boldsymbol{p}\}$

a single-system trajectory

initial conditions

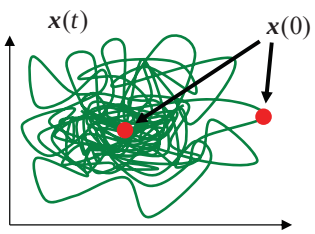
$\boldsymbol{x}(0)$



equilibrium phase-space probability density, nowhere zero !

$$\bar{\rho}(\boldsymbol{x}) \sim \frac{\exp\left(-\beta \mathcal{H}(\boldsymbol{x})\right)}{\int d\boldsymbol{x} \exp\left(-\beta \mathcal{H}(\boldsymbol{x})\right)}$$

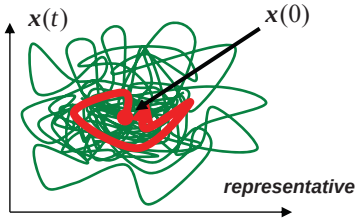
→ So, the choice of initial conditions should be irrelevant



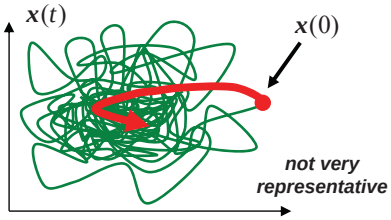
possible initial conditions

→ But: this is not true for **finite trajectory segments** !

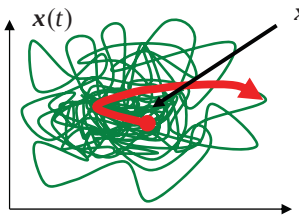
end of previous, reverted velocities



representative



not very representative



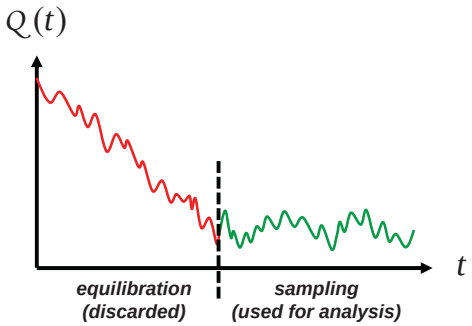
neither

Equilibration vs sampling

- For this reason, we normally discard the initial period of a simulation, called **equilibration**
  → try to **lose memory** of **non-representative initial conditions**
  → better than nothing – but **still no guarantee** !
  → the proper way would be to perform **multiple simulations** from **different initial conditions** *e.g. different initial random velocities*

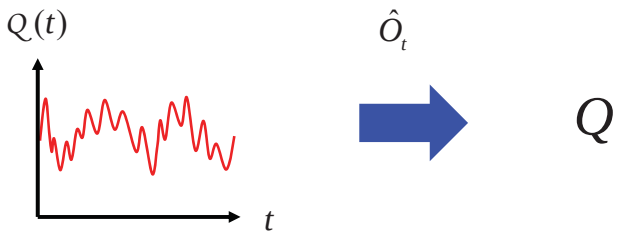
Seldom done in practice: Expensive simulations → people are already happy with one !

- Equilibration is monitored by following **a set of relevant observable** and waiting for **stabilization**



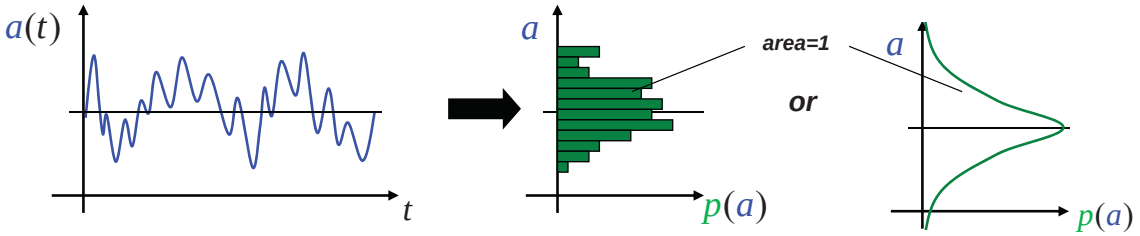
- For complex systems, equilibration is only possible within the available computer time if we start from an already very reasonable initial configuration (*e.g.* protein structure: from X-ray or NMR [evtl. modeling])

Statistical analysis of time series



Probability distribution

- The basic result of monitoring an instantaneous observable Q is a **time series**
- If you want to discard the time information, but preserve everything else, the most relevant quantity is the **probability distribution** of Q over the simulation



time series | *histogram (computer)* | *function (math)*

$\langle \ldots \rangle$ *average over time*

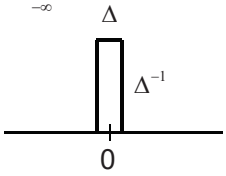
$$< f(t) > = \lim_{t\to\infty} t^{-1} \int_0^\infty dt \;\; f(t)$$

$$p(a') = < h(a(t) - a', \Delta a) >$$

with (binning/windowing function)

$$h(x,\Delta) = \begin{cases} \Delta^{-1} & if \quad -\Delta/2 < x \le \Delta/2 \\ 0 & otherwise \end{cases}$$

$$\int_{-\infty}^{\infty} dx\, h(x,\Delta) = 1$$

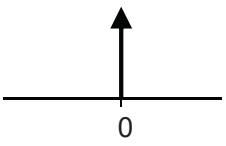


$$p(a') = < \delta(a(t) - a') >$$

with (Dirac delta function)

$$\delta(x) = \begin{cases} \infty & if \quad x = 0 \\ 0 & otherwise \end{cases}$$

$$\int_{-\infty}^{\infty} dx\, \delta(x) = 1$$

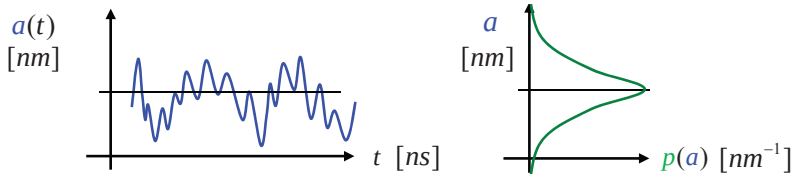


Probability distribution

- A probability distribution function should always be **normalized** $\int_{-\infty}^{\infty} da\; p(a) = 1$

  → often forgotten in the simulation literature and this is **very bad** practice !

- The units of the probability distribution are the **inverse of the units** of the quantity



  → often labelled “arbitrary units” in the simulation literature and this is **very bad** practice !

- If you have *p(a)* you can easily calculate the distribution of any function *f(a)* of *a*

$$p(f') = < \delta(f(a(t)) - f') > = \lim_{t\to\infty} t^{-1} \int_0^\infty dt \;\; \delta(f(a(t)) - f')$$

$$= \lim_{t\to\infty} t^{-1} \int_0^\infty dt \int_0^\infty da' \;\; \delta(a(t) - a') \delta(f(a') - f')$$

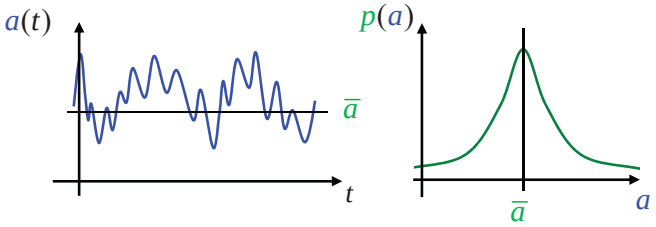
$$= \int_0^\infty da' \, \delta(f(a') - f') \left[\lim_{t\to\infty} t^{-1} \int_0^\infty dt \, \delta(a(t) - a') \right]$$

$$= \int_0^\infty da' \, \delta(f(a') - f') \, p(a')$$

a convolution of the two functions

Moments of the probability distribution

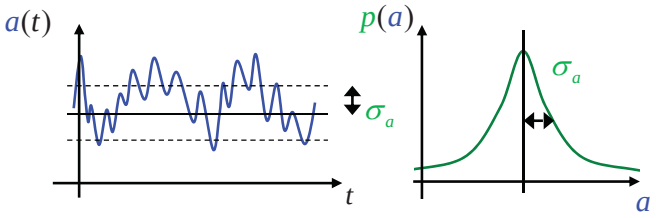
- Often, the probability distribution itself contains too much information, and the most relevant are its **first two moments**
- The **first moment** is the **average** (mean value)



follows from the previously derived convolution principle

$$\bar{a} = \langle a \rangle = \int_{-\infty}^{\infty} da\, a\, p(a)$$

- The **second moment** is the **variance**



follows from the previously derived convolution principle

$$\sigma_a^2 = \left\langle (a-\bar{a})^2 \right\rangle = \int_{-\infty}^{\infty} da\, (a-\bar{a})^2\, p(a)$$

→ the square-root of the variance is called the **standard deviation** (loosely: the [root-mean-square or rms] **fluctuations**)

→ it is easily shown that

$$\sigma_a^2 = \left\langle (a-\bar{a})^2 \right\rangle = \left\langle a^2 \right\rangle - \bar{a}^2$$

often computationally more convenient (single-sweep over the data [instead of two successive sweeps])

Moments of the probability distribution

- One should be a bit careful when calculating the moments of the probability distribution associated with a **periodic coordinate** (*e.g.* dihedral angle)

  → ... more details next year

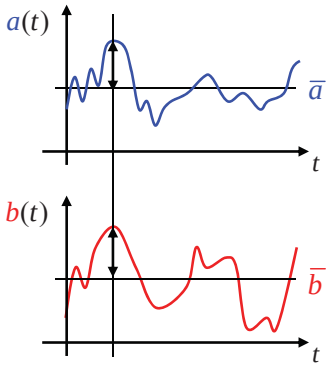
- **Higher-order moments** can also be calculated, but it is seldom done

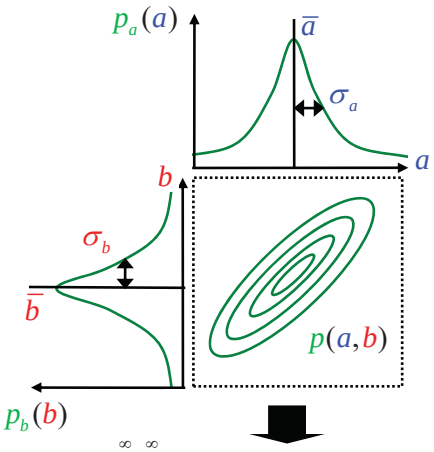
  → typically the first two moment have the highest physical significance (*e.g.* related to macroscopic observables / properties or their temperature/pressure/composition derivatives)

  → based on a finite simulation, the higher the moment, the lower the accuracy of the calculated value (average converges quickly, fluctuations less, higher-order even less)

Correlations between observables

- Still discarding the time information, one may wish to establish the extent correlation in the **probability distributions** of two observables Q and R over the simulation
- This can be done by means of a **two-dimensional probability distributions**





$$\int_{-\infty}^{\infty} db\; p(a,b) = p_a(a)$$

$$\int_{-\infty}^{\infty} da\; p(a,b) = p_b(b)$$

$$p(a,b) \neq p_a(a)\,p_b(b)$$

$$C_{ab} = <(a-\bar{a})(b-\bar{b})> = <ab> - \bar{a}\bar{b}$$

Covariance

$$C_{ab} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} da\,db\,(a-\bar{a})(b-\bar{b})\,p(a,b)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} da\,db\,ab\;p(a,b) \;-\; \bar{a}\bar{b}$$

→ the **cross-correlation** is a normalized covariance

$$c_{ab} = \frac{C_{ab}}{\sigma_a \sigma_b}$$

- ***+1 perfect correlation***
- ***0 no correlation***
- ***-1 perfect anticorrelation***

Time correlations

- When the time information (relaxation) is important, one may look at **time-correlation functions**

→ **Time-autocorrelation function**

$$C_Q(\tau) = \langle Q(t)Q(\tau+t)\rangle$$

autocorrelation function

$$c_Q(t) = \frac{C_Q(t)}{C_Q(0)}$$

normalized autocorrelation function

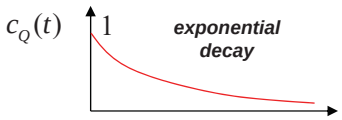
Questions:
- what is $C_Q(0)$ when <Q>=0 ?
- what is $c_Q(0)$ when <Q>=0 ?

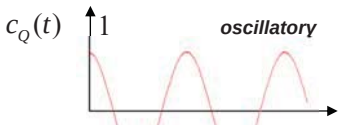
if $\langle Q\rangle = 0$:

$$C_Q(0) = \sigma_Q^2$$

$$c_Q(0) = 1$$

Extreme behaviors





Note:
The quantity Q can be a scalar or a vector

→ **Spectral density**

Fourier transform of a time-correlation function

$$J_Q(\omega) = \int_0^{\infty} dt\; e^{-i\omega t} c(t)$$

(complex)

Oscillations: ω peaks

Decay: Lorenzian broadening

→ **Time-crosscorrelation** function between two observables

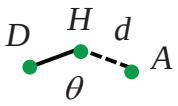
$$C_{QR}(\tau) = \langle Q(t)R(\tau+t)\rangle$$

$$c_{QR}(t) = \frac{C_{QR}(t)}{C_Q^{1/2}(0)\,C_R^{1/2}(0)}$$

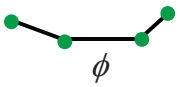
Statistical characteristics

- The time series can be used to derive specific **statistical characteristics**; the following apply to on-off coordinates defining **events**

E.g.



H-bond existence



dihedral well

$0° < \phi < 120°$ *gauche+*

$120° < \phi < 240°$ *trans*

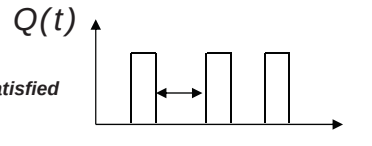
$240° < \phi < 360°$ *gauche-*

→ **Occurrence**

fraction of the trajectory configurations where the event condition is satisfied

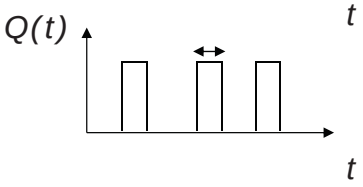
→ **Visiting time**

average time separating periods during which the event condition is satisfied



→ **Residence time**

average time during which the event condition is satisfied



Statistical characteristics

- The time series can be used to derive specific **statistical characteristics**; the following apply to changes resulting from **variations of systems parameters**

→ **Finite-difference** change in average observable upon varying a system property

E.g. NVT ensemble

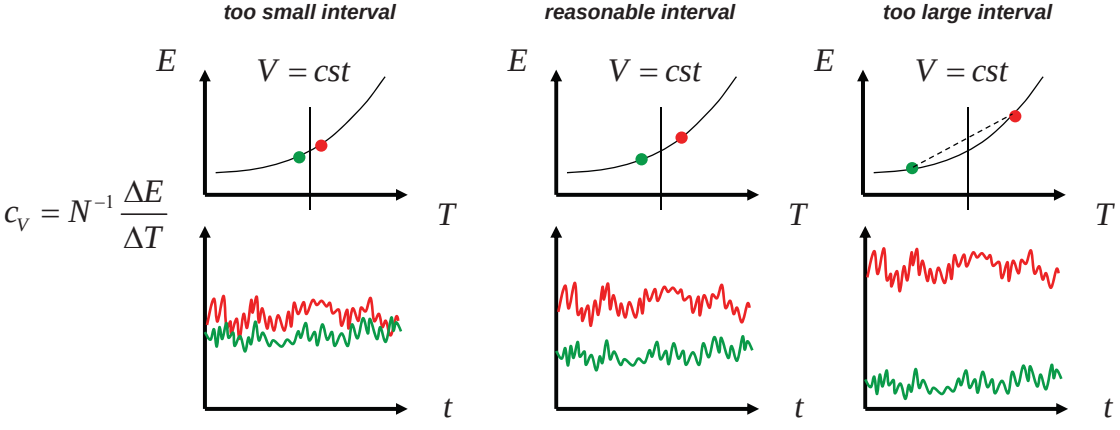
$$\left(\frac{\partial\langle Q\rangle}{\partial V}\right)_{N,T} \qquad \left(\frac{\partial\langle Q\rangle}{\partial T}\right)_{N,V} \qquad \left(\frac{\partial\langle Q\rangle}{\partial N}\right)_{V,T}$$

E.g. NPT ensemble

$$\left(\frac{\partial\langle Q\rangle}{\partial P}\right)_{N,T} \qquad \left(\frac{\partial\langle Q\rangle}{\partial T}\right)_{N,P} \qquad \left(\frac{\partial\langle Q\rangle}{\partial N}\right)_{P,T}$$

→ The choice of the finite-difference **interval** is crucial!

$$c_V = N^{-1}\frac{\Delta E}{\Delta T}$$

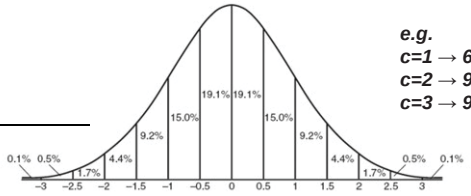


Statistical error

- When you perform *N* **independent experiments** with results following (at least approximately) a **normal distribution** (Gaussian), you can estimate the **error on the mean** as

$$\varepsilon = c\frac{\sigma}{N^{1/2}}$$

- N *number of experiments*
- σ *variance*
- c *confidence factor*

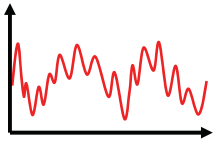


e.g.
c=1 → 68%
c=2 → 95%
c=3 → 99.7%

→ Interpretation: if I perform many *N*-experiments sets, the mean I find has *e.g.* a 95% chance to be within ±ε with *c=2* of the true mean (the one I would get with $N\to\infty$)

- When you analyze a time series, the data is **correlated** in time; so, only averages over "long enough" blocks can be considered independent

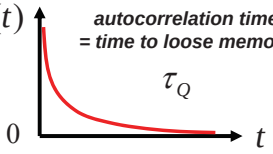
$Q(t)$



t

Correlation analysis

$c_Q(t)$



autocorrelation time = time to loose memory

τ_Q

$$\varepsilon = c\frac{\sigma_Q}{N_{eff}^{1/2}}$$

"effective" number of independent samples

Independent repeats

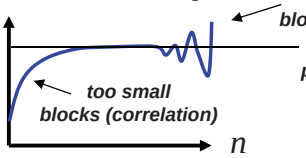
do independent (long enough) repeats of the simulation

Block averaging

divide time series of N points in M blocks of length n

$N = M \cdot n$

$c\frac{\sigma_M}{M^{1/2}}$ *standard deviation of the M block averages*

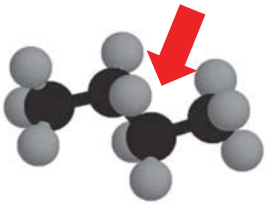


ε

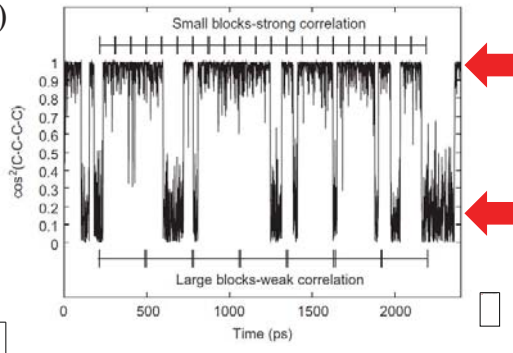
Statistical error

- Block averaging: example with butane dihedral time series

$\cos^2\phi_{CCCC}$



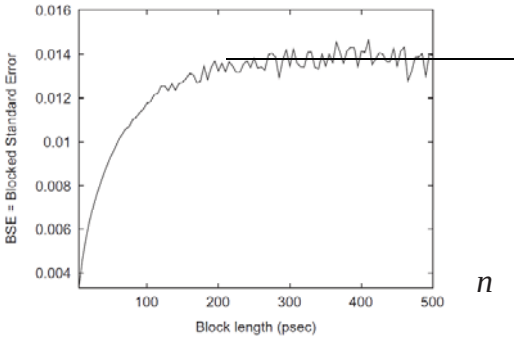
$Q(t)$



trans

gauche-, gauche+

here with c=1



plateau value is true error

ε

n *block length, mapped here to block durations*

Figure 6 The block-averaging procedure considers a full range of block sizes. The upper panel shows the time series for the squared cosine of the central dihedral of butane, with two different block sizes annotated. The lower panel shows the block-averaged standard error for that times series, as a function of block size.

Examples of property calculations

1. Structural properties

- Atomic positions/fluctuations → B-factors
- Radius of gyration
- Solvent accessible surface area
- Radial distribution function
- Orientational correlation function

Structural properties

• **Average atomic positions** (of a solute molecule)

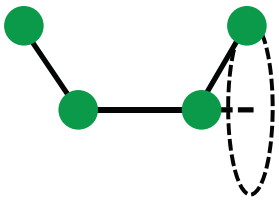
$$\langle \vec{r}_i \rangle = t^{-1} \int_0^t \vec{r}_i(t')dt' = \frac{1}{N_t} \sum_{n=1}^{N_t} \vec{r}_i(t_n)$$

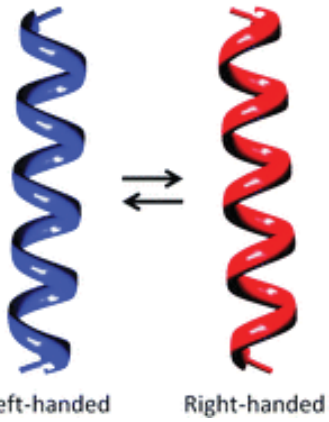
N_t = number of trajectory configurations

After periodicity gathering and rototranslational least-squares fitting (based on a reference structure)

These have often little meaning per se (distorted structure !)

→ Example: imagine what the average structure in the presence of the following equilibria





Structural properties

• **Atomic positional root-mean-square fluctuations** (RMSF; of an atom in a solute molecule)

$$\left\langle \left[\vec{r}_i - \langle \vec{r}_i \rangle\right]^2 \right\rangle^{1/2} = \left[\frac{1}{N_t}\sum_{n=1}^{N_t}\left[\vec{r}_i(t_n) - \langle \vec{r}_i \rangle\right]^2\right]^{1/2}$$

$$= \left[\frac{1}{N_t}\sum_{n=1}^{N_t}\left[\vec{r}_i(t_n)\right]^2 - \langle \vec{r}_i \rangle^2\right]^{1/2}$$

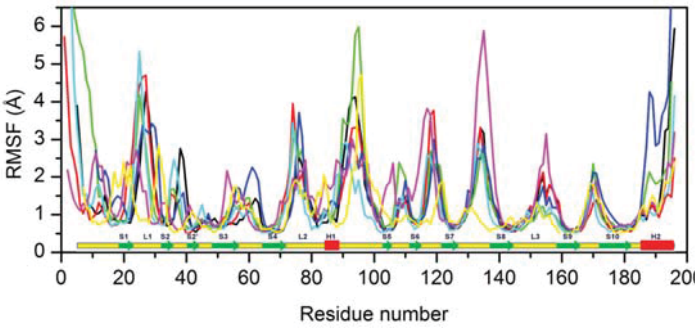
$$= \left[\frac{3B_i}{8\pi^2}\right]^{1/2}$$

After periodicity gathering and roto-translational least-squares fitting (based on a reference structure)

B-factors are used in crystallography to account for the spread in the electron density [but: may have many different causes than atomic fluctuations + there is crystal packing !]

B_i = isotropic crystallographic B-factor of atom i

→ Example: residue-averaged RMSF as a function of residue number



RMSF of Cα atoms of p53 proteins from MD simulations

p53_human (black line) p63 (red line) and p73 (green line), p53_mouse (blue line), p53_chicken (cyan line), p53_fly (purple line) and p53_worm (yellow line). The residue number refers to p53_human. Secondary structure of p53_human is displayed along the sequence (bottom panel): α-helices and β-strands are shown by red rectangles and green arrows, respectively.

[doi:10.1371/journal.pone.0076014.g006]

Structural properties

• **Radius of gyration** (of a solute molecule)

$$R_{gyr} = \left[\frac{1}{N_a}\sum_{i=1}^{N_a}\left[\vec{r}_i - \vec{R}_{cm}\right]^2\right]^{1/2}$$

N_a = number of atoms in the molecule

$\vec{R}_{cm}$ = centre of mass of molecule

$$\vec{R}_{cm} = \frac{1}{M}\sum_{i=1}^{N_a} m_i \vec{r}_i$$

$$M = \text{mass of molecule} = \sum_{i=1}^{N_a} m_i$$

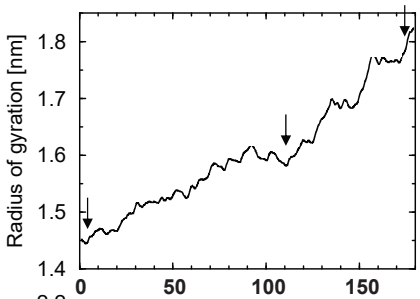
After periodicity gathering (does not depend on translation or rotation; internal coordinate)

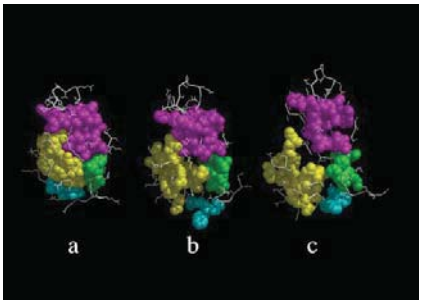
RADIUS OF GYRATION

$$R_{gyr}(t) = \left\{ \frac{1}{N_{at}} \sum_{i=1}^{N_{at}} [\boldsymbol{r}_i(t) - \boldsymbol{r}_{cm}(t)]^2 \right\}^{1/2} \quad \textit{with} \quad \boldsymbol{r}_{cm}(t) = \frac{\sum_{i=1}^{N_{at}} m_i \boldsymbol{r}_i(t)}{\sum_{i=1}^{N_{at}} m_i}$$

Example (ulfolding lysozyme in water):

Thermal unfolding 500K

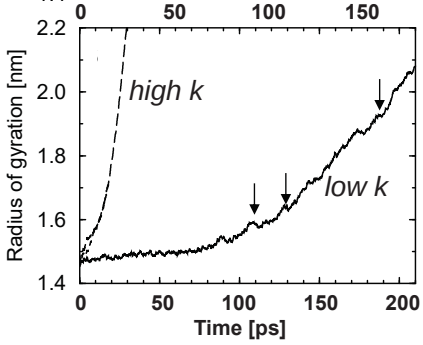


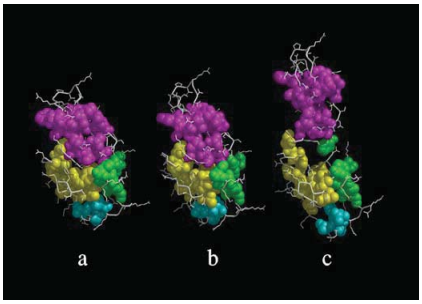


Mean-radius-driven unfolding (300K) with

$$V_{unf} = -\frac{k_{unf}}{N_{at}} \sum_{i=1}^{N_{at}} \left\| \boldsymbol{r}_i(t) - \boldsymbol{r}_{cm}(t) \right\|$$

⇒ constant radial outwards-directed force on all atoms





Structural properties

- **Root-mean-square atomic positional deviation** (RMSD; of a solute molecule relative to a reference structure, often from experiment, like X-ray or NMR)

for two configurations *m* and *n* of N_a atoms

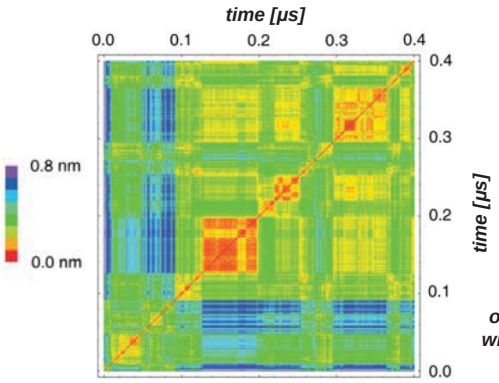
$$\mathrm{RMSD}(m,n) = \left[\frac{1}{N_a} \sum_{i=1}^{N_a} [\vec{r}_i(m) - \vec{r}_i(n)]^2 \right]^{\frac{1}{2}}$$

After periodicity gathering and roto-translational least-squares fitting (based on a reference structure)

Depends on which set of atoms is used for the **translational and rotational superposition of structures** *m* and *n* ***and*** on which atoms are included in the sum $1..N_a$

e.g. all-atom RMSD based on backbone C_α fitting

→ Example: RMSD matrix



Comparison of a trajectory of a 14-residue peptide in water with itself (you can also compare different trajectories !)

GROMOS++ program rmsd

arguments:

@topo	<molecular topology file>
@pbc	<boundary type> [<gathermethod>]
@time	<**time and dt**>
@atomsrmsd	<**atoms** to consider for rmsd>
[@atomsfit	<atoms to consider for fit>]
[@ref	<reference coordinates (if absent, the first frame of @traj is reference)>]
@traj	<trajectory files>

GROMOS++ programs rmsd_cmp or rmsdmat

Structural properties

- **Covariance** or **cross-correlations** of **atomic positions**
- **Solvent-accessible surface area**
- **Occurrence** of intramolecular and solute-solvent **H-bonds**
- **Occurrence** of **secondary structure** patterns
- **Radial distribution functions**
- **Orientational correlation functions**

Example: simulations of α-lactalbumin in water

MD of α-lactalbumin in H_2O

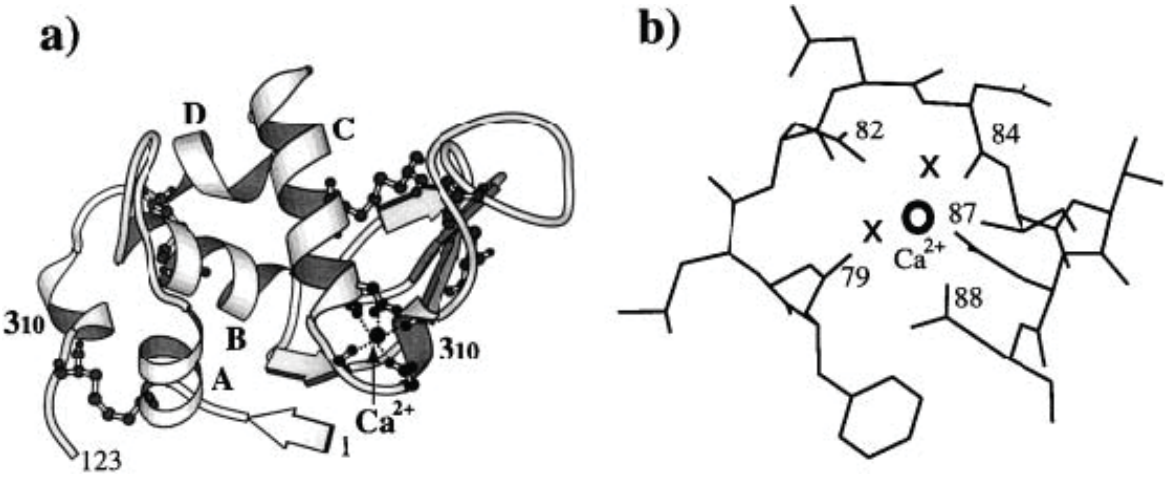
(example is a bit «pedestrian» - I'll change it next year...)

Charge

-8e at ***normal*** **pH = 6.5** with Ca^{2+}

+16e at ***low*** **pH = 2.0** ⟹ ASP, GLU, HIS } protonated, no Ca^{2+}

in ≈ 5600 H_2O, periodic



L. Smith et al., Proteins, 36 (1999) 77-86

Example: simulations of α-lactalbumin in water

Differences between the systems

Human α-lactalbumin at high and low pH

Both systems:

Force field:
Protein GROMOS96 43A1
Solvent SPC water

Starting Structure:
X-Ray at pH = 6.5 with Ca^{2+}

Simulation length:
700 ps

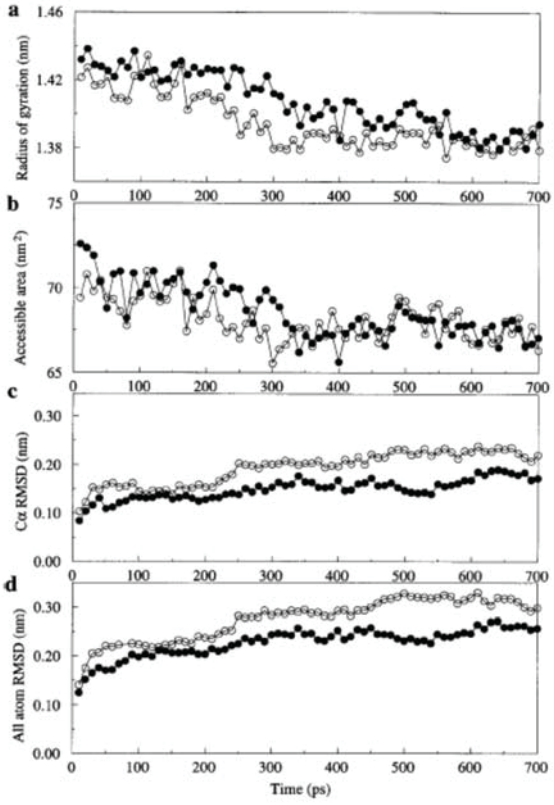
truncated octahedron →

	High pH (pH 8)	Low pH (pH 2)
Aspartate side chains[a]	not protonated	protonated
Glutamate side chains[b]	not protonated	protonated
Histidine side chains[c]	not protonated	protonated
C-terminus	not protonated	protonated
Overall protein charge (in e)	-8	+16
Calcium ion	present	absent
Crystallographic waters	used	not used
Box dimensions (initial)	7.1832 nm	7.1832 nm
Total number of:		
protein atoms	1243 + Ca^{2+}	1267
water molecules	5574	5582

[a] Aspartate residues are 14, 16, 37, 45, 74, 78, 82, 83, 84, 87, 88, 97, 102.
[b] Glutamate residues are 7, 25, 43, 46, 49, 113, 116, 121.
[c] Histidine residues are 32, 107.

Example: simulations of α-lactalbumin in water

α-Lactalbumin: Structural properties



pH = 8 filled circles ●
pH = 2 open circles ○ no Ca^{2+}

a Radius of gyration

b Solvent accessible surface area

c C_α-positional RMSD from X-ray at pH=6.5

d All-atom positional RMSD from X-ray at pH=6.5

- At low pH and no Ca^{2+}:
  → Protein is (?) more compact
  → Structure deviates more from X-ray (at pH 6.5 with Ca^{2+})

Example: simulations of α-lactalbumin in water

Occurrences of main-chain hydrogen bonds

Hydrogen bonds present in the X-ray structure

NH-CO	High pH	Low pH
Helix A		
8-4	83	99
9-5	93	97
10-6	63	57
11-7	95	92
12-8	91	90
13-10	2	5
15-12	7	63
16-13	44	62
Helix B		
27-23	99	99
28-24	82	96
29-25	99	88
30-26	78	86
31-27	98	98
32-28	90	75
33-29	92	35
34-30	72	0
β-sheet		
42-49	72	56
44-47	53	86
49-42	98	97
51-40	59	54
50-55	98	98
54-51	54	77
55-50	98	99
57-48	56	86

NH-CO	High pH	Low pH
3_{10} helix 1		
79-76	38	61
80-77	97	13
81-78	68	14
Helix C		
89-85	98	49
90-86	95	91
91-87	93	98
92-88	99	99
93-89	98	96
94-90	86	57
95-91	92	98
96-92	68	98
97-93	94	91
98-94	94	74
99-95	60	87
Helix D		
107-104	13	40
109-105	95	0
110-106	43	1
111-107	40	7
3_{10} helix 2		
118-115	72	0
119-116	65	0

near calcium

(indicated by residue numbers) in regions of secondary structure in the human α-Lactalbumin simulations

The populations listed are percentages over the simulation time 300-700 ps.

Number of H-bonds with occurrence ≥ 10% :

High pH 81
Low pH 56

- At low pH and no Ca^{2+}:
  → Helices D and (second) 3_{10} disrupted

Example: simulations of α-lactalbumin in water

Root-mean-square fluctuations about torsion angles

(in degrees)[a] in the human α-Lactalbumin simulations[b]

	Time period (ps)						
	100-200	200-300	300-400	400-500	500-600	600-700	300-700
φ							
High pH	14.5	13.8	14.3	13.2	13.8	13.2	15.6
Low pH	15.3	15.3	14.0	14.8	13.7	14.6	15.7
ψ							
High pH	13.6	13.4	14.0	12.6	13.3	12.6	15.7
Low pH	15.0	15.3	13.7	14.7	13.5	14.4	15.9
χ_1							
High pH	20.1	17.0	16.7	13.7	15.7	15.6	21.3
Low pH	19.9	21.8	18.3	20.5	21.3	20.3	30.0
χ_2							
High pH	38.1	29.4	31.4	32.2	25.9	27.4	52.8
Low pH	44.3	35.3	32.8	39.0	38.3	35.0	59.0
χ_3							
High pH	57.1	47.1	44.1	40.9	38.5	46.2	67.3
Low pH	83.7	57.9	61.7	48.9	66.6	56.9	122.9

Backbone:

similar fluctuations

Side chains:

larger fluctuations at low pH

- At low pH and no Ca^{2+}:
  → More sidechain conformational fluctuations

Remark: calculating the average and moments of periodic coordinates (e.g. dihedral angles) is dangerous ! – I don't know how it was done here !!!

[a] The fluctuations for χ_4 are not listed due to the small number of residues with this torsion angle
[b] The side chains of proline and cysteine residues are excluded from the analysis.

Example: simulations of α-lactalbumin in water

The total number of torsion angle transitions

in the human α-lactalbumin simulations[c]

	Time period (ps)						
	100-200	200-300	300-400	400-500	500-600	600-700	300-700
All main chain							
High pH	616	670	589	639	627	587	2416
Low pH	808	839	749	727	730	631	2806
All side chain							
High pH	843	848	677	663	684	680	2710
Low pH	1330	1095	1063	970	965	1013	4011
≥120° side chain[d]							
High pH	149	129	100	70	109	98	380
Low pH	193	160	161	137	153	103	563

Increased { motion / disorder } at lower pH

in agreement with experiment

- At low pH and no Ca^{2+}:

  → More sidechain torsional angle transitions

[c] The total number of torsional angle transitions of 60° or greater (All) and of 120° or greater (≥120°) are listed.
[d] There are no main chain torsion angle transitions of 120° or greater in either of the simulations.

Example: simulations of α-lactalbumin in water

Conclusions

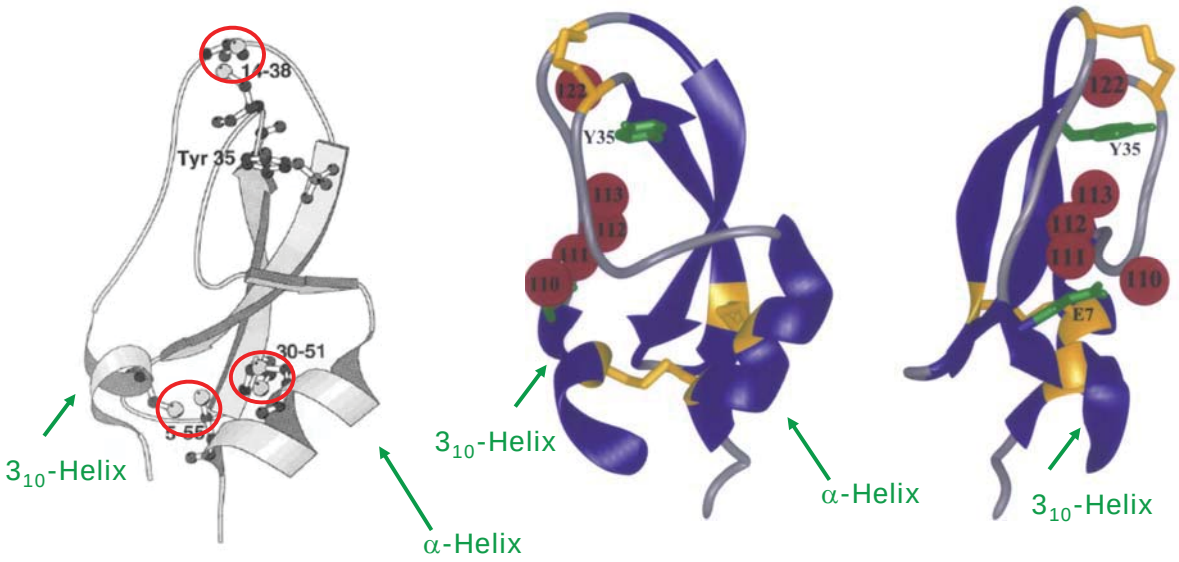
- **Structural properties of α-lactalbumin are well reproduced at pH=8 (native)**
- **Upon lowering pH to 2.0 and removing Ca^{2+}-ion**
  - **Protein becomes more compact**
  - **Deviations from X-ray structure increase**
  - **D-helix, 3_{10}-helix (2nd) } disrupted — in agreement with X-ray at pH = 4.2**
  - **Greater side chain mobility — in agreement with NMR at pH = 2.0**

Example: simulations of BPTI in water

Bovine pancreatic trypsin inhibitor (BPTI)

58 amino acid residues
3 S-S bridges

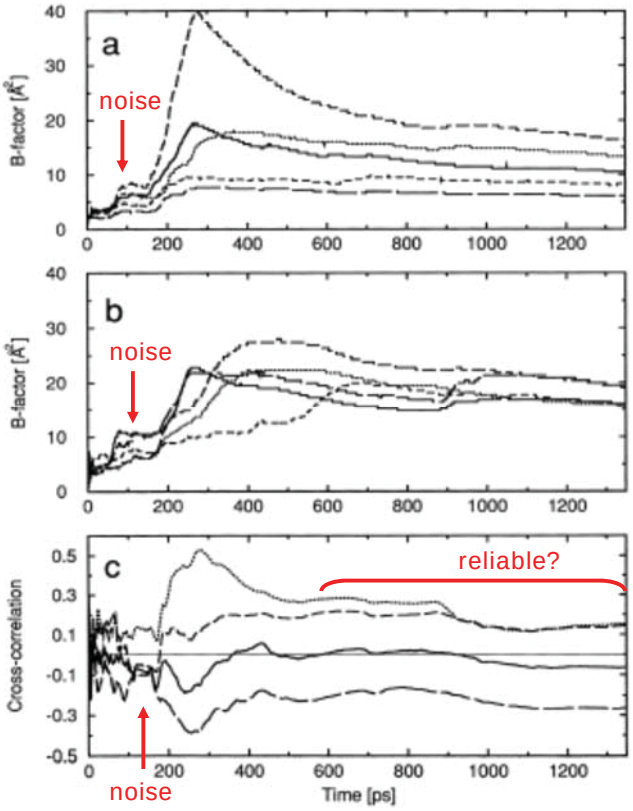
5 internal water molecules



C. Schiffer et al., Proteins, 26 (1996) 66-71

Example: simulations of BPTI in water

Atomic B-factors: protein BPTI



$$B_i = \left(8\pi^2/3\right)\left\langle\left[\vec{r}_i - \left\langle\vec{r}_i\right\rangle\right]^2\right\rangle$$

~ mean square fluctuation of position of atom i

a C_α atoms in α-helix (50-55)

b C_α atoms in 3_{10}-helix (3-7)

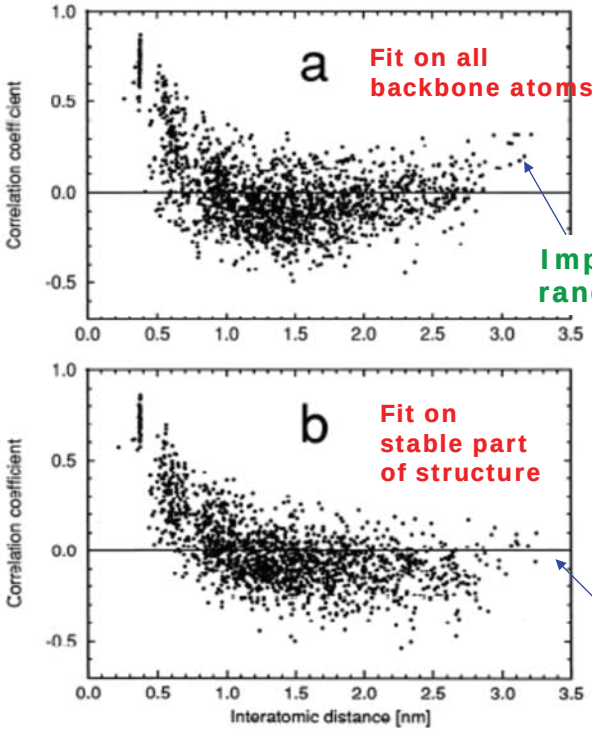
c Cross correlation

Mean-square atomic fluctuations converge only after at least a ns of simulation

P. Hünenberger et al., J. Mol. Biol., 252 (1995) 492

Example: simulations of BPTI in water

BPTI: atom-positional cross-correlations as a function of the distance between atoms



Is it just an artefact?

Fitting to remove translational and/or rotational motion can induce spurious fluctuations and correlations.

Important long range correlations?

No long range correlations

P. Hünenberger et al., J. Mol. Phys., 252 (1995) 492

Structural properties

- **Radial distribution function** $g(r)$

r : minimum-image distance !

g(r) : unitless function

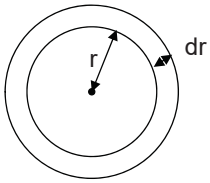
Yields radial structure of liquid, solid

$\rho\, g(r)\, 4\pi r^2\, dr$ is the probability or frequency of observing an atom in a spherical shell between r and $r+dr$ around an atom

or rather: the local density relative to the bulk !

ρ = number density

$= \frac{N}{V}$



Normalization:

or N if solvent around a solute

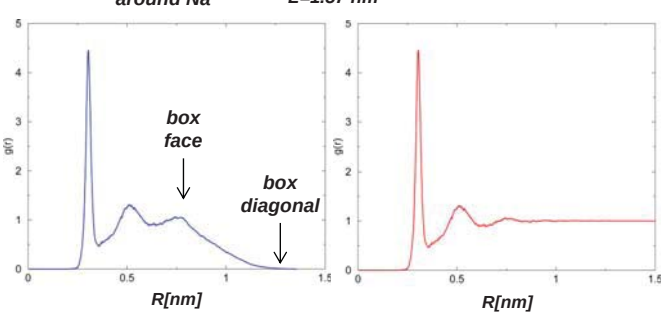
$$\int_0^{volume} \underbrace{\rho g(r) 4\pi r^2 dr}_{\text{number of atoms in shell between r and r+dr around a central atom}} = N - 1$$

Limits:

$$\lim_{r \to 0} g(r) = 0$$

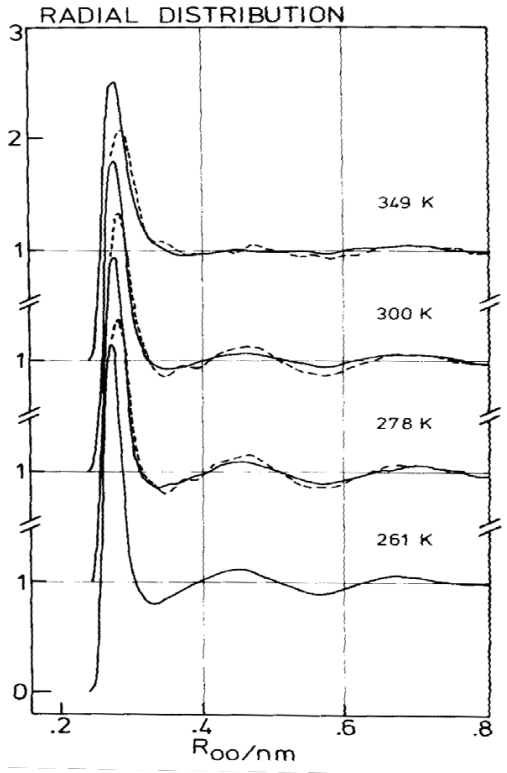
$$\lim_{r \to \infty} g(r) = 1$$

e.g.: water oxygen around Na^+ *N=128, L=1.57 nm* *corrected*



Example: simulations of water

Radial distribution function g (R_{o-o})



for the SPC model at four *different temperatures* and constant volume compared with experimental results

Solid lines: g (R) for SPC model

Dashed lines: X-ray data

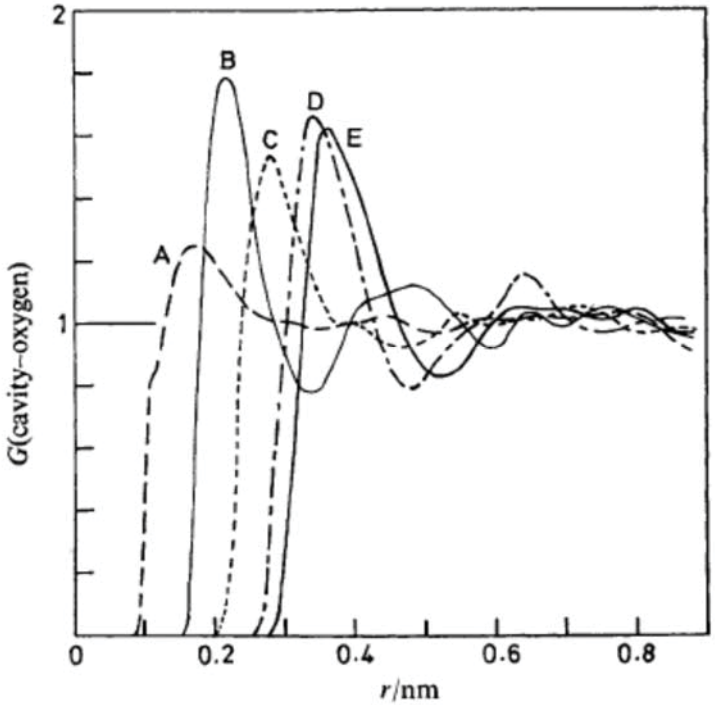
First neighbour peak becomes lower at higher temperature and the second one disappears in agreement with the (not very precise) experimental results

J. Postma, A molecular dynamics study of water, PhD-thesis (1985), Univ. Groningen

Example: simulations of a cavity in water

Cavity-oxygen radial distribution function g(r)

for five values of the thermal radius of the cavity



Cavity-H_2O g(r)

Curves are obtained by spline smoothing with deviation from the data not exceeding 0.05. No smoothing was applied for distance below the first maximum

Radius of cavity:

A: 0.100 nm
B: 0.178 nm
C: 0.238 nm
D: 0.299 nm
E: 0.317 nm

Solvation structure varies with cavity radius

J. Postma et al., Faraday Symp. Chem. Soc., 17 (1982) 55-67

Structural properties

- **Orientational correlation function** *c(r)*

→ Ion-dipole

$$\left\langle \frac{\vec{r}_{ij} \cdot \vec{\mu}_j}{r_{ij}\mu} \right\rangle_r = \langle \cos\theta(r) \rangle_r$$

$\vec{\mu}_i$ = vector at position $\vec{r}_i$ (size=μ)

r_{ij} = distance i-j = $\left(\left[\vec{r}_i - \vec{r}_j \right]^2 \right)^{1/2}$

$$\cos\theta = \frac{\vec{r}_{ij} \cdot \vec{\mu}_j}{r_{ij}\mu}$$

→ Dipole-dipole

$$\left\langle \frac{\vec{\mu}_i \cdot \vec{\mu}_j}{\mu^2} \right\rangle_r = \langle \cos\theta(r) \rangle_r$$

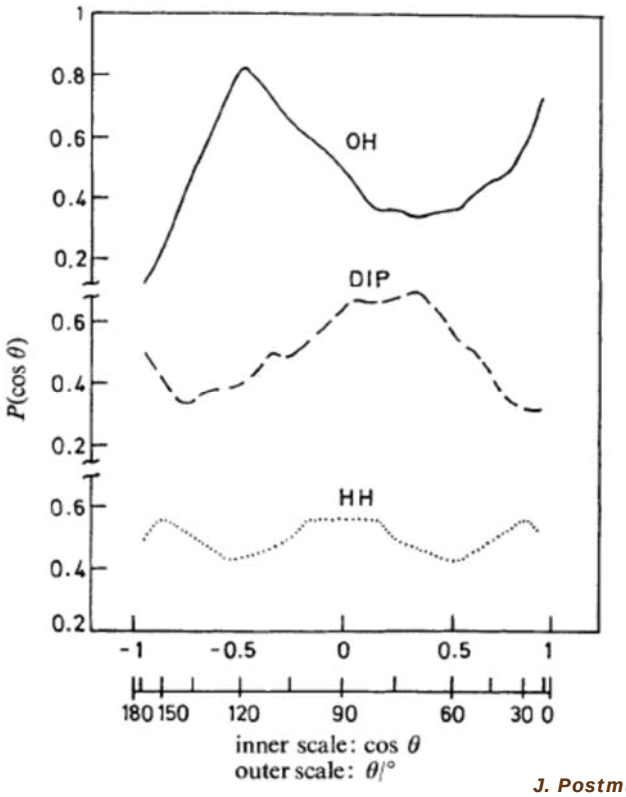
$\vec{\mu}_i$ = vector at position $\vec{r}_i$ (size=μ)

r_{ij} = distance i-j = $\left(\left[\vec{r}_i - \vec{r}_j \right]^2 \right)^{1/2}$

$$\cos\theta = \frac{\vec{\mu}_i \cdot \vec{\mu}_j}{\mu^2}$$

Example: simulations of a cavity in water

Orientation of H_2O molecules around a cavity



Simulation of cavity formation

Probability density of orientation of OH, dipole and HH direction with respect to radius from cavity center to oxygen, expressed as distribution over cosθ.

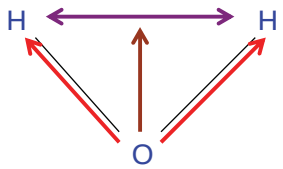
Data applied to molecules in first shell (r < 0.475 nm) from simulation with r_{th} = 0.299 nm.

Angle between vector from cavity center to water oxygen and the vectors

OH:

DIP:

HH:



J. Postma et al., Faraday Symp. Chem. Soc., 17 (1982) 55-67

Example: simulations of a cavity in water

Orientation of H_2O Molecules around a cavity

Two possible orientations of water molecules consistent with the orientational distributions shown before

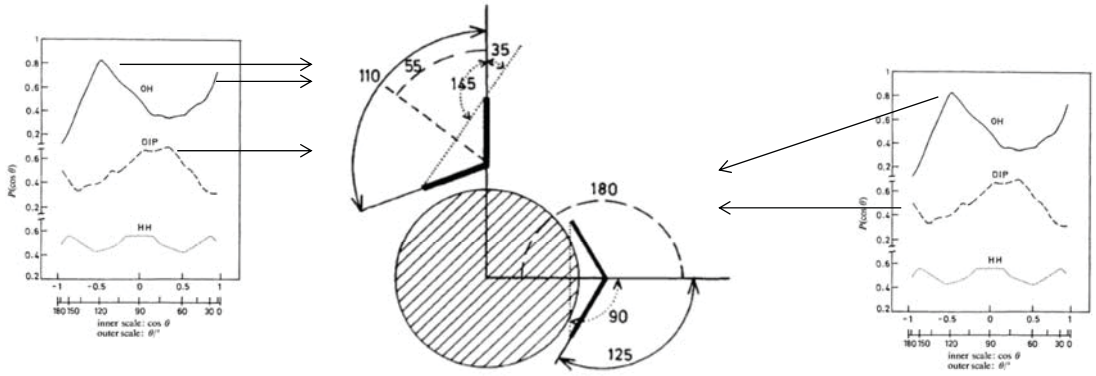


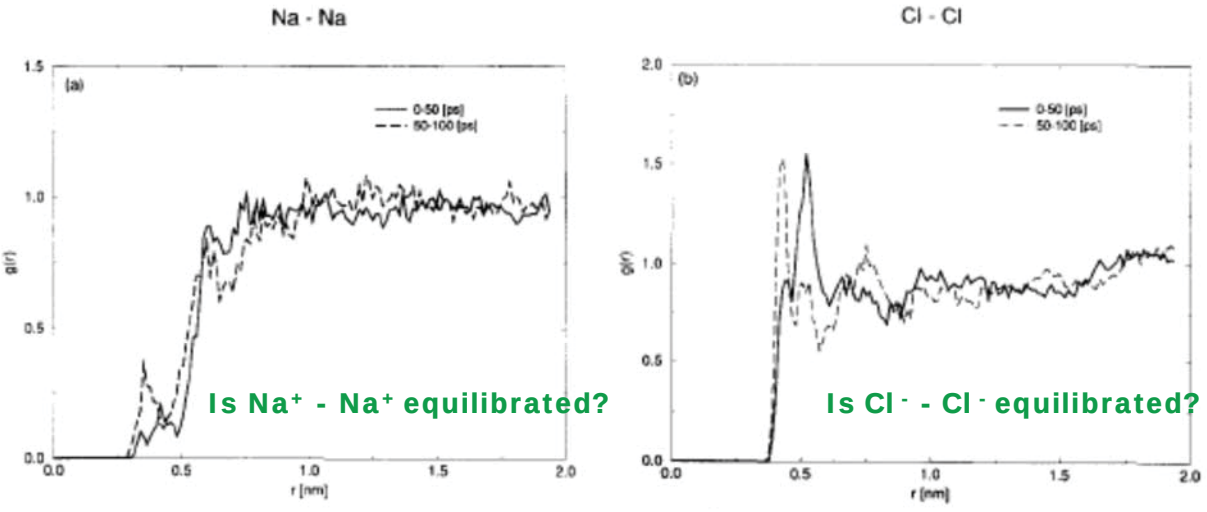
FIG. 5.—Two possible orientations of water consistent with the orientational distributions of fig. 4.

J. Postma et al., Faraday Symp. Chem. Soc., 17 (1982) 55-67

Example: simulations of NaCl in water

System equilibration

Compare different parts of the same simulation.



Radial distribution functions *g(r)* obtained from an MD simulation of a 1 molar sodium chloride solution (40 Na^+, 40 Cl^-, 2127 H2O)
averaged over different 50 ps of the simulation.

(A) Na^+ - Na^+ **(B) Cl^- - Cl^-**

W. F. van Gunsteren et al., Comput. Phys. Commun., 91 (1995) 305-319

Examples of property calculations

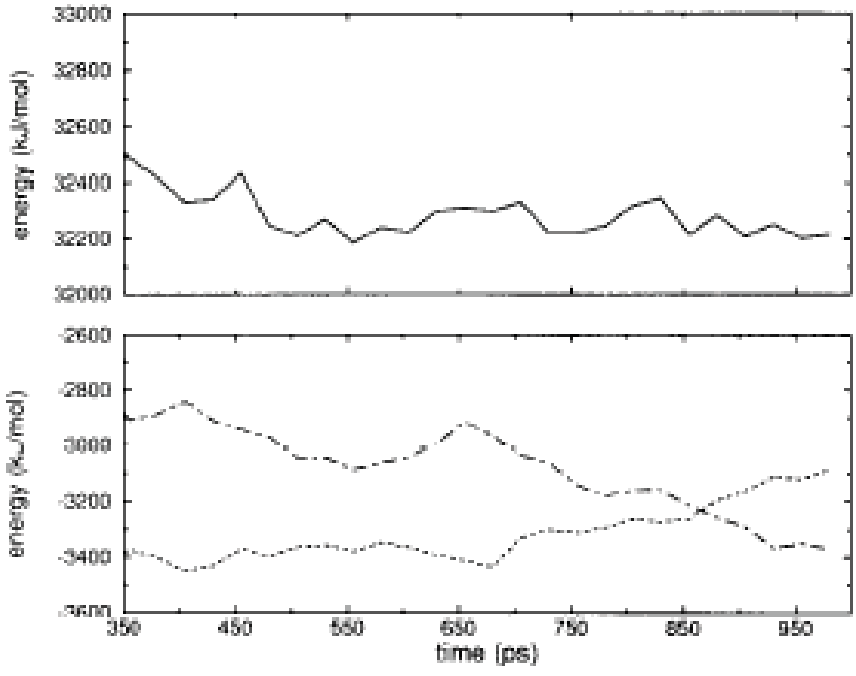
2. Thermodynamic properties

→ Thermodynamic properties
- Energy, pressure, heat capacity
- Compressibility, thermal expansion coefficient
- Free energy, entropy → *discussed in lectures about free energy !*

Example: simulations of lysozyme in water

System equilibration

Van der Waals non-bonded energy of lysozyme in 5000 water molecules as a function of time.



The data points represent 25 ps averages.

Upper panel:
total energy
(solid line)

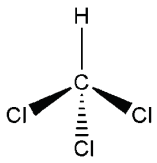
Lower panel:
intra-protein energy
(dotted line),
protein-water energy
(dot-dashed line)

Considering global properties is not sufficient:
Total energy is constant, but components still drift.

W. F. van Gunsteren et al., Comput. Phys. Commun., 91 (1995) 305-319

Example: simulations of liquid chloroform

Chloroform models



Model	C—C	C—Cl	C—H	Cl—Cl	Cl—H	H—H
			$C_6 \times 10^{-3}$/kJ mol^{-1} nm^6			
Dietz and Heinzinger [27]	2·6309	4·6754	0·3622	8·3067	0·6493	0·0377
Kovacs *et al.* [25]	1·8212	4·4421	0·6027	10·699	1·4997	0·1927
Jorgensen *et al.* [29]	4·0340	5·9482	0·0000	8·7708	0·0000	0·0000
			$C_{12} \times 10^{-6}$/kJ mol^{-1} nm^{12}			
Dietz and Heinzinger	4·0642	7·4813	0·1745	13·765	0·3266	0·0043
Kovacs *et al.*	1·9555	6·2786	0·4178	19·667	1·3967	0·0834
Jorgensen *et al.*	12·146	13·637	0·0000	15·312	0·0000	0·0000

	Q_C/e	Q_{Cl}/e	Q_H/e	μ/D
Dietz and Heinzinger	0·179	−0·087	0·082	1·10
Kovacs *et al.*	0·320	−0·140	0·100	1·60
Jorgensen *et al.*	0·420	−0·140	0·000	1·07

	d_{CCl}/nm	d_{CH}/nm	θ_{ClCCl}	θ_{HCCl}
	0·1758	0·1100	111·30	107·57

Geometry: rigid — bonds d, angles θ (SHAKE: tolerance=10^{-4})

Interaction:

$$\frac{C_{12}}{r_{ij}^{12}} - \frac{C_6}{r_{ij}^6} + \frac{Q_i Q_j}{4\pi\varepsilon_0}\left[\frac{1}{r_{ij}} + \frac{(\varepsilon_{RF}-1)}{2\varepsilon_{RF}+1}\frac{r_{ij}^2}{R_c^3}\right]$$

Cut-off radius
R_c = 1.4 nm
Dielectric constant
ε_{RF} = 5

Simulation:
216 molecules in cubic box

(10 ps equilibrium, 50 ps sampling, every 50 fs)

Δt = 2 fs
τ_T = 0.1 ps
T = 293 K
τ_P = 0.2 ps
P = 1 atm
β_T = 10^{-9} m^{2}N^{-1}

I. Tironi et al., Mol. Phys., 83 (1994) 381-403

Example: simulations of liquid chloroform

Chloroform models

Energy, density, pressure, temperature

	Dietz and Heinzinger model		Kovacs *et al.* model		Jorgensen *et al.* model		Exp.
	This work	Ref. [28]	This work	Ref. [25]	This work	Ref. [29]	
T/K	293 ± 5	295	293 ± 5	288	293 ± 5	298	293
pressure p/bar	0·3 ± 84		1·8 ± 90		1·1 ± 85		1·0
density ρ/kg m^{-3}	1520 ± 12	1484	1466 ± 10	1400	1456 ± 10	1480	1489[a]
E_{pot}/kJ mol^{-1}	−28·6 ± 0·3	−28·6	−34·1 ± 0·3	−29·5	−27·7 ± 0·3	−31·3	−31·4[b]
E_{el}/kJ mol^{-1}	−0·53 ± 0·07		−1·6 ± 0·1	−1·78	−0·58 ± 0·08		
E_{LJ}/kJ mol^{-1}	−28·1 ± 0·3		−32·5 ± 0·3	−27·8	−27·1 ± 0·3		

[a] From reference [42].
[b] $-\Delta H_v$ interpolated for 293 K from [41].

Example: simulations of liquid chloroform

Chloroform model: thermodynamic properties

Isothermal compressibility: $\kappa_T = -\frac{1}{V}\left(\frac{\partial V}{\partial p}\right)_T = \left(\frac{\partial \ln \rho}{\partial p}\right)_T \approx \left(\frac{\ln(\rho_2 / \rho_1)}{p_2 - p_1}\right)_T$

Thermal expansion coefficient: $\alpha = \frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_p \approx -\left(\frac{\ln(\rho_2 / \rho_1)}{T_2 - T_1}\right)_p$

Heat capacity: $C_V = \left(\frac{\partial E}{\partial T}\right)_V \approx \left(\frac{U_2 - U_1}{T_2 - T_1}\right)_V + 3R + C_V^{vib}$ *Why 3R ?*

		20–40 ps	40–60 ps	MD	exp
T/K	ρ/kg m^{-3}	$P/10^5$ Pa		$\kappa_T/10^{-10}$ Pa^{-1}	
298	1489	7	−10	8·9	9·7[a]
298	1560	542	511		
$P/10^5$ Pa	T/K	V/nm^3		$\alpha/10^{-3}$ K^{-1}	
1	293	28·5	28·4	1·23	1·27[b]
1	320	29·7	29·4		
ρ/kg m^{-3}	T/K	$-U$/kJ mol^{-1}		C_v/J K^{-1} mol^{-1}	
1489	293	6115·6	6118·9	67	74·7[b]
1489	320	6066·9	6062·1	73	
1489	340	6020·7	6007·6		

[a] Reference [40].
[b] Reference [42].

Examples of property calculations

3. Transport properties

→ Transport properties
- Diffusion (translational/rotational), viscosity

Transport coefficient: the example of diffusion

⇒ Self-diffusion (of particles within identical particles)

→ Fick's law

$$\boldsymbol{j} = -D\,\nabla c$$

$\boldsymbol{j}(\boldsymbol{r},t)$	*particle flux*
$c(\boldsymbol{r},t)$	*particle concentration*
D	*self-diffusion coefficient*

→ Conservation law

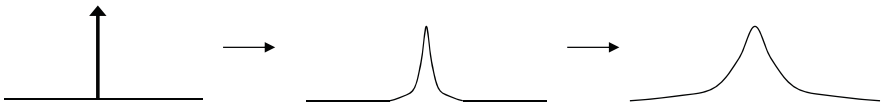
$$\frac{\partial c}{\partial t} + \nabla \cdot \boldsymbol{j} = 0$$

→ Diffusion equation (combining the two former equations)

$$\frac{\partial c}{\partial t} - D\nabla^2 c = 0$$

⇒ One particular solution

$$c(\boldsymbol{r},0) = \delta(\boldsymbol{r}) \quad \Rightarrow \quad c(\boldsymbol{r},t) = (4\pi D t)^{-3/2} e^{-r^2/(4Dt)}$$



Diffusion coefficient from simulation: The Einstein relation

⇒ The Einstein relation

$$\frac{d}{dt}\left\langle \boldsymbol{r}^2(t)\right\rangle = \frac{d}{dt}\int_\Omega d^3\boldsymbol{r}\, r^2 c(\boldsymbol{r},t) = D\int_\Omega d^3\boldsymbol{r}\, r^2\nabla^2 c(\boldsymbol{r},t)$$

$$= D\left[\int_\Omega d^3\boldsymbol{r}\,\nabla\cdot\left(r^2\nabla c(\boldsymbol{r},t)\right) - \int_\Omega d^3\boldsymbol{r}\,\nabla r^2\cdot\nabla c(\boldsymbol{r},t)\right]$$

=0 because gradient vanishes at ρ→∞

$$= D\left[\int_\Sigma d^2\sigma\cdot r^2\nabla c(\boldsymbol{r},t) - 2\int_\Omega d^3\,\boldsymbol{r}\cdot\nabla c(\boldsymbol{r},t)\right]$$

$$= -2D\left[\int_\Omega d^3\boldsymbol{r}\,\nabla\cdot\left(\boldsymbol{r}\,c(\boldsymbol{r},t)\right) - \int_\Omega d^3\boldsymbol{r}\left(\nabla\cdot\boldsymbol{r}\right)c(\boldsymbol{r},t)\right]$$

=0 because conc. vanishes at ρ→∞

$$= -2D\left[\int_\Sigma d^3\sigma\cdot\boldsymbol{r}\,c(\boldsymbol{r},t) - 3\int_\Omega d^3\boldsymbol{r}\,c(\boldsymbol{r},t)\right] = 6D$$

Ω: sphere of radius ρ→∞ around r(0)

Σ: surface of Ω

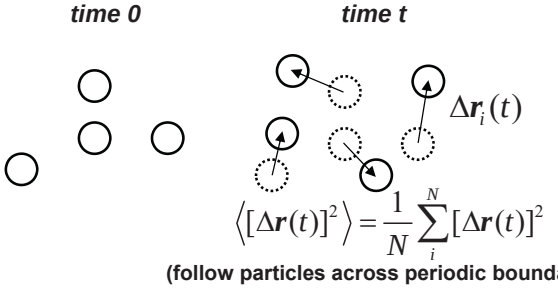
Convention:

$$\int_\Omega d^3\boldsymbol{r}\,c(\boldsymbol{r},t) = 1$$

Einstein relation

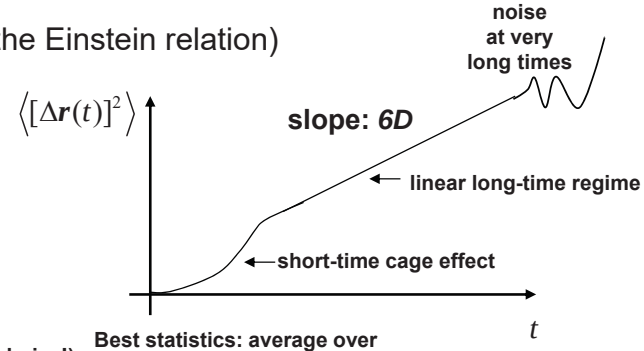
$$D = (1/6)\lim_{t\to\infty}\frac{d}{dt}\left\langle \boldsymbol{r}^2(t)\right\rangle$$

⇒ Computing *D* from simulations (via the Einstein relation)



$$\left\langle[\Delta\boldsymbol{r}(t)]^2\right\rangle = \frac{1}{N}\sum_i^N[\Delta\boldsymbol{r}(t)]^2$$

(follow particles across periodic boundaries!)



Best statistics: average over particles and time origins!

Diffusion coefficient from simulation: The Green-Kubo relation

⇒ The Green-Kubo relation

$$2D = \frac{d}{dt}\left\langle x^2(t)\right\rangle = \frac{d}{dt}\left\langle \left[\int_0^t dt'\, v_x(t')\right]^2 \right\rangle = \frac{d}{dt}\int_0^t dt' \int_0^t dt''\, \left\langle v_x(t')v_x(t'')\right\rangle$$

Einstein, isotropic

$$= 2\frac{d}{dt}\int_0^t dt' \int_{t'}^t dt''\, \left\langle v_x(t')v_x(t'')\right\rangle$$

$$= 2\frac{d}{dt}\int_0^t dt' \int_0^{t-t'} d\tau\, \left\langle v_x(t')v_x(t'+\tau)\right\rangle$$

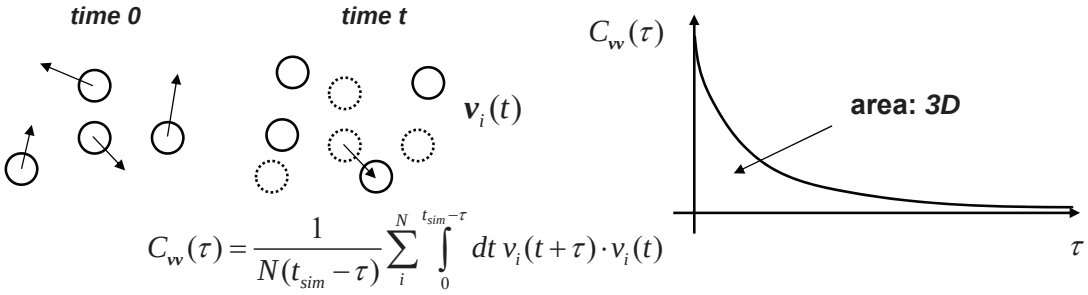
$$= 2\frac{d}{dt}\int_0^t dt' \int_0^t d\tau\, C_{v_x v_x}(\tau)$$

$$= 2\int_0^t d\tau\, C_{v_x v_x}(\tau) = (2/3)\int_0^t d\tau\, C_{\boldsymbol{vv}}(\tau)$$

Green-Kubo relation

$$D = (1/3)\lim_{\tau\to\infty}\int_0^\tau d\tau\, C_{\boldsymbol{vv}}(\tau)$$

⇒ Computing *D* from simulations (via the Green-Kubo relation)



$$C_{\boldsymbol{vv}}(\tau) = \frac{1}{N(t_{sim}-\tau)}\sum_i^N \int_0^{t_{sim}-\tau} dt\, \boldsymbol{v}_i(t+\tau)\cdot\boldsymbol{v}_i(t)$$

Other transport coefficients

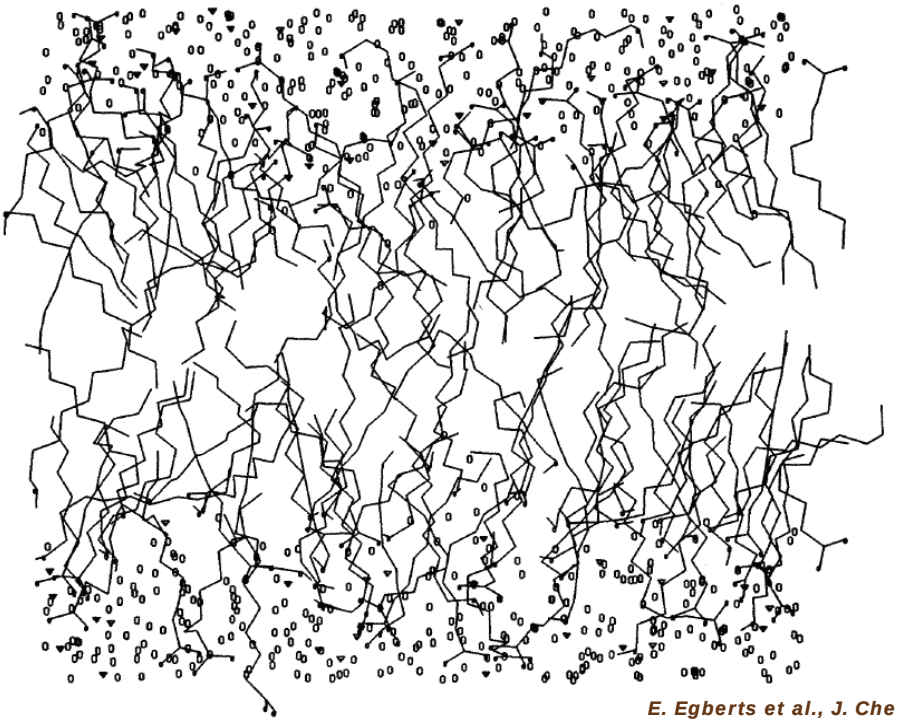
⇒ can be calculated via appropriate Einstein-like or Green-Kubo-like equations (the latter ones being more commonly used)

Examples:	***Autocorrelation function:***
• Diffusion coefficient (see above)	particle velocity
• Viscosity	stress tensor
• Thermal conductivity	heat current
• Electrical conductivity	electrical current
• Ionic conductivity	ionic current
• Frequency-dependent dielectric permittivity	box dipole moment

Example: simulations of a model bilayer

Snapshot of a simple bilayer of decanoate / decanol molecules, water and ions

Projection on the xz-plane of a snapshot of the system after 30 ps simulation:



Decanoate ions and decanol molecules (——)

Sodium ions (Δ)

Water oxygens (O)

Oxygens in the lipids (o)

E. Egberts et al., J. Chem. Phys., 89 (1988) 3718

Example: simulations of a model bilayer

Diffusion constants from mean square displacements

Lateral diffusion constants were calculated for **Na^+ ions**, and for the centers of mass of **decanoate ions**, **decanol molecules** and **water molecules** from mean square displacements using the relation:

$$\lim_{t\to\infty} \left\langle r^2(t) \right\rangle = 4Dt, \text{ where } r^2(t) = (x(t) - x(0))^2 + (y(t) - y(0))^2$$

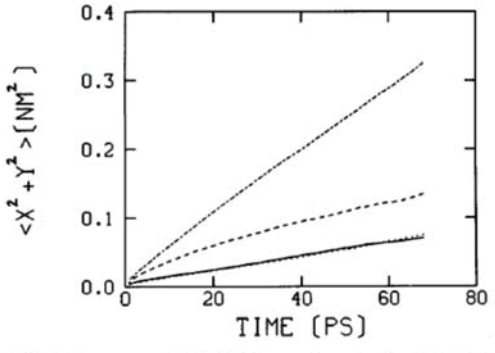


FIG. 12. Average squared lateral displacements as a function of time for the centers of mass of water molecules (- - - -), decanol molecules (– –), decanoate ions (—), and sodium ions (···).

Calculated diffusion constants:

Sodium

D = (2.7±0.3) x 10^{-6} cm^2s^{-1}

Decanoate

D = (2.7±0.3) x 10^{-6} cm^2s^{-1}

Decanol

D =(5.2±0.4) x 10^{-6} cm^2s^{-1}

Water

D=(1.2±0.05) x 10^{-5} cm^2s^{-1}

Fig 12: Average squared lateral displacements as a function of time for the centers of mass of decanoate ions (——), decanol molecules (– – –), water molecules (– · – ·) and sodium ions (·····).

Calculating diffusion constants

Translational and rotational diffusion

Dynamic properties **Chloroform**

C-H: z-direction

	Dietz and Heinzinger model		Kovacs *et al.* model		Jorgensen *et al.* model		
	This work	Ref. [28]	This work	Ref. [25]	This work	Ref. [29]	Exp.
$D_{tr}/10^{-9}$ m^2 s^{-1}	2·5	2·6	1·7	3·38	2·5		2·32[a]
$\tau_1(z)$/ps	3·6	3·1	4·9		3·1		
$\tau_1(xy)$/ps	3·8	3·6	5·3		3·6		
$\tau_2(z)$/ps	1·2	1·2	1·7	0·96	1·1		1·3[b]
$\tau_2(xy)$/ps	1·3	1·4	2·1		1·4		1·5[b]

[a] From reference [60].
[b] Rotational correlation times from [24].

Translation: $D_{tr} = \lim_{t\to\infty} \left\langle \left[\bar{r}_i(t'+t) - \bar{r}_i(t')\right]^2 \right\rangle_{t',i} \Big/ \left[6t\right]$

Rotation: $\left\langle \hat{\mu}_i(t') \cdot \hat{\mu}_i(t'+t) \right\rangle_{t',i} = e^{-t/\tau_1}$ $\quad \hat{\mu} = \mu_z$

$$\left\langle \frac{1}{2}\left[3\left(\hat{\mu}_i(t') \cdot \hat{\mu}_i(t'+t)\right)^2 - 1\right] \right\rangle_{t',i} = e^{-t/\tau_2} \qquad \hat{\mu} = \mu_x, \mu_z$$

I. Tironi et al., Mol. Phys., 83 (1994) 381-403

Examples of property calculations

4. Electromagnetic properties

→ Electromagnetic properties
- Dielectric permittivity, relaxation
- NMR parameters, relaxation → *discussed in lectures about refinement !*

Example: simulations of liquid chloroform

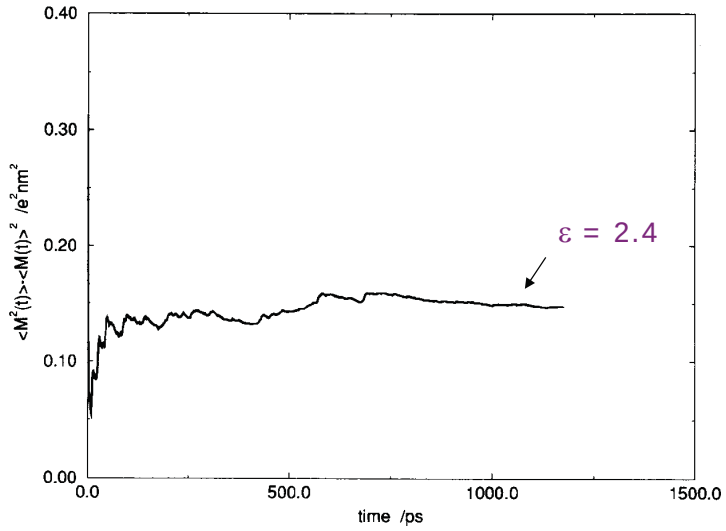
Liquid chloroform

Dielectric properties:

Dielectric **permittivity** ε:

$$(\varepsilon - 1)\left(\frac{2\varepsilon_{RF}+1}{2\varepsilon_{RF}+\varepsilon}\right) = \frac{\left\langle \left(\vec{M} - \langle \vec{M} \rangle\right)^2 \right\rangle}{3\varepsilon_0 V k_B T} = \frac{\langle \vec{M}^2 \rangle - \langle \vec{M} \rangle^2}{3\varepsilon_0 V k_B T}$$

$$\vec{M} = \sum_{i=1}^{N} \vec{\mu}_i = \text{total dipole moment}$$



$$\langle \vec{M}(t) \cdot \vec{M}(0) \rangle \propto e^{-t/\tau}$$

τ = 3.4 ps

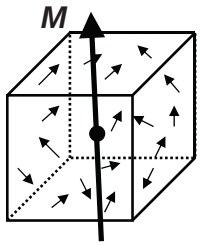


Figure 4. Cumulative average of the total dipole moment fluctuation of the system (Dietz model) as a function of time.

I. Tironi et al., Mol. Phys., 83 (1994) 381-403

Example: simulations of liquid chloroform

Liquid chloroform

Dielectric properties:

Debye **relaxation** time τ_D

$$\frac{\left\langle \vec{M}(t') \cdot \vec{M}(t'+t) \right\rangle_{t'}}{\left\langle \vec{M}(t') \cdot \vec{M}(t') \right\rangle_{t'}} = e^{-t/\tau} \qquad \tau \approx \frac{2\varepsilon_{RF}+1}{2\varepsilon_{RF}+\varepsilon}\tau_D$$

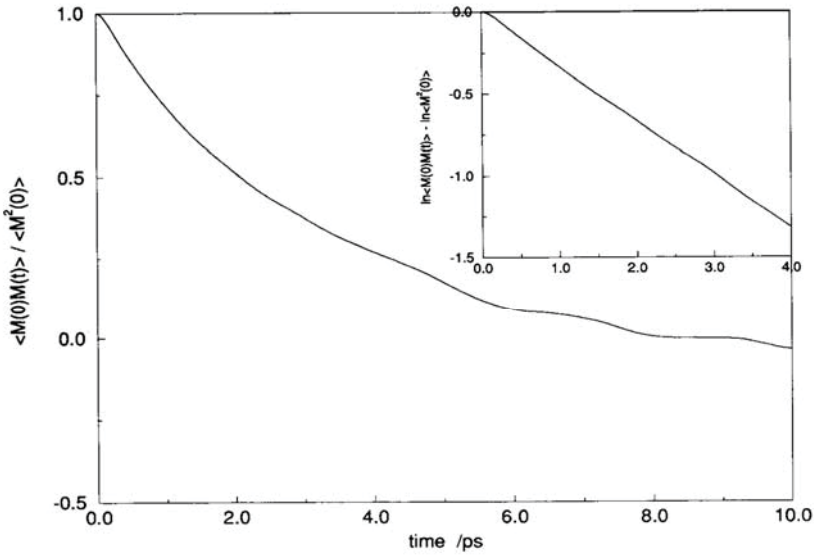


Figure 5. Normalized autocorrelation function of the total dipole moment of the sy (Dietz model). The inset shows a semilogarithmic plot of the correlation function the first 5 ps.

I. Tironi et al., Mol. Phys., 83 (1994) 381-403

Example: simulations of liquid chloroform

Chloroform: various properties

	MD	Exp.
Dielectric properties		
ϵ	2·4	4·81 [40]
τ_D/ps^{-1}	3·8	5·4[a]
Helmholtz energy/kJ mol^{-1}		
over 50 ps: forward	15·6	15·3
backward	−14·3	
over 80 ps: forward	15·4	
backward	−14·5	
Viscosity/cP		
η via Einstein	0·6	0·568[b]
via Green–Kubo	0·53	

[a] Reference [37].
[b] Reference [40].

I. Tironi et al., Mol. Phys., 83 (1994) 381-403

The art of analysis

- My recommended sequence for simulation analysis

  → (A) Make a **movie** and watch it carefully — *This won't tell you anything "sharp" but may give you an idea of (1) what is possibly going wrong and (2) what might be particularly interesting looking at*

  → (B) Do **standard analyses** — *E.g. energy components, temperatures, pressure, RMSD, specific internal coordinates... These are easy (programs available) and may give you information on (1) what is possibly going wrong; (2) whether you are well equilibrated; (3) what might be particularly interesting looking at*

  → (C) Do **tailored analyses** — *Define the observable(s) that are most likely to characterize simply and precisely the process you observe; here, don't be lazy: if the program does not exist, write it!*

  → (D) Make **well-designed** graphs — *Plot things in a way that the main message is easy/immediate to grasp from the figures/tables*

  → (E) If possible, report **error estimates** — *Not always easy/possible --- can be cumbersome; also: what is in the error --- usually only statistical error*

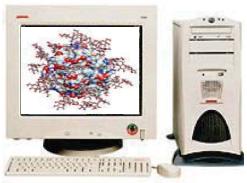
- Good simulation studies **start** with a **good hypothesis/question**; keep this question in mind when analyzing the results; find the observables and make the graphs that will answer your initial question in the most clear-cut way
- Most studies that stop at (B) end up being **totally dull**; and if you stop at (C), you may **fail to convey** clearly your conclusions to other people

The art of analysis

- Some usual problems may cost you time for nothing; so check for them early:
  - → SHAKE failures *They are usually an «alarm bell» for something else !*
  - → System was not equilibrated long/carefully enough
  - → Temperature is not what it should be (or is inhomogeneous in the system) *In particular: check Tsolute vs Tsolvent !*
  - → Center-of-mass motion builds up unexpectedly
  - → Periodic gathering is not done appropriately
- Also remember that the apparent stability of standard observables in time is a necessary, not a sufficient condition for equilibrium

P.H. Hünenberger

COMPUTER SIMULATION OF MOLECULAR SYSTEMS

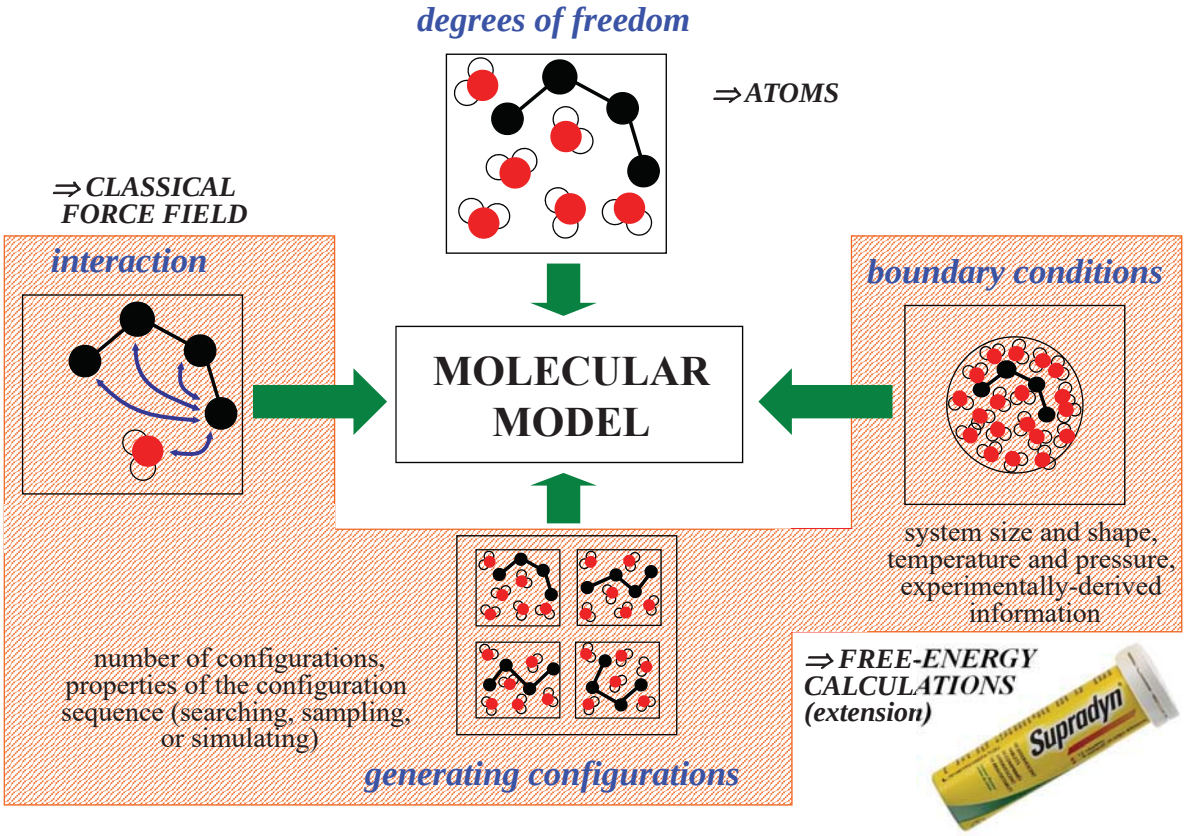


Herbstsemester 2019
Tuesday 9:45-11:30 a.m.
HCI D2

Lecture 529-0004-00
www.csms.ethz.ch/education/CSCBP

LECTURE 7 (WEEK 8):
Free energy calculations I

★ **Four basic choices defining a molecular model**



Free energy in thermodynamics

- In **phenomenological thermodynamics**, the **free energy** is a **thermodynamic potential** (state function with an energetic component), with the definition

  → Helmholtz free energy $F = E - TS$

  (also sometimes noted A; appropriately called the free energy)

  → Gibbs free enthalpy $G = F + PV = E + PV - TS$

  (often also called Gibbs free energy or simply free energy)

E *internal (=total) energy*
P *pressure*
V *volume*
T *(absolute) temperature*
S *(absolute) entropy*

- These quantities are **state functions** (solely depending on the present state of the system, not on its history), **extensive** and defined for **any system** (irrespective of its boundary conditions)

  microcanonical

  → For a system under **closed isochoric adiabatic** boundary conditions (NVE), any **spontaneous** change in the system is accompanied by an **increase in *S***

$$dS\big|_{NVE} \geq 0$$ *for any spontaneous change in the system (zero: system at equilibrium)*

2nd law of thermodynamics ⇓ *derived formulation*

  canonical

  → For a system under **closed isochoric isothermal** boundary conditions (NVT), any spontaneous change in the system is accompanied by a **decrease in *F***

$$dF\big|_{NVT} \leq 0$$ *for any spontaneous change in the system (zero: system at equilibrium)*

derived formulation

isothermal-isobaric (Gibbs)

  → For a system under **closed isobaric isothermal** boundary conditions (NPT), any spontaneous change in the system is accompanied by a **decrease in *G***

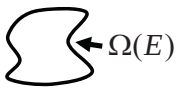
$$dG\big|_{NPT} \leq 0$$ *for any spontaneous change in the system (zero: system at equilibrium)*

derived formulation

Free energy in statistical mechanics

- In statistical mechanics, the **free energy** is related to the **partition function**

  → For a system under **closed isochoric adiabatic** boundary conditions (NVE), the entropy *S* is related to the **microcanonical** partition function



area of an energy isosurface at E in 6N-dimensional phase space

Boltzmann equation (a priori equiprobability) + normalization

$$S = k_B \ln\left[\xi \iint d\boldsymbol{r}\, d\boldsymbol{p}\, \delta\left(E^{-1}\mathcal{H}(\boldsymbol{r},\boldsymbol{p}) - 1\right)\right]$$

Density of states $\Omega(E)$

given N,V

$$S(N,V,E)$$

N *number particles* h *Planck constant*
δ *Dirac delta function* $\xi = \left(h^{3N} N!\right)^{-1}$
k_B *Boltzmann constant* *[indistinguishable particles]*
$\mathcal{H}(\boldsymbol{r},\boldsymbol{p}) = \sum_{i=1}^{N} \frac{\boldsymbol{p}_i^2}{2m_i} + \mathcal{V}(\boldsymbol{r})$ *Hamiltonian*

  → For a system under **closed isochoric isothermal** boundary conditions (NVT), the free energy *F* is related to the **canonical** partition function

$$F = -\beta^{-1} \ln Z_{NVT} = -\beta^{-1} \ln\left[\xi \iint d\boldsymbol{r}\, d\boldsymbol{p} \exp\left(-\beta \mathcal{H}(\boldsymbol{r},\boldsymbol{p})\right)\right]$$

canonical partition function

T *(absolute) temperature*
$\beta = \left(k_B T\right)^{-1}$

  → For a system under **closed isobaric isothermal** boundary conditions (NPT), the free enthalpy *G* is related to the **Gibbs** partition function

$$G = -\beta^{-1} \ln Z_{NPT} = -\beta^{-1} \ln\left[\xi \int d\mathcal{V} \iint d\boldsymbol{r}\, d\boldsymbol{p} \exp\left(-\beta[\mathcal{H}(\boldsymbol{r},\boldsymbol{p},\mathcal{V}) + P\mathcal{V}]\right)\right]$$

Gibbs partition function

$\mathcal{V}$ *volume*
P *(reference) pressure*

- For simplicity, we will only further consider the **free energy** and the **canonical ensemble**

Free energy and molecular dynamics

- In the canonical (NVT) ensemble, a given **Hamiltonian** defines a unique canonical **probability distribution**

$$\mathcal{H}(\boldsymbol{r},\boldsymbol{p}) \quad \Rightarrow \quad P(\boldsymbol{r},\boldsymbol{p}) = \frac{\exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p})\right)}{\iint d\boldsymbol{r}\, d\boldsymbol{p} \exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p})\right)}$$

normalized

→ Many observables can be written as canonical **ensemble averages**

$$Y = \left\langle \mathcal{Y}(\boldsymbol{r},\boldsymbol{p}) \right\rangle = \iint d\boldsymbol{r}\, d\boldsymbol{p}\, P(\boldsymbol{r},\boldsymbol{p}) \mathcal{Y}(\boldsymbol{r},\boldsymbol{p})$$

Y — *ensemble-average observable*

$\mathcal{Y}(\boldsymbol{r},\boldsymbol{p})$ — *corresponding instantaneous observable*

→ The free energy can also **formally** be written based on an ensemble average

$$F = \beta^{-1}\left[\ln\left\langle \exp\left(+\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p})\right)\right\rangle + C\right]$$

reason:

$$\left\langle \exp\left(+\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p})\right)\right\rangle = \frac{\iint d\boldsymbol{r}\, d\boldsymbol{p}}{\iint d\boldsymbol{r}\, d\boldsymbol{p} \exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p})\right)} = \tilde{C}\xi Z^{-1}$$

- **Molecular dynamics** at **constant volume** (periodic boundary conditions) and with an appropriate **thermostat** samples the **canonical** (NVT) ensemble

→ This method automatically generates configuration according to the canonical probability distribution

$$Y = \left\langle \mathcal{Y}(\boldsymbol{r},\boldsymbol{p}) \right\rangle_{MD}$$

i.e. a straight average over the sampled configurations is already a Boltzmann-weighted ensemble average

→ This method does **not** allow to calculate **absolute free energies**

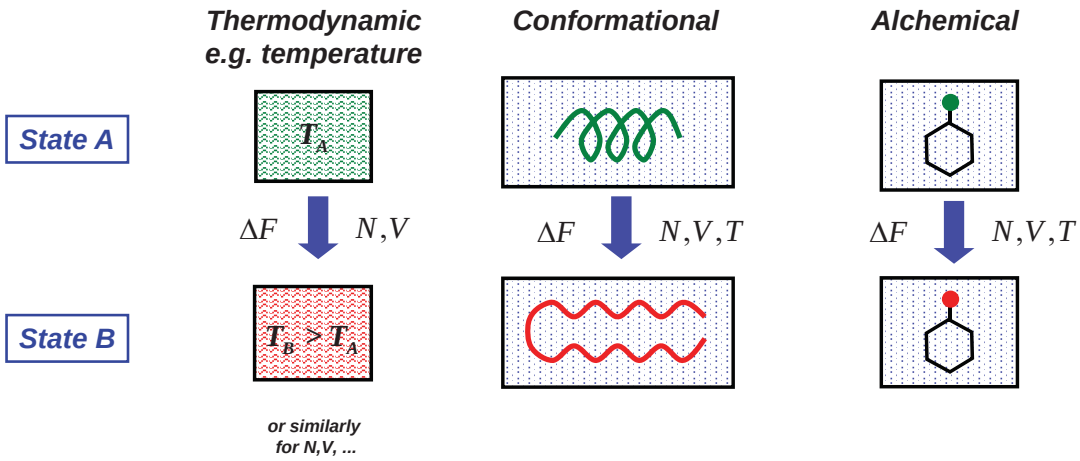
the average $\left\langle \exp\left(+\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p})\right)\right\rangle_{MD}$ *will never converge...*

the averaged quantity is highest where the sampling probability is lowest !

→ But: we can calculate **free-energy differences** !

Three types of free-energy differences

- Restricting the discussion to state **pairs**, one may distinguish **three basic types** of free-energy changes

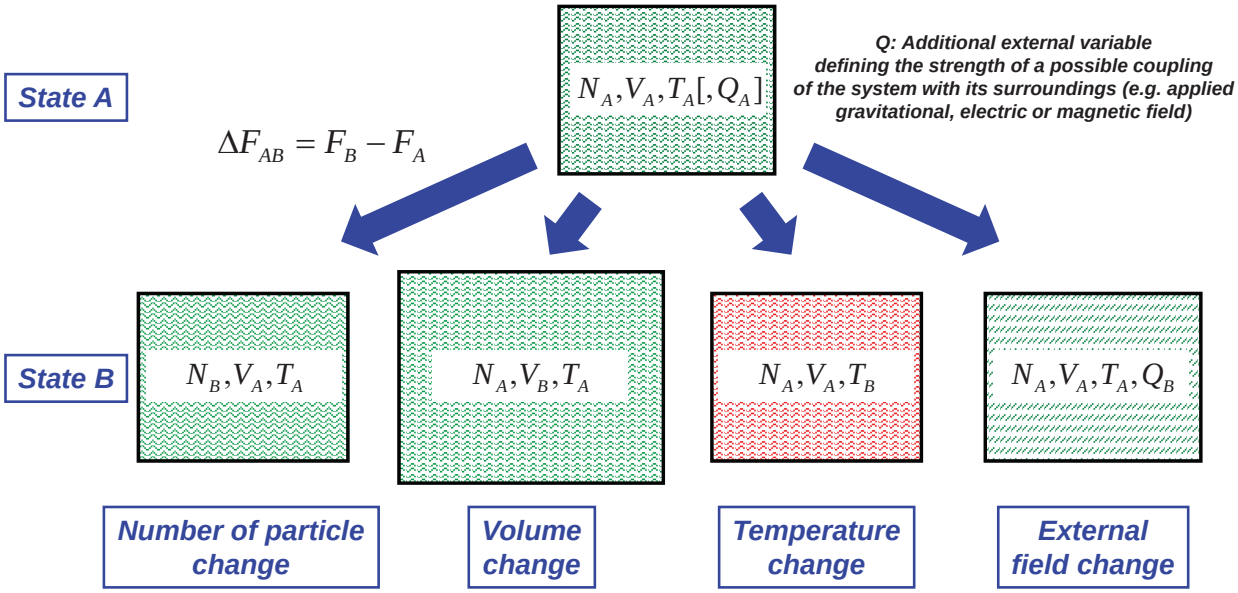


Three types of free-energy differences

- Free-energy change associated with a **thermodynamic state change**

  → The change is **physical** and **real** (obvious experimental counterpart), and involves one of the parameters defining the **thermodynamic boundary conditions** of the system

$$F_X = -\beta_X^{-1} \ln\left[\xi_X \iint d\boldsymbol{r}\, d\boldsymbol{p} \exp\left(-\beta_X \mathcal{H}(\boldsymbol{r}, \boldsymbol{p}; V_X[, Q_X])\right)\right] \qquad \boldsymbol{r}, \boldsymbol{p}: \; 3N_X \text{ - dimensional vectors}$$

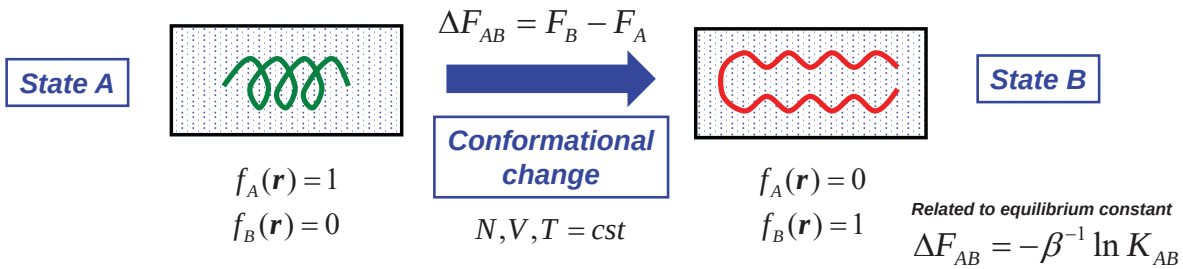


Three types of free-energy differences

- Free-energy change associated with a **conformational state change**

  → The change is **physical** but **virtual** (indirect experimental counterpart), and involves a partitioning of the conformational space of the system into distinct **conformational states**

$$F_X = -\beta^{-1} \ln\left[\xi \iint d\boldsymbol{r}\, d\boldsymbol{p}\; f_X(\boldsymbol{r}) \exp\left(-\beta \mathcal{H}(\boldsymbol{r}, \boldsymbol{p})\right)\right] \qquad f_X(\boldsymbol{r}): \text{ conformational-state indicator function (no dependence on momenta [conformational])}$$



Related to equilibrium constant

$$\Delta F_{AB} = -\beta^{-1} \ln K_{AB}$$

→ The definition of states can be

Thermodynamic
A state gathers the conformations within a common free-energy basin (but in which reduced space ?)

Structural
A state is a collection of configurations with low mutual coordinate deviations

Kinetic
A state is a collection of configurations with rapid interconversion rates (compared to conversion to others)

Experimental
A state is a collection of configurations with specific spectroscopic or functional properties

Would one of the most fundamental concepts of chemistry be... undefined ?

States are largely "in the eye of the beholder" → a device to bring a complex configurational ensemble dynamics into a simplified form, amenable to human reasoning

Watch out for possible discrepancies between definitions

(e.g. unfolded state = high rmsd or dark in NMR ?)

Three types of free-energy differences

- Free-energy change associated with a **conformational state change**

→ One may also be interested in the free-energy change along a **reaction coordinate**, which is called a **potential of mean force** (PMF)

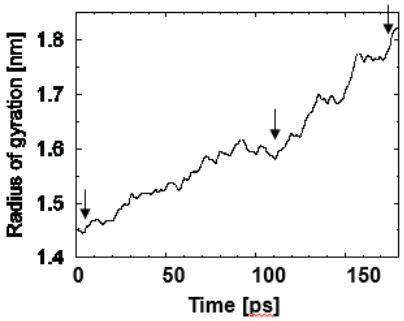
$$F(s) = -\beta^{-1} \ln\left[\xi \iint d\boldsymbol{r}\, d\boldsymbol{p}\; f(\boldsymbol{r};s) \exp\left(-\beta \mathcal{H}(\boldsymbol{r},\boldsymbol{p})\right)\right] \qquad f(\boldsymbol{r},s): \quad \textit{reaction coordinate indicator function}$$

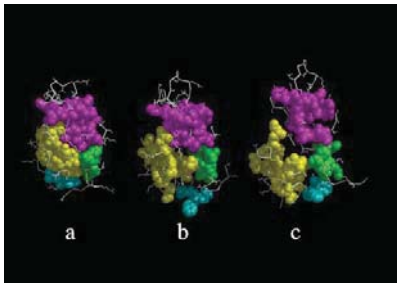
→ Example: lysozyme thermal denaturation in water (500K)

$$R_{gyr}(\boldsymbol{r}) = \left[\frac{1}{M}\sum_{i=1}^{N} m_i (\boldsymbol{r}_i - \boldsymbol{R}_{cm})^2\right]^{1/2}$$

center-of-mass position $\boldsymbol{R}_{cm} = \frac{1}{M}\sum_{i=1}^{N} m_i \boldsymbol{r}_i$

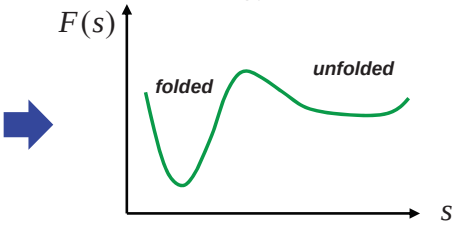
total mass $M = \sum_{i=1}^{N} m_i$





→ Try to establish a PMF as a function of the radius of gyration at 300K (difficult task !)

$$f(\boldsymbol{r};s) = \delta(R_{gyr}(\boldsymbol{r}) - s)$$

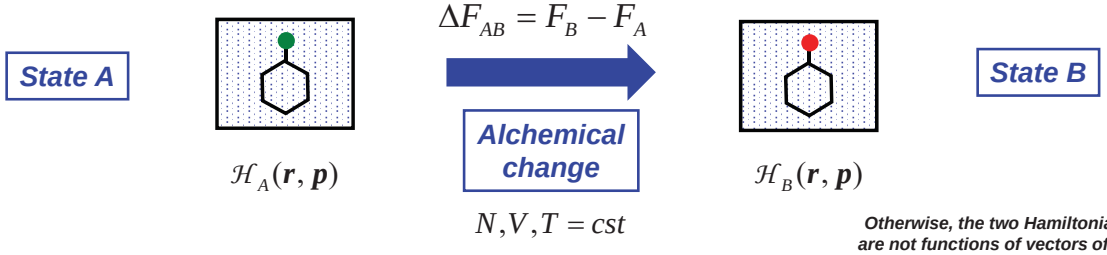


Three types of free-energy differences

- Free-energy change associated with an **alchemical state change**

→ The change is **unphysical** (no experimental counterpart), and involves a change in the form of the **potential energy function** *via* the associated molecular topology

$$F_X = -\beta^{-1} \ln\left[\xi \iint d\boldsymbol{r}\, d\boldsymbol{p} \exp\left(-\beta \mathcal{H}_X(\boldsymbol{r},\boldsymbol{p})\right)\right] \qquad \mathcal{H}_X(\boldsymbol{r},\boldsymbol{p}): \quad \textit{state Hamiltonian}$$



Otherwise, the two Hamiltonians are not functions of vectors of the same dimensionality! Also: deleting a point mass does not make sense classically, and non-interacting point-mass statistical mechanics (ideal gas) is analytical

→ The **number of particles** must be **identical** in the two states, *i.e.* only atom **mutations** are possible

→ The **creation** or **deletion** of an atom corresponds to its mutation from or to a **dummy atom** (point mass without any interactions to its environment, or only intramolecular ones)

Example: ethane→methane (united-atom)

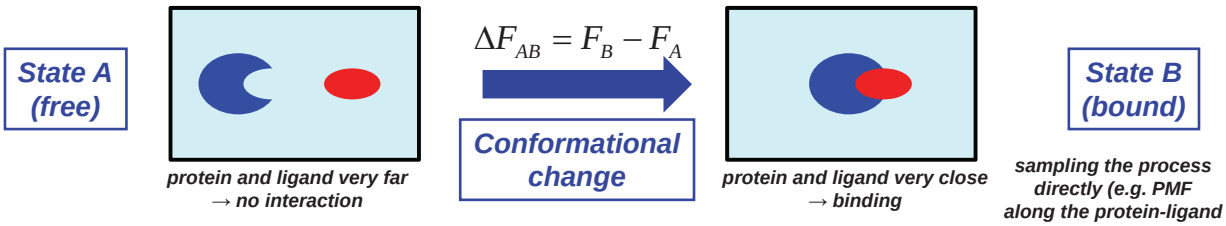
No covalent interaction is often unwise: particle is a gas particle not seeing the rest of the system (sampling issue)

→ The calculated free-energy change cannot be directly compared to any experimental number; only the difference between two closely-related alchemical calculations, differing in the **environment** of the molecule, are compared in a **thermodynamic cycle**

Thermodynamic cycle

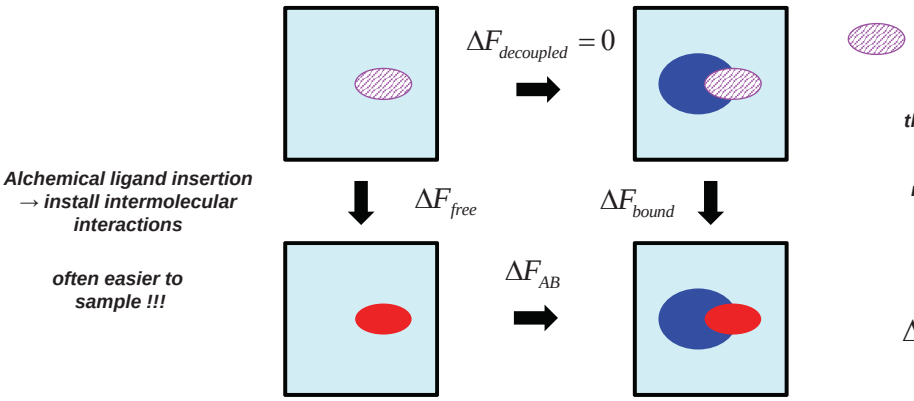
- The interest of **alchemical** transformations is that they offer an interesting (often computationally more efficient [easier sampling]) alternative to the monitoring of **conformational** changes; this is achieved by using a **thermodynamic cycle**

→ Example of conformational process of interest



sampling the process directly (e.g. PMF along the protein-ligand distance) is difficult

→ Corresponding thermodynamic cycle



Decoupled state of the ligand → no intermolecular interactions

⇒ Many choices; generally, the intramolecular interactions will be either kept unaltered (or only the intramolecular non-bonded will be removed)

Alchemical ligand insertion → install intermolecular interactions

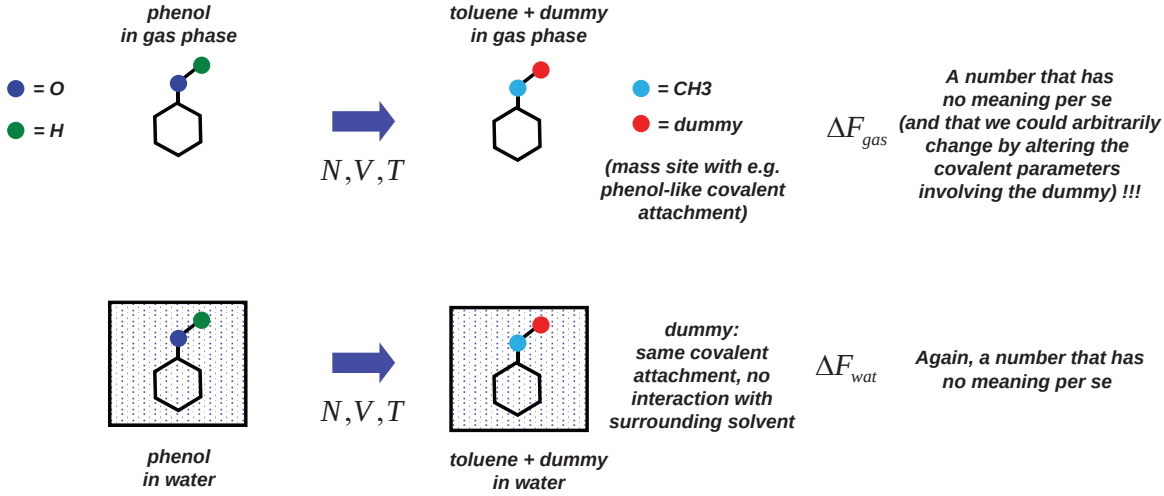
often easier to sample !!!

calculated via cycle:

$$\Delta F_{AB} = \Delta F_{bound} - \Delta F_{free}$$

Thermodynamic cycle

- Another example (cf exercise 5)



(mass site with e.g. phenol-like covalent attachment)

ΔF_{gas}

A number that has no meaning per se (and that we could arbitrarily change by altering the covalent parameters involving the dummy) !!!

dummy: same covalent attachment, no interaction with surrounding solvent

ΔF_{wat}

Again, a number that has no meaning per se

→ Physical quantity

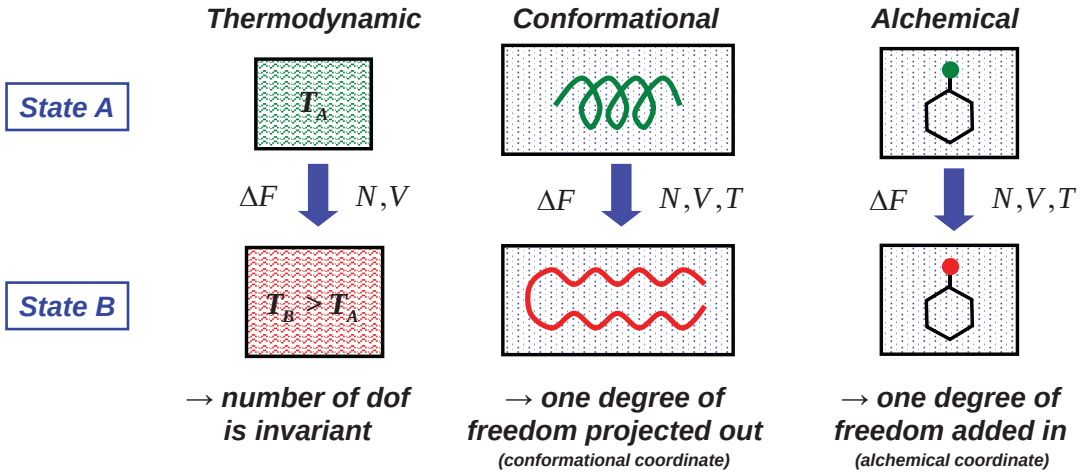
$$\Delta F_{wat} - \Delta F_{gas} = \Delta F_{slv}(tol) - \Delta F_{slv}(phe)$$

The meaningful (physically measurable) quantity !

[the effect of a specific covalent attachment of the dummy becomes irrelevant (cancels out)]

Three types of free-energy differences

- The three types of processes mainly differ by the **number of degrees of freedom** (dof) they consider



→ The free-energy methods applied to the three types of problems are **historically** quite different; but if one looks more closely, any type of method can be applied to any of the three types of problems

→ It is possible to transform *thermodynamic* and *alchemical* problems into *pseudo-conformational* problems using an **extended-system** approach

→ It is possible to transform *conformational* problems into *pseudo-alchemical* problems using a **reduced-system** (constraint) approach

Extended-system pseudo-conformational free-energy changes

- A **thermodynamic** problem can be converted to a **pseudo-conformational** problem by adding the varied parameter as a dynamic variable in the MD simulation

→ Examples

Box volume	$V \to \mathcal{V}(t)$	$\ddot{\mathcal{V}}(t) = -m_V^{-1} \left. \frac{\partial \mathcal{H}(\boldsymbol{r}, \boldsymbol{p}, \mathcal{V})}{\partial \mathcal{V}} \right\|_t$	m_V *fictitious (piston) mass*	*= Andersen barostat*
Thermostat reference temperature	$T \to \mathcal{T}(t)$	*[more complicated]*		*= multicanonical sampling or SPEED*
Number of particles	$N \to \mathcal{N}(t)$	$\ddot{\mathcal{N}}(t) = -m_N^{-1} \left. \frac{\partial \mathcal{H}(\boldsymbol{r}, \boldsymbol{p}, \mathcal{N})}{\partial \mathcal{N}} \right\|_t$	m_N *fictitious mass*	*= grand-canonical MD*

- An **alchemical** problem can be converted to a **pseudo-conformational** problem by introducing a Hamiltonian coupling parameter, and treating it as a dynamic variable in the MD simulation

→ We first need to define a **coupling scheme**

λ: Hamiltonian coupling parameter

$$\mathcal{H}(\boldsymbol{r}, \boldsymbol{p}; \lambda) = \begin{cases} \mathcal{H}_A(\boldsymbol{r}, \boldsymbol{p}) & \text{if } \lambda=0 \\ \mathcal{H}_B(\boldsymbol{r}, \boldsymbol{p}) & \text{if } \lambda=1 \end{cases}$$

this condition is actually compatible with many alternative coupling schemes !

→ Then we use so-called **λ-dynamics** *λ becomes a dynamic variable !*

Coupling parameter	$\lambda \to \lambda(t)$	$\ddot{\lambda}(t) = -m_\lambda^{-1} \left. \frac{\partial \mathcal{H}(\boldsymbol{r}, \boldsymbol{p}, \lambda)}{\partial \lambda} \right\|_t$	m_λ *fictitious mass*	*= λ-dynamics*

Reduced-system pseudo-alchemical free-energy changes

- A **conformational** problem can be converted to a **pseudo-alchemical** problem by removing one (collective) dynamical variable in the MD simulation

  → Collective coordinate constraint

dynamics using $\mathcal{H}(\boldsymbol{r},\boldsymbol{p})+\gamma(t)\delta(S(\boldsymbol{r})-s)$

- $S(\boldsymbol{r})$ — *collective coordinate*
- s — *reference value* → *pseudo-alchemical variable*
- $\gamma(t)$ — *Lagrange multiplier (determined at each timestep so that the constraint is satisfied; e.g. SHAKE)*
- δ — *Dirac delta function*

Overview of free-energy methods

- The available methods to calculate free-energy changes can be classified as historically applied to the different types of changes

  → But remember that extended- or reduced-system approaches allow to interconvert the problem types, so that all methods are in principle applicable to all problem types !

  → Still, some combinations are more **practical** or/and **efficient** than others...

Thermodynamic	*Conformational*	*Alchemical*
Temperature integration	**Direct counting (DC)**	**Thermodynamic integration (TI)**
Pressure integration	**Umbrella sampling (US)**	**Free-energy perturbation (FEP)**
Grand-canonical integration		**λ-dynamics local elevation umbrella sampling (λ-LEUS)**
		Enveloping distribution sampling (EDS)
		Fast growth (FG)

$$\Delta F_{AB} = -\beta^{-1} \ln \frac{\langle f_B(\boldsymbol{r}) \rangle}{\langle f_A(\boldsymbol{r}) \rangle}$$

Direct counting (DC)

- The *a priori* simplest method to calculate a free-energy difference is **direct counting** (DC); it applies to *conformational problems*, as well as to thermodynamic and alchemical problems reformulated as *pseudo-conformational* problems
- Consider a (pseudo-)conformational Hamiltonian and the associated canonical probability distribution

$$\mathcal{H}(\boldsymbol{r},\boldsymbol{p},Q) \Rightarrow P(\boldsymbol{r},\boldsymbol{p},Q)=\frac{\exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p},Q)\right)}{\iint d\boldsymbol{r}\,d\boldsymbol{p}\,dQ\,\exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p},Q)\right)}$$

Q *possible extended-system variable (only present if «pseudo»)*

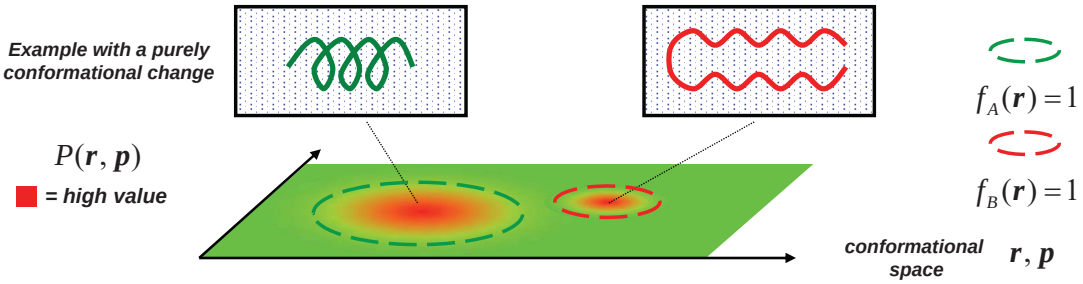
- The free energy of a (pseudo-)conformational state X is

$$F_X=-\beta^{-1}\ln\left[\iint d\boldsymbol{r}\,d\boldsymbol{p}\,dQ\;P(\boldsymbol{r},\boldsymbol{p},Q)\,f_X(\boldsymbol{r},Q)\right]+C=-\beta^{-1}\ln\left\langle f_X(\boldsymbol{r},Q)\right\rangle+C$$

- Thus, in MD, the free-energy difference between two (pseudo-)conformational states A and B can in principle be obtained by monitoring the occurrences of the two states

$$\Delta F_{AB}=-\beta^{-1}\ln\frac{\left\langle f_B(\boldsymbol{r},Q)\right\rangle}{\left\langle f_A(\boldsymbol{r},Q)\right\rangle}$$

We just to count the number of trajectory frames in the MD where f_A=1 and where f_B=1 !



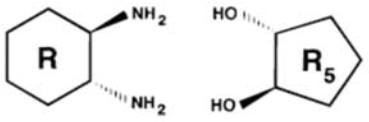
Direct counting (DC)

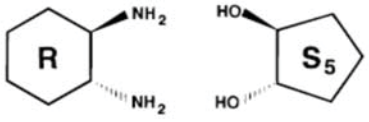
- Example of application to enantioselective complexation

An (unusual) example where it works...

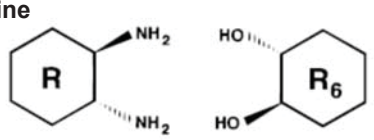
$T=298.15K$

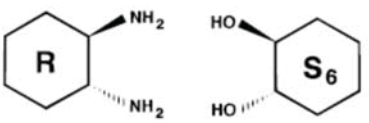
1,2-R/R- or S/S-cyclopentanediol





1,2-R/R-cyclohexanediamine





Experimental values:

		In Benzene	in CCl_4
K_b	[1/M]	40	97
ΔG_b	[kJ/mol]	-9.3	-11.5
ΔH_b	[kJ/mol]	-20.2	-17.6
ΔS_b	[J/(K·mol)]	-36.4	-20.3
$T\Delta S_b$	$[kJ/mol]$	−10.9	−6.1
K_b	[1/M]	15	53
ΔG_b	[kJ/mol]	-6.8	-10.0
ΔH_b	[kJ/mol]	-41.7	-34.3
ΔS_b	[J/(K·mol)]	-116.3	-81.2
$T\Delta S_b$	$[kJ/mol]$	−34.9	−24.3
K_b	[1/M]	19	53
ΔG_b	[kJ/mol]	-7.4	-10.0
ΔH_b	[kJ/mol]	-21.8	-15.9
ΔS_b	[J/(K·mol)]	-47.9	-19.8
K_b	[1/M]	14	38
ΔG_b	[kJ/mol]	-6.6	-9.1
ΔH_b	[kJ/mol]	-40.6	-36.0
ΔS_b	[J/(K·mol)]	-113.2	-89.6

$$\Delta G_b=\Delta H_b-T\Delta S_b=-RT\ln K_b$$

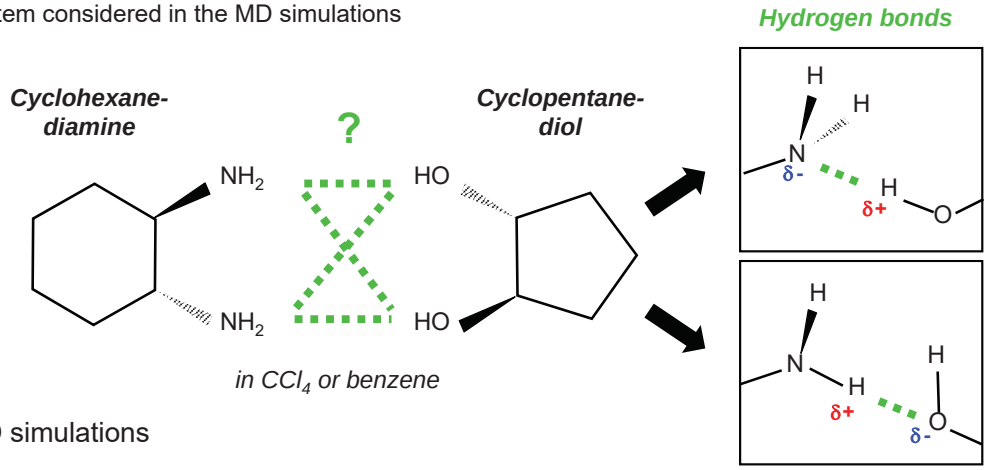
1,2-R/R- or S/S-cyclohexanediol

Hünenberger et al., JACS 119 (1997) 7533-7544

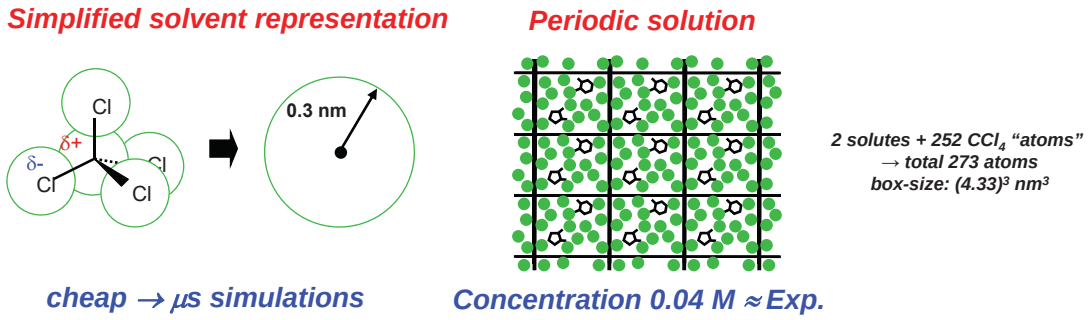
Direct counting (DC)

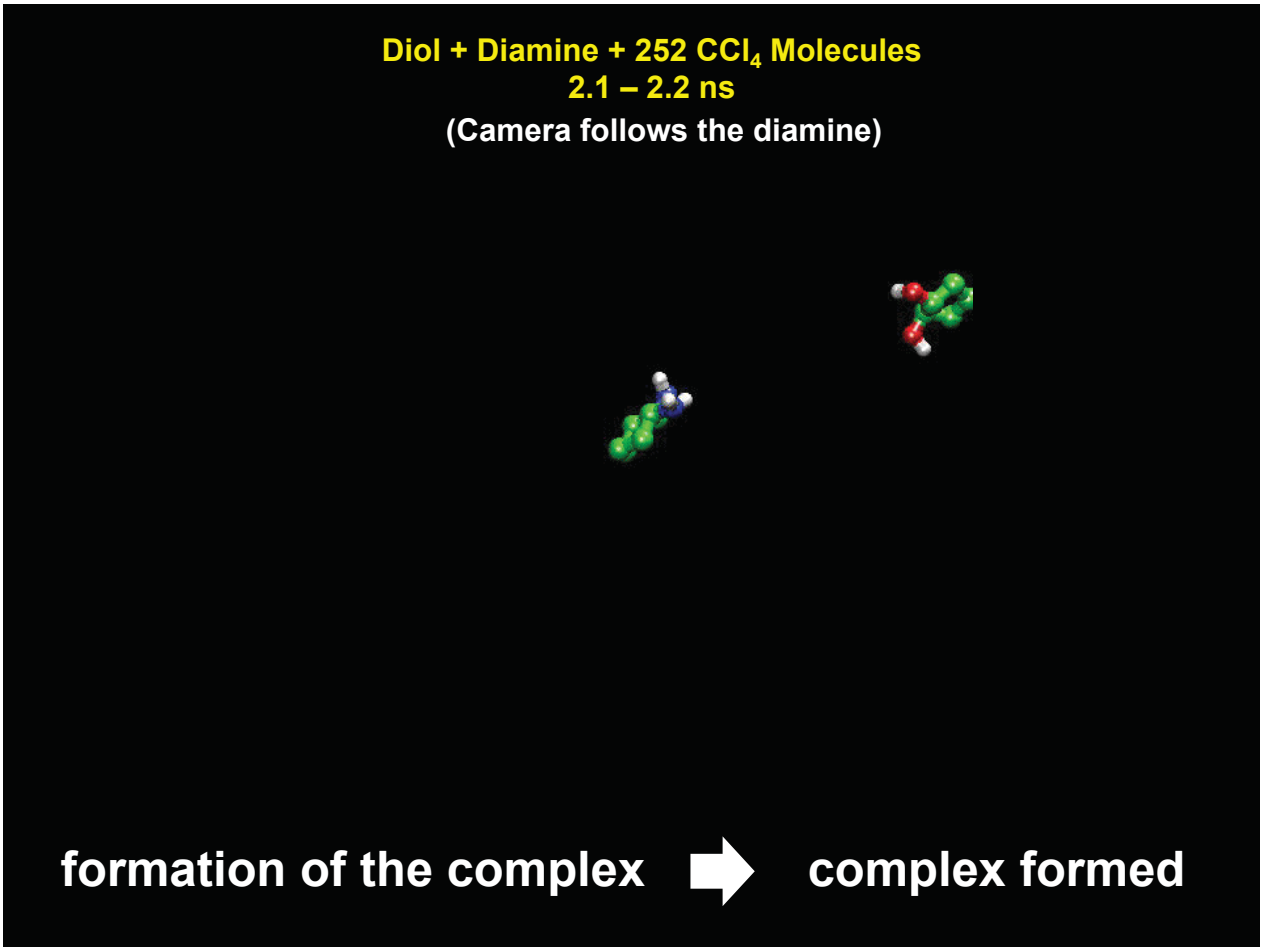
- Example of application to enantioselective complexation (continued)

→ System considered in the MD simulations



→ MD simulations



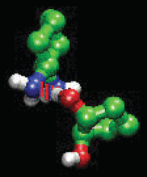


Diol + Diamine + 252 CCl_4 Molecules
3.2 – 4.0 ns
(Camera follows the diamine)

Hydrogen bonds

O → N

N → O

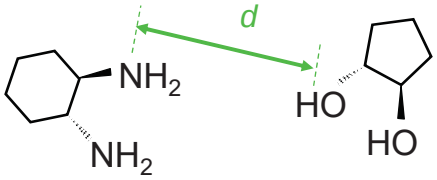


… and a nanosecond afterwards … ➡ **the molecules are free again…**

Direct counting (DC)

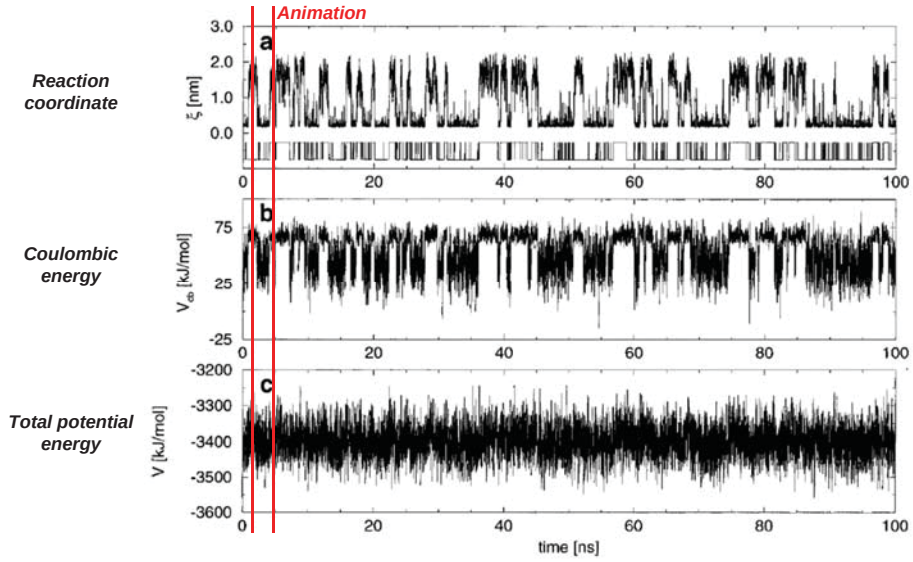
- Example of application to enantioselective complexation (continued)

→ Reaction coordinate



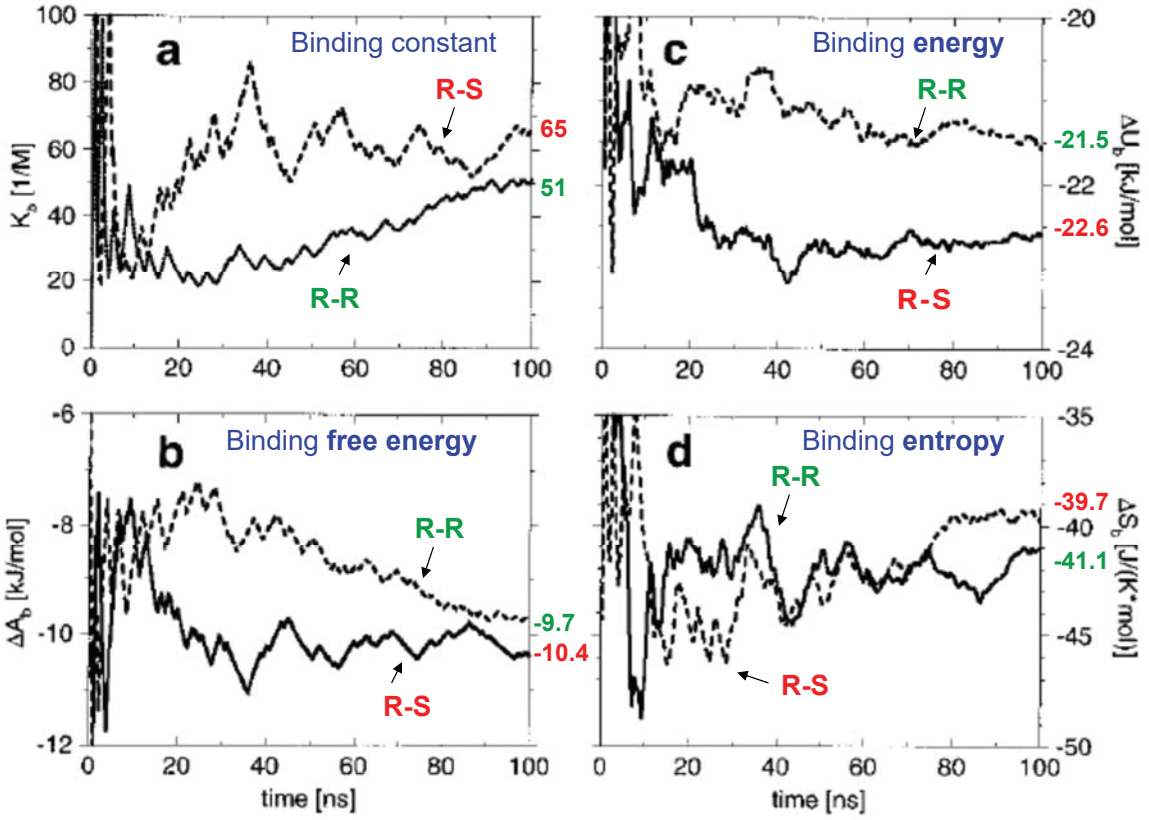
Reaction coordinate ξ(t): Minimum O(diol) to N(diamine) distance

→ Time series of the reaction coordinate and energy during the simulation



Direct counting (DC)

- Example of application to enantioselective complexation (continued)



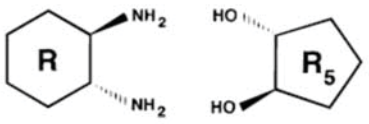
Direct counting (DC)

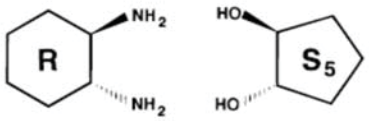
- Example of application to enantioselective complexation

An (unusual) example where it works...

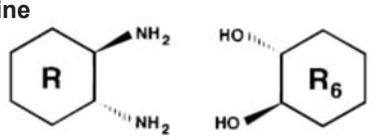
$T = 298.15K$

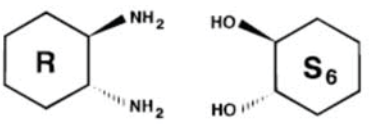
1,2-R/R- or S/S-cyclopentanediol





1,2-R/R-cyclohexanediamine





1,2-R/R- or S/S-cyclohexanediol

		Experimental values: In Benzene	in CCl_4	Simul.
K_b	[1/M]	40	97	51
ΔG_b	[kJ/mol]	-9.3	-11.5	-9.7
ΔH_b	[kJ/mol]	-20.2	-17.6	-21.5
ΔS_b	[J/(K·mol)]	-36.4	-20.3	-39.7
$T\Delta S_b$	$[kJ/mol]$	−10.9	−6.1	-11.8
K_b	[1/M]	15	53	65
ΔG_b	[kJ/mol]	-6.8	-10.0	-10.4
ΔH_b	[kJ/mol]	-41.7	-34.3	-22.6
ΔS_b	[J/(K·mol)]	-116.3	-81.2	-41.1
$T\Delta S_b$	$[kJ/mol]$	−34.9	−24.3	-12.2
K_b	[1/M]	19	53	
ΔG_b	[kJ/mol]	-7.4	-10.0	
ΔH_b	[kJ/mol]	-21.8	-15.9	
ΔS_b	[J/(K·mol)]	-47.9	-19.8	
K_b	[1/M]	14	38	
ΔG_b	[kJ/mol]	-6.6	-9.1	
ΔH_b	[kJ/mol]	-40.6	-36.0	
ΔS_b	[J/(K·mol)]	-113.2	-89.6	

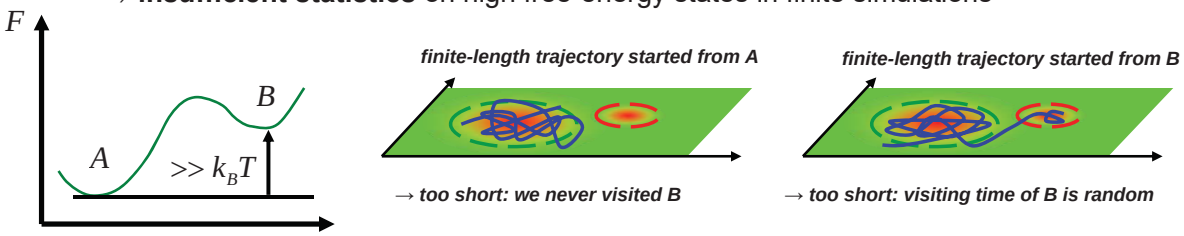
$$\Delta G_b = \Delta H_b - T\Delta S_b = -RT \ln K_b$$

Hünenberger et al., JACS 119 (1997) 7533-7544

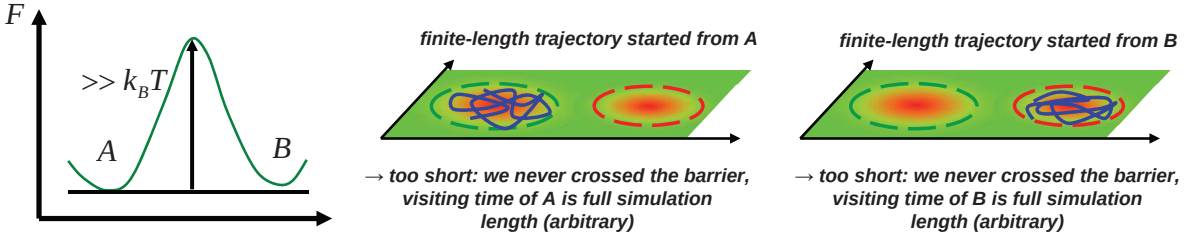
Direct counting (DC)

- In practice, direct counting **almost never works** based on **finite trajectories**, for two reasons (often present simultaneously)

→ One state may have a much lower free energy than the other
⇒ **insufficient statistics** on high free-energy states in finite simulations



→ The two states are of comparable free energies but separated by a high barrier
⇒ **insufficient number of transitions** between the states in finite simulations



- In other words, the challenge in free-energy calculations is to design schemes that

→ Enforce **good statistics** on **all relevant states** irrespective of their relative free energies

→ Enforce a sufficient **number of transitions** between these states irrespective of the barriers

Overview of free-energy methods

- The available methods to calculate free-energy changes can be classified as historically applied to the different types of changes

→ But remember that extended- or reduced-system approaches allow to interconvert the problem types, so that all methods are in principle applicable to all problem types !

→ Still, some combinations are more **practical** or/and **efficient** than others...

Thermodynamic	*Conformational*	*Alchemical*
Temperature integration	⇨ **Direct counting (DC)**	**Thermodynamic integration (TI)**
Pressure integration	➡ **Umbrella sampling (US)**	**Free-energy perturbation (FEP)**
Grand-canonical integration		**λ-dynamics local elevation umbrella sampling (λ-LEUS)**
		Enveloping distribution sampling (EDS)
		Fast growth (FG)

$$\Delta F_{AB} = -\beta^{-1} \ln \frac{\left\langle f_B(\boldsymbol{r}) \exp\left(+\beta \mathcal{U}_b(\boldsymbol{r})\right)\right\rangle_b}{\left\langle f_A(\boldsymbol{r}) \exp\left(+\beta \mathcal{U}_b(\boldsymbol{r})\right)\right\rangle_b}$$

Umbrella sampling (US)

- The idea of **umbrella sampling** (US) is to perform the MD simulation with a biased Hamiltonian

$$\mathcal{H}_b(\boldsymbol{r},\boldsymbol{p}) = \mathcal{H}(\boldsymbol{r},\boldsymbol{p}) + \mathcal{U}_b(\boldsymbol{r}) \qquad \mathcal{U}_b(\boldsymbol{r}) \quad \textit{biasing potential}$$

→ The biasing potential should be designed in such a way that the biased trajectories sample **all the relevant states** with a **sufficient number of interconversion transitions**

→ In the physical ensemble (not simulated !), direct counting would give

$$\Delta F_{AB} = -\beta^{-1} \ln \frac{\langle f_B(\boldsymbol{r}) \rangle}{\langle f_A(\boldsymbol{r}) \rangle}$$

$$\langle f_X(\boldsymbol{r}) \rangle = \iint d\boldsymbol{r}\, d\boldsymbol{p}\, P(\boldsymbol{r},\boldsymbol{p})\, f_X(\boldsymbol{r})$$

$$P(\boldsymbol{r},\boldsymbol{p}) = \frac{\exp\left(-\beta \mathcal{H}(\boldsymbol{r},\boldsymbol{p})\right)}{\iint d\boldsymbol{r}\, d\boldsymbol{p} \exp\left(-\beta \mathcal{H}(\boldsymbol{r},\boldsymbol{p})\right)}$$

→ In the biased ensemble, this translates to

$$\Delta F_{AB} = -\beta^{-1} \ln \frac{\left\langle f_B(\boldsymbol{r}) \exp\left(+\beta \mathcal{U}_b(\boldsymbol{r})\right) \right\rangle_b}{\left\langle f_A(\boldsymbol{r}) \exp\left(+\beta \mathcal{U}_b(\boldsymbol{r})\right) \right\rangle_b}$$

$$\langle f_X(\boldsymbol{r}) \rangle_b = \iint d\boldsymbol{r}\, d\boldsymbol{p}\, P_b(\boldsymbol{r},\boldsymbol{p})\, f_X(\boldsymbol{r})$$

$$P_b(\boldsymbol{r},\boldsymbol{p}) = \frac{\exp\left(-\beta\left(\mathcal{H}(\boldsymbol{r},\boldsymbol{p}) + \mathcal{U}_b(\boldsymbol{r},\boldsymbol{p})\right)\right)}{\iint d\boldsymbol{r}\, d\boldsymbol{p} \exp\left(-\beta\left(\mathcal{H}(\boldsymbol{r},\boldsymbol{p}) + \mathcal{U}_b(\boldsymbol{r},\boldsymbol{p})\right)\right)}$$

- Main difficulty: how to **design** a biasing potential with the desired properties !

Umbrella sampling (US)

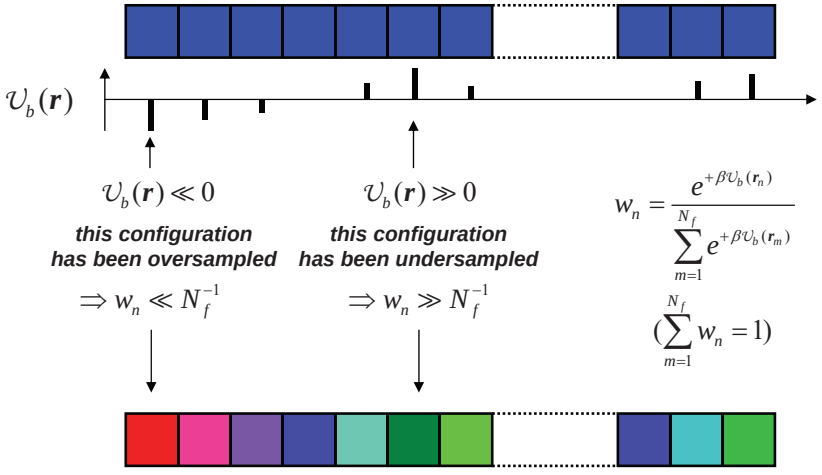
- Interpretation of the reweighting

→ Trajectory in the **original ensemble** (physical, unbiased) ***Boltzmann sampling: e.g. MD+thermostat***

all frames have equal weight in the ensemble average

$$w_n = N_f^{-1}$$

→ Trajectory in the **biased ensemble** (unphysical, biased)



sampling has been biased, ensemble averages are incorrect for the physical ensemble

$$w_n = \frac{e^{+\beta \mathcal{U}_b(\boldsymbol{r}_n)}}{\sum_{m=1}^{N_f} e^{+\beta \mathcal{U}_b(\boldsymbol{r}_m)}}$$

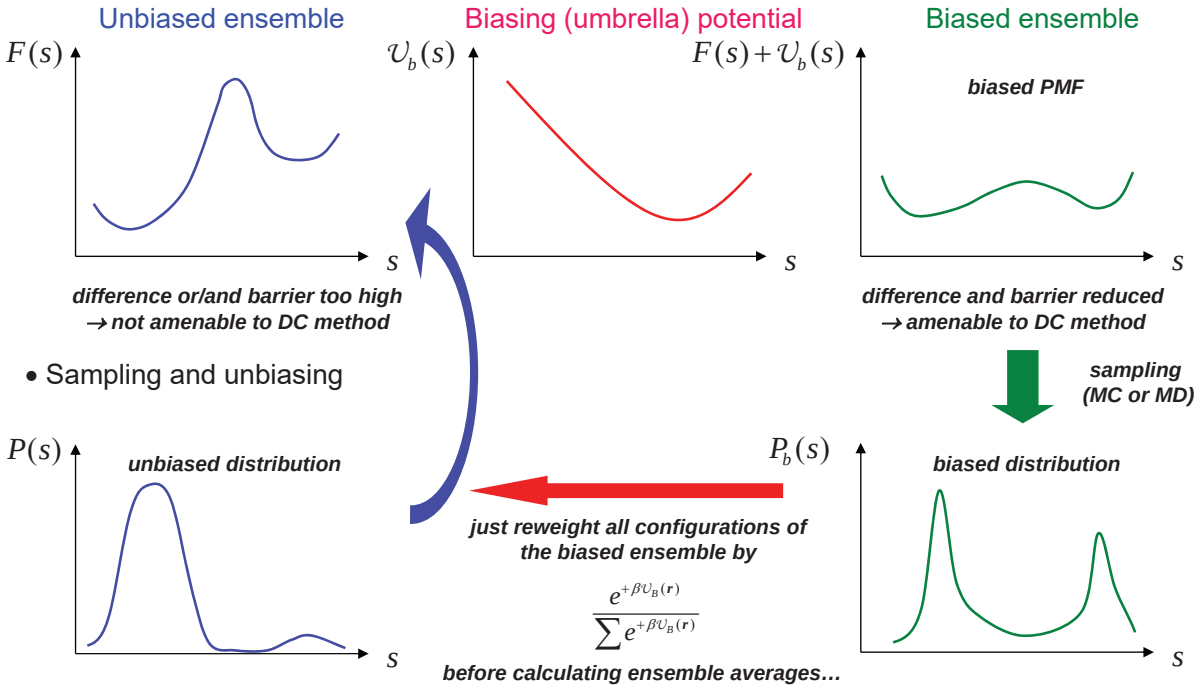
$$\left(\sum_{m=1}^{N_f} w_n = 1\right)$$

Reweighting

bias has been removed !

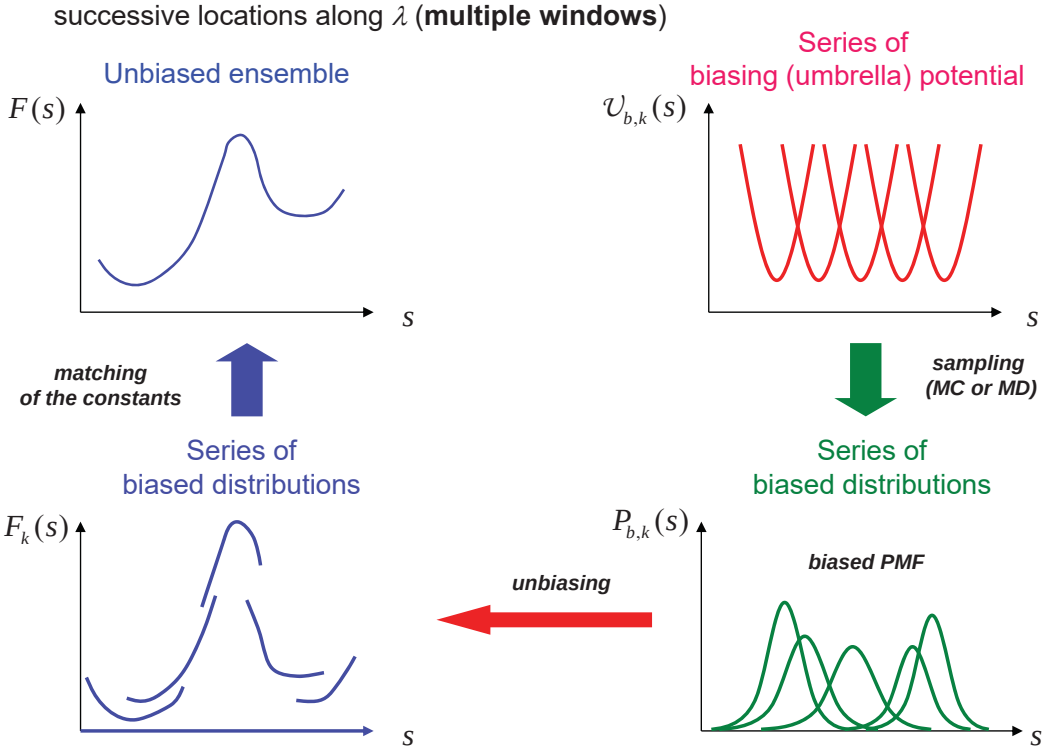
Umbrella sampling (US)

- With reaction coordinate



Umbrella sampling (US)

- The choice of a biasing potential in US is not obvious. One solution is to use a series of simulations with a local biasing potential (*e.g.* harmonic) that is placed at different successive locations along λ (**multiple windows**)



Umbrella sampling (US)

• Beyond the multiple-window scheme: **generalized umbrellas**

A (Torrie&Valleau) umbrella

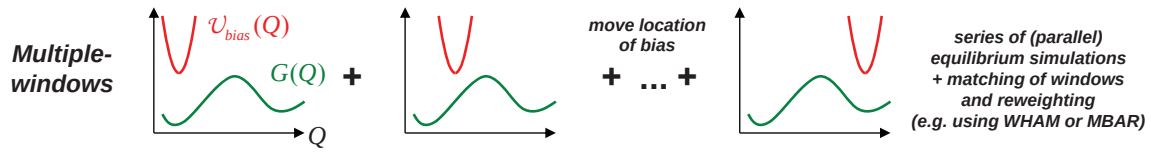
Robert Doisneau
"Un Musicien Sous La Pluie"
Paris - 1957

A (generalized) umbrella

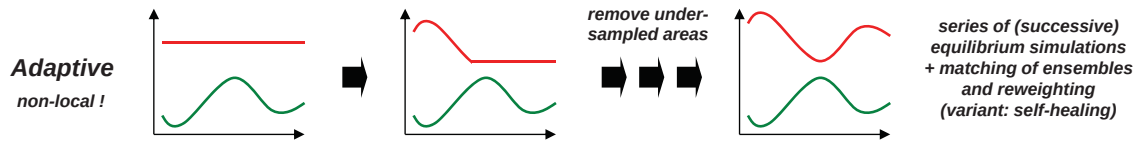
Umbrella sampling (US)

• The multiple windows approach is tedious in terms of analysis. It has now been superseded by other (iterative, thus more automatable) methods; these lead to **optimal biasing potentials**

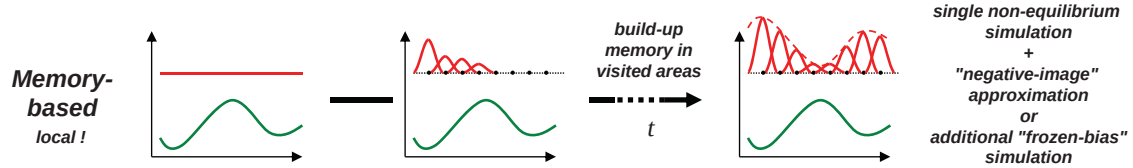
→ **Window-based** methods



→ In **adaptive** methods, successive biasing potential are produced **globally**



→ In **memory-based** methods, successive biasing potential are produced **locally**



Local elevation umbrella sampling (LEUS)

• The *basic idea* underlying memory-based US is known under many names

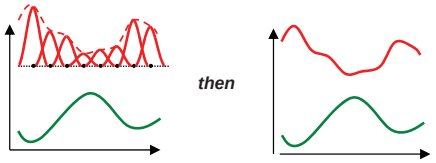
***deflation* (1969)**	***conformational flooding* (1995)**	***metadynamics* (2002)**	***gaussian-mixture US* (2009)**
***tunneling* (1985)**	***Engkvist-Karlström* (1996)**	***filling potential* (2003)**	***basin paving* (2010)**
***tabu search* (1989)**	***Wang-Landau* (2001)**	***adaptive reaction coordinate force* (2009)**	***LEUS* (2010)**
***local elevation* (1994)**	***adaptive biasing force* (2001)**		

→ but: implementation choices may affect a lot the *applicability* and *accuracy* in practice !

• Local elevation umbrella sampling (LEUS)

(our favorite flavor of this principle)
Hansen & Hünenberger
J. Comput. Chem., 31, 1 (2010).

Two-steps implementation



then

duration t_{LE}

LE BUILD-UP PHASE
non-equilibrium
→ rough biasing potential

$$G(Q) \approx -\mathcal{U}_{bias}(Q)$$

→ not very accurate or requires slow build-up

duration t_{US}

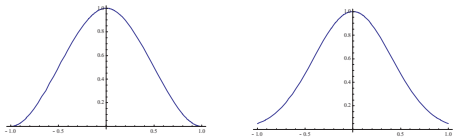
US SAMPLING PHASE
frozen biasing potential

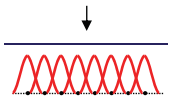
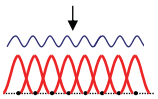
reweighting
→ "irons out" the roughness of the biasing potential

→ results less sensitive to build-up protocol

→ systematic error reduction upon increasing t_{US}

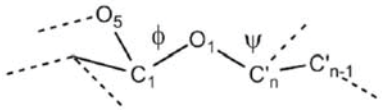
Truncated polynomial basis functions



	TRUNCATED POLYNOMIAL	***A COMPARABLE GAUSSIAN***
	$f(x) = (1-3x^2+2\lvert x^3\rvert)h(\lvert x\rvert-1)$	$f(x) = \exp(-3x^2)$
computation	***cheap***	***expensive***
range	***finite (next grid point)***	***formally infinite***
continuity	***yes (+derivative)***	***formally no (if cutoff)***
„wiggling"	***no***	***yes***
even better: spline of order 2		 ***increases with build-up magnitude !***

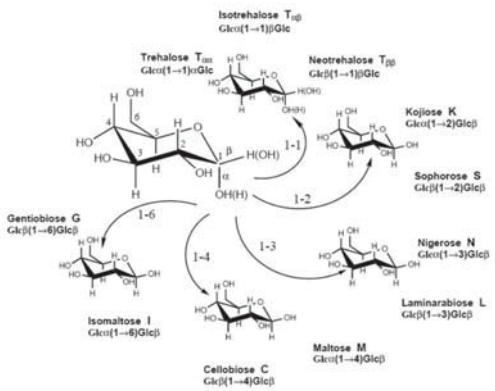
LEUS: Glucose-based disaccharides in water

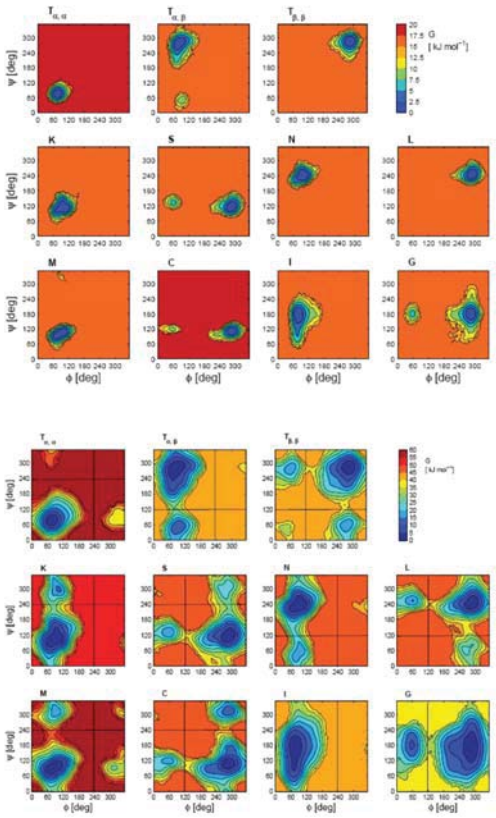
50 ns Plain MD
(initiated from X-ray structure)



2D subspace

(rotation timescales ~ 10 ns – 1 μs)



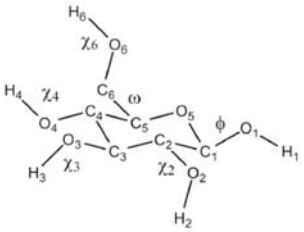


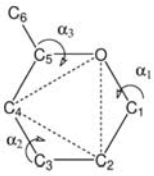
50+50 ns LEUS

Perić-Hassler, Hansen, Baron & Hünenberger
Carbohydr. Res. 345, 1781 (2010)

LEUS: Glucose in water

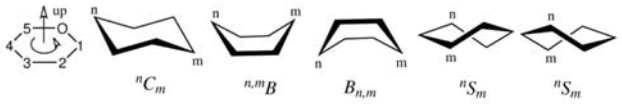
Hansen & Hünenberger
J. Comput. Chem. 31, 1 (2010)
[use for desing of new carbohydrate force field GROMOS 56A$_{CARBO}$:
J. Comput. Chem. 32, 998 (2011)]



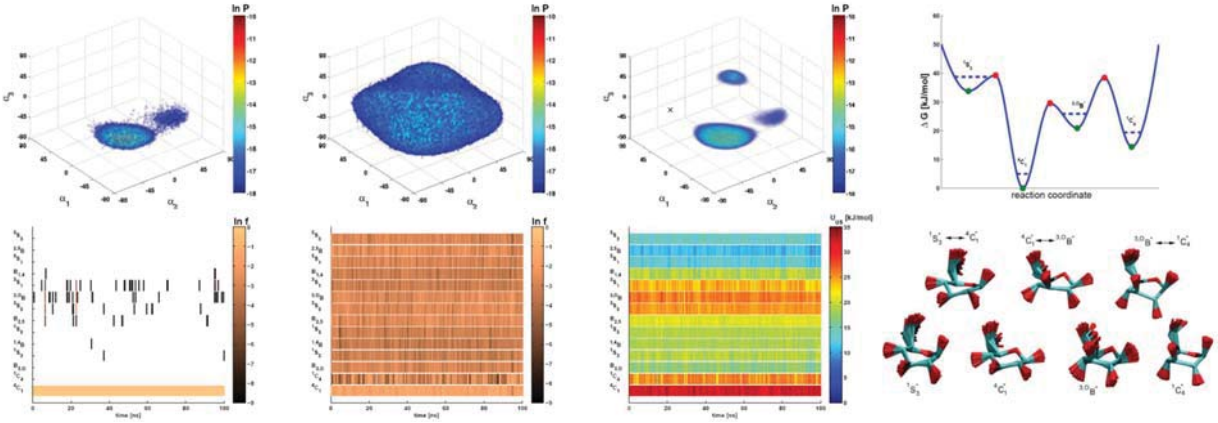


Pickett & Strauss ring puckering coordinates

3D subspace



(pseudorotation timescales ~ 50 ns – 1 μs)



100 ns Plain MD
(initiated from $^4C_1$)

biased

after unbiasing

analysis of pathways and barriers

50+100 ns LEUS

Overview of free-energy methods

- The available methods to calculate free-energy changes can be classified as historically applied to the different types of changes

  → But remember that extended- or reduced-system approaches allow to interconvert the problem types, so that all methods are in principle applicable to all problem types !

  → Still, some combinations are more **practical** or/and **efficient** than others...

Thermodynamic	*Conformational*	*Alchemical*
Temperature integration	**Direct counting (DC)**	**Thermodynamic integration (TI)**
Pressure integration	**Umbrella sampling (US)**	**Free-energy perturbation (FEP)**
Grand-canonical integration		**λ-dynamics local elevation umbrella sampling (λ-LEUS)**
		Enveloping distribution sampling (EDS)
		Fast growth (FG)

$$\frac{d(T^{-1}F(T))}{d(T^{-1})} = \left\langle \mathcal{H}(\mathbf{r},\mathbf{p}) \right\rangle_T$$

Temperature integration

- **Temperature integration** aims at calculating the **free energy change** upon **changing the temperature** (a thermodynamic change)

→ Principle

$$F(T) = -\beta^{-1} \ln Z = -\beta^{-1} \ln\left[\xi \iint d\boldsymbol{r}\, d\boldsymbol{p} \exp\left(-\beta \mathcal{H}(\boldsymbol{r}, \boldsymbol{p})\right)\right] \qquad \beta = (k_B T)^{-1}$$

$$\frac{d(T^{-1}F(T))}{d(T^{-1})} = \frac{d(\beta F(T))}{d\beta} = -\frac{d \ln Z}{d\beta} = -Z^{-1}\frac{dZ}{d\beta}$$

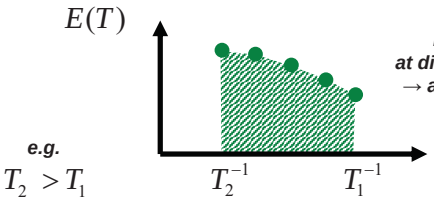
$$= -Z^{-1}\xi \iint d\boldsymbol{r}\, d\boldsymbol{p} \frac{\partial \exp\left(-\beta \mathcal{H}(\boldsymbol{r}, \boldsymbol{p})\right)}{\partial \beta}$$

$$= Z^{-1}\xi \iint d\boldsymbol{r}\, d\boldsymbol{p}\, \mathcal{H}(\boldsymbol{r}, \boldsymbol{p}) \exp\left(-\beta \mathcal{H}(\boldsymbol{r}, \boldsymbol{p})\right)$$

$$= \frac{\iint d\boldsymbol{r}\, d\boldsymbol{p}\, \mathcal{H}(\boldsymbol{r}, \boldsymbol{p}) \exp\left(-\beta \mathcal{H}(\boldsymbol{r}, \boldsymbol{p})\right)}{\iint d\boldsymbol{r}\, d\boldsymbol{p} \exp\left(-\beta \mathcal{H}(\boldsymbol{r}, \boldsymbol{p})\right)} = \left\langle \mathcal{H}(\boldsymbol{r}, \boldsymbol{p}) \right\rangle_T = E(T)$$

→ This is called the Gibbs-Helmholtz equation and you can also derive it easily based on the laws of thermodynamics

→ Procedure



NVT simulations at different temperatures → average total energy

e.g. $T_2 > T_1$

numerical integration (quadrature)

$$T_2^{-1}F(T_2) = T_1^{-1}F(T_1) + \int_{T_1^{-1}}^{T_2^{-1}} dT^{-1} E(T)$$

Note: E is defined within an arbitrary constant; but it is easily checked that E→E+C leads to F→F+C

Overview of free-energy methods

- The available methods to calculate free-energy changes can be classified as historically applied to the different types of changes

→ But remember that extended- or reduced-system approaches allow to interconvert the problem types, so that all methods are in principle applicable to all problem types !

→ Still, some combinations are more **practical** or/and **efficient** than others...

Thermodynamic	*Conformational*	*Alchemical*
Temperature integration	**Direct counting (DC)**	**Thermodynamic integration (TI)**
Pressure integration	**Umbrella sampling (US)**	**Free-energy perturbation (FEP)**
Grand-canonical integration		**λ-dynamics local elevation umbrella sampling (λ-LEUS)**
		Enveloping distribution sampling (EDS)
		Fast growth (FG)

$$\frac{dF(V)}{dV} = -\left\langle \mathcal{P}(\boldsymbol{r}, \boldsymbol{p}) \right\rangle_V$$

Pressure integration

- **Pressure integration** aims at calculating the **free energy change** upon **changing the pressure** (a thermodynamic change)

  → Principle

→ Actually, we should probably rather call it volume integration

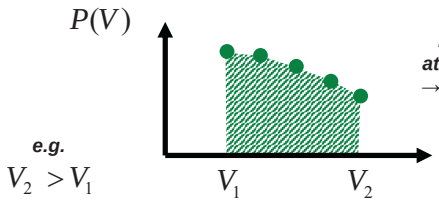
$$\frac{dF(V)}{dV} = -P(V) = -\left\langle \mathcal{P}(\boldsymbol{r}, \boldsymbol{p}) \right\rangle_V$$

→ This one is easily derived based on the laws of thermodynamics

(stat mech derivation is possible, but more tricky !)

$$dF = -PdV - SdT$$

→ Procedure



e.g. $V_2 > V_1$

NVT simulations at different volumes → average pressure

numerical integration (quadrature)

$$F(V_2) = F(V_1) - \int_{V_1}^{V_2} dV\ P(V)$$

Overview of free-energy methods

- The available methods to calculate free-energy changes can be classified as historically applied to the different types of changes

  → But remember that extended- or reduced-system approaches allow to interconvert the problem types, so that all methods are in principle applicable to all problem types !

  → Still, some combinations are more **practical** or/and **efficient** than others...

Thermodynamic	*Conformational*	*Alchemical*
Temperature integration	**Direct counting (DC)**	**Thermodynamic integration (TI)**
Pressure integration	**Umbrella sampling (US)**	**Free-energy perturbation (FEP)**
Grand-canonical integration		**λ-dynamics local elevation umbrella sampling (λ-LEUS)**
		Enveloping distribution sampling (EDS)
		Fast growth (FG)

$$\Delta F_{AB} = \int_{\lambda_A}^{\lambda_B} d\lambda' \left\langle \frac{\partial \mathcal{H}(\boldsymbol{r}, \boldsymbol{p}; \lambda)}{\partial \lambda} \right\rangle_{\lambda'}$$

Thermodynamic integration (TI)

- **Thermodynamic integration** aims at calculating the **free energy change** upon **changing the topology** of a molecule (an alchemical change)
- It requires the introduction of a Hamiltonian **coupling parameter** λ

$$\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)=\begin{cases}\mathcal{H}_A(\boldsymbol{r},\boldsymbol{p}) & \text{if } \lambda=0\\ \mathcal{H}_B(\boldsymbol{r},\boldsymbol{p}) & \text{if } \lambda=1\end{cases}$$

this condition is actually compatible with many alternative coupling schemes !

specific λ-dependence determines pathway from A to B

→ Example: simple linear coupling scheme (not necessarily the best choice!)

$$s(\lambda)=(1-\lambda)s_A+\lambda s_B$$

where "s" is any force-field parameter

mass, charge, pairwise LJ coefficient, covalent reference value and force constant, ...

mass
$$m_i(\lambda)=\left[1-\lambda\right]m_i{}^A+\lambda m_i{}^B$$

bond stretching
$$V(b;\lambda)=\frac{1}{4}\left\{\left[1-\lambda\right]K_b{}^A+\lambda K_b{}^B\right\}\left\{b^2-\left(\left[1-\lambda\right]b_0{}^A+\lambda b_0{}^B\right)^2\right\}^2$$

bond-angle bending
$$V(\theta;\lambda)=\frac{1}{2}\left\{\left[1-\lambda\right]K_\theta{}^A+\lambda K_\theta{}^B\right\}\left\{\cos\theta-\left(\left[1-\lambda\right]\cos\theta_0{}^A+\lambda\cos\theta_0{}^B\right)\right\}^2$$

improper-dihedral distortion
$$V(\xi;\lambda)=\frac{1}{2}\left\{\left[1-\lambda\right]K_\xi{}^A+\lambda K_\xi{}^B\right\}\left\{\xi-\left(\left[1-\lambda\right]\xi_0{}^A+\lambda\xi_0{}^B\right)\right\}^2$$

proper-dihedral torsion
$$V(\varphi;\lambda)=\left[1-\lambda\right]K_\varphi{}^A\left[1+\cos\delta^A\cos m^A\varphi\right]+\lambda K_\varphi{}^B\left[1+\cos\delta^B\cos m^B\varphi\right]$$

non-bonded interactions
$$V(r_{ij};\lambda)=\left[1-\lambda\right]^n\left\{\frac{C_{12}^A}{\left[\alpha_{LJ}\lambda^2+r_{ij}{}^6\right]^2}-\frac{C_6^A}{\left[\alpha_{LJ}\lambda^2+r_{ij}{}^6\right]}+\frac{1}{4\pi\varepsilon_0\varepsilon_1}\frac{q_i^Aq_j^A}{\left[\alpha_c\lambda^2+r_{ij}{}^2\right]^{1/2}}\right\}$$
$$+\lambda^n\left\{\frac{C_{12}^B}{\left[\alpha_{LJ}\left[1-\lambda\right]^2+r_{ij}{}^6\right]^2}-\frac{C_6^B}{\left[\alpha_{LJ}\left[1-\lambda\right]^2+r_{ij}{}^6\right]}+\frac{1}{4\pi\varepsilon_0\varepsilon_1}\frac{q_i^Bq_j^B}{\left[\alpha_c\left[1-\lambda\right]^2+r_{ij}{}^2\right]^{1/2}}\right\}$$

+ RF terms

Note: n≠1, α parameters → non-linear coupling

Thermodynamic integration (TI)

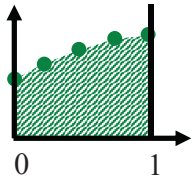
- The **thermodynamic integration formula** is as follows

$$F(\lambda)=-\beta^{-1}\ln Z=-\beta^{-1}\ln\left[\xi\iint d\boldsymbol{r}\,d\boldsymbol{p}\exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)\right)\right]$$

$$F'(\lambda)=\frac{dF(\lambda)}{d\lambda}=-\beta^{-1}\frac{d\ln Z}{d\lambda}=-Z^{-1}\beta^{-1}\frac{dZ}{d\lambda}$$
$$=-Z^{-1}\beta^{-1}\xi\iint d\boldsymbol{r}\,d\boldsymbol{p}\frac{\partial\exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)\right)}{\partial\lambda}$$
$$=Z^{-1}\xi\iint d\boldsymbol{r}\,d\boldsymbol{p}\frac{\partial\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)}{\partial\lambda}\exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)\right)$$
$$=\frac{\iint d\boldsymbol{r}\,d\boldsymbol{p}\frac{\partial\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)}{\partial\lambda}\exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)\right)}{\iint d\boldsymbol{r}\,d\boldsymbol{p}\exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)\right)}=\left\langle\frac{\partial\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)}{\partial\lambda}\right\rangle_\lambda$$

→ Procedure

$$F'(\lambda)=\left\langle\frac{\partial\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)}{\partial\lambda}\right\rangle_\lambda$$



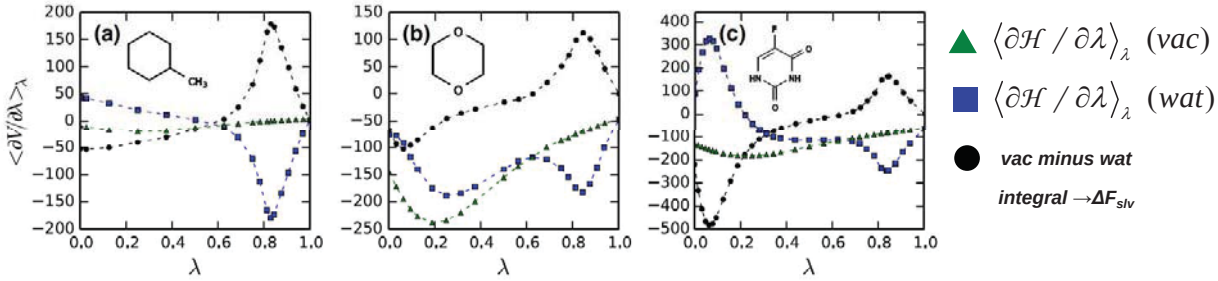
NVT simulations at different λ values → average Hamiltonian derivative

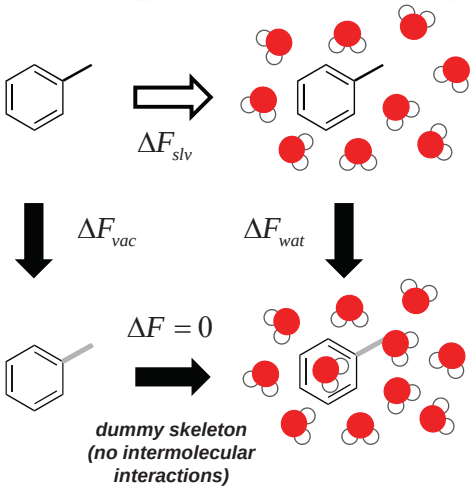
numerical integration (quadrature)

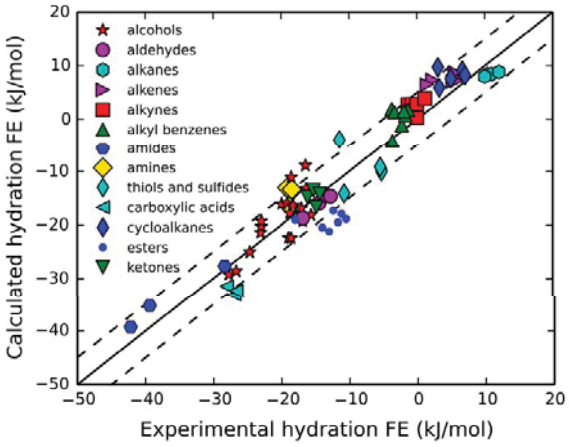
$$F(1)=F(0)+\int_0^1 d\lambda\,F'(\lambda)$$

Thermodynamic integration (TI)

- Example: **solvation free energy** of small organic molecules





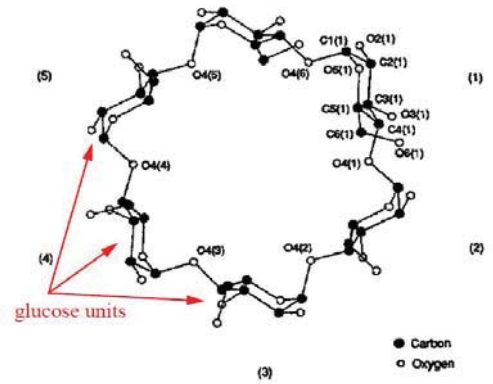


J. Comput.-Aided Mol. Des. **28**, 221 (2014).

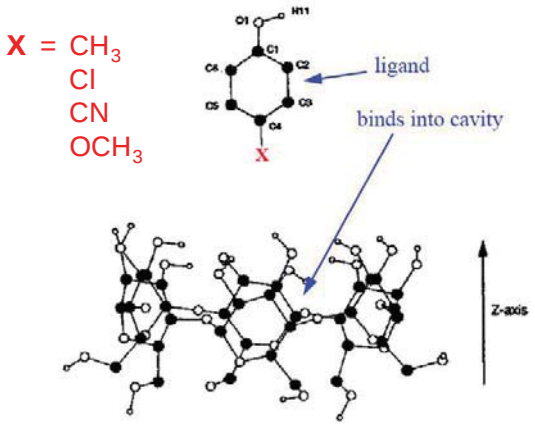
Thermodynamic integration (TI)

- Example: **relative binding free energies** of p-substituted phenols to cyclodextrin

  → Small host-guest system: α-cyclodextrin consists of 6 sugar (glucose) units (cyclic)



α-cyclodextrin



p-methoxyphenol showing the orientation of the guest when inserted.

→ Simulations of one α-cyclodextrin + 500 H_2O, NPT ensemble

→ 3x2 mutations (from OCH_3 to the 3 others, in free and bound forms)

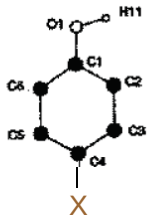
→ TI with ~20 lambda points per mutation

Mark et al., JACS ***116*** *(1994) 6293-6302*

Thermodynamic integration (TI)

- Example: **relative binding free energies** of p-substituted phenols to cyclodextrin

→ Results



Note: CH_3 and Cl compounds have an attached dummy !

Binding parameters:

Guest X=	MD Enthalpy ΔH(~ΔE)	MD Free energy ΔG=ΔH-TΔS	Experiment kJmol^{-1} ΔG
-O-CH_3	0	0	0
-C≡N	-9.9	0.8	-2.8
-Cl	2.5	-2.6	-3.8
-CH_3	13.7	4.2	0.1

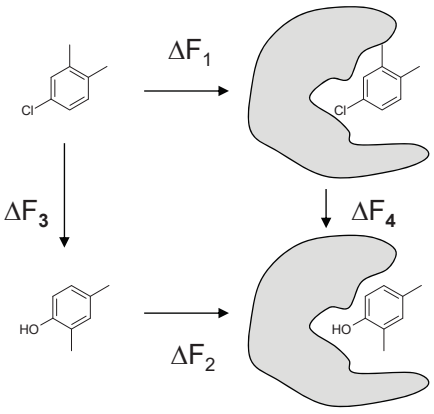
→ The relative binding free energies cannot be interpreted in terms of enthalpy only: **entropic effects** (solute and solvent) plays a **significant role** here!

*Mark et al., JACS **116** (1994) 6293-6302*

Thermodynamic integration (TI)

- Example: **relative binding free energies** of multiple ligands to a protein

→ We use a thermodynamic cycle



$$\Delta\Delta F_{binding} = \Delta F_2 - \Delta F_1 = \Delta F_4 - \Delta F_3$$

→ Bound state must be meta-stable

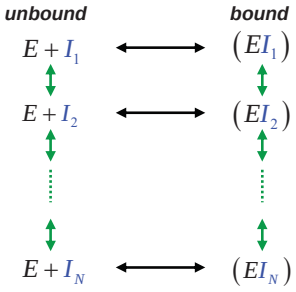
If the ligand tends to escape, you can add restraints – but then you have to correct for them later !

→ We need many simulations... N inhibitors, M λ-points

$$F_2 - F_1 = \sum_{j=1}^{M} \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda_j} \Delta\lambda_j$$

$2\,M \quad N$ simulations

$10 \quad 10 \quad \underline{200}$



Thermodynamic integration (TI)

- The **thermodynamic integration formula**

  → is **accurate** if *sufficient sampling* <…>
  and *sufficient number of λ-points* λ_i

  Important:
  accuracy can be systematically improved
  by sampling longer / adding more λ-point

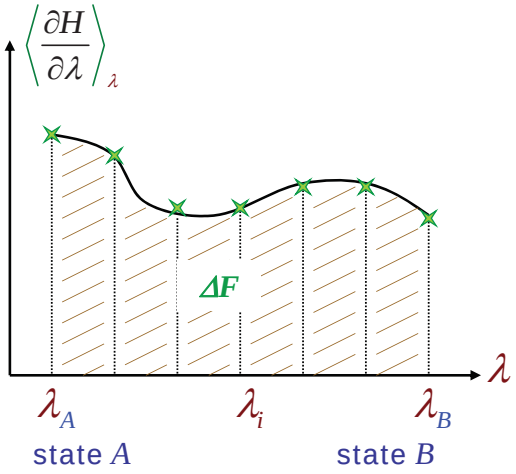
$$\Delta F = F_B - F_A = \int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial H(\vec{p},\vec{r};\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda$$

  → is **time consuming** in the sense that many (10-100) simulations are required for *each leg* and there are as many legs as there are *pairs of states* A and B considered

  and in particular:

$$\lambda_i \neq \lambda_A \text{ or } \lambda_B$$

  are unphysical, so irrelevant for anything else but the free-energy calculation



  → is **robust** – the "workhorse" of free-energy calculations

- Note that

  - every point needs some equilibration time (discarded)

  - orthogonal barriers may be difficult to overcome at fixed lambda

Thermodynamic integration (TI)

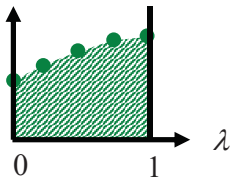
TO BE ADDED TO FREE-ENERGY I

- **Slow growth** (SG) is an "old-fashioned" way to do TI

  → Standard TI → set of independent equilibrium simulations at different λ-points

$$\left\langle \frac{\partial \mathcal{H}(\mathbf{r},\mathbf{p};\lambda)}{\partial \lambda} \right\rangle_\lambda$$

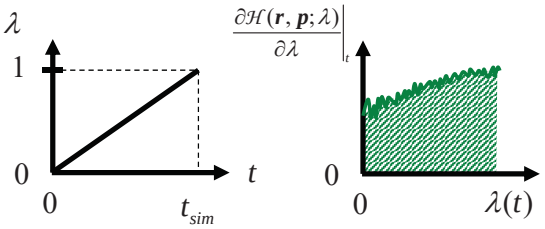
NVT simulations at different λ values → average Hamiltonian derivative



numerical integration (quadrature)

$$\Delta F_{AB} = \int_0^1 d\lambda' \left\langle \frac{\partial \mathcal{H}(\mathbf{r},\mathbf{p};\lambda)}{\partial \lambda} \right\rangle_{\lambda'}$$

  → SG → slowly "sweep" the λ-variable in a single simulation



N timesteps n=1..N

$$\lambda(t_n) = \frac{n}{N}$$

$$\Delta F_{AB} \approx \frac{1}{N} \sum_{n=1}^{N} \left. \frac{\partial \mathcal{H}(\mathbf{r},\mathbf{p};\lambda)}{\partial \lambda} \right|_{t_n}$$

- **Advantage** of SG: single simulation!

- **Shortcomings** of SG – which is no longer so much used...

  → Time-dependent Hamiltonian, system is formally never at equilibrium

  → Sampling can only be improved by redoing the full simulation in a longer time t_{sim}
  (≠ TI, where you just add more time for [all or specific] λ-points)

Overview of free-energy methods

- The available methods to calculate free-energy changes can be classified as historically applied to the different types of changes

  → But remember that extended- or reduced-system approaches allow to interconvert the problem types, so that all methods are in principle applicable to all problem types !

  → Still, some combinations are more **practical** or/and **efficient** than others...

Thermodynamic	*Conformational*	*Alchemical*
⇨ **Temperature integration**	⇨ **Direct counting (DC)**	⇨ **Thermodynamic integration (TI)**
⇨ **Pressure integration**	⇨ **Umbrella sampling (US)**	➡ **Free-energy perturbation (FEP)**
Grand-canonical integration		**λ-dynamics local elevation umbrella sampling (λ-LEUS)**
		Enveloping distribution sampling (EDS)
		Fast growth (FG)

$$F(\lambda+\Delta\lambda)-F(\lambda) = -\beta^{-1}\ln\left\langle\exp\left(-\beta[\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda+\Delta\lambda)-\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)]\right)\right\rangle_\lambda$$

Free energy perturbation (FEP)

- The **free energy perturbation** formula is as follows

$$F(\lambda) = -\beta^{-1}\ln Z = -\beta^{-1}\ln\left[\xi\iint d\boldsymbol{r}\,d\boldsymbol{p}\exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)\right)\right]$$

➡ $$F(\lambda+\Delta\lambda)-F(\lambda) = -\beta^{-1}\ln\left[\frac{\xi\iint d\boldsymbol{r}\,d\boldsymbol{p}\exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda+\Delta\lambda)\right)}{\xi\iint d\boldsymbol{r}\,d\boldsymbol{p}\exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)\right)}\right]$$

$$= -\beta^{-1}\ln\left[\frac{\iint d\boldsymbol{r}\,d\boldsymbol{p}\exp\left(-\beta[\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda+\Delta\lambda)-\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)]\right)\exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)\right)}{\iint d\boldsymbol{r}\,d\boldsymbol{p}\exp\left(-\beta\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)\right)}\right]$$

$$= -\beta^{-1}\ln\left\langle\exp\left(-\beta[\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda+\Delta\lambda)-\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)]\right)\right\rangle_\lambda$$

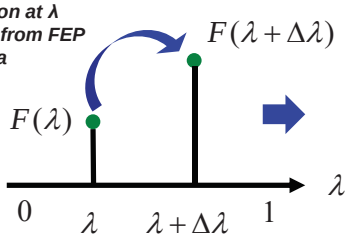
Note: the TI formula is recovered by evaluating

$$F'(\lambda) = \lim_{\Delta\lambda\to 0}\frac{F(\lambda+\Delta\lambda)-F(\lambda)}{\Delta\lambda}$$

So, TI is the linearization of FEP (omitting higher-order derivatived of F)

→ Procedure

NVT simulation at λ → calculate ΔF from FEP formula



$$F(\lambda+\Delta\lambda) = F(\lambda)-\beta^{-1}\ln\left\langle\exp\left(-\beta[\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda+\Delta\lambda)-\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)]\right)\right\rangle_\lambda$$

Free energy perturbation (FEP)

- The **daring** implementation of the FEP is called **one-step perturbation** (OSP)

→ Procedure

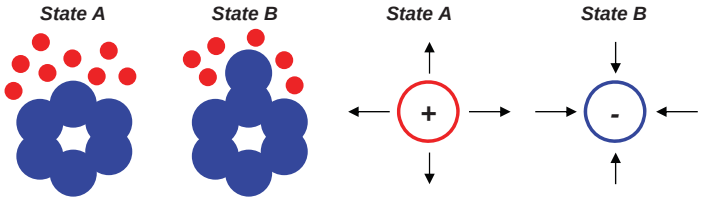
Set Δλ to one !

→ Huge advantages over TI: (1) we can calculate the free energy of **multiple end states** from a simulation at a **single reference state**; (2) the calculation is performed in one go, *i.e.* **without any intermediate step**

→ But: we may have a serious problem in terms or **reliability** and **accuracy** !

$$-\beta^{-1}\ln\left\langle\exp\left(-\beta[\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda+\Delta\lambda)-\mathcal{H}(\boldsymbol{r},\boldsymbol{p};\lambda)]\right)\right\rangle_{\lambda}$$

Example

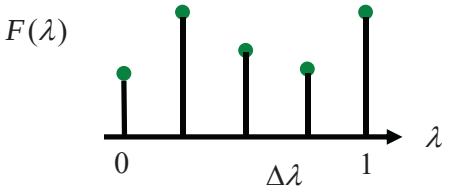


State A contains very few configurations with low energy in state B → at finite sampling time, the least bad will dominate the ensemble average and the result will be crap !

→ **Tentative remedy**: choose the **reference state** very carefully !

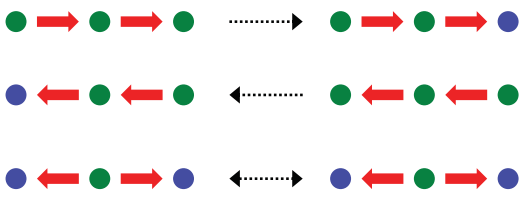
Free energy perturbation (FEP)

- The **conservative** implementation of the FEP is called **multi-configuration free energy perturbation** (MCFEP)



Do many intermediate steps with a small spacing Δλ (K intervals, K+1 points)

→ Forward, backward or double-wide



forward perturbations (K simulations)

backward perturbations (K simulations [same +1 new])

hysteresis is a measure of the error

double-wide sampling (K/2 simulations)

no hysteresis can be determined

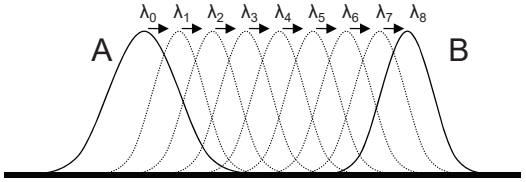
→ We gain a lot of **accuracy and reliability** !

→ But: we loose the advantage compared to TI because we **cannot freely extrapolate** to any end state based on a single simulation

Free energy perturbation (FEP)

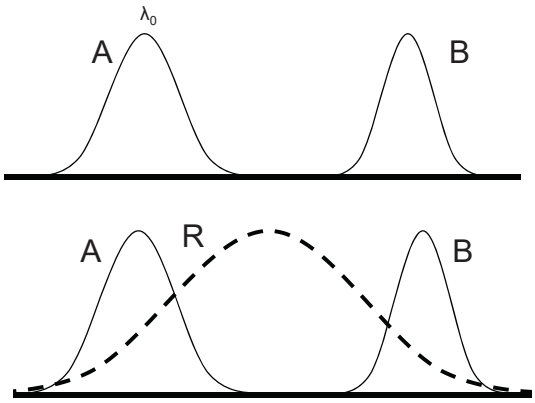
- How to use OSP efficiently?

MCFEP → works but needs *many simulations*



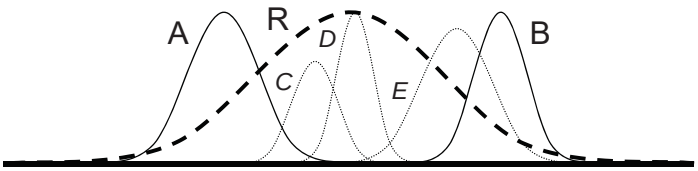
x = conformational coordinate

OSP(initial state) → won't work



OSP(smart reference state) → may work

→ In addition, if the reference state "covering enough ground", you can extrapolate to many states from a single reference simulation



Free energy perturbation

- Use of OFP to predict multiple relative binding free energies from a single (pair of) simulation

$$F_i - F_R = -k_B T \ln \left\langle \exp\left\{ -\frac{H_i - H_R}{k_B T} \right\} \right\rangle_R$$

$$E + I_R \rightarrow I_1, I_2, \ldots, I_N \qquad (EI_R) \rightarrow (EI_1), (EI_2), \ldots, (EI_N)$$

2 simulations of an ***unphysical state*** which is chosen to optimise sampling for entire set of N inhibitors

- Design of the reference state

Idea: use ***soft-core atoms*** for each site where the inhibitors possess different (or no) atoms

The reference state simulation (R) should produce an ensemble that contains low-energy configurations for all of the Hamiltonians (inhibitors)

$$H_1, H_2, \ldots, H_N$$

Free energy perturbation

- **Soft-core** non-bonded interactions

→ Functional form

van der Waals

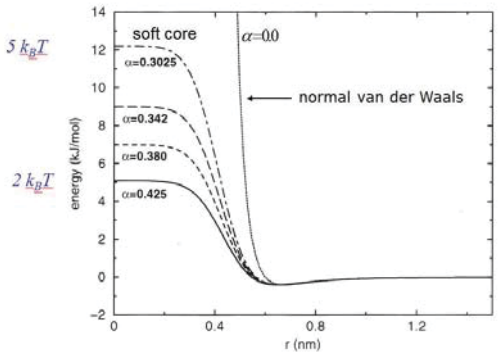
$$V(r) = 4\varepsilon\left[\frac{1}{\left[\alpha + (r/\sigma)^6\right]^2} - \frac{1}{\alpha + (r/\sigma)^6}\right]$$

electrostatics

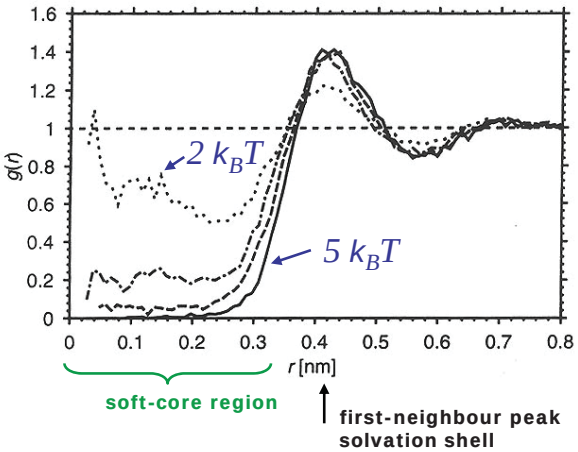
$$V(r) = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_1}\left[\frac{1}{\left[\alpha_c + (r)^2\right]^{\frac{1}{2}}} - \frac{0.5C_{rf}r^2}{\left[\alpha_c + R_{rf}^2\right]^{\frac{1}{2}}} - \frac{1-0.5C_{rf}}{R_{rf}}\right]$$

→ Effect

interaction energy



Distribution of water molecules around and in the soft-core cavity of 0.6 nm diameter



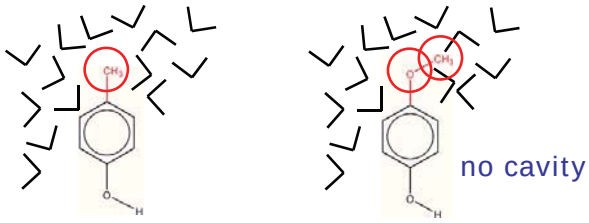
Free energy perturbation

- **Soft-core** atoms in OSP

$$F(\underset{\boldsymbol{B}}{\text{final}}) - F(\underset{\boldsymbol{A}}{\text{initial}}) = -k_B T \ln\left\{\left\langle \exp\left\{-\left[H(\underset{\boldsymbol{B}}{\text{final}}) - H(\underset{\boldsymbol{A}}{\text{initial}})/k_B T\right]\right\}\right\rangle_{\underset{\text{A}}{\text{initial}}}\right\}$$

→ Normal non-bonded interactions

initial state *A* **final state *B***



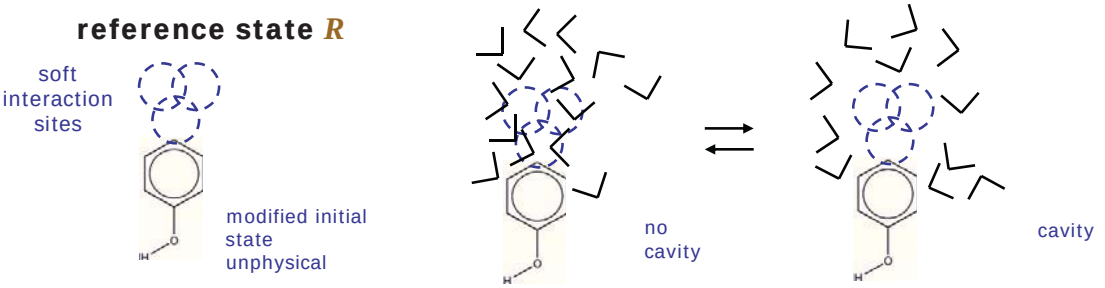
***H*(final) very large**

no contribution to *ΔF*

modified initial (reference) state ensemble contains configurations with low energy (= H) for various final state Hamiltonians

→ Soft-core non-bonded interactions

reference state *R*

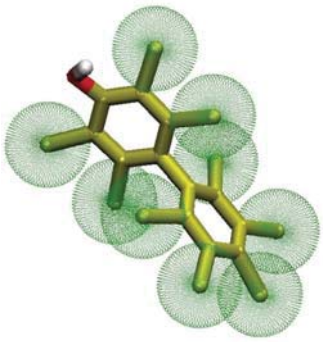


Free energy perturbation

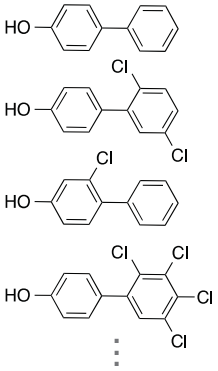
- Example: relative binding free energy of ligands to the oestrogen receptor

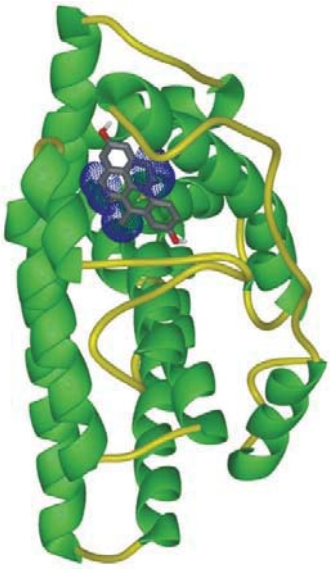
  → Relative binding free energy for **16 ligands** from **two simulations** only

Unphysical reference ligand (soft-core atoms):



Physical ligands:
16 polychlorinated biphenyls





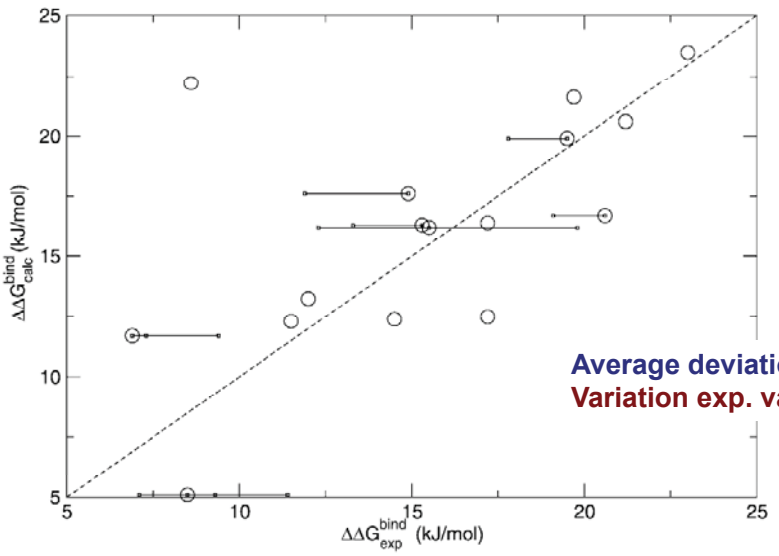
Two simulations:
- Reference ligand in water
- Reference ligand bound to protein in water

Proteins **54**, 237 (2004).

Free energy perturbation

- Example: relative binding free energy of ligands to the oestrogen receptor

  → Relative binding free energy for **16 ligands** from **two simulations** only



Average deviation: 2.5 kJ mol^{-1}
Variation exp. values: 4.2 kJ mol^{-1}

→ Warning: combinatorial problem!

N sites that can be water-covered or not → 2^N combination (for one ligand with a specific substitution, only one of these 2^N combinations will be relevant)

Proteins **54**, 237 (2004).

Overview of free-energy methods

- The available methods to calculate free-energy changes can be classified as historically applied to the different types of changes

  → But remember that extended- or reduced-system approaches allow to interconvert the problem types, so that all methods are in principle applicable to all problem types !

  → Still, some combinations are more **practical** or/and **efficient** than others...

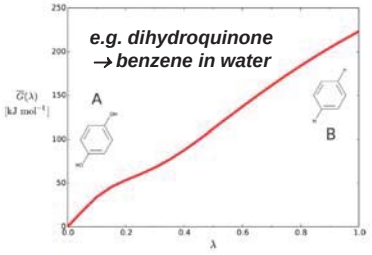
Thermodynamic	*Conformational*	*Alchemical*
⇨ **Temperature integration**	⇨ **Direct counting (DC)**	⇨ **Thermodynamic integration (TI)**
⇨ **Pressure integration**	⇨ **Umbrella sampling (US)**	⇨ **Free-energy perturbation (FEP)**
Grand-canonical integration		**λ-dynamics local elevation umbrella sampling (λ-LEUS)**
		Enveloping distribution sampling (EDS)
		Fast growth (FG)

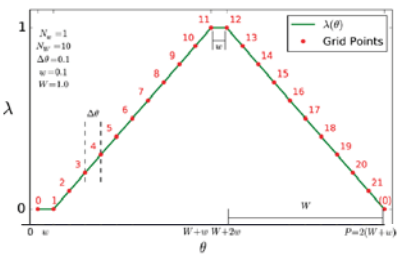
λ-dynamics local elevation umbrella sampling (λ-LEUS)

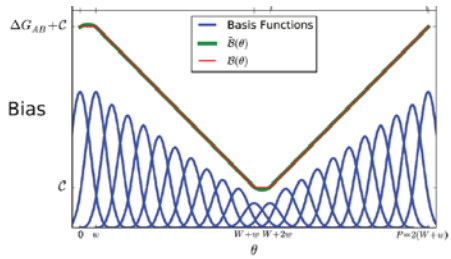
→ For an alchemical change, we use so-called **λ-dynamics** *λ becomes a dynamic variable !*

Coupling parameter $\lambda \rightarrow \lambda(t)$ $\ddot{\lambda}(t) = -m_\lambda^{-1} \left. \frac{\partial \mathcal{H}(\boldsymbol{r}, \boldsymbol{p}, \lambda)}{\partial \lambda} \right|_t$ m_λ *fictitious mass* *= λ-dynamics*

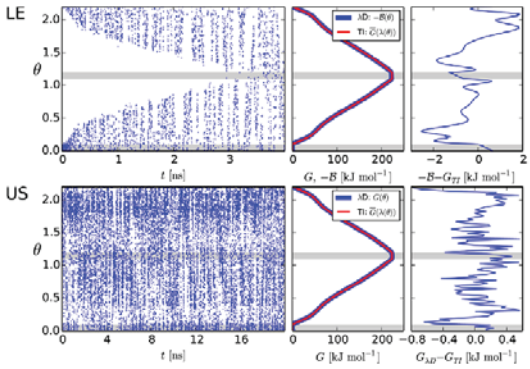
→ We apply a smart coordinate transformation λ(θ)







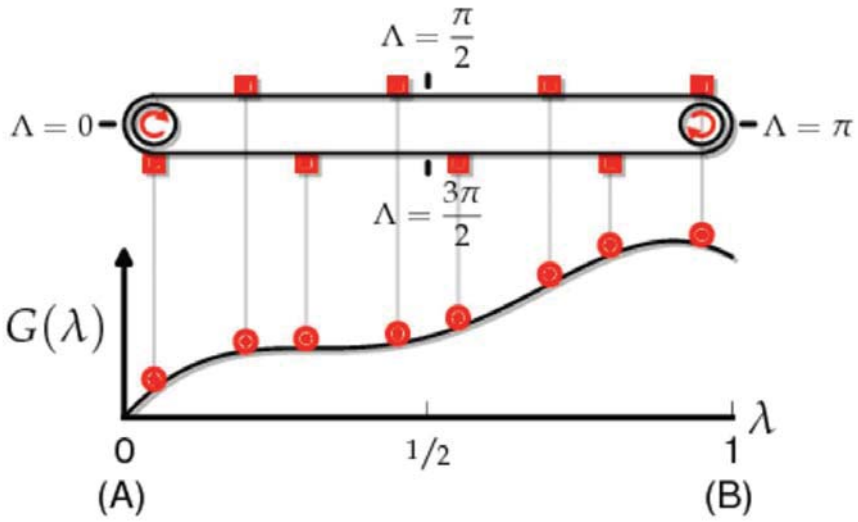
→ We apply a bias on θ using LEUS



Bieler, N.S., Häuselmann, R. & Hünenberger, P.H. J. Chem. Theory. Comput. 10 (2014) 3006

Conveyor Belt Thermodynamic Integration

→ An alternative with multiple replicas & no biasing potential!



→ See: talk of David Hahn on in three weeks (4.12)

Overview of free-energy methods

- The available methods to calculate free-energy changes can be classified as historically applied to the different types of changes

  → But remember that extended- or reduced-system approaches allow to interconvert the problem types, so that all methods are in principle applicable to all problem types !

  → Still, some combinations are more **practical** or/and **efficient** than others...

Thermodynamic	*Conformational*	*Alchemical*
⇨ **Temperature integration**	⇨ **Direct counting (DC)**	⇨ **Thermodynamic integration (TI)**
⇨ **Pressure integration**	⇨ **Umbrella sampling (US)**	⇨ **Free-energy perturbation (FEP)**
Grand-canonical integration		⇨ **λ-dynamics local elevation umbrella sampling (λ-LEUS)**
		➡ **Enveloping distribution sampling (EDS)**
		Fast growth (FG)

Enveloping distribution sampling (EDS)

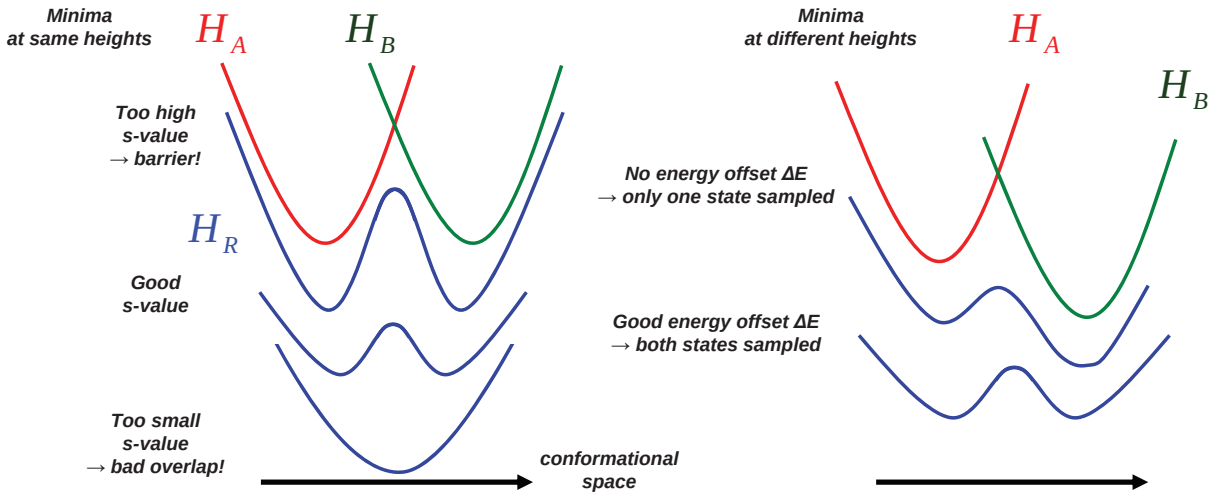
- Idea: we know the Hamiltonians H_A and H_B corresponding to two alchemical states, and we want to sample using a constructed a Hamiltonian H_R so that

  → we get configurations relevant to both A and B *i.e. some where H_A is low and others where H_B is low*

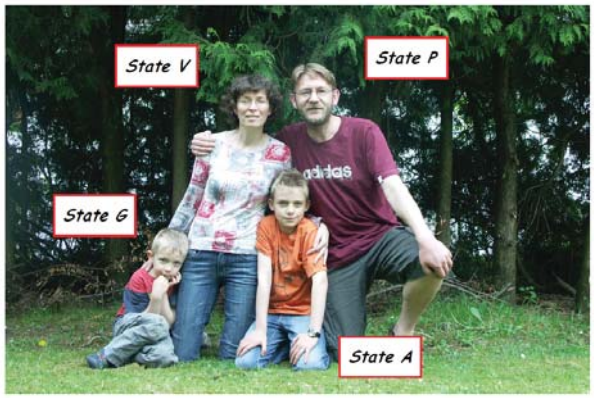
  → we get many transitions between these two types of configurations *i.e. the relative weighting can be determined accurately*

- A smart choice for H_R is an **enveloping Hamiltonian** generated by exponential averaging

$$H_R = -(\beta s)^{-1} \ln\left(e^{-\beta s H_A} + e^{-\beta s (H_B + \Delta E)}\right)$$



Enveloping distribution sampling (EDS)



Enveloping distribution sampling (EDS)

- Procedure:
  - → preoptimize the *s* and *ΔE* parameters using an empirical automated procedure
  - → sample using H_R
  - → use reweighting to H_A or H_B to get the free-energy difference

This is not absolutely guaranteed to work, as there may be no solution at all (but normally works for "simple enough" problems)

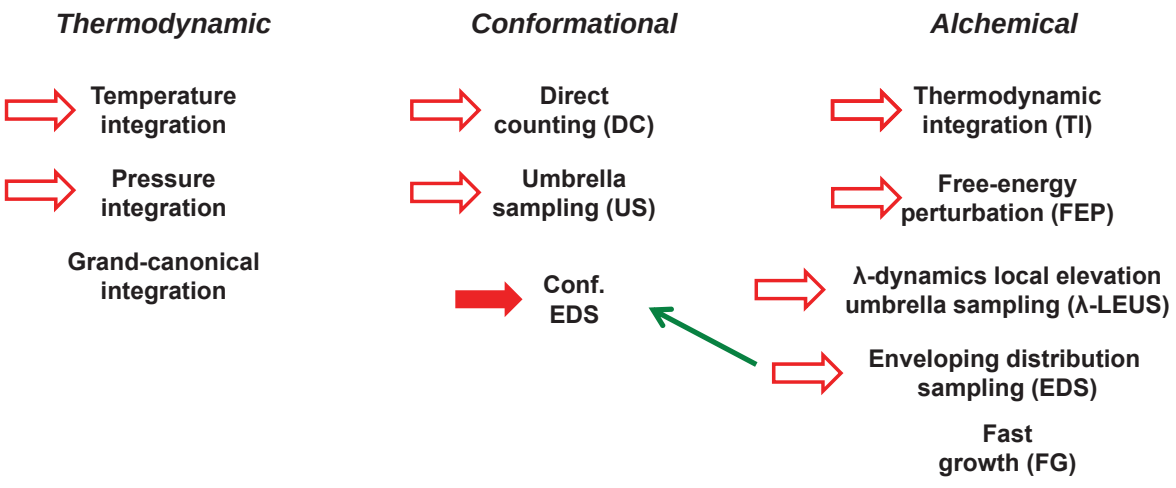
And difference in other properties if you like...

- Idea can be expanded to more than two states
- No interconversion pathway (coupling scheme) is defined between A and B
  - → is often good because the system finds its "own good pathway"
  - → may be bad because there may be no "good pathway" or because the unphysical space opened may be big
- Compared to OSP
  - → generally more accurate & accuracy can be evaluated
  - → but needs to know the end states in advance!

- predict end-state energy distributions and compare with end-state simulations (→ sufficient sampling of the end states?)
- monitor end-state energy time series (→ sufficient number of transitions between the end states?)

Overview of free-energy methods

- The available methods to calculate free-energy changes can be classified as historically applied to the different types of changes
  - → But remember that extended- or reduced-system approaches allow to interconvert the problem types, so that all methods are in principle applicable to all problem types !
  - → Still, some combinations are more **practical** or/and **efficient** than others...



EDS for conformational changes

- Use a biasing potential constructed using the EDS principle !

- Example:

$U_{bias,A}$ $U_{bias,B}$

$$U_{bias} = -(\beta s)^{-1} \ln\left(e^{-\beta s U_{bias,A}} + e^{-\beta s (U_{bias,B} + \Delta E)}\right)$$

→ with good choices of *s* and *ΔE* (may be difficult!), we should see frequent interconversions between the two equipopulated states – from which we can calculate the relative free energy

P.H. Hünenberger

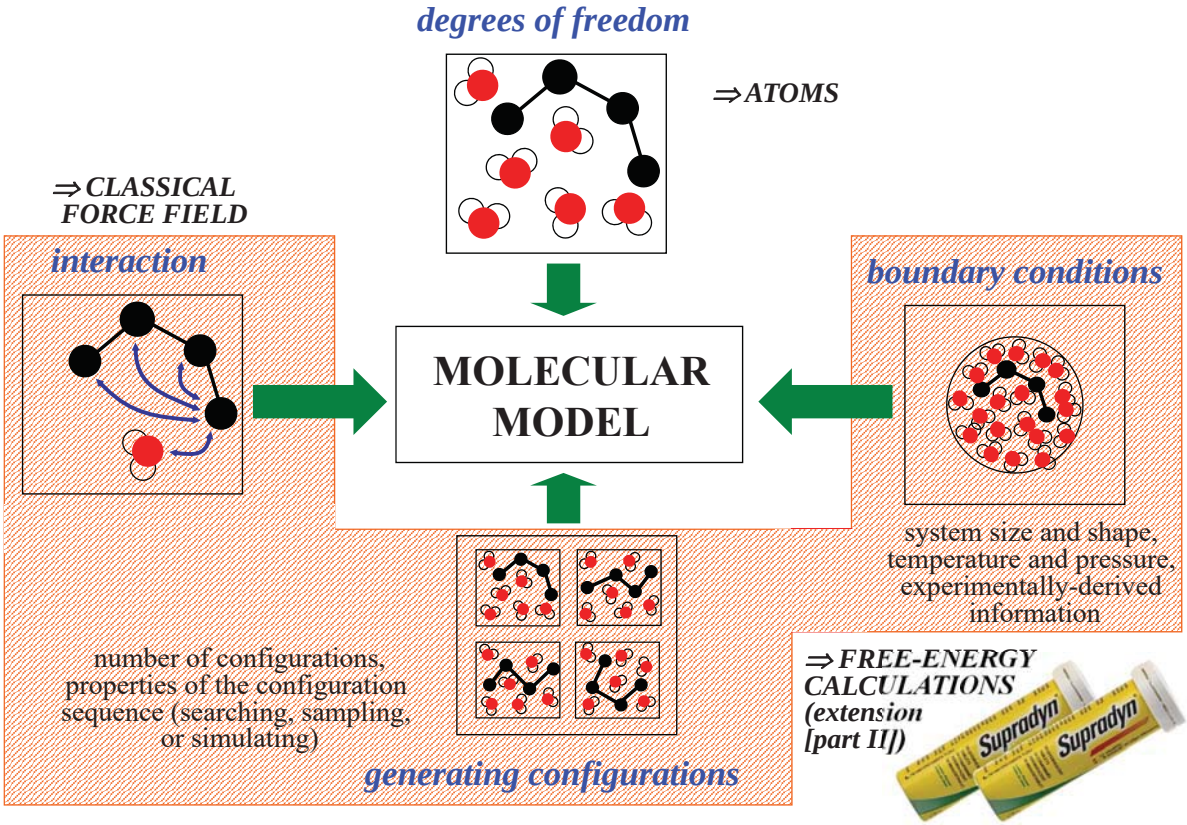
COMPUTER SIMULATION OF MOLECULAR SYSTEMS

Herbstsemester 2019
Tuesday 9:45-11:30 a.m.
HCI D2

Lecture 529-0004-00
www.csms.ethz.ch/education/CSCBP

LECTURE 8 (WEEK 9):
Free energy calculations II

Four basic choices defining a molecular model



Calculating free energies and entropies

(overview)

- The **free-energy differences** calculated based on **molecular simulations** are affected by the following **sources of errors**

→ Force field errors (functional form, parameters)

→ Finite-size errors (finite size of the simulated system, approximate electrostatics)

See other lectures (not re-discussed specifically here)

→ Sampling errors (finite timescale of the simulation)

→ Quadrature errors (TI)

→ Methodological issues

- Jacobian factor
- Metric tensor effects
- Contribution of constraints
- Contribution of restraints
- Singularities upon atom creation or deletion
- Free-energy components
- Standard-state corrections

→ Depend on

- Choice of pathway/cycle

→ Can be assed (in part) by

- Hysteresis
- Cycle closure
- Error estimation (*e.g.* block averaging)

- Calculation of entropies based on molecular simulations

Calculating free energies and entropies

(overview)

- The calculation of free energies based on molecular simulations are affected by the following practical issues

→ Equilibrium, finite sampling

→ Numerical quadrature (TI)

→ Cycle closure

→ Atom creation or deletion

→ Contribution of constraints

→ Free energy components ?

→ Choice of thermodynamic cycle and pathway

- Calculation of entropies based on molecular simulations

Error estimation in free-energy calculations

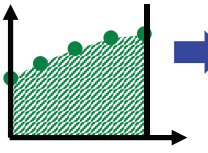
- There are many different ways to assess the reliability or/and estimate errors in free-energy calculations; a number of **common ways** are

→ Estimate **statistical error** *via* independent repeats
(alternative: time-correlation analysis or block-averaging on one simulation [or bootstrapping])

N independent repeats n=1..N

$$\Delta F = \frac{1}{N}\sum_{n=1}^{N}\Delta F_n \qquad \varepsilon = c\frac{\sigma}{N^{1/2}}$$

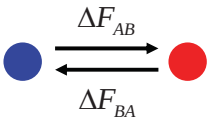
→ Estimate **statistical error** *via* error propagation (for TI)



$$\Delta F = \sum_{k=1}^{K} w_k \left\langle \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_{\lambda_k} \qquad \varepsilon = \left[\sum_{k=1}^{K}\left(w_k \varepsilon_k\right)^2\right]^{1/2}$$

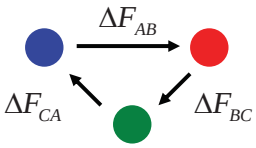
w_k *quadrature weights* ε_k *e.g. from block-averaging*

→ Check **hysteresis** (for "directional" methods)
e.g. FEP, SG or TI (with specific setup protocol)



$\Delta F_{AB} + \Delta F_{BA}$ *= hysteresis (should ideally be zero!)*

→ Check **cycle closure** (when you have ≥3 states)



$\Delta F_{AB} + \Delta F_{BC} + \Delta F_{CA}$ *= closure error (should ideally be zero!)*

→ Check **extrapolation accuracy** (for "extrapolative" methods)
e.g. FEP, OSP, EDS

e.g. predict end-state (A or B) potential-energy distribution and compare with real one from a separate plain simulation of this state

Error estimation in free-energy calculations

- These "indicators" generally give "**optimistic/lower bound**" information; additional error sources may easily be overlooked!

→ Estimate **statistical error** *via* independent repeats
(alternative: time-correlation analysis or block-averaging on one simulation)

Did we generate truly independent repeats (e.g. is changing the initial velocities enough?)

Can there be systematic errors in the methodology?

→ Estimate **statistical error** *via* error propagation (for TI)

Did we equilibrate and simulate long enough at each lambda points? (were there states we completely missed?)

Was the numerical quadrature good enough?

Can there be systematic errors in the methodology?

→ Check **hysteresis** (for "directional" methods)
e.g. FEP, SG or TI (with specific setup protocol)

Low hysteresis may be due to too short sampling

→ Check **cycle closure** (when you have ≥3 states)

Errors on each leg may be larger than cycle-closure error when they coincidentally compensate each other

→ Check **extrapolation accuracy** (for "extrapolative" methods)
e.g. FEP, OSP, EDS

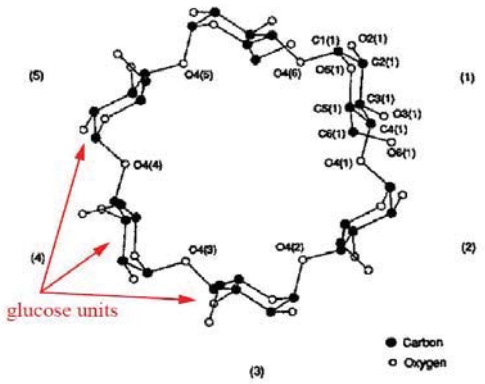
Difficult to convert an extent of overlap with a free-energy error estimate

Equilibration and sampling

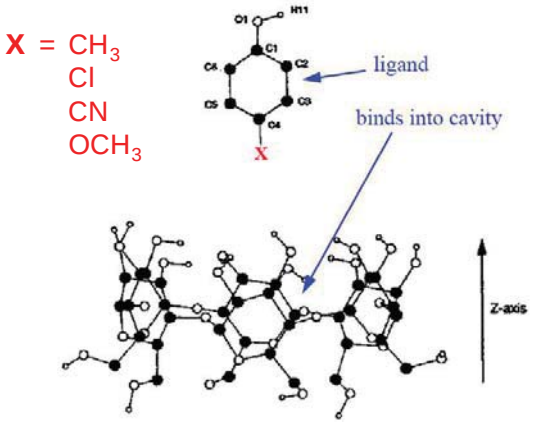
MA96.12: Mark et al., JACS 116 (1994) 6293

- Example: **relative binding free energies** of *p*-substituted phenols to cyclodextrin

  → Small host-guest system: α-cyclodextrin consists of 6 sugar (glucose) units (cyclic)



α-cyclodextrin



p-methoxyphenol showing the orientation of the guest when inserted.

→ Simulations of one α-cyclodextrin + 500 H_2O, NPT ensemble

→ 6x2 mutations (any of the 4 to any of the 4, in free and bound forms)

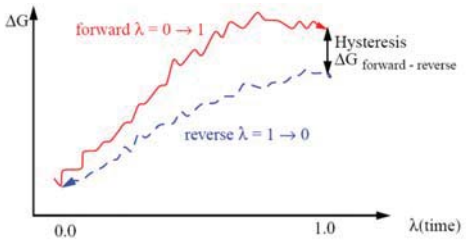
→ Slow growth (SG) *vs.* thermodynamic integration (TI) with 12 lambda points per mutation

Equilibration and sampling

- The **hysteresis** of a free-energy calculation is the sum of the results obtained when performing the calculation forward and then backward

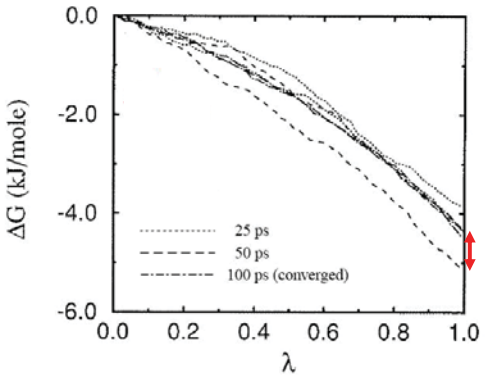
  → for a "perfect" calculation, it should be zero (free energy is a state function)

  → but a value close to zero does not prove that the result is correct!



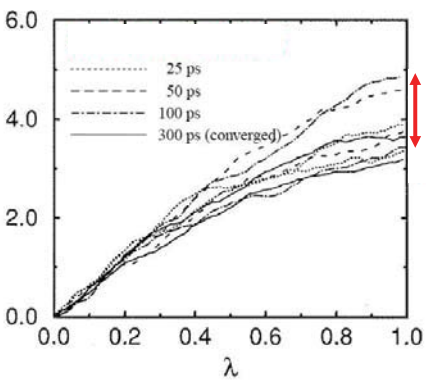
- Example: hysteresis of a SG calculation for *p*-chloro- to *p*-methyl-phenol

in water



largest for 50 ps

in cyclodextrin

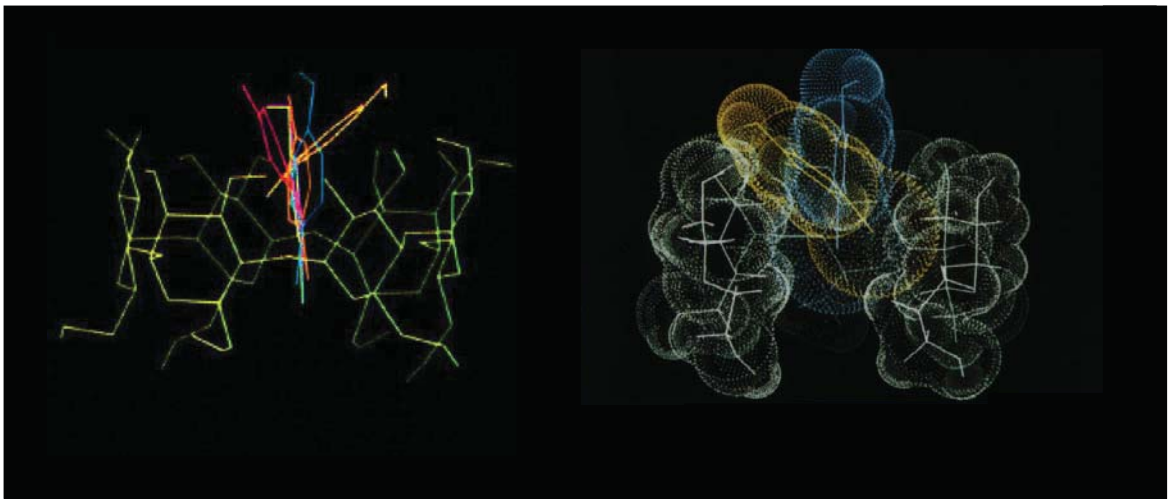


largest for 100 ps

→ Hysteresis **first increases then decreases** with longer sampling!

Equilibration and sampling

- Explanation (for the bound case): the ligand evidences slow motions in the host

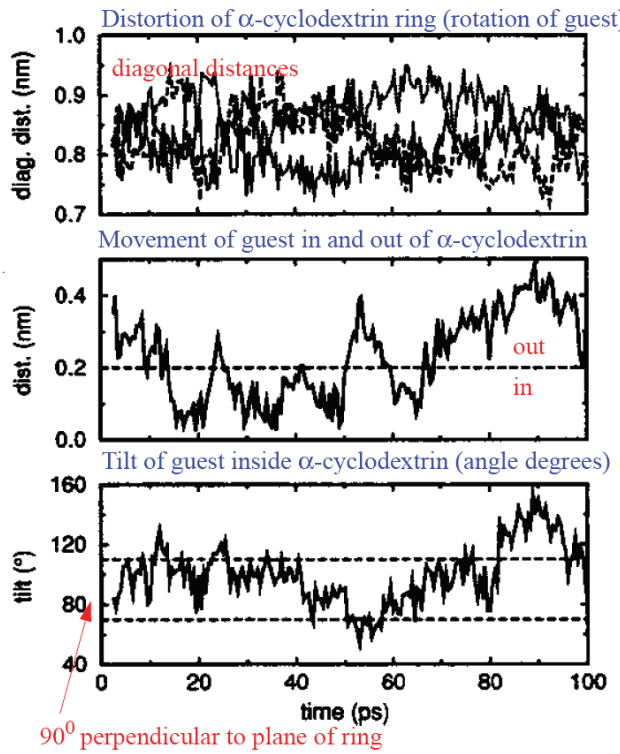


Side view of α-CD with *p*-Cl-phenol showing motion of the guest inside the cavity (water not shown).

αCD with *p*-Cl-phenol showing van der Waals contacts.

Equilibration and sampling

- Timescales of the motions of *p*-Cl-phenol in α-cyclodextrin



Relaxation times (guest)

a) Rotational averaging	20-40ps
b) In/out motion	60-80ps
c) Tilt averaging	>80ps

Equilibration and sampling

- Observations with SG

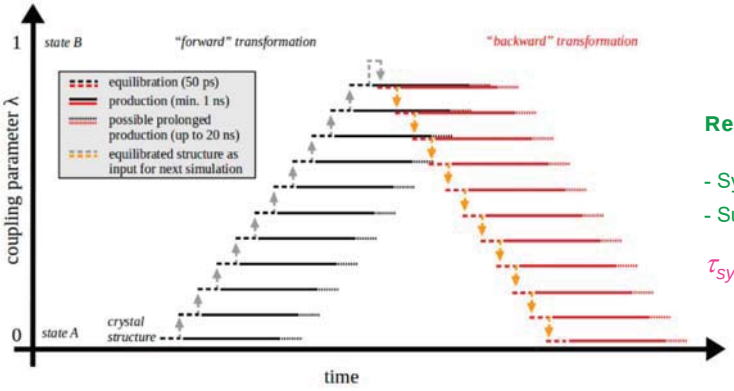
→ if you go **very fast**, the system has no time to sample motions of the system ➡ *Low hysteresis BUT incorrect free energy* — *FROZEN SYSTEM*

→ if you go **moderately fast**, the system only partially samples motions of the system ➡ *High hysteresis AND incorrect free energy* — *PARTIAL SAMPLING*

→ if you go **very slowly**, the system has enough time to relax (equilibrate) and sample at each λ ➡ *Low hysteresis AND correct free energy* — *ADEQUATE SAMPLING*

- There can also be hysteresis in standard TI

→ *via* the way you typically setup your systems at the different λ-points, *e.g.*



The same considerations apply

Requirements: (for each λ):

- System in equilibrium: $\tau_{equil.} > \tau_{system}$
- Sufficient sampling : $\tau_{sample} >> \tau_{system}$

τ_{system} = *relaxation time of the system*

Equilibration and sampling

- Example: sampling at each λ-point for *p*-chloro- to *p*-methyl-phenol in α-cyclodextrin

→ Ensemble average $\left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda}$ must converge at each value of λ

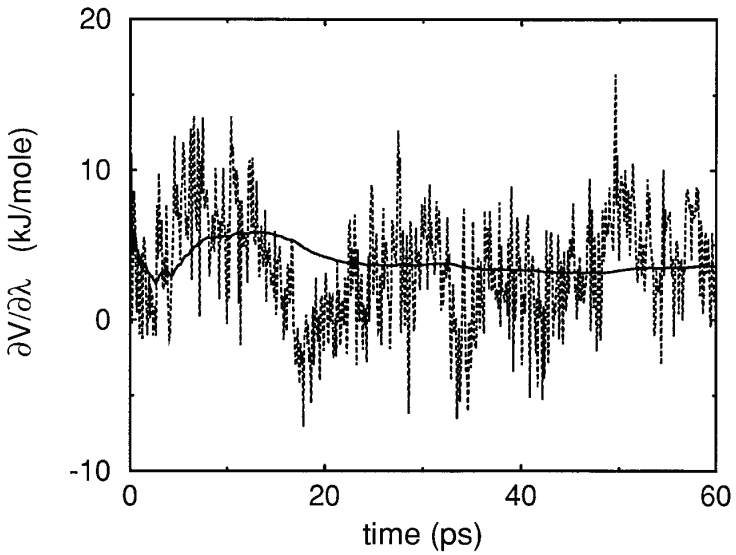


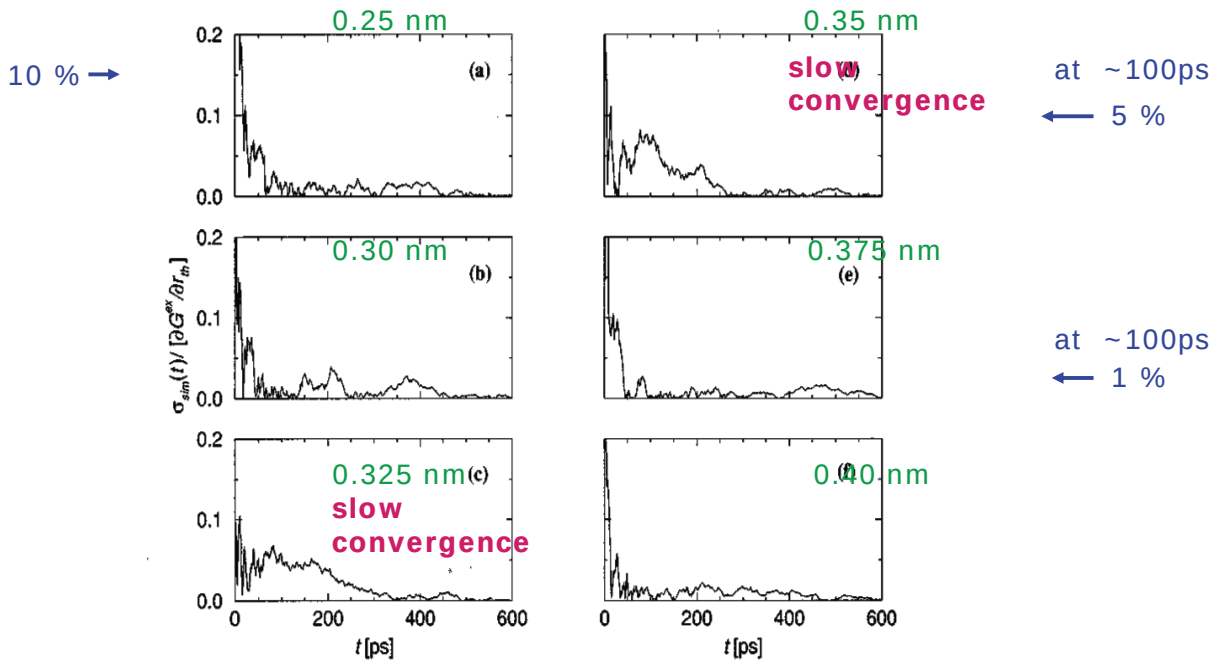
Figure 5. $\partial V/\partial\lambda$ (kJ mol^{-1}) at $\lambda \approx 0.5$ as a function of simulation time for the mutation of *p*-chlorophenol to *p*-methylphenol complexed to α-cyclodextrin. The dashed line corresponds to the instantaneous derivative, and the solid line, to the accumulative average.

Equilibration and sampling

J. Chem. Phys.
102 (1995) 3787

- Example: Relative accuracy of the excess Gibbs free energy of cavity formation in water
  → Cavities of different size (radius) in H_2O

Error (10^2 %)

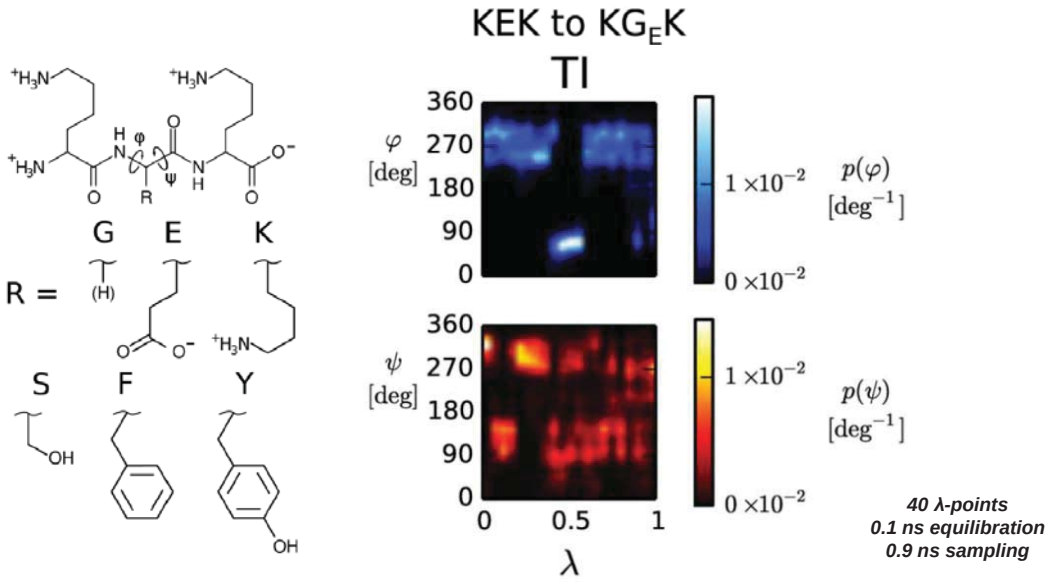


Equilibration and sampling

Bieler & Hünenberger
J. Comput. Chem., 36 (2015) 1686.

- In TI, the convergence at each λ-point may be affected by slow **orthogonal-relaxation** processes

  → Example: mutations in the central residue of a KXK tripeptide



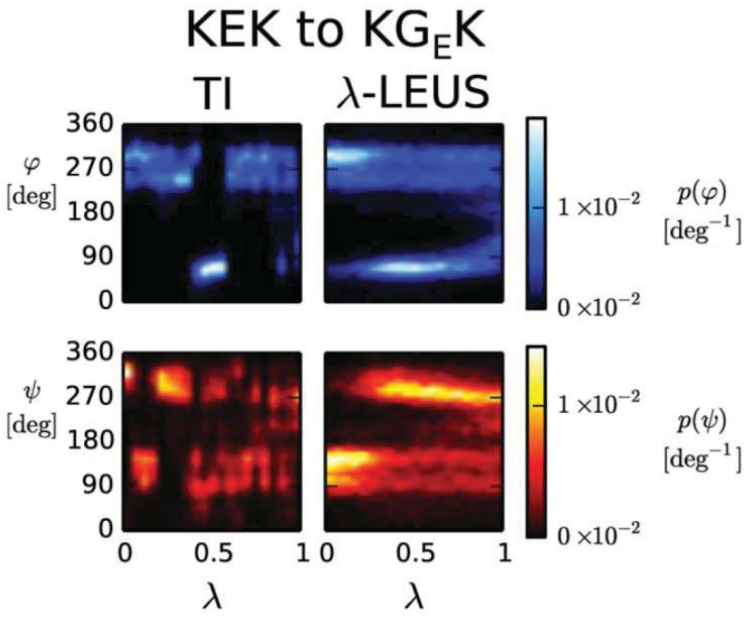
40 λ-points
0.1 ns equilibration
0.9 ns sampling

→ The results will depend a lot how you started the simulations at each λ-point and how long you equilibrated/simulated

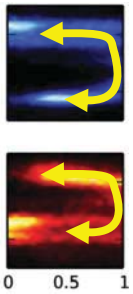
(the statistical error may also be a clear underestimate of the true error)

Equilibration and sampling

- To improve the orthogonal sampling
  - → Apply orthogonal sampling enhancement (*e.g.* parallel tempering) *(next lecture)*
  - → Make system "hop" between λ-values (*e.g.* Hamiltonian replica-exchange, λ-LEUS)
- E.g. with λ-LEUS



λ varies over the course of the simulation → system finds λ-values where barriers are easier to cross (in this case: in the glycine state where ϕ,ψ-rotation is easy!)



Calculating free energies and entropies

(overview)

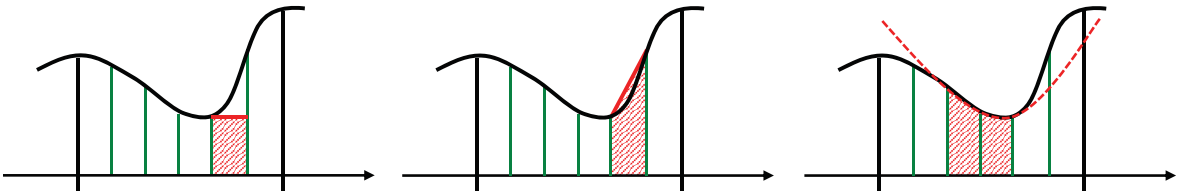
- The calculation of free energies based on molecular simulations are affected by the following practical issues
  - → Equilibrium, finite sampling
  - → Numerical quadrature (TI)
  - → Cycle closure
  - → Atom creation or deletion
  - → Contribution of constraints
  - → Free energy components ?
  - → Choice of thermodynamic cycle and pathway
- Calculation of entropies based on molecular simulations

Quadrature integration

- Choice of a quadrature method

→ Standard quadrature scheme $$I_{ab} = h\sum_{n=0}^{N} w_n f(x_n)$$

$\boldsymbol{w}^Q = (\ 1\ 1\ \ldots\ 1\ 0\ )$ $\boldsymbol{w}^T = \left(\ \frac{1}{2}\ 1\ 1\ \ldots\ 1\ 1\ \frac{1}{2}\ \right)$ $\boldsymbol{w}^S = \left(\ \frac{1}{3}\ \frac{4}{3}\ \frac{2}{3}\ \ldots\ \frac{2}{3}\ \frac{4}{3}\ \frac{1}{3}\ \right)$



Rectangle *Trapezoid* *Simpson*

→ You can also fit a polynomial or spline and integrate it ← *Recommended*

→ The choice of the number of points may be crucial! *Romberg = overkill!*

Quadrature integration

- Number/choice of λ-points

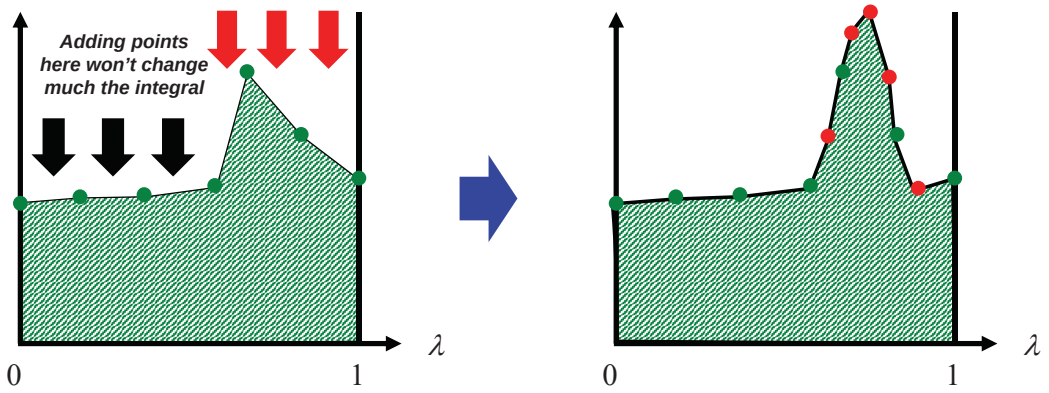
→ Look at the curves!

→ If needed, add new points in areas of high curvature

ΔF = integral = area under the curve

$$\Delta F_{AB} = \int_0^1 d\lambda' \left\langle \frac{\partial \mathcal{H}(\boldsymbol{r}, \boldsymbol{p}; \lambda)}{\partial \lambda} \right\rangle_{\lambda'}$$

$\left\langle \frac{\partial \mathcal{H}(\boldsymbol{r}, \boldsymbol{p}; \lambda)}{\partial \lambda} \right\rangle_{\lambda}$

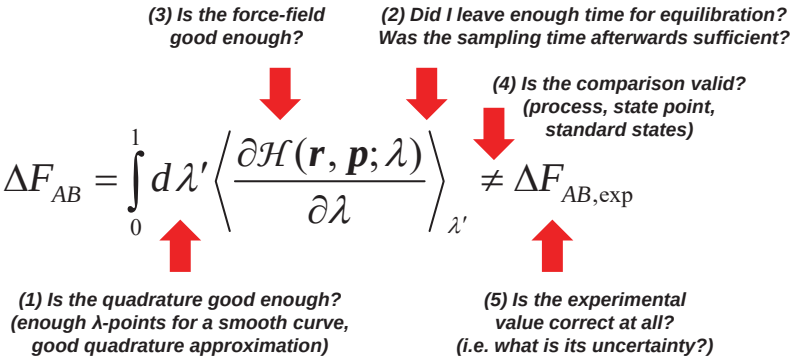


Choice: how many λ-values, distribution of λ-values } **integrate a smooth curve?**

Sources of error in TI

- Say you do a TI calculation and compare the result to experiment and... it does not agree!

  → The five questions to ask are:



Calculating free energies and entropies

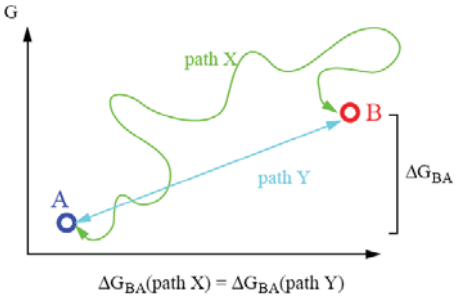
(overview)

- The calculation of free energies based on molecular simulations are affected by the following practical issues
  - → Equilibrium, finite sampling
  - → Numerical quadrature (TI)
  - → Cycle closure
  - → Atom creation or deletion
  - → Contribution of constraints
  - → Free energy components ?
  - → Choice of thermodynamic cycle and pathway
- Calculation of entropies based on molecular simulations

Cycle closure

- Free energy is a **state function**

  → Difference in free energy between two states is independent of the path used to go from the one to the other state

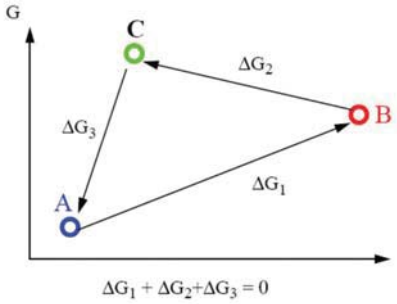


- **Thermodynamic cycle** → a closed path for which ΔG should be zero

- The **cycle closure error** of a free-energy calculation is the sum of the results obtained when summing free-energy differences over the three (or more) legs of a closed path

  → for a "perfect" calculation, it should be zero (free energy is a state function)

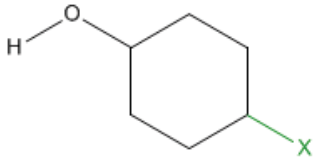
  → but a value close to zero does not prove that the result is correct!



Cycle closure

Residual free energy for a closed thermodynamic cycle

Binding of p-substituted phenols to α-cyclodextrin



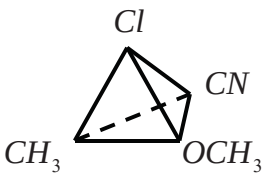
Four possible 3-membered closed cycles:

in water ↓ in α-CD ↓ X = Cl, CH_3, CN, OCH_3

mutation	ΔG_{water}	$\Delta G_{\alpha\text{-CD}}$	$\Delta\Delta G_{water\text{-}\alpha\text{-CD}}$
Cl → CH_3 → CN → Cl	-1.0	5.2	6.2
Cl → CN → OCH_3 → Cl	17.2	12.8	**-4.4**
Cl → CH_3 → OCH_3 → Cl	2.7	6.0	**3.3**
CH_3 → OCH_3 → CN → CH_3	-13.5	-12.0	**1.5**

gives wrong error estimate

X_1 -> X_2 -> X_3 -> X_1



Ideally, all values should be zero

Actual error much bigger

use *single cycle* values to estimate lower bound of error, *not ΔΔ G* values

Calculating free energies and entropies

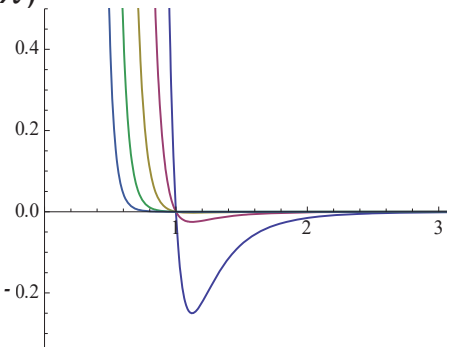
(overview)

- The calculation of free energies based on molecular simulations are affected by the following practical issues
  - → Equilibrium, finite sampling
  - → λ integration
  - → Cycle closure
  - → Atom creation or deletion
  - → Contribution of constraints
  - → Free energy components ?
  - → Choice of thermodynamic cycle and pathway
- Calculation of entropies based on molecular simulations

Singularity upon atom creation or deletion

- The problem arises when we **create** or **delete** an atom in a mutation, *i.e.* when we mutate between a **dummy** site and a site that has **non-bonded interactions**
  - → Example: we create an atom between states A and B, and consider the LJ interactions with the solvent

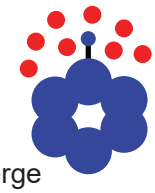
$V_{LJ}(r;\lambda)$



$$V_{LJ}(\lambda) = (1-\lambda)V_{LJ,A} + \lambda V_{LJ,B}$$

$$V_{LJ,A} = 0 \qquad V_{LJ,B} = 4\varepsilon\left(\frac{\sigma^{12}}{r_{ij}^{12}} - \frac{\sigma^{6}}{r_{ij}^{6}}\right)$$

- → When λ is very small, $V_{LJ}(r;\lambda)$ becomes very narrow and steep
- → Simulation crash (SHAKE failure), because r ≈ 0 will be accessed at finite timestep, and the force (slope of $V_{LJ}(r;\lambda)$) is then very large
- → Even if they don't crash, the B-state energy will be super-noisy and the ensemble-averaged Hamiltonian derivative will not converge
- → Atom deletion: same problem close to λ = 1



Singularity upon atom creation or deletion

- Can we do something using **soft-core** non-bonded interactions?

→ Functional form

van der Waals

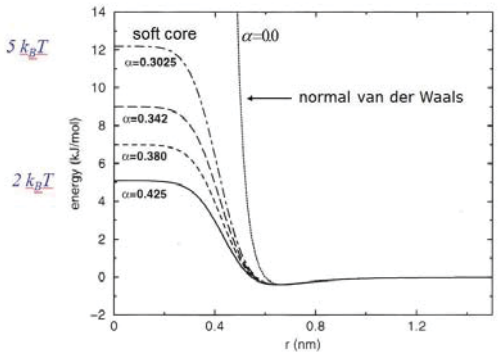
$$V(r) = 4\varepsilon\left[\frac{1}{\left[\alpha + (r/\sigma)^6\right]^2} - \frac{1}{\alpha + (r/\sigma)^6}\right]$$

electrostatics

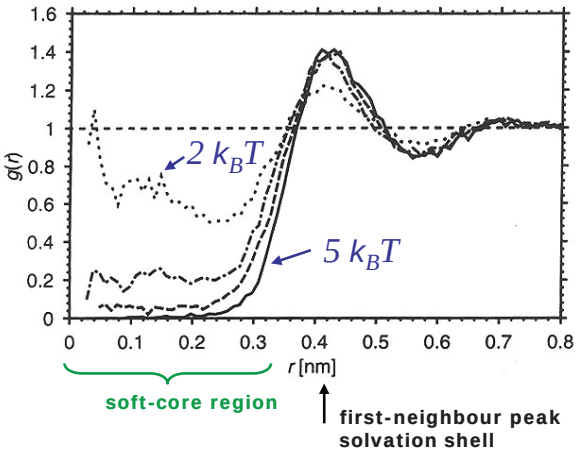
$$V(r) = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_1}\left[\frac{1}{\left[\alpha_c + (r)^2\right]^{\frac{1}{2}}} - \frac{0.5C_{rf}r^2}{\left[\alpha_c + R_{rf}^2\right]^{\frac{1}{2}}} - \frac{1-0.5C_{rf}}{R_{rf}}\right]$$

→ Effect

interaction energy



Distribution of water molecules around and in the soft-core cavity of 0.6 nm diameter



Singularity upon atom creation or deletion

- The solution to this problem is to define a better coupling scheme

→ The only requirement on a coupling scheme is

$$\mathcal{H}(\mathbf{r},\mathbf{p};\lambda) = \begin{cases} \mathcal{H}_A(\mathbf{r},\mathbf{p}) & \text{if } \lambda=0 \\ \mathcal{H}_B(\mathbf{r},\mathbf{p}) & \text{if } \lambda=1 \end{cases}$$

this condition is actually compatible with many alternative coupling schemes !

specific λ-dependence determines pathway from A to B

→ This leaves a lot of freedom!

- A good solution is to use a **soft-core coupling scheme**

→ For LJ interactions

$$V_{LJ} = 4\varepsilon\left(\frac{\sigma^{12}}{r_{ij}^{12}} - \frac{\sigma^6}{r_{ij}^6}\right)$$

$$V_{LJ}(r_{ij};\lambda) = [1-\lambda]^n\left\{\frac{C_{12}^A}{\left[\alpha_{LJ}\lambda^2 + r_{ij}^6\right]^2} - \frac{C_6^A}{\left[\alpha_{LJ}\lambda^2 + r_{ij}^6\right]}\right\} + \lambda^n\left\{\frac{C_{12}^B}{\left[\alpha_{LJ}[1-\lambda]^2 + r_{ij}^6\right]^2} - \frac{C_6^B}{\left[\alpha_{LJ}[1-\lambda]^2 + r_{ij}^6\right]}\right\}$$

→ And similarly for Coulomb interactions

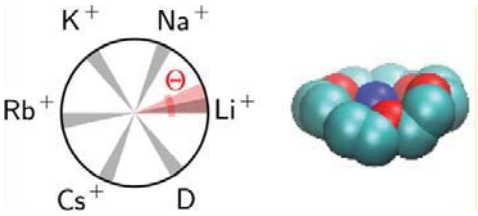
$$V_{CB}(r_{ij};\lambda) = [1-\lambda]^n\left\{\frac{1}{4\pi\varepsilon_0\varepsilon_1}\frac{q_i^A q_j^A}{\left[\alpha_c\lambda^2 + r_{ij}^2\right]^{1/2}}\right\} + \lambda^n\left\{+\frac{1}{4\pi\varepsilon_0\varepsilon_1}\frac{q_i^B q_j^B}{\left[\alpha_c[1-\lambda]^2 + r_{ij}^2\right]^{1/2}}\right\}$$

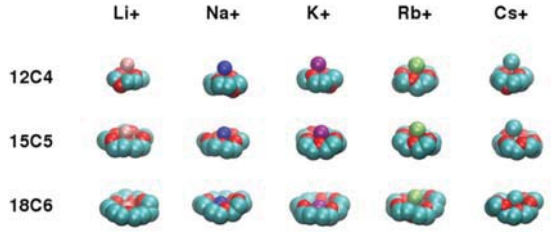
For GROMOS: also add RF terms

Singularity upon atom creation or deletion

Bieler, Tschopp & Hünenberger, P.H. J. Chem. Theory. Comput. 11 (2015) 2575

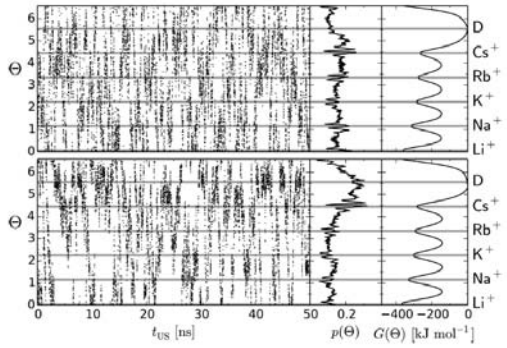
- Watch out: soft-core may create rather unphysical intermediate state!



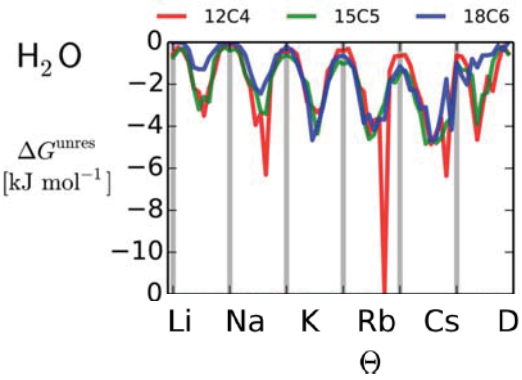


ion in water

ion in 15C5 host



correction for the positional restraints of the ion in the host



→ in the intermediate soft-core state, the ion would love to exit the host and be freely hydrated!

Calculating free energies and entropies

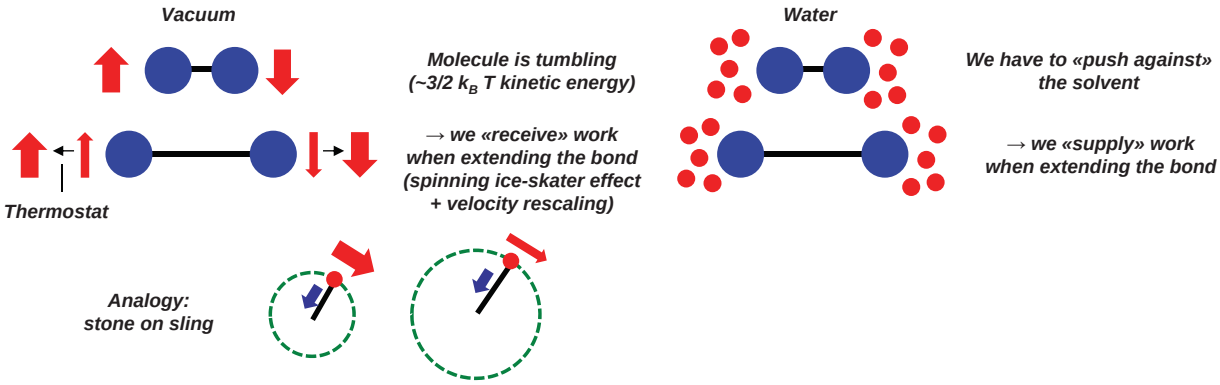
(overview)

- The calculation of free energies based on molecular simulations are affected by the following practical issues
  - → Equilibrium, finite sampling
  - → Numerical quadrature (TI)
  - → Cycle closure
  - → Atom creation or deletion
  - → Contribution of constraints
  - → Free energy components ?
  - → Choice of thermodynamic cycle and pathway
- Calculation of entropies based on molecular simulations

Contribution of constraints

- Consider a diatomic molecule where we **change the bond length** over a simulation

→ We supply work to / receive work from the system



→ Exchanging work reversibly means that there is a **change in free energy**

→ For flexible bonds (*e.g.* harmonic potential), this contribution arises automatically $\frac{\partial V_{bond}(\lambda)}{\partial \lambda}$

→ Problem: we use SHAKE in the simulation, so there is by default no contribution of this effect to the free energy; *e.g.* in TI

$$V_{SHAKE}(\lambda) = ??? \qquad \frac{\partial V_{SHAKE}(\lambda)}{\partial \lambda} = ???$$

Contribution of constraints

Free energy as a function of a changing constraint

1. Assume *R* is a constraint ("reaction coordinate") defined by

$R(\vec{r}_1, \vec{r}_2, ..., \vec{r}_N)$ e.g. distance r_{ij} *or* torsional angle φ

2. Free energy change with respect to *R*:

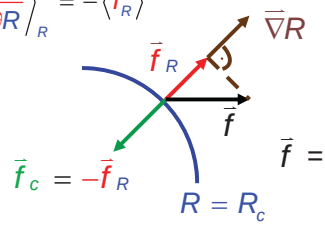
$$F(R) = -\beta^{-1} \ln Z(R)$$

$$Z(R) = \xi(\text{momentum part}) \int e^{-\beta V(\vec{r}_1, \vec{r}_2, ..., \vec{r}_N; R)} d\vec{r}_1 d\vec{r}_2 ... d\vec{r}_N$$

$$\frac{\partial F}{\partial R} = -\beta^{-1} \frac{\int (-\beta) \frac{\partial V}{\partial R} e^{-\beta V(\vec{r}_1, \vec{r}_2, ..., \vec{r}_N; R)} d\vec{r}_1 d\vec{r}_2 ... d\vec{r}_N}{\int e^{-\beta V(\vec{r}_1, \vec{r}_2, ..., \vec{r}_N; R)} d\vec{r}_1 d\vec{r}_2 ... d\vec{r}_N} = \left\langle \frac{\partial V}{\partial R} \right\rangle_R = -\langle f_R \rangle$$

$$\vec{f}_R = \frac{(\vec{f} \cdot \vec{\nabla} R) \vec{\nabla} R}{(\vec{\nabla} R)^2}$$

$$= -\vec{f}_{constraint}$$



$\vec{f}$ = free force

= force to be subtracted from $\vec{f}$ to satisfy $R(\vec{r}_1, \vec{r}_2, ..., \vec{r}_N) = R_c$

= known in a constraint simulation

e.g. Lagrange multipliers of SHAKE!

3. Free energy difference

$$F(R_2) - F(R_1) = \int_{R_1}^{R_2} \frac{\partial F}{\partial R} dR = \int_{R_1}^{R_2} \langle f_c \rangle dR$$

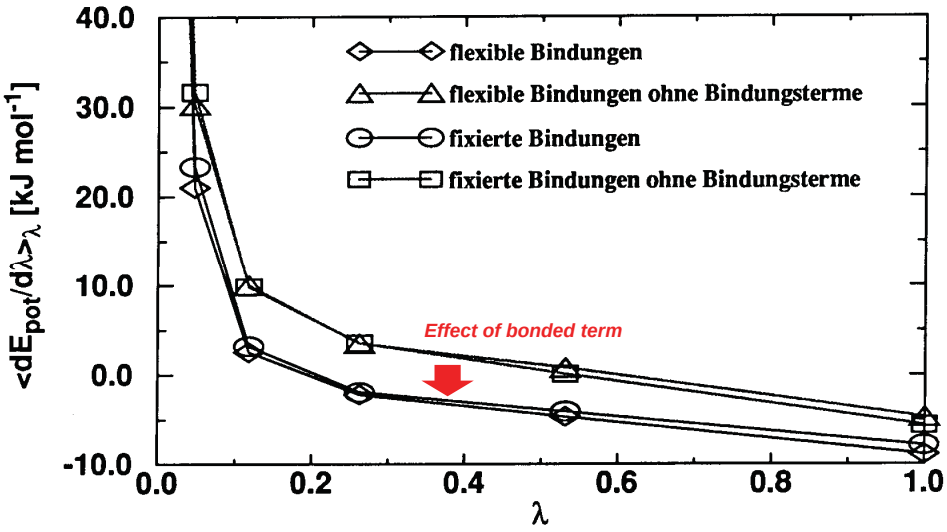
e.g. if solvent "compresses" bond, $f_c > 0$ → F increases if we extend the bond (we give work to push against the solvent)

Contribution of constraints

Contribution of bond-stretching term in V(r) to ΔG

Process: change 512 H_2O into 512 CH_3OH

	ΔG_{bonds}	ΔG_{total} (kJ / mol)	
gas phase	≈ 0	≈ 0	*vacuum contribution is small*
liquid phase 1. flexible model	-6.1	8.4	
2. rigid model I	-4.8	9.1	
rigid model II	-4.7	11.3	



H—O—H

H—O—CH_3

negative → here, the surroundings want to "stretch" the bond (probably because of H-bonds)

Bond contribution cannot be neglected

M. Rüedi, master thesis ETH (1992)

Calculating free energies and entropies

(overview)

- The calculation of free energies based on molecular simulations are affected by the following practical issues
  - → Equilibrium, finite sampling
  - → Numerical quadrature (TI)
  - → Cycle closure
  - → Atom creation or deletion
  - → Contribution of constraints
  - → Free energy components ?
  - → Choice of thermodynamic cycle and pathway
- Calculation of entropies based on molecular simulations

Free energy components

Decomposition of a free energy change ΔF into components ?

Question:
Is a decomposition of a free energy difference $\Delta F = \Delta F_1 + \Delta F_2$ possible, if the potential energy is decomposable, so if $V = V_1 + V_2$?

Free energy: only configurational integral

$$F(\lambda) = -\beta^{-1} \ln\left\{\int e^{-\beta V(\vec{r}^N;\lambda)} d\vec{r}^N\right\} + constant$$

Free energy difference, perturbation formula:

$$\Delta F(\lambda) \equiv F(\lambda) - F(0)$$
$$= -\beta^{-1} \ln\left\{\left\langle e^{-\beta\Delta V(\lambda)}\right\rangle_{\lambda=0}\right\} \qquad \text{where} \quad \Delta V(\vec{r}^N;\lambda) \equiv V(\vec{r}^N;\lambda) - V(\vec{r}^N;0)$$

Decompose potential energy:

$$V(\vec{r}^N;\lambda) = V_1(\vec{r}^N;\lambda) + V_2(\vec{r}^N;\lambda)$$

So:

$$\Delta F(\lambda) = -\beta^{-1} \ln\left\langle e^{-\beta\Delta V_1(\lambda)} e^{-\beta\Delta V_2(\lambda)}\right\rangle_{\lambda=0}$$

= expand in powers of β up till β^3 use: $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots$

$= \Delta F_1(\lambda) + \Delta F_2(\lambda)$??? $\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots$

Free energy components

Can one write $\Delta F(\lambda) = \Delta F_1(\lambda) + \Delta F_2(\lambda)$?

$$\Delta F(\lambda) = F(\lambda) - F(0)$$
$$= -\beta^{-1} \ln\left\langle e^{-\beta\Delta V_1(\lambda)} e^{-\beta\Delta V_2(\lambda)}\right\rangle_{\lambda=0}$$

$e^x \approx 1 + x + \frac{x^2}{2}$

$$= -\beta^{-1} \ln\left\{\left\langle\left[1 - \beta\Delta V_1 + \beta^2\frac{(\Delta V_1)^2}{2} - O(\beta^3)\right]\left[1 - \beta\Delta V_2 + \beta^2\frac{(\Delta V_2)^2}{2} - O(\beta^3)\right]\right\rangle_{\lambda=0}\right\}$$

$$= -\beta^{-1} \ln\left\{1 - \beta\left[\langle\Delta V_1\rangle_{\lambda=0} + \langle\Delta V_2\rangle_{\lambda=0}\right] + \frac{\beta^2}{2}\left[\left\langle(\Delta V_1)^2\right\rangle_{\lambda=0} + \left\langle(\Delta V_2)^2\right\rangle_{\lambda=0} + 2\langle\Delta V_1\Delta V_2\rangle_{\lambda=0}\right] + O(\beta^3)\right\}$$

$\ln(1+x) \approx x - \frac{x^2}{2}$

$$= -\beta^{-1}\left\{-\beta\left[\langle\Delta V_1\rangle_{\lambda=0} + \langle\Delta V_2\rangle_{\lambda=0}\right] + \frac{\beta^2}{2}\left[\left\langle(\Delta V_1)^2\right\rangle_{\lambda=0} + \left\langle(\Delta V_2)^2\right\rangle_{\lambda=0} + 2\langle\Delta V_1\Delta V_2\rangle_{\lambda=0}\right] + O(\beta^3)\right.$$
$$\left. - \frac{\beta^2}{2}\left[\langle\Delta V_1\rangle^2_{\lambda=0} + \langle\Delta V_2\rangle^2_{\lambda=0} + 2\langle\Delta V_1\rangle_{\lambda=0}\langle\Delta V_2\rangle_{\lambda=0}\right] + O(\beta^3)\right\}$$

$$= \langle\Delta V_1\rangle_{\lambda=0} - \frac{\beta}{2}\left[\left\langle(\Delta V_1)^2\right\rangle_{\lambda=0} - \langle\Delta V_1\rangle^2_{\lambda=0}\right] + O(\beta^2)$$
$$+ \langle\Delta V_2\rangle_{\lambda=0} - \frac{\beta}{2}\left[\left\langle(\Delta V_2)^2\right\rangle_{\lambda=0} - \langle\Delta V_2\rangle^2_{\lambda=0}\right] + O(\beta^2) - \beta\left[\langle\Delta V_1\Delta V_2\rangle_{\lambda=0} - \langle\Delta V_1\rangle_{\lambda=0}\langle\Delta V_2\rangle_{\lambda=0}\right] + O(\beta^2)$$

$$= \Delta F_1(\lambda) + \Delta F_2(\lambda) - \beta\left[\left\langle\Delta V_1(\lambda)\Delta V_2(\lambda)\right\rangle_{\lambda=0} - \left\langle\Delta V_1(\lambda)\right\rangle_{\lambda=0}\left\langle\Delta V_2(\lambda)\right\rangle_{\lambda=0}\right] + O(\beta^2)$$

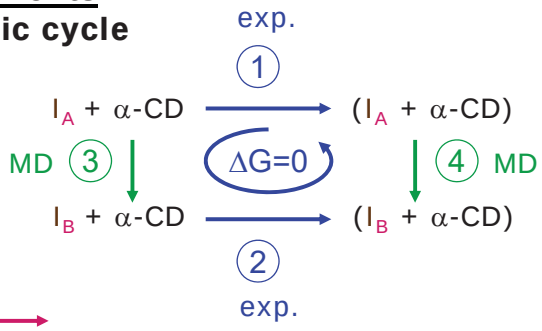
$\neq \Delta F_1(\lambda) + \Delta F_2(\lambda)$ **unless $\Delta V_1(\lambda)$ and $\Delta V_2(\lambda)$ are uncorrelated**

Smith and van Gunsteren, J. Phys. Chem. ***98*** *(1994) 13735-13740*

Free energy components

Use of thermodynamic cycle

Not for components of the Hamiltonian



Free enthalpy G is state function →

→ **ΔG is independent of path chosen** →

→ **ΔG (around cycle) = 0** → $\Delta\Delta G_{12} \equiv \Delta G_1 - \Delta G_2 = \Delta G_3 - \Delta G_4 \equiv \Delta\Delta G_{34}$

exp. ↑ MD

only correct for state function, **not for components!!**

$$\int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial H}{\partial \lambda} \right\rangle d\lambda \quad = \textit{in}\text{dependent of } \lambda \text{ path}$$

For components H_1 and H_2 of H this is not the case:

$$\int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial (H_1 + H_2)}{\partial \lambda} \right\rangle d\lambda = \underbrace{\int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial (H_1)}{\partial \lambda} \right\rangle d\lambda}_{\text{path } \textit{de}\text{pendent}} + \underbrace{\int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial (H_2)}{\partial \lambda} \right\rangle d\lambda}_{\text{path } \textit{de}\text{pendent}}$$

sum is path ***in***dependent → **Decomposition is meaningless !**

Free energy components

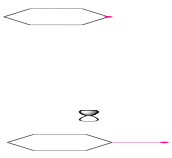
Decomposition of ΔΔG in free enthalpy components ?

Example:

Grow atom X on a benzene ring in solvent

Path I:

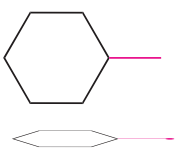
1. **grow van der Waals interaction:** $\Delta G_{vdW} = 0$



2. **grow bond-length interaction:** $\Delta G_{bond} = \textit{finite}$

Path II:

1. **grow bond-length interaction:** $\Delta G_{bond} = 0$



2. **grow van der Waals interaction:** $\Delta G_{vdW} = \textit{finite}$

Path I: bond forces determine *ΔG*
Path II: van der Waals forces determine *ΔG*

A.E. Mark et al., J. Mol. Biol. ***240*** *(1994) 167-176*

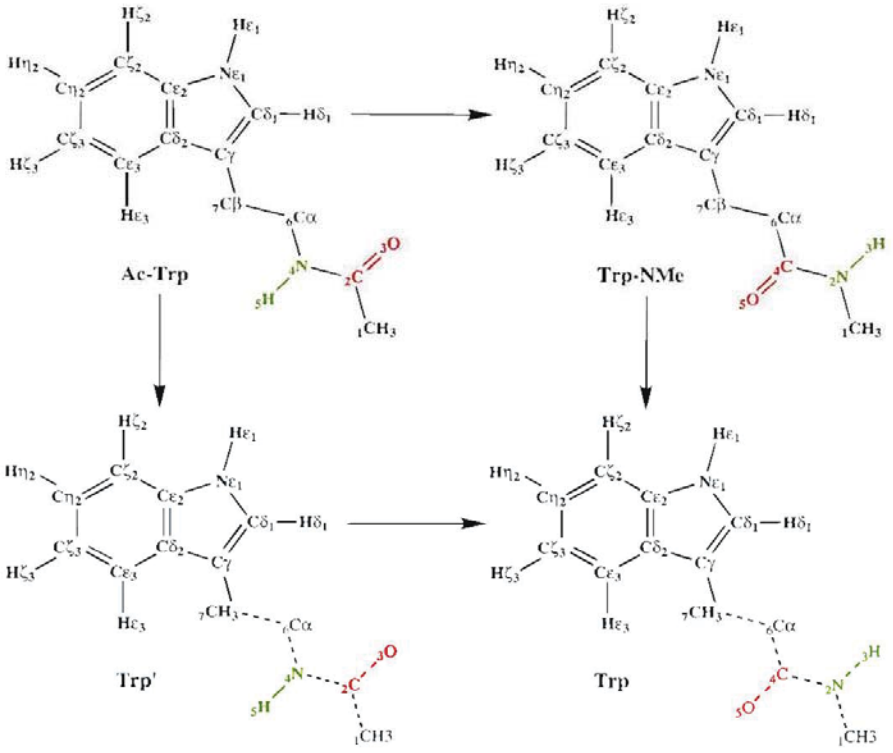
Calculating free energies and entropies

(overview)

- The calculation of free energies based on molecular simulations are affected by the following practical issues
  - → Equilibrium, finite sampling
  - → Numerical quadrature (TI)
  - → Cycle closure
  - → Atom creation or deletion
  - → Contribution of constraints
  - → Free energy components ?
  - → Choice of thermodynamic cycle and pathway
- Calculation of entropies based on molecular simulations

Choice of pathways and cycles

Choice of pathway: which is the most efficient ?



*X. Daura et al., JACS **118** (1996) 6285-6294*

Choice of pathways and cycles

$\langle \partial V / \partial \lambda \rangle_{\lambda'}$

Small "dummy interconversion" component !

Table 6. Results from Set 5 Calculations

	water						chloroform						vacuum	
	Trp-NMe → Trp		Ac-Trp → Trp′		Ac-Trp → Trp-NMe		Trp-NMe → Trp		Ac-Trp → Trp′		Ac-Trp → Trp-NMe		Trp′ → Trp[c]	
λ	simulation time (ps)	⟨∂V/∂λ⟩[a] (kJ/mol)	simulation time (ps)	⟨∂V/∂λ⟩[a] (kJ/mol)	simulation time (ps)	⟨∂V/∂λ⟩[a] (kJ/mol)	simulation time (ps)	⟨∂V/∂λ⟩[a] (kJ/mol)	simulation time (ps)	⟨∂V/∂λ⟩[a] (kJ/mol)	simulation time (ps)	⟨∂V/∂λ⟩[a] (kJ/mol)	simulation time (ps)	⟨∂V/∂λ⟩[a] (kJ/mol)
0.00	50	128.5 ± 1.9	50	131.7 ± 1.1	50	74.6 ± 1.1	50	115.9 ± 1.1	50	115.8 ± 1.0	50	46.2 ± 0.9	100	−0.14 ± 0.05
0.05	100	195.3 ± 2.3	100	188.2 ± 2.3	100	138.2 ± 2.2	100	124.6 ± 1.0	100	136.0 ± 1.2	50	58.9 ± 2.0	100	0.05 ± 0.07
0.10	100	194.4 ± 2.6	100	196.7 ± 2.8	125	148.9 ± 2.5	100	145.0 ± 1.4	100	137.6 ± 1.2	50	69.9 ± 1.6	100	−0.07 ± 0.06
0.20	100	147.2 ± 1.8	100	148.8 ± 1.5	100	100.7 ± 1.7	100	127.2 ± 1.7	100	138.5 ± 1.4	50	67.8 ± 3.1	100	−0.05 ± 0.04
0.30	50	106.8 ± 1.5	50	112.4 ± 1.1	50	62.3 ± 1.5	50	110.3 ± 1.4	50	112.5 ± 1.3	50	51.5 ± 2.2	100	0.02 ± 0.05
0.40	50	79.4 ± 1.5	50	87.9 ± 1.3	50	34.2 ± 1.1	50	85.5 ± 2.4	50	92.5 ± 2.0	50	30.9 ± 1.9	100	−0.15 ± 0.03
0.50	50	58.5 ± 1.4	50	67.4 ± 1.9	50	3.6 ± 1.4	50	71.5 ± 1.4	50	73.7 ± 2.3	50	2.5 ± 1.8	100	−0.11 ± 0.04
0.60	50	45.3 ± 2.0	50	48.1 ± 2.1	50	−24.0 ± 1.2	50	57.1 ± 1.6	50	59.5 ± 2.1	50	−15.8 ± 1.8	100	−0.16 ± 0.04
0.70	100	20.9 ± 2.4	100	27.3 ± 2.0	50	−58.6 ± 2.4	100	41.7 ± 2.1	100	42.5 ± 1.7	50	−46.2 ± 1.8	100	−0.17 ± 0.04
0.80	200	−1.5 ± 3.4	200	−10.5 ± 3.7	100	−96.6 ± 1.5	150	17.5 ± 3.2	150	9.1 ± 3.1	50	−64.0 ± 2.0	100	−0.49 ± 0.11
0.90	100	−10.3 ± 2.7	100	−12.5 ± 2.2	125	−142.1 ± 2.3	150	−1.9 ± 3.1	150	−6.4 ± 2.9	50	−68.6 ± 1.8	100	−0.55 ± 0.15
0.95	100	14.4 ± 1.6	100	9.4 ± 1.7	125	−145.2 ± 1.7	125	13.5 ± 2.3	125	9.3 ± 2.1	50	−64.0 ± 1.3	100	−0.43 ± 0.14
1.00	50	34.3 ± 1.2	50	29.1 ± 1.8	50	−77.7 ± 1.2	50	38.2 ± 2.0	50	32.5 ± 2.2	50	−45.6 ± 0.8	100	−1.02 ± 0.21
ΔG (kJ/mol)[b]		74.0 ± 2.2		75.9 ± 2.1		2.3 ± 1.8		72.6 ± 2.1		73.6 ± 2.0		2.5 ± 2.0		−0.21 ± 0.09

[a] The first 10 ps of simulation at each λ′ point is not considered for the calculation of the average. [b] ΔG is calculated by trapezoidal integration. [c] This column shows 13 of the 21 λ′ points used for the calculation of the integral.

1. Ac-Trp → Trp-NMe (direct): 2.3 ± 1.8 kJ/mol 1025 ps ← **More efficient pathway**

2. Ac-Trp → Trp': 75.9 ± 2.1 — 1100 ps
   Trp' → Trp: -0.21 ± 0.09 — 1.7 ± 3.0 kJ/mol — - — 2200 ps
   Trp → Trp-NMe: -74.0 ± 2.2 — 1100 ps

*X. Daura et al., JACS **118** (1996) 6285-6294*

Calculating free energies and entropies

(overview)

- The calculation of free energies based on molecular simulations are affected by the following practical issues
  - → Equilibrium, finite sampling
  - → Numerical quadrature (TI)
  - → Cycle closure
  - → Atom creation or deletion
  - → Contribution of constraints
  - → Free energy components ?
  - → Choice of thermodynamic cycle and pathway
- Calculation of entropies based on molecular simulations

Entropy calculations

Four Ways to Compute Entropy Differences

Coupling parameter λ approach

Hamiltonian is made function of λ: $H_a(\vec{p},\vec{r}) = H(\vec{p},\vec{r};\lambda_a) \rightarrow$ state a

$H_b(\vec{p},\vec{r}) = H(\vec{p},\vec{r};\lambda_b) \rightarrow$ state b

Free energy depends on λ: $A_{NVT}(\lambda) = -k_b T \ln\left(\left[h^{3N} N!\right]^{-1} \iint \exp\left(-H(\vec{p},\vec{r};\lambda)/k_b T\right) d\vec{p} d\vec{r}\right)$

Four methods to calculate entropy differences

- **METHOD 1**:

→ Calculate ΔF using **thermodynamic integration**

→ Calculate ΔU as a difference between the **end-state energies**

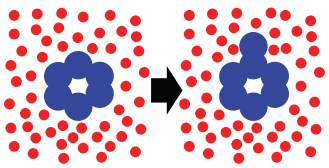
→ Apply the **Gibbs equation**

$$\Delta A_{ba}^{TI} = A(\lambda_b) - A(\lambda_a) = \int_{\lambda_a}^{\lambda_b} \frac{dA}{d\lambda} d\lambda = \int_{\lambda_a}^{\lambda_b} \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda} d\lambda$$

Can be calculated accurately (TI)

$$\Delta U_{ba}^{end} = U(\lambda_b) - U(\lambda_a) = \langle H \rangle_{\lambda_b} - \langle H \rangle_{\lambda_a}$$

Difficult to calculate accurately (end)



You drown in solvent-solvent interaction «noise» !

$$\Delta S_{ba}^{TI+end} = \frac{\Delta U_{ba}^{end} - \Delta A_{ba}^{TI}}{T}$$

Will also not be very accurate !

Four methods to calculate entropy differences

- **METHOD 2**:

→ Calculate ΔS directly using **thermodynamic integration**

→ From statistical mechanics (canonical ensemble)

$$F(\lambda) = -\beta^{-1} \ln Z = -\beta^{-1} \ln\left[\xi \iint \exp\left(-\beta \mathcal{H}\right)\right] \qquad \textit{free energy}$$

$$E(\lambda) = Z^{-1} \iint \mathcal{H} \exp\left(-\beta \mathcal{H}\right) = \left\langle \mathcal{H} \right\rangle_{\lambda} \qquad \textit{energy}$$

$$S(\lambda) = -k_B \beta \left[F(\lambda) - E(\lambda)\right] \qquad \textit{entropy}$$

→ TI formula for the entropy

$$\frac{\partial S(\lambda)}{\partial \lambda} = \ldots = -k_B \beta^2 \left[\left\langle \mathcal{H} \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_{\lambda} - \left\langle \mathcal{H} \right\rangle_{\lambda} \left\langle \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_{\lambda}\right]$$

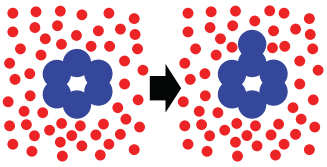
→ So

$$\Delta S_{ba}^{TI} = \frac{1}{k_b T^2} \int_{\lambda_a}^{\lambda_b} \left[\left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda} \left\langle H \right\rangle_{\lambda} - \left\langle \frac{\partial H}{\partial \lambda} H \right\rangle_{\lambda}\right] d\lambda$$

correlation between $\frac{\partial H}{\partial \lambda}$ and H

only λ-dependent terms — all terms

Difficult to calculate accurately (TI for S)



You drown again in solvent-solvent interaction «noise» !

Four methods to calculate entropy differences

- **METHOD 3**:

→ Calculate ΔS from ΔF at two temperatures using **finite difference**

→ From thermodynamics

$$S = -\left(\frac{\partial A}{\partial T}\right)_{N,T}$$

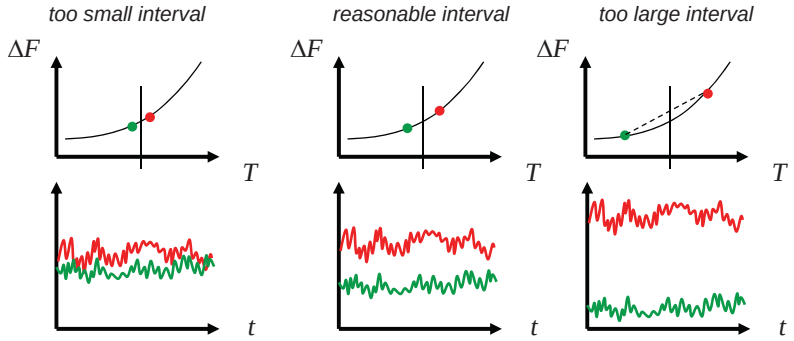
→ So

$$\Delta S_{ba}^{TI+FD} = -\frac{\Delta A_{ba}^{TI}(T + \Delta T) - \Delta A_{ba}^{TI}(T - \Delta T)}{2\Delta T}$$

Difference between equally accurate values

→ can work, but you need very long simulations and a careful choice of ΔT

→ The choice of the finite-difference interval is crucial!



Four methods to calculate entropy differences

- **METHOD 4**:

→ Calculate ΔS_{uv} (solute-solvent **«partial-»entropy** change) using thermodynamic integration

→ Calculate ΔU_{vv} (solvent-solvent energy change) from **end-state simulations**

$$\Delta S_{ba} = \frac{1}{k_B T^2}\int_{\lambda_a}^{\lambda_b}\left[\left\langle\frac{\partial H_{uv}}{\partial\lambda}\right\rangle_\lambda\left\langle H_{uv}\right\rangle_\lambda - \left\langle\frac{\partial H_{uv}}{\partial\lambda}H_{uv}\right\rangle\right]d\lambda + \frac{1}{T}\left[\left\langle H_{vv}\right\rangle_{\lambda_b} - \left\langle H_{vv}\right\rangle_{\lambda_a}\right]$$

$$= \Delta S_{ba}^{TI,uv} + \frac{\Delta U_{ba}^{end,vv}}{T}$$

accurate ↑ **only solute-solvent terms**

not so accurate ↑ **all solvent terms**

solvent: *v*
solute: *u*

→ You normally stop at ΔS_{uv} and call it an «altervative form» of entropy
(I am skeptical! – see later)

Entropy calculations

Comparison of

1. Excess free energy, entropy of water

2. Hydration free energy, entropy of water

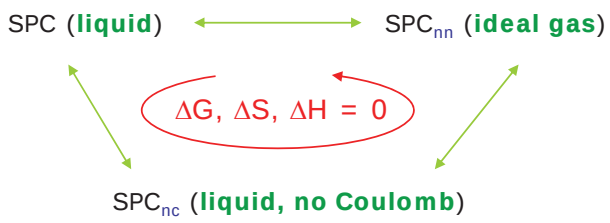
using four different methods

C. Peter et al. J. Chem. Phys. ***120*** *(2004) 2652-2661*

Three models or Hamiltonians:

1. SPC Model: ***Coulomb*** plus ***van der Waals*** interaction
2. SPC_{nc} Model: ***no Coulomb*** interaction
3. SPC_{nn} Model: ***no*** (non-bonded) ***interaction***

Thermodynamic cycle



System

1000 H_2O molecules
periodic boundary conditions
T = 280K, 300K, 320K
simulations = 100-600ps
NVT ↔ NPT

Change:
1 H_2O → hydration
all H_2O → excess **more accurate**

Entropy calculations

Free Energy and Entropy of Water

1. Entropy difference via *free energy* (TI) and *energy*
2. Entropy difference *directly* via TI

4. via *solute-solvent*

m	transition	tbc	T [K]	t [ps]	$<V>$ [nm^3]	ΔA^{TI} (ΔG^{TI}) [kJ mol^{-1}]	ΔU^{end}_{pot} (ΔH^{end}_{pot}) [kJ mol^{-1}]	ΔS^{TI} [A] [J K^{-1} mol^{-1}] (method 2)	ΔS^{end} [B] [J K^{-1} mol^{-1}] (method 1)	$\Delta U^{end,vv}_{pot}$ ($\Delta H^{end,vv}_{pot}$) [kJ mol^{-1}]	$\Delta S^{TI,uuv}$ [J K^{-1} mol^{-1}] (method 4)	$\Delta S^{TI,all}$ [C] [J K^{-1} mol^{-1}] (method 4)
all	SPC → SPC$_{nn}$	NVT	300	100	30.73	23.2	41.3	47.0	60.3			
	SPC → SPC$_{nc}$		300	100	30.73	28.7	39.4	22.3	35.7			
	SPC$_{nc}$ → SPC$_{nn}$		300	100	30.73	-5.3	1.9	22.9	24.0			
single	SPC → SPC$_{nn}$	NVT	300	600	30.73	23.1	40.7	10.7	58.7	-41.4	162.5	24.5
			300	200	30.73	23.0	45.3	22.1	74.3	-37.0	157.8	34.5
			280	200	30.73	25.0	29.6	29.7	16.4	-56.4	175.2	-26.2
			320	200	30.73	21.8	48.7	40.3	84.1	-32.1	149.2	48.9
		NPT	300	600	30.80	23.4	59.5	92.4	120.3	-23.6	161.7	83.0
			300	200	30.81	23.3	76.2	62.9	176.3	-8.0	166.5	139.8
			280	200	30.36	24.7	39.7	51.7	53.6	-45.9	171.5	7.6
			320	200	31.34	22.4	59.6	63.7	116.2	-21.3	144.8	78.2
single	SPC → SPC$_{nc}$	NVT	300	600	30.73	31.3	52.1	27.8	69.3	-23.1	117.5	40.5
			300	200	30.73	31.5	64.6	69.4	110.3	-10.7	119.0	83.3
		NPT	300	600	30.82	30.9	52	12.9	70.3	-24.2	117.5	36.8
			300	200	30.82	31.1	58.4	5.6	91.0	-18.8	115.4	52.7
			280	200	30.39	31.8	75	24.9	154.3	-3.3	125.6	113.8
			320	200	31.35	30.6	61.7	19.7	97.2	-12.5	106.4	67.3
single	SPC$_{nc}$ → SPC$_{nn}$	NVT	300	600	30.73	-7.7	9.7	27.2	58.0	2.7	46.0	55.0
			300	200	30.73	-7.4	-1.7	65.5	19.0	-8.7	45.3	16.3
		NPT	300	600	30.82	-7.8	17	33.9	82.7	10.0	45.9	79.2
			300	200	30.82	-8.0	15.1	25.6	77.0	8.0	45.9	72.6
			280	200	30.37	-6.4	23.8	37.0	107.9	16.8	42.4	102.4
			320	200	31.34	-7.7	-14.4	24.7	-20.9	-21.1	41.0	-24.9

Table I:

Reference: J.Chem.Phys. ***120*** *(2004) 2652-2661*

Entropy calculations

Free Energy and Entropy of Water

3. Entropy difference via finite temperature difference

transition	tbc	$\Delta S^{\Delta T}$ [J K^{-1} mol^{-1}]	
SPC → SPC$_{nn}$	NVT	80	
SPC → SPC$_{nn}$	NPT	58	close
SPC → SPC$_{nc}$	NPT	30	63
SPC$_{nc}$ → SPC$_{nn}$	NPT	33	

Table II:

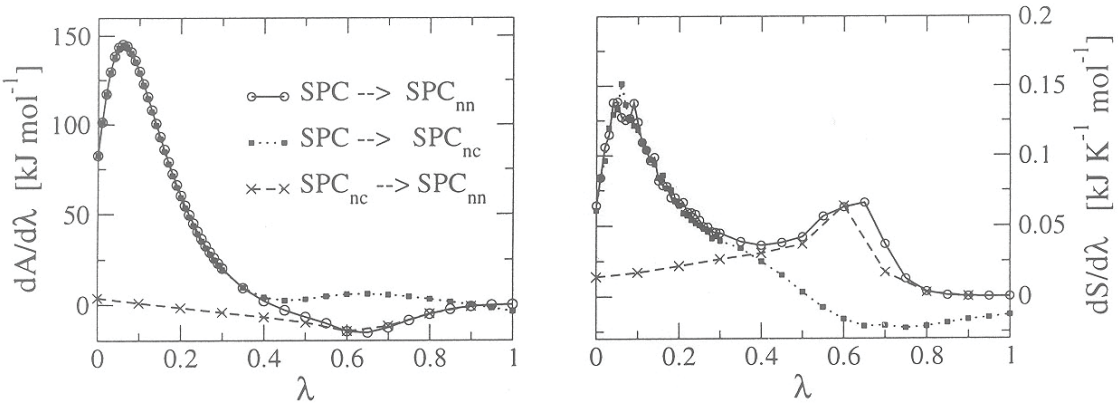
... and close to experimental value of 51

Entropy calculations

All 1000 H_2O Molecules Changed

Excess free energy **Entropy**

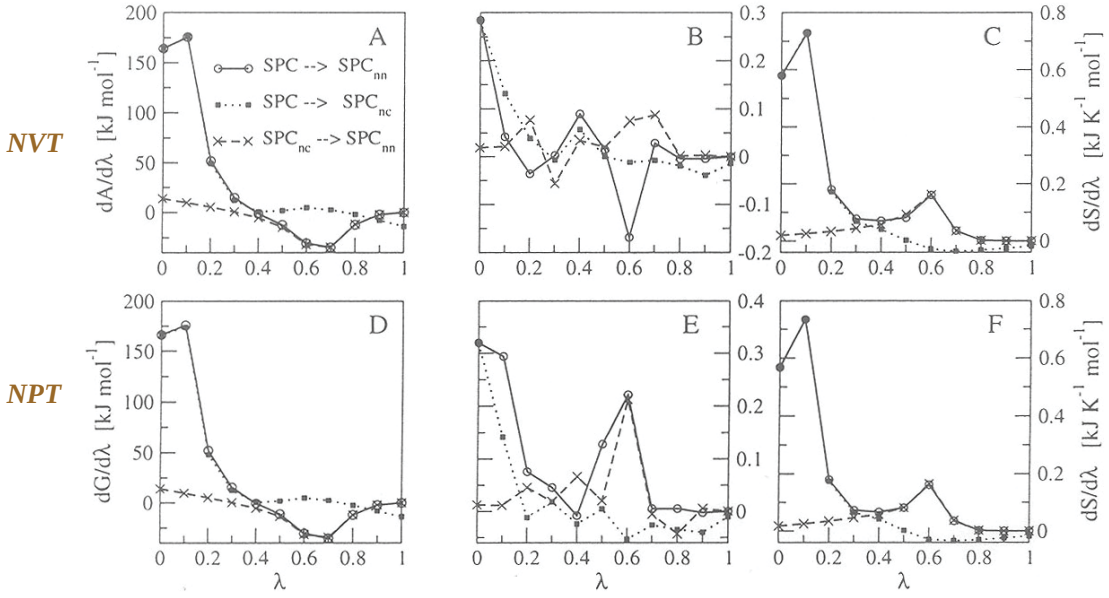
TI Method 2 **TI**



Entropy calculations

A Single H_2O Molecule Changed

ΔA via TI ΔS via TI ΔS^{uv} via TI



same pattern as for 1000 H_2O being changed

erratic *not converged*

same pattern as for 1000 H_2O being changed

Four methods to calculate entropy differences

- **METHOD 4**:

  → Calculate ΔS_{uv} (solute-solvent **«partial-»entropy** change) using thermodynamic integration

  → Calculate ΔU_{vv} (solvent-solvent energy change) using from **end-state simulations**

$$\Delta S_{ba} = \frac{1}{k_B T^2}\int_{\lambda_a}^{\lambda_b}\left[\left\langle\frac{\partial H_{uv}}{\partial\lambda}\right\rangle_\lambda\left\langle H_{uv}\right\rangle_\lambda - \left\langle\frac{\partial H_{uv}}{\partial\lambda}H_{uv}\right\rangle\right]d\lambda + \frac{1}{T}\left[\left\langle H_{vv}\right\rangle_{\lambda_b} - \left\langle H_{vv}\right\rangle_{\lambda_a}\right]$$

$$= \Delta S_{ba}^{TI,uv} + \frac{\Delta U_{ba}^{end,vv}}{T}$$

accurate ↑ only solute-solvent terms

not so accurate ↑ all solvent terms

solvent: *v*
solute: *u*

→ You normally stop at ΔS_{uv} and call it an «altervative form» of entropy
(I am skeptical! – see later)

Partial energies and entropies

- Consider a λ-dependent Hamiltonian $\mathcal{H} = \mathcal{H}(\boldsymbol{r}, \boldsymbol{p}; \lambda)$

  → From statistical mechanics (canonical ensemble)

$$\Rightarrow F(\lambda) = -\beta^{-1}\ln Z = -\beta^{-1}\ln\left[\xi\iint\exp\left(-\beta\mathcal{H}\right)\right] \qquad \textit{free energy}$$

$$\Rightarrow E(\lambda) = Z^{-1}\iint\mathcal{H}\exp\left(-\beta\mathcal{H}\right) = \left\langle\mathcal{H}\right\rangle_\lambda \qquad \textit{energy}$$

$$\Rightarrow S(\lambda) = -k_B\beta\left[F(\lambda) - E(\lambda)\right] \qquad \textit{entropy}$$

- The thermodynamic integration (TI) equations for the corresponding λ-derivatives are

$$\Rightarrow \frac{\partial F(\lambda)}{\partial\lambda} = -\beta^{-1}Z^{-1}\,\xi\iint(-\beta)\left(\frac{\partial\mathcal{H}}{\partial\lambda}\right)\exp\left(-\beta\mathcal{H}\right) = \left\langle\frac{\partial\mathcal{H}}{\partial\lambda}\right\rangle_\lambda$$

$$\Rightarrow \frac{\partial E(\lambda)}{\partial\lambda} = Z^{-2}\left\{Z\left[\iint\left(\frac{\partial\mathcal{H}}{\partial\lambda}\right)\exp\left(-\beta\mathcal{H}\right) + \iint(-\beta)\mathcal{H}\left(\frac{\partial\mathcal{H}}{\partial\lambda}\right)\exp\left(-\beta\mathcal{H}\right)\right]\right.$$
$$\left. -\left[\iint\mathcal{H}\exp\left(-\beta\mathcal{H}\right)\right]\left[\iint(-\beta)\left(\frac{\partial\mathcal{H}}{\partial\lambda}\right)\exp\left(-\beta\mathcal{H}\right)\right]\right\}$$
$$= \left\langle\frac{\partial\mathcal{H}}{\partial\lambda}\right\rangle_\lambda - \beta\left\langle\mathcal{H}\frac{\partial\mathcal{H}}{\partial\lambda}\right\rangle_\lambda + \beta\left\langle\mathcal{H}\right\rangle_\lambda\left\langle\frac{\partial\mathcal{H}}{\partial\lambda}\right\rangle_\lambda$$

$$\Rightarrow \frac{\partial S(\lambda)}{\partial\lambda} = -k_B\beta\left[\frac{\partial F(\lambda)}{\partial\lambda} - \frac{\partial E(\lambda)}{\partial\lambda}\right] = -k_B\beta\left[\left\langle\frac{\partial\mathcal{H}}{\partial\lambda}\right\rangle_\lambda - \left\langle\frac{\partial\mathcal{H}}{\partial\lambda}\right\rangle_\lambda + \beta\left\langle\mathcal{H}\frac{\partial\mathcal{H}}{\partial\lambda}\right\rangle_\lambda - \beta\left\langle\mathcal{H}\right\rangle_\lambda\left\langle\frac{\partial\mathcal{H}}{\partial\lambda}\right\rangle_\lambda\right]$$
$$= -k_B\beta^2\left[\left\langle\mathcal{H}\frac{\partial\mathcal{H}}{\partial\lambda}\right\rangle_\lambda - \left\langle\mathcal{H}\right\rangle_\lambda\left\langle\frac{\partial\mathcal{H}}{\partial\lambda}\right\rangle_\lambda\right]$$

Partial energies and entropies

- For a λ-change from zero (A-state) to one (B-state), one has thus

$$\Delta F = \int_0^1 d\lambda' \left\langle \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_{\lambda'} \qquad \Delta U = -\beta \int_0^1 d\lambda' \left[-\beta^{-1} \left\langle \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_{\lambda'} + \left\langle \mathcal{H} \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_{\lambda'} - \left\langle \mathcal{H} \right\rangle_{\lambda} \left\langle \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_{\lambda'} \right] = \left\langle \mathcal{H} \right\rangle_1 - \left\langle \mathcal{H} \right\rangle_0$$

$$\Delta S = -k_B \beta^2 \int_0^1 d\lambda' \left[\left\langle \mathcal{H} \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_{\lambda'} - \left\langle \mathcal{H} \right\rangle_{\lambda'} \left\langle \frac{\partial \mathcal{H}}{\partial \lambda} \right\rangle_{\lambda'} \right]$$

- Now consider a Hamiltonian of the form $\mathcal{H}(\lambda) = \mathcal{H}_d(\lambda) + \mathcal{H}_i$

all λ-dependent terms: $\mathcal{H}_d(\lambda)$

all λ-independent terms: $\mathcal{H}_i$

→ The above expressions become

$$\Delta F = \int_0^1 d\lambda' \left\langle \frac{\partial \mathcal{H}_d}{\partial \lambda} \right\rangle_{\lambda'} \qquad \Delta U_d = -\beta \int_0^1 d\lambda' \left[-\beta^{-1} \left\langle \frac{\partial \mathcal{H}_d}{\partial \lambda} \right\rangle_{\lambda'} + \left\langle \mathcal{H}_d \frac{\partial \mathcal{H}_d}{\partial \lambda} \right\rangle_{\lambda'} - \left\langle \mathcal{H}_d \right\rangle_{\lambda'} \left\langle \frac{\partial \mathcal{H}_d}{\partial \lambda} \right\rangle_{\lambda'} \right] = \left\langle \mathcal{H}_d \right\rangle_1 - \left\langle \mathcal{H}_d \right\rangle_0$$

$$\Delta U = \Delta U_d + \Delta U_i$$

$$\Delta S = \Delta S_d + \Delta S_i \qquad \Delta S_d = -k_B \beta^2 \int_0^1 d\lambda' \left[\left\langle \mathcal{H}_d \frac{\partial \mathcal{H}_d}{\partial \lambda} \right\rangle_{\lambda'} - \left\langle \mathcal{H}_d \right\rangle_{\lambda'} \left\langle \frac{\partial \mathcal{H}_d}{\partial \lambda} \right\rangle_{\lambda'} \right]$$

$$\Delta U_i = T\Delta S_i = -\beta \int_0^1 d\lambda' \left[\left\langle \mathcal{H}_i \frac{\partial \mathcal{H}_d}{\partial \lambda} \right\rangle_{\lambda'} - \left\langle \mathcal{H}_i \right\rangle_{\lambda'} \left\langle \frac{\partial \mathcal{H}_d}{\partial \lambda} \right\rangle_{\lambda'} \right] = \left\langle \mathcal{H}_i \right\rangle_1 - \left\langle \mathcal{H}_i \right\rangle_0$$

→ Observations

The energy and entropy contributions deriving from the λ-independent terms (in correlation with the λ-dependent ones) cancel out in the Gibbs equation

$$\Delta U_i = T\Delta S_i \quad \Rightarrow \quad \Delta F = \Delta U - T\Delta S = \Delta U_d - T\Delta S_d$$

This is nice in situations where we have few λ-dependent terms (e.g. solute-solvent interactions) and many λ-independent ones (e.g. solvent-solvent interactions)

→ in these situations, the cancelling terms would be very difficult to calculate !

$\mathcal{H}_i$ ***very large***

Partial energies and entropies

- Based on these considerations

→ We define

$$\Delta U_d = \left\langle \mathcal{H}_d \right\rangle_1 - \left\langle \mathcal{H}_d \right\rangle_0$$

partial-energy change of the process

$$\Delta S_d = -k_B \beta^2 \int_0^1 d\lambda' \left[\left\langle \mathcal{H}_d \frac{\partial \mathcal{H}_d}{\partial \lambda} \right\rangle_{\lambda'} - \left\langle \mathcal{H}_d \right\rangle_{\lambda'} \left\langle \frac{\partial \mathcal{H}_d}{\partial \lambda} \right\rangle_{\lambda'} \right]$$

partial-entropy change of the process

I add the word «partial» although it is often omitted in the literature

→ We then write the Gibbs equation as

$$\Delta F = \Delta U_d - T\Delta S_d$$

and try to interpret it in this form...

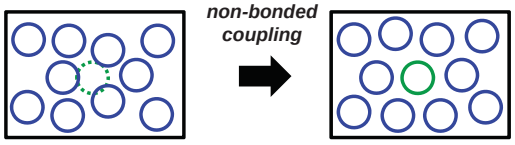
→ And we forget about the other terms, which are cancelling and hard to calculate

$$\Delta U_i = T\Delta S_i = ?$$

- Typical example: solvation free energy

solute (u) with intramolecular but no intermolecular interactions

solvent (v)



non-bonded coupling

$$\mathcal{H}(\lambda) = \mathcal{H}_{uv}(\lambda) + \mathcal{H}_{uu} + \mathcal{H}_{vv}$$

$$\Delta F = \Delta U_{uv} - T\Delta S_{uv}$$

"solute-solvent" partial solvation energy

"solute-solvent" partial solvation entropy

Partial energies and entropies

- From Wilfred's original slides

$$\Delta F = \Delta U_{uv} - T\Delta S_{uv}$$

They yield insight into *enthalpic* and *entropic* driving forces, are computable, but *not* measurable

1. **Measurable quantities**: ΔG^{assoc}, ΔH^{assoc} , ΔS^{assoc}

   ΔH^{assoc} and ΔS^{assoc} contain ***exactly*** compensating terms

   $$\Delta H^{assoc} = \Delta H_{uv}^{assoc} + \Delta H_{vv}^{assoc}$$
   $$\text{equal}$$
   $$T\Delta S^{assoc} = T\Delta S_{uv}^{assoc} + T\Delta S_{vv}^{assoc}$$
   $$\Delta G^{assoc} = \Delta H^{assoc} - T\Delta S^{assoc}$$

   These compensating terms **may *mask* real driving forces of association**

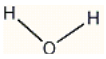
2. **Real driving terms** ΔH_{uv}^{assoc} *and* $T\Delta S_{uv}^{assoc}$ are *not* measurable, but are computable and **do explain driving forces**

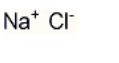
Seems like a paradox, right ? A non-measurable quantity cannot have an influence on a measurable property ! So, either these quantites are actually measurable, or they cannot have any predictive power whatsoever in the real world... [my opinion: the latter is true !]

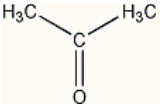
Entropy calculations

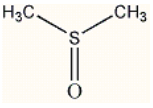
Nico van der Vegt

mole fraction









	N_C	N_{H2O}	x_C	ρ_C (M)	t_{MD} (ns)	ε_{RF}	T (K)	V (nm^3)	ρ (g/cm^3)
H_2O (SPC)	0	739	0.0	0.0	21.0	54	302.3	22.73	0.973
H_2O (SPC/E)	0	739	0.0	0.0	1.0	54	302.6	22.22	0.994
$NaCl/H_2O$ (SPC/E)	8	1000	0.008	0.44	1.6	54	302.9	30.22	1.016
	16	1000	0.016	0.87	1.6	54	302.9	30.43	1.034
	32	1000	0.031	1.72	1.6	54	303.0	30.94	1.067
	64	1000	0.060	3.26	1.6	54	302.7	32.57	1.109
	125	1000	0.111	5.80	1.6	54	302.2	35.80	1.174
Acetone/H_2O (SPC/E)	168	1512	0.10	4.36	1.0	50	299.3	63.93	0.961
	275	1100	0.20	7.14	1.0	39	298.4	63.96	0.929
	396	594	0.40	10.36	1.0	25	297.7	63.49	0.881
	432	432	0.50	11.32	1.0	28	297.5	63.36	0.862
	462	308	0.60	12.07	1.0	23	297.4	63.54	0.846
	500	125	0.80	13.06	1.0	18	297.4	63.55	0.818
	513	57	0.90	13.43	1.0	16	297.3	63.42	0.807
	520	0	1.00	13.71	1.0	15	297.3	62.97	0.796
$DMSO/H_2O$ (SPC)	43	812	0.05	2.37	1.3	77.4	299.7	30.14	0.991
	91	812	0.10	4.22	1.3	75.7	299.2	35.78	1.009
	188	812	0.19	6.63	1.2	72.4	298.7	47.08	1.034
	241	651	0.27	8.30	1.3	70.1	298.5	48.22	1.052
	349	651	0.35	9.54	2.2	67.2	298.3	60.75	1.066
	478	522	0.48	11.05	2.2	63.0	298.3	71.83	1.081
	331	186	0.64	12.37	2.3	57.9	298.2	44.43	1.092
	814	186	0.81	13.31	2.1	52.0	298.3	101.54	1.095
	512	0	1.00	14.00	2.2	46.0	298.5	60.73	1.094

«Solvent» considered: mixtures at various compositions

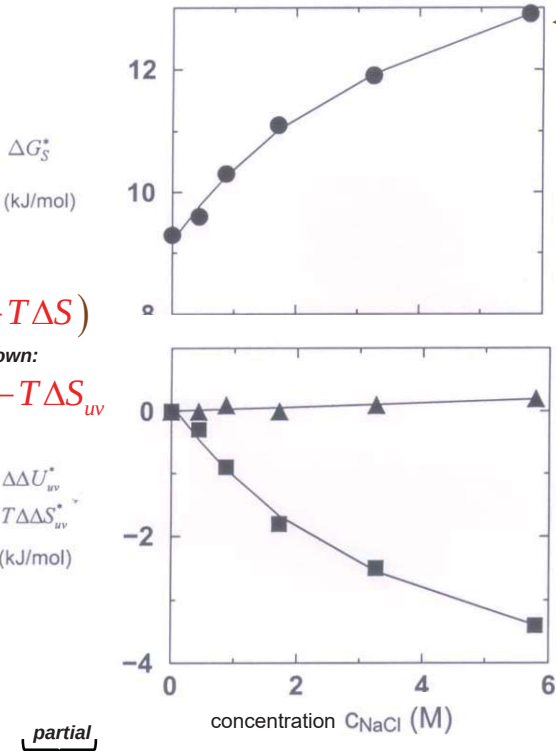
«Solute» considered: methane

Reference: J.Phys.Chem.B. 108 (2004) 1056

Entropy calculations

Solvation of Methane in Na^+Cl^- Solutions

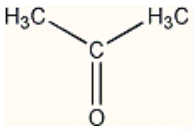
Na^+Cl^-

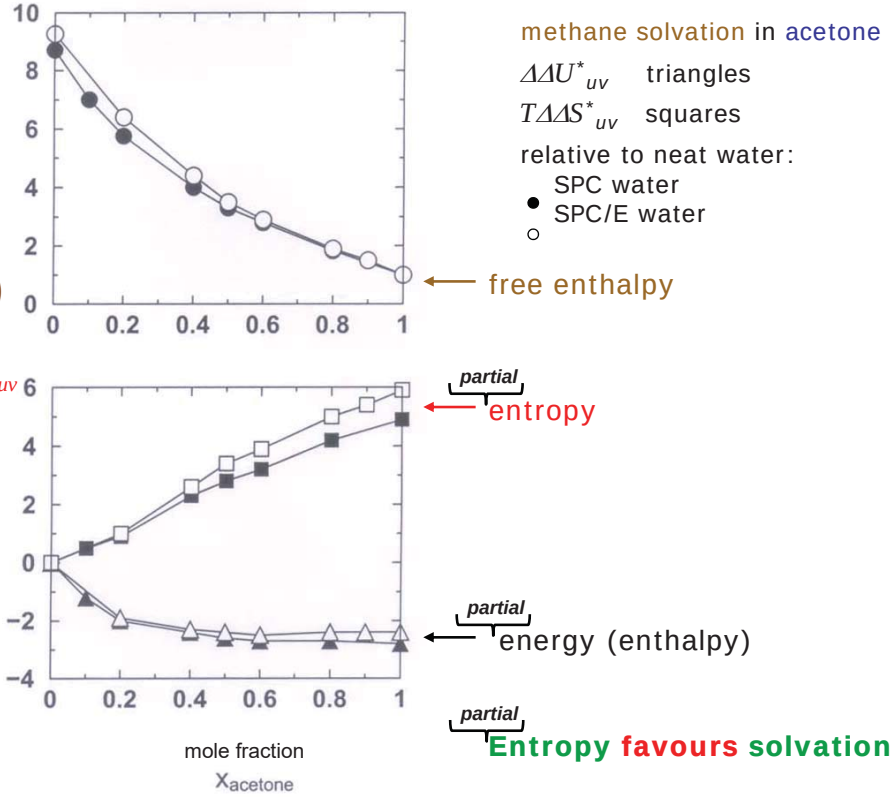


Entropy disfavours solvation increasingly with salt concentration (non-linear)

Entropy calculations

Solvation of Methane in Acetone Solution

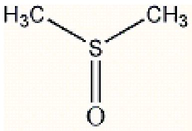




Entropy favours solvation

Entropy calculations

Solvation of Methane in Dimethylsulfoxide (DMSO) Solutions



$(\Delta G = \Delta U - T\Delta S)$ *Actually shown:* $\Delta G = \Delta U_{uv} - T\Delta S_{uv}$

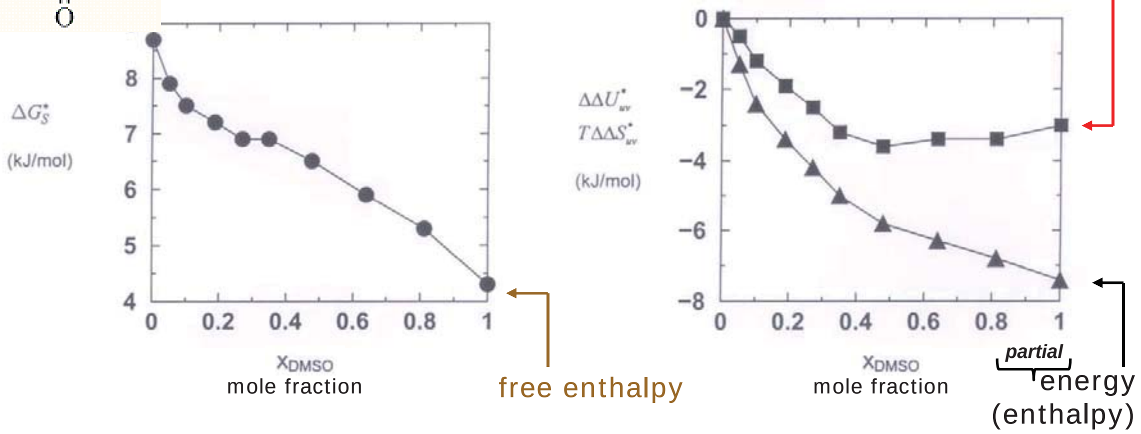


Figure 5. Methane solvation Gibbs energy (upper panel), and solute-solvent energy $\Delta\Delta U_{uv}^*$ (triangles) and entropy $T\Delta\Delta S_{uv}^*$ (squares) relative to neat water (lower panel) versus the dimethyl sulfoxide mole fraction of the solution.

partial **Energy favours solvation (non-linearly)**

Reference: J. Chem. Phys. B. 108 (2004) 1056

Entropy calculations

ΔG_S ΔU_{uv} $T\Delta S_{uv}$ Relative to Solvation in Pure Water

mole fraction ↓ — different models (Acetone(I), Acetone(II))

Solute	NaCl (11%)			Urea (15%)			DMSO (10%)#			Acetone(I) (10%)			Acetone(I) (50%)			Acetone(II) (10%)		
	$\Delta\Delta G_s$	$\Delta\Delta U_{uv}$	$T\Delta\Delta S_{uv}$	$\Delta\Delta G_s$	$\Delta\Delta U_{uv}$	$T\Delta\Delta S_{uv}$	$\Delta\Delta G_s$	$\Delta\Delta U_{uv}$	$T\Delta\Delta S_{uv}$	$\Delta\Delta G_s$	$\Delta\Delta U_{uv}$	$T\Delta\Delta S_{uv}$	$\Delta\Delta G_s$	$\Delta\Delta U_{uv}$	$T\Delta\Delta S_{uv}$	$\Delta\Delta G_s$	$\Delta\Delta U_{uv}$	$T\Delta\Delta S_{uv}$
Helium	2.0	−0.4	−2.4	0.8	−0.1	−0.9	0.0	−0.2	−0.2	−0.4	−0.2	0.2	−1.7	−0.6	1.1	−0.1	−0.2	−0.1
Neon	2.8	0.2	−2.6	0.6	−0.6	−1.2	−0.3	−0.7	−0.4	−0.7	−0.4	0.3	−2.4	−1.0	1.4	−0.4	−0.6	−0.2
Argon	3.7	−0.1	−3.8	0.3	−1.8	−2.1	−0.9	−1.8	−0.9	−1.3	−1.0	0.3	−4.4	−2.1	2.3	−1.1	−1.7	−0.6
Krypton	4.0	−0.2	−4.2	−0.2	−2.6	−2.4	−1.2	−2.3	−1.1	−1.6	−1.2	0.4	−5.2	−2.6	2.6	−1.5	−2.3	−0.8
Xenon	4.4	−1.2	−5.6	−0.7	−3.5	−2.8	−1.5	−2.7	−1.2	−2.2	−1.7	0.5	−6.8	−3.6	3.2	−2.1	−3.4	−1.3
Methane	4.2	−0.2	−4.4	0.1	−2.3	−2.4	−1.1	−2.1	−1.0	−1.7	−1.2	0.5	−5.4	−2.6	2.8	−1.4	−2.2	−0.8
Ethane	5.6	−0.8	−6.4	−0.7	−4.2	−3.5	−2.8	−4.4	−1.6	−2.9	−2.4	0.5	−8.4	−4.6	3.8	−2.6	−4.3	−1.7
Propane	5.4	−0.3	−5.7	−2.2	−5.1	−2.9	−4.8	−5.8	−1.0	−5.8	−3.9	1.9	−13.0	−6.5	6.5	−4.8	−6.2	−1.4
n-Butane	7.4	−1.2	−8.6	−2.6	−7.6	−5.0	−5.8	−7.9	−2.1	−6.4	−5.6	0.8	−15.7	−9.3	6.4	−4.3	−6.0	−1.7
iso-Butane	5.7	−1.2	−6.9	−2.2	−4.3	−2.1	−6.5	−6.7	−0.2	−6.7	−4.9	1.8	−15.2	−7.7	7.5	−5.9	−7.7	−1.8
neo-Pentane	8.2	−0.6	−8.8	−2.4	−8.4	−6.0	−6.0	−8.3	−2.3	−6.7	−4.3	2.4	−15.8	−7.1	8.7	−5.3	−9.7	−4.4

dominant — counteracts enthalpy — changes sign — co-act — enthalpy and entropy — counteract

relative and absolute contributions *(partial)* **enthalpy** / *(partial)* **entropy** **do vary**

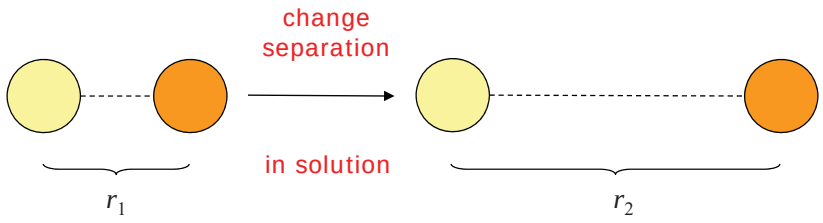
relative values of $\Delta\Delta U_{uv}$, $T\Delta\Delta S_{uv}$ change, $\Delta\Delta G_s$ not so much

Reference: Chem. Phys. Chem. 5 (2004) 144

Entropy calculations

Free energy, enthalpy and entropy of molecular association

Calculation of a free energy as function of a distance (r)



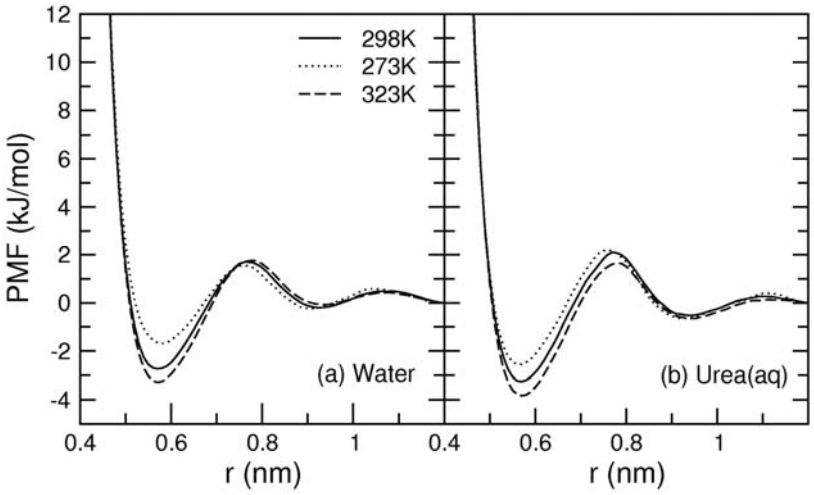
Two neo-pentane molecules in water

Association can be considered as ***solvation*** at different distances,

because gas-phase ΔG, ΔH, ΔS are easily calculated.

Entropy calculations

Free energy of a *neo*-pentane pair as function of distance in water and in 6.9 mol/l urea solution



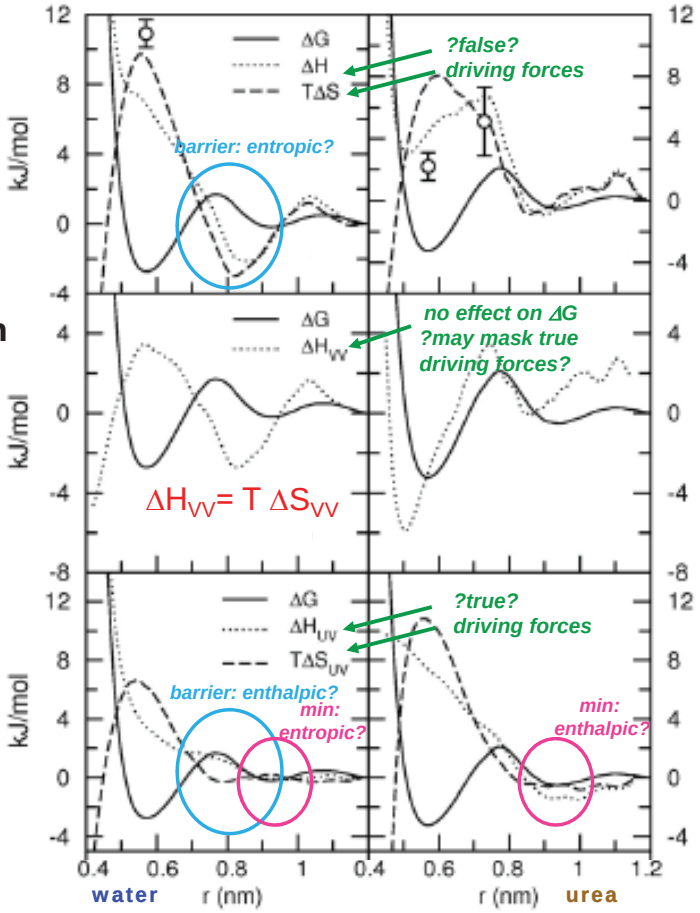
The contact minimum becomes deeper with increasing temperature

➡ ***neo*-pentane association is entropic in both systems**

Now, we talk about the real entropy (negative derivative of the free energy with respect to temperature)

Entropy calculations

Thermodynamics of a *neo*-pentane pair as function of distance in water and in 6.9 mol/l urea solution



The *free energies* (ΔG) are ***similar*** in water and in urea, whereas the

partial *enthalpy* (ΔH_{uv}) and

partial *entropy* ($T\Delta S_{uv}$) contributions are ***different***.

The *solvent reorganisation enthalpies* ($\Delta H_{vv}=T\Delta S_{vv}$) are ***quite different***.

J. Phys. Chem. B 110 (2006) 12852-12855

Partial energies and entropies

- From Wilfred's original slides

$$\Delta F = \Delta U_{uv} - T\Delta S_{uv}$$

They yield insight into *enthalpic* and *entropic* driving forces, are computable, but *not* measurable

1. **Measurable quantities**: ΔG^{assoc}, ΔH^{assoc}, ΔS^{assoc}

ΔH^{assoc} and ΔS^{assoc} contain ***exactly*** compensating terms

$$\Delta H^{assoc} = \Delta H^{assoc}_{uv} + \Delta H^{assoc}_{vv}$$
$$\updownarrow \text{ equal}$$
$$T\Delta S^{assoc} = T\Delta S^{assoc}_{uv} + T\Delta S^{assoc}_{vv}$$

$$\Delta G^{assoc} = \Delta H^{assoc} - T\Delta S^{assoc}$$

These compensating terms **may *mask* real driving forces of association**

2. **Real driving terms** ΔH^{assoc}_{uv} *and* $T\Delta S^{assoc}_{uv}$ are *not* measurable, but are computable and **do explain driving forces**

Examples:

1. Barrier to neo-pentane self association in water is:
   - *measurement*: entropic — ***based on real quantities***
   - *driving forces*: enthalpic — ***based on partial quantities***

Again, what is meant here ? Is a driving force a «dream in the mind of the chemist», or an experimentally measurable quantity ?

2. Solvent separated minimum free energy configuration of two neo-pentanes in water or in 6.9 mol/l urea is:
   - entropic in *water*
   - enthalpic in *urea*

based on partial quantities

based on «measured» quantities: enthalpy in urea, balance between both distance dependences in water

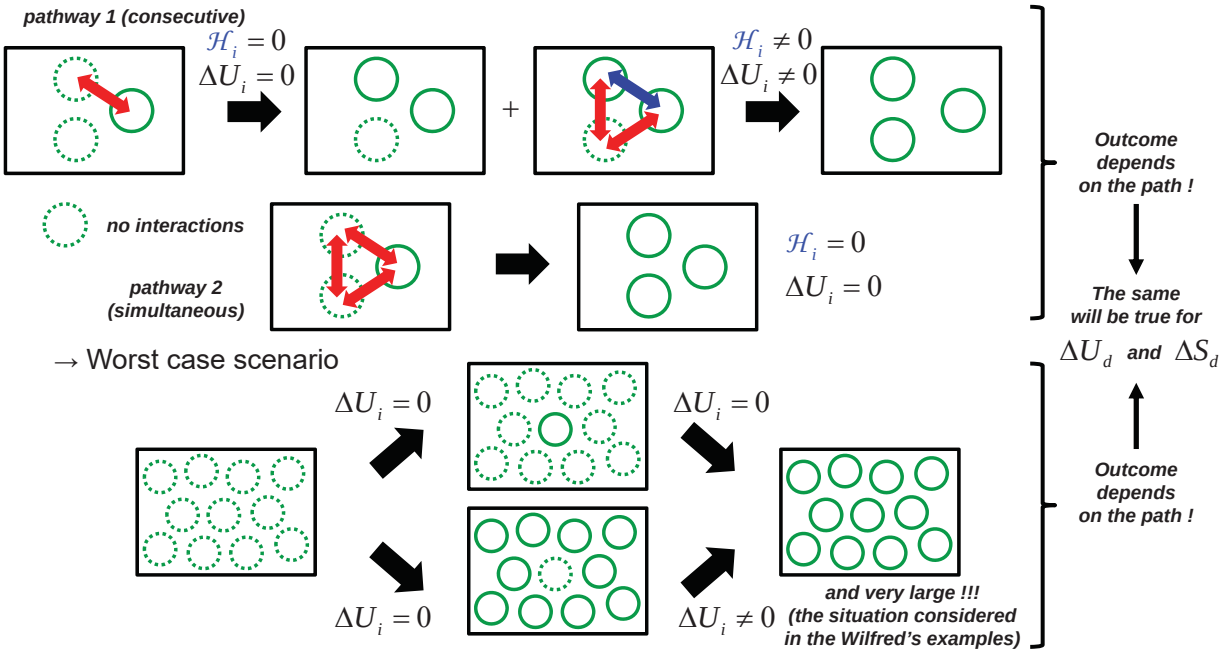
But wait a minute...

- We said a moment ago that **free-energy components** are **meaningless** because **pathway-dependent**; isn't it the same for the partial energies and entropies ?

$$\Delta U_i = T\Delta S_i = -\beta\int_0^1 d\lambda' \left[\left\langle \mathcal{H}_i \frac{\partial \mathcal{H}_d}{\partial \lambda} \right\rangle_{\lambda'} - \left\langle \mathcal{H}_i \right\rangle_{\lambda} \left\langle \frac{\partial \mathcal{H}_d}{\partial \lambda} \right\rangle_{\lambda'} \right]$$

Interaction in $\mathcal{H}_d$
Interaction in $\mathcal{H}_i$

→ Consider the following process for monoatomic species (*e.g.* argon)



→ Worst case scenario

Partial energies and entropies

- Phil's objections to "partial" quantities

→ Energy and entropy have well-defined meanings in thermodynamics

$\Delta F = \Delta U - T\Delta S$ Gibbs equation

$\left.\begin{matrix}\Delta U \\ \Delta S\end{matrix}\right\}$ are path-independent (*U* and *S* are state functions)

$\left(\frac{\partial \Delta F}{\partial T}\right)_{N,V} = -\Delta S$ entropy tells us how equilibria shift with temperature

$\Delta S\big|_{N,V,Q=W_n=0} \geq 0$ in a closed, isochoric, adiabatic and uncoupled (no non-volume work) system, the entropy increases along any spontaneous process

$\Delta U\big|_{N,V,T,W_n=0} = Q$ in a closed, isochoric, isothermal and uncoupled (no non-volume work) system, the energy change is equal to the heat supplied along a process

$-\Delta F\big|_{N,V,T,rev} = -W_n$ in a closed, isochoric and isothermal system, the reversible non-volume (*e.g.* electric) work $-W_n$ that can be exported along a process is equal to minus the free energy change

$-(\Delta U - T\Delta S)\big|_{N,V,T,rev} = -W_n$

energetic driving force: due to entropy change in the surroundings (via heat exchange)

entropic driving force: due to entropy change in the system (the only one left if we replace isothermal by adiabatic)

⇒ **the <u>real</u> driving forces, and both are <u>measurable</u> (reversible-work measurements – *e.g.* electrochemistry) !**

→ Partial energy and entropy do not satisfy most of these equations

$\Delta F = \Delta U_d - T\Delta S_d$ **but** $\left.\begin{matrix}\Delta U_d \\ \Delta S_d\end{matrix}\right\}$ are path-dependent

$\left(\frac{\partial \Delta F}{\partial T}\right)_{N,V} \neq -\Delta S_d$ unless ΔS_i=0

$\left.\begin{matrix}\Delta S_d\big|_{N,V,W_n=Q=0} \geq 0 \\ \Delta U_d\big|_{N,V,T,W_n=0} = Q\end{matrix}\right\}$ incorrect unless ΔS_i=0

→ In other words

$\Delta F = \Delta U - T\Delta S$ *can be transformed in many ways to* $\Delta F = \Delta U' - T\Delta S'$ *with* $\begin{cases}\Delta U' = \Delta U + CT \\ \Delta S' = \Delta S + C\end{cases}$

→ But: these derived variables do not have much meaning...

P.H. Hünenberger

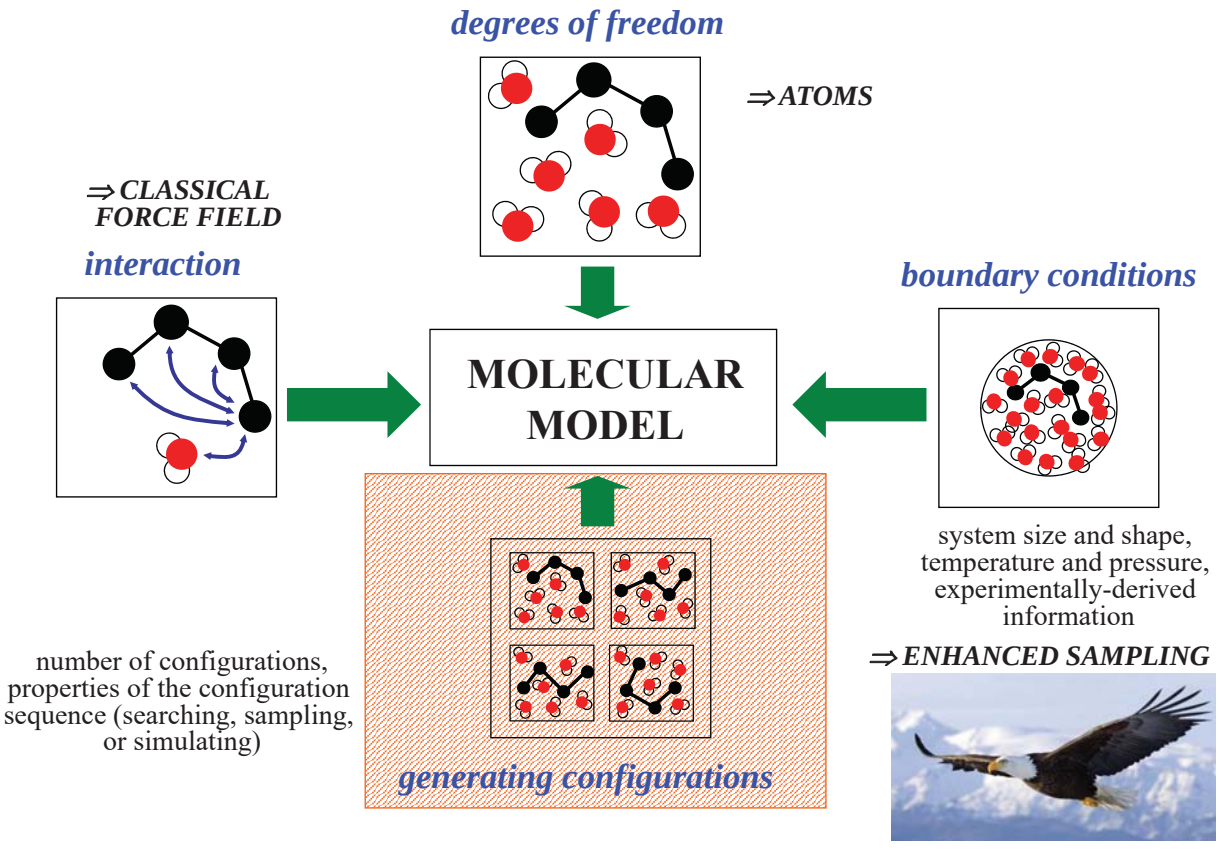
COMPUTER SIMULATION OF MOLECULAR SYSTEMS

Herbstsemester 2019
Tuesday 9:45-11:30 a.m.
HCI D2

Lecture 529-0004-00
www.csms.ethz.ch/education/CSCBP

LECTURE 9 (WEEK 10): Enhanced sampling

Four basic choices defining a molecular model



Enhanced-sampling

- The problem of **generating the configurations** of **molecular systems** is plagued by the **combinatorial explosion** (exponential increase) in the number of configurations (volume elements, local basins/minima and barriers of the potential energy) with the number of degrees of freedom
- For this reason, there is a huge effort in the field to devise schemes generating **preferentially relevant configurations** (and skipping automatically the irrelevant ones)

  → Massive literature, plethora of methods/combinations/acronyms !
- The goal maybe to efficiently

	SEARCH	***SAMPLE***	***SIMULATE***
Generated configurations	As diverse as possible Relevant No well-defined statistics No dynamical connection	As diverse as possible Relevant Well-defined statistics No dynamical connection ***(reweightable to Boltzmann)***	Boltzmann distributed Dynamically connected → MD (or SD) ***(no room to improve)***

→ Relevant usually means "low potential energy" but there may be other conditions, *e.g.*

- if I want a ΔG or PMF (conformational or alchemical), I want all the curve sampled ***even if some states have a high free energy***
- if I have experimental information (refinement), I may want it satisfied
- ...

→ Many (not all) **enhanced-searching** methods can also be implemented as **enhanced-sampling** methods if one is smart; this is usually the historical path

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods

Systematic search

Random search

Stepwise build-up

Genetic algorithm

Multicopy "sampling"

Distance geometry

Homology modelling

...

MD-based schemes for enhanced searching

Altered parameters	***Altered potential energy***
Altered masses	**Use of soft-core atoms**
Adiabatic decoupling	**Diffusion-equation search**
Temperature annealing	**Local-elevation search**
High-temperature sampling	**Biasing (US)**
Parallel tempering	**incl. LEUS (+λ,FB,B&S), metadyn or EDS**
	Hamiltonian replica exchange
Altered dimensionality ***(and potential energy)***	***Altered prescription of motion***
Four-dimensional MD	**PEACS**
Coarse-grained MD	**SPEED**
Multigraining	**Monte Carlo sampling**
Essential dynamics	**Markov-state modeling**
	SWARM MD

Systematic search

- *A priori*, the **simplest way** to **search** the configurational space of a molecular system would be to **vary systematically all Cartesian coordinates** by small increments (grid search)

  → If you have sufficiently small grid cells, this is even a form of **sampling**

  grid configurations $\{\boldsymbol{r}\}$
  weights $\sim \exp\left(-\beta \mathcal{U}(\boldsymbol{r})\right)$
  ⟹ canonical ensemble

  → Problem: combinatorial explosion!

  N coordinates
  n grid spacings per coordinate
  ⟹ n^N grid points

- So, you can only do this for **very small systems** (small molecule in vacuum)
- Or you have to restrict the search to a **small subset of conformational coordinates**

  → Exclude **solvent** coordinates from the search, or use an implicit-solvent model (common!)

  → Exclude **hard (bonds, angles)** degrees of freedom

  → Exclude **sidechain dihedrals** in a polymer

  → Use low-frequency (soft) **normal-mode coordinates**

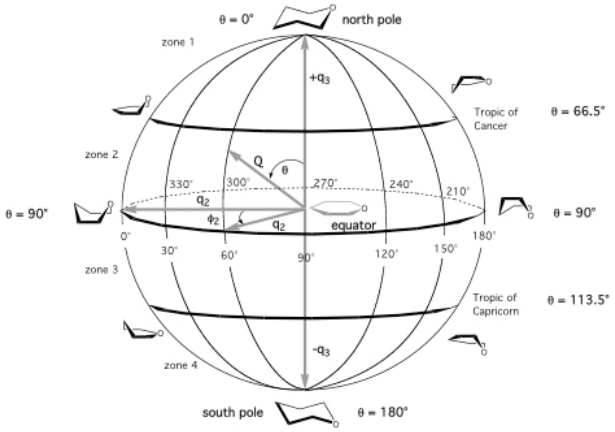
  → ...

Searched (Q)	*Omitted (q)*
Solute	*Solvent (unless implicit)*
Dihedrals	*+ Bonds, angles (standard values)*
Backbone dihedrals	*+ Sidechain dihedrals (standard values (?))*
Soft coordinates	*Hard coordinates (constrained (?))*

Systematic search

- Examples

  → **Pseudo-rotation** angle for rings



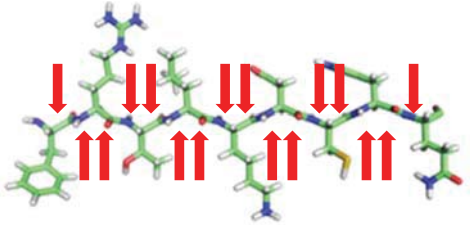
Cremer-Pople coordinates for six-membered rings

→ Q,θ,φ

If bonds and angles are fixed and all the same

→ θ,φ

  → Peptide **backbone dihedral search**



Nonapeptide
Planar peptide group

→ 16 dihedral angles

60° resolution

→ 10^{12} grid points

Systematic search

- The problem is that now, each grid point ***Q*** matches a whole ensemble of values of the omitted ("orthogonal") degrees of freedom ***q***

  → You can **energy minimize** for the ***q*** coordinates, but

  - You cannot cross barriers — ***So, this makes sense only if q is very simple (one minimum)***

  - What is the meaning of the resulting energy? — ***The weight of the grid point Q would actually depend on a free energy!***

- There exist a wealth of **heuristic schemes** to sample as well as possible a high-dimensional ***Q*** while dealing as well as possible with the orthogonal ***q***, *e.g.*

  Random search *Stepwise build-up* *Genetic algorithm* ...

  Multicopy "sampling" *Distance geometry* *Homology modelling*

  → They are **seldom statistical-mechanically rigorous** (→searching and not sampling)

  → They are usually **cheap** (especially with implicit solvent) and used to get a quick "feeling"

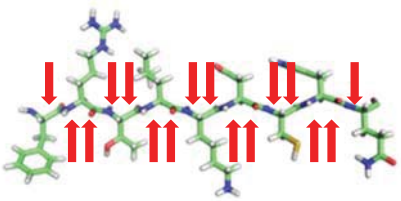
- The alternative is to use **MD-based schemes**, *i.e.* to let MD take care of the ***q*** variables

  → Often, the best (especially with explicit solvent) — ***Discussed a bit later...***

Heuristic search methods

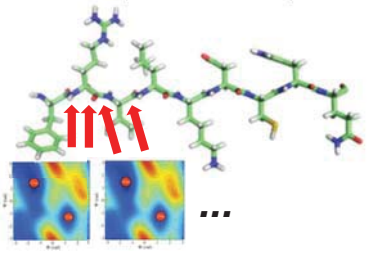
- Illustrative **examples**

Random search



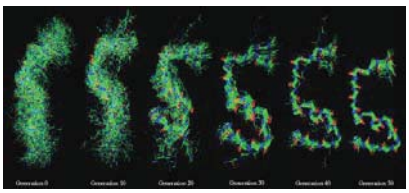
Space Q is too big, so we search it by making random moves rather than systematically

Stepwise build-up



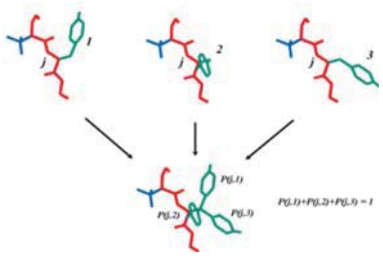
We generate conformations by taking successive (ϕ,ψ) pairs at random, with a probability (Boltzmann) depending on the dimer free-energy map; and we «prune» configurations with bad overlaps

Genetic algorithm



We evolve a population of conformations, «breeding» (i.e. combining the coordinates of distinct segments) the lowest-energy ones and letting the others «die»

Multicopy "sampling"

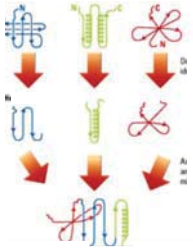


Space q is modelled by including multiple conformations with fractional weights; as we search Q, we optimize the weights in q

Distance geometry

We use NOE information to fix inter-proton distances and try to find the best solution for these distances (along with a primitive force field)

Homology modelling



We use the 3D coordinates of proteins in the PDB as a template to explore the conformational space of of a new sequence (threading)

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods	***MD-based schemes for enhanced searching***	
	Altered parameters	***Altered potential energy***
Systematic search	**Altered masses**	**Use of soft-core atoms**
Random search	**Adiabatic decoupling**	**Diffusion-equation search**
Stepwise build-up	**Temperature annealing**	**Local-elevation search**
Genetic algorithm	**High-temperature sampling**	**Biasing (US)**
Multicopy "sampling"	**Parallel tempering**	**incl. LEUS (+λ,FB,B&S), metadyn or EDS**
		Hamiltonian replica exchange
Distance geometry	***Altered dimensionality*** *(and potential energy)*	***Altered prescription of motion***
Homology modelling	**Four-dimensional MD**	**PEACS**
...	**Coarse-grained MD**	**SPEED**
	Multigraining	**Monte Carlo sampling**
	Essential dynamics	**Markov-state modeling**
		SWARM MD

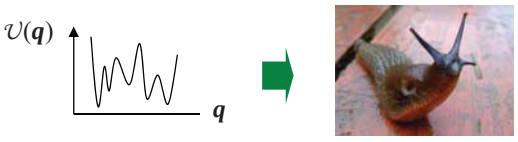
MD-based schemes for enhanced sampling

- MD (+thermostat) is a **great simulation method**, but a **poor sampling/searching method**

→ MD explores configuration space at a "sluggish" pace imposed by the "natural" system dynamics

energy barriers, *narrow passes* } *trapping*

limited diffusivity *recrossing*

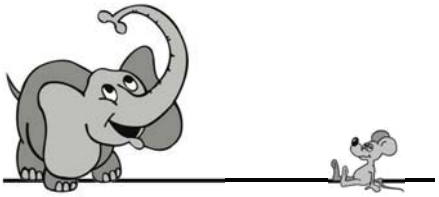


→ computers themselves work at a "sluggish" pace compared to nature (factor ~10^{-13} for 10^4-atom system)

also: nature actually uses QM in "real-time" and with a "scaling" of O[N]

- The **scaling of Nature's "computer"** – some considerations for fun...

Elephas Maximus
(about 10^{30} atoms)



Mus Musculus
(about 10^{24} atoms)

→ It does not take Nature more "CPU time" to propagate the elephant forward in time than it takes for the mouse ⇒ Scaling of O[1] in the number of particles ?

→ But to be fair, it takes an elephant-volume of Nature to propagate the elephant, and only a mouse-volume of it for the mouse ⇒ Scaling of O[N] at fixed CPU volume !

→ Nature is a *linear scaling*, *amazingly fast* and *massively parallel* computer with perfect scalability *(and its true equations of motion must be local!)*

Natural volume also tends to be cheaper than computer-processor volume (there are exceptions !)

MD-based schemes for enhanced sampling

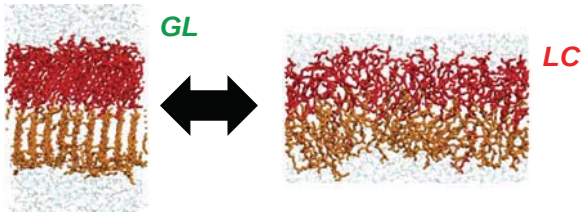
- So, there are processes that can be simulated with MD, and some that cannot

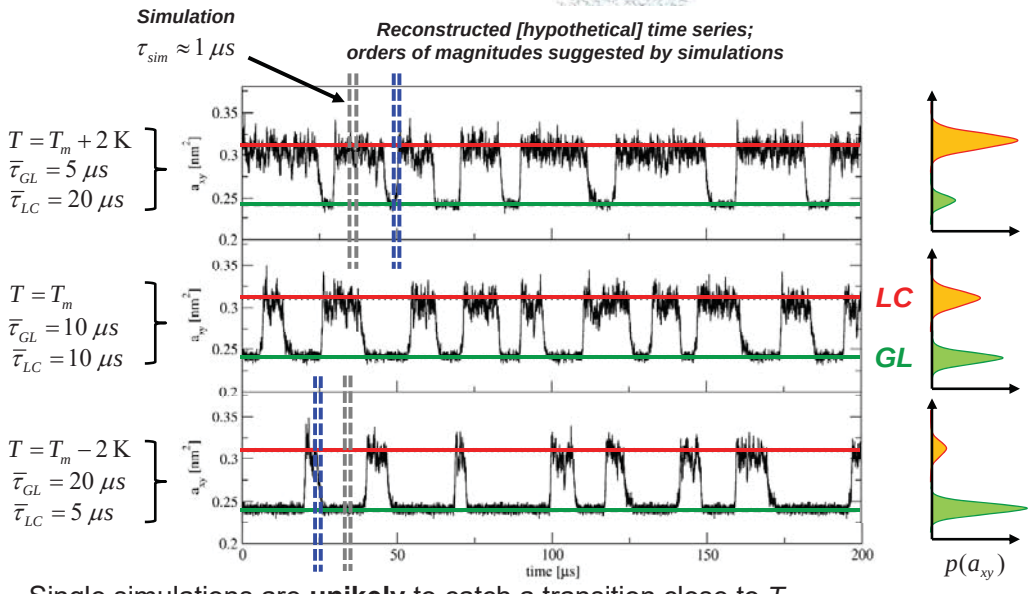
System type	*Relaxation time (indicative)*	
gas	~ 1 ps	
pure liquid	~ 10 – 100 ps	*acceptable*
small organic molecule in solution	~ 10 ps – 1 ns	
short peptide in solution	~ 10 – 100 ns	
lipid aggregation in solution	~ 10 – 100 ns	
protein folding	~ 1 ms – 1000 s	
	50 years	*intractable*

- Treasures of ingenuity have been invested into the design of **clever algorithms** (in particular many MD variants) that lead to **more efficient searching or sampling**

MD-based schemes for enhanced sampling

- Example: phase transitions in a GMP bilayer patch are **two-state** (GL↔LC), **fast** (~ns) but **infrequent** (~10 µs at T_m) events



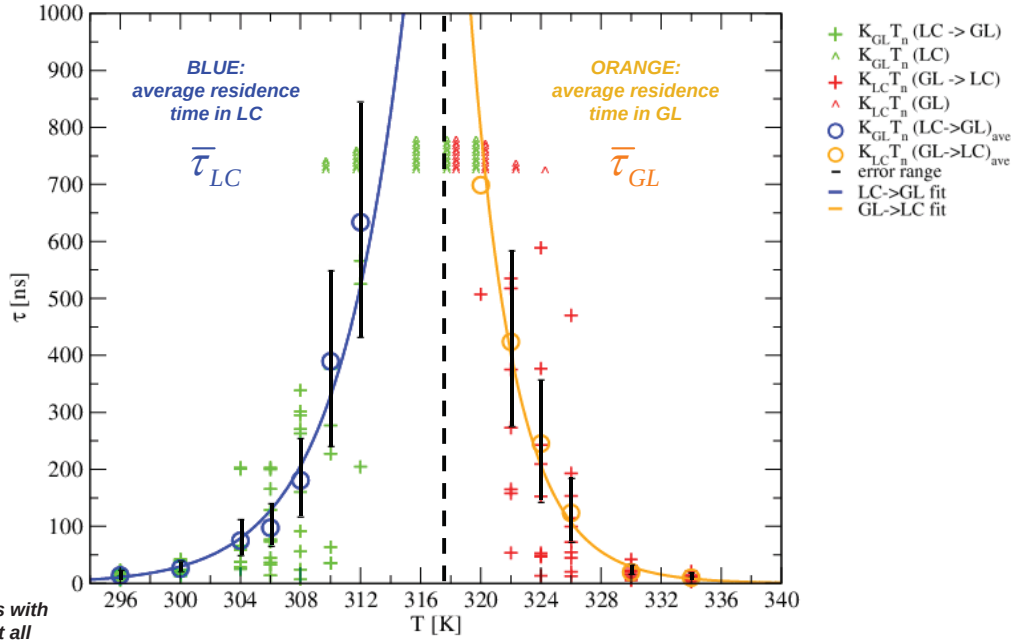


→ Single simulations are **unikely** to catch a transition close to T_m

→ Transitions observable **only in one direction** far away from T_m

MD-based schemes for enhanced sampling

- Average **residence-times** as a function of temperature



Fit (temperatures with no transition at all are excluded)

$$\overline{\tau}(T) = C \exp\left(\frac{A}{RT}\right)$$

$C_{GL} = 4.2 \cdot 10^{-45}$ ns $\quad A_{GL} = 289.6$ kJ mol^{-1}

$C_{LC} = 7.0 \cdot 10^{33}$ ns $\quad A_{LC} = -185.9$ kJ mol^{-1}

[values per mol "system" = 128 lipids]

MD-based schemes for enhanced sampling

- A (tentative) classification of methods:

→ Altered parameters
- altered masses *(sampling)* — $\boldsymbol{M}$ — *→ faster dynamics*
- altered temperature *(isothermal dynamics; sampling or searching)* — $k_B T$ — *→ faster dynamics; → but higher entropy regions !; → possible variants: energy, volume, pressure, number of particles or chemical potential*

→ Altered potential energy
- memory-based bias *(searching)* — $\mathcal{V}_{bias}(\boldsymbol{q},t)$ — *→ avoid continuous "revisiting" of known configurations*
- *freezing*
- time-independ bias *(sampling)* — $\mathcal{V}_{bias}(\boldsymbol{q})$ — *→ make minima less deep; → make barriers less high; → focus on relevant regions*

→ Altered dimensionality (and potential energy)
- increased dimensionality *(sampling or searching)* — $dim[\Gamma]\nearrow$ — *→ pathways around barriers*
- reduced dimensionality *(sampling or searching)* — $dim[\Gamma]\searrow$ — *→ faster dynamics; → fewer particles; → longer timestep*

→ Altered prescription of motion *(sampling or searching)*

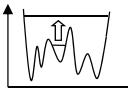
→ Time-saving techniques

(e.g. multiple time-step methods, cutoffs, pairlists, grid-based neighbor search, ...)

...and probably many more...

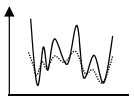
MD-based schemes for enhanced sampling

M *(sampling)* — MD with altered masses[1]



$k_B T$ *(sampling or searching)* — As a searching method (*e.g.* Simulated annealing[2] or J-walking[3]); As a sampling method (*e.g.* high-temperature MD or replica exchange[4])

$\mathcal{V}_{bias}(\boldsymbol{q},t)$ *(searching)* — "local-elevational adaptative deflational metadynamical flooding" + stochastic tunneling[10]; deflation[5], local elevation[6], conformational flooding[7], adaptive biasing force[8], metadynamics[9]



$\mathcal{V}_{bias}(\boldsymbol{q})$ *(sampling)* — Umbrella sampling[11] (arbitrary biasing potential); Solute potential scaling[12]; Diffusion equation method[13]; Soft-core atoms[14]; Fluctuating potential[15]; "hyper" MD[16], "accelerated" MD[17],...

$dim[\Gamma]\searrow$ *(sampling or searching)* — Constraints (*e.g.* bond lengths[18]); Implicit solvation; Reduced space (e.g. essential dynamics[19]); Coarse-graining

For sampling: $\tilde{\mathcal{E}}(\Gamma,\lambda)=\lambda\mathcal{E}_{red}(\Gamma_{red}(\Gamma))+(1-\lambda)\mathcal{E}(\Gamma)$ — *e.g.* resolution exchange/ multigraining[20]

$dim[\Gamma]\nearrow$ *(sampling or searching)* — Cartesian dimension (*e.g.* 4D-MD[21]); Multiple copies (*e.g.* LES[22]); Multiple systems (*e.g.* SWARM-MD[23])

For sampling: $\tilde{\mathcal{E}}(\Gamma_{ext},\lambda)=\lambda\mathcal{E}_{ext}(\Gamma_{ext})+(1-\lambda)\mathcal{E}(\Gamma(\Gamma_{ext}))$ — *e.g.* 4D free energy[24]

1: Jacucci & Rahman, 1974; Bennett, 1975
2: Kirkpatrick, Gelatt & Vecchi, 1983
3: Frantz, Freeman & Doll, 1990
4: Sugita & Okamoto, 1999
5: Crippen & Scheraga, 1969
6: Huber & van Gunsteren, 1994
7: Grubmüller, 1995
8: Darve & Pohorille, 2002
9: Laio & Parrinello, 2003
10: Levy & Montalvo, 1985
11: Torrie & Valleau, 1977
12 Tsujishita et al., 1993
13 Piela, Kostrowicki & Scheraga, 1989
14: Huber et al., 1997
15: Liu & Berne, 1993
16: Voter, 1997
17: Steiner et al., 1998; Fichthorn, 1999; Gong & Wilkins, 1999; Hamelberg et al., 2002; Mongan & McCammon, 2004
18: Ryckaert et al., 1977
19: Amadei et al., 1993
20: Christen & van Gunsteren, 2006; Lyman et al., 2006
21: van Schaik et al., 1993
22: Elber & Karplus, 1990
23: Huber & van Gunsteren, 1998
24: Beutler & van Gunsteren, 1994

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods

Systematic search

Random search

Stepwise build-up

Genetic algorithm

Multicopy "sampling"

Distance geometry

Homology modelling

...

MD-based schemes for enhanced searching

Altered parameters

Altered masses

Adiabatic decoupling

Temperature annealing

High-temperature sampling

Parallel tempering

Altered potential energy

Use of soft-core atoms

Diffusion-equation search

Local-elevation search

Biasing (US)

incl. LEUS (+λ,FB,B&S), metadyn or EDS

Hamiltonian replica exchange

Altered dimensionality *(and potential energy)*

Four-dimensional MD

Coarse-grained MD

Multigraining

Essential dynamics

Altered prescription of motion

PEACS

SPEED

Monte Carlo sampling

Markov-state modeling

SWARM MD

Altered masses

- The **configurational distribution** of a molecular system at equilibrium is **independent of the atomic masses**

Depends on the masses, but is the same for all configurations!

$$P(\boldsymbol{r}) \sim \int d\boldsymbol{p}\, \exp\left(-\beta \mathcal{H}(\boldsymbol{r},\boldsymbol{p})\right) = \int d\boldsymbol{p}\, \exp\left(-\beta \mathcal{K}(\boldsymbol{p})\right) \exp\left(-\beta \mathcal{U}(\boldsymbol{r})\right) = C \exp\left(-\beta \mathcal{U}(\boldsymbol{r})\right)$$

→ but the **kinetics** (*e.g.* oscillations, diffusion, viscosity^{-1}) becomes **faster** for lower masses

- So, what about scaling all the masses by a factor α<1 to make everything faster?

→ nice try, but there is **no free lunch**

Newton $\dfrac{d^2\boldsymbol{r}}{dt^2} = \dfrac{\boldsymbol{F}}{m}$

Given T $\dfrac{d\boldsymbol{r}}{dt} \sim \left(\dfrac{T}{m}\right)^{1/2}$

Scale m by α
Scale t by $\alpha^{1/2}$

Same dynamics!

But also requires for the same accuracy

$$\Delta t \rightarrow \alpha^{1/2} \Delta t$$

- In practice, you can play with mass in two ways

→ **homogenize** all the system masses (*e.g.* to mass of hydrogen)

Especially for the solvent, it will make it less "viscous"

→ make **subset of atoms heavier** (more inertia) and **other subset lighter** (faster response)

Situation of "adiabatic decoupling", reminiscent of the Born-Oppenheimer approximation

[the configurational ensembles are the same, but you get more motion (sampling) in the loop in 5ns time]

X-ray structure → *5ns MD* → *5ns MD (mass ×5 in green loop)*

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

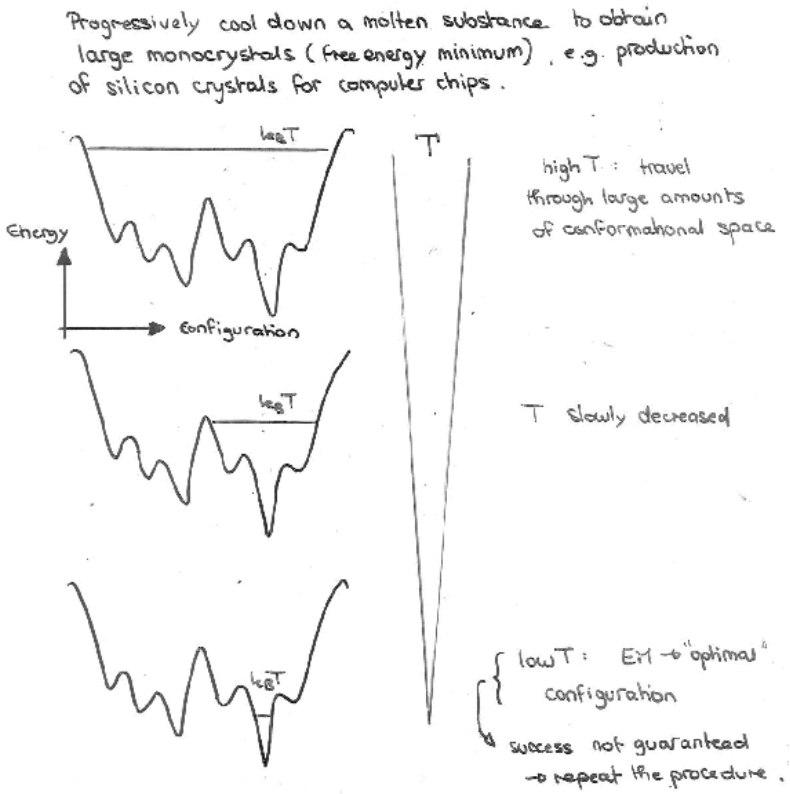
Systematic and heuristic search methods	*MD-based schemes for enhanced searching*	
	Altered parameters	***Altered potential energy***
Systematic search		
Random search	**Altered masses**	**Use of soft-core atoms**
	Adiabatic decoupling	**Diffusion-equation search**
Stepwise build-up	**Temperature annealing**	**Local-elevation search**
Genetic algorithm	**High-temperature sampling**	**Biasing (US)**
	Parallel tempering	**incl. LEUS (+λ,FB,B&S), metadyn or EDS**
Multicopy "sampling"		**Hamiltonian replica exchange**
	Altered dimensionality *(and potential energy)*	***Altered prescription of motion***
Distance geometry	**Four-dimensional MD**	**PEACS**
Homology modelling	**Coarse-grained MD**	**SPEED**
...	**Multigraining**	**Monte Carlo sampling**
	Essential dynamics	**Markov-state modeling**
		SWARM MD

Temperature annealing

Also called "simulated annealing"

- An "old" method (1983), for **searching only**

Initially, a cool and efficient solution to the "traveling salesman" problem – and other "non-molecular" optimization problems



The barriers that can be crossed are on the order of k_BT

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods	*MD-based schemes for enhanced searching*	
	Altered parameters	***Altered potential energy***
Systematic search	**Altered masses**	**Use of soft-core atoms**
Random search	**Adiabatic decoupling**	**Diffusion-equation search**
Stepwise build-up	**Temperature annealing**	**Local-elevation search**
Genetic algorithm	**High-temperature sampling**	**Biasing (US)**
Multicopy "sampling"	**Parallel tempering**	**incl. LEUS (+λ,FB,B&S), metadyn or EDS**
		Hamiltonian replica exchange
Distance geometry	***Altered dimensionality*** *(and potential energy)*	***Altered prescription of motion***
Homology modelling	**Four-dimensional MD**	**PEACS**
...	**Coarse-grained MD**	**SPEED**
	Multigraining	**Monte Carlo sampling**
	Essential dynamics	**Markov-state modeling**
		SWARM MD

High-temperature sampling

- If you simulate at a **higher temperature** than your normal temperature, you can **reweight**

$$\langle Q(\boldsymbol{r})\rangle_{\beta} = \frac{\int d\boldsymbol{r}\, Q(\boldsymbol{r})\exp\left(-\beta\mathcal{U}(\boldsymbol{r})\right)}{\int d\boldsymbol{r}\exp\left(-\beta\mathcal{U}(\boldsymbol{r})\right)} = \frac{\int d\boldsymbol{r}\, Q(\boldsymbol{r})\exp\left(-(\beta-\beta')\mathcal{U}(\boldsymbol{r})\right)\exp\left(-\beta'\mathcal{U}(\boldsymbol{r})\right)}{\int d\boldsymbol{r}\exp\left(-(\beta-\beta')\mathcal{U}(\boldsymbol{r})\right)\exp\left(-\beta'\mathcal{U}(\boldsymbol{r})\right)} = \frac{\left\langle Q(\boldsymbol{r})\exp\left(-(\beta-\beta')\mathcal{U}(\boldsymbol{r})\right)\right\rangle_{\beta'}}{\left\langle \exp\left(-(\beta-\beta')\mathcal{U}(\boldsymbol{r})\right)\right\rangle_{\beta'}}$$

→ the **kinetics** will become **faster** (nice!)

→ but you have to decrease the timestep

Given T $\quad \frac{d\boldsymbol{r}}{dt} \sim \left(\frac{T}{m}\right)^{1/2}$ $\quad$ *scale T by α* ➡ *For similar integration accuracy* $\quad \Delta t \rightarrow \alpha^{-1/2}\Delta t$

→ and already for small temperature increases, the **statistical efficiency** of the reweighting will become very bad (because you favor higher entropy regions)

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods	*MD-based schemes for enhanced searching*	
	Altered parameters	***Altered potential energy***
Systematic search		
Random search	**Altered masses**	**Use of soft-core atoms**
	Adiabatic decoupling	**Diffusion-equation search**
Stepwise build-up	**Temperature annealing**	**Local-elevation search**
Genetic algorithm	**High-temperature sampling**	**Biasing (US)**
	⇨ **Parallel tempering**	**incl. LEUS (+λ,FB,B&S), metadyn or EDS**
Multicopy "sampling"		**Hamiltonian replica exchange**
	Altered dimensionality *(and potential energy)*	***Altered prescription of motion***
Distance geometry	**Four-dimensional MD**	**PEACS**
Homology modelling	**Coarse-grained MD**	**SPEED**
	Multigraining	**Monte Carlo sampling**
...	**Essential dynamics**	**Markov-state modeling**
		SWARM MD

Parallel tempering

Also called "replica-exchange (in temperature)"

- Simulate N **independent replicas** of the system in parallel, each at a **different temperature**
- At **regular intervals** τ_{exc}, we attempt to **swap configurations** (coordinates and velocities) between two adjacent systems

Probability of the initial two-system state

$$p_i \sim \exp\left(-\beta\mathcal{H}(\boldsymbol{x}) - \beta'\mathcal{H}(\boldsymbol{x}')\right) \quad \Rightarrow \quad p_f \sim \exp\left(-\beta\mathcal{H}(\boldsymbol{x}') - \beta'\mathcal{H}(\boldsymbol{x})\right)$$

Probability of the swapped two-system state

$$\boldsymbol{x} = (\boldsymbol{r}, \boldsymbol{p})$$

- To ensure that the system pair remains Boltzmann-distributed, we **accept or reject** the swap according to a **Monte Carlo criterion**

$$P_{acc}(i \to f) = \begin{cases} 1 & if \quad p_f > p_i \\ p_f / p_i & otherwise \end{cases} \quad \textit{or} \quad = \begin{cases} 1 & if \quad \Delta \le 0 \\ \exp(-\Delta) & otherwise \end{cases}$$

with $p_f / p_i = \exp\left((\beta' - \beta)(\mathcal{H}(\boldsymbol{x}') - \mathcal{H}(\boldsymbol{x}))\right)$ *with* $\Delta = -(\beta' - \beta)(\mathcal{H}(\boldsymbol{x}') - \mathcal{H}(\boldsymbol{x}))$

→ one may also swap the coordinates only and apply the test in terms of potential energy only (in this case, one usually also swaps the velocities and rescales them after the swap)

- It is easily seen that this procedure results in a set of N **canonical ensembles,** each at one of the selected temperatures

This means in particular that the dynamics is nonsense...

→ if you look at **one temperature**, the trajectory is **discontinuous** (systems come and go)

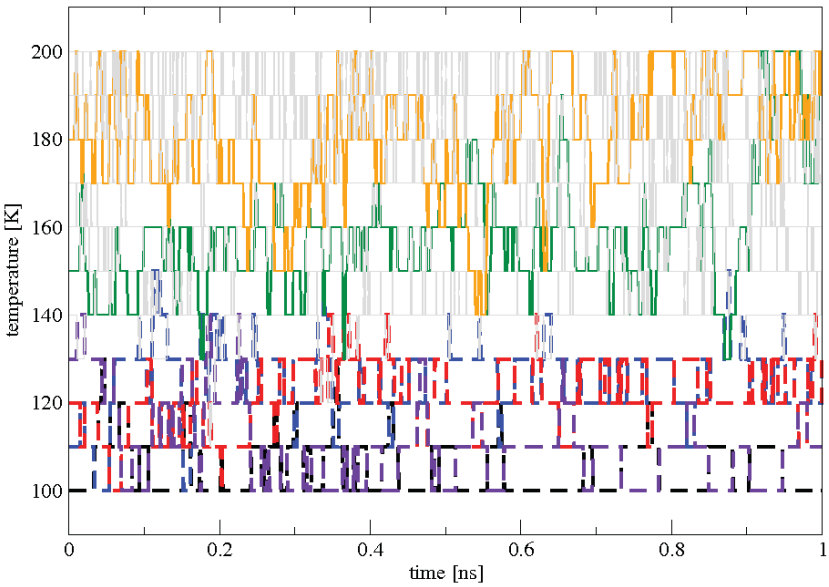
→ if you look at **one system**, the trajectory **hops across temperatures** (enabling enhanced sampling at the highest temperatures!)

→ this is a bit like simulated annealing (searching) converted to a sampling method!

Parallel tempering

- Example: parallel tempering SD-simulation of 512 *n*-butane molecules $N = 11$

$\tau_{exc} = 1\,ps$



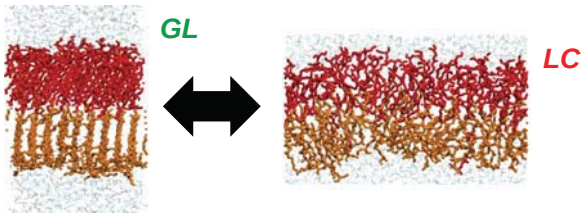
→ Note that the choice of the time interval τ_{exc} between exchange attempts and of the T-ladder may have a large impact on the achieved sampling enhancement (and is not trivial to optimize!)

→ Note also that we are in principle not limited to pairwise exchanges (swaps) between adjacent replicas (any permutation would work as well; but this is the most common choice)

Parallel tempering

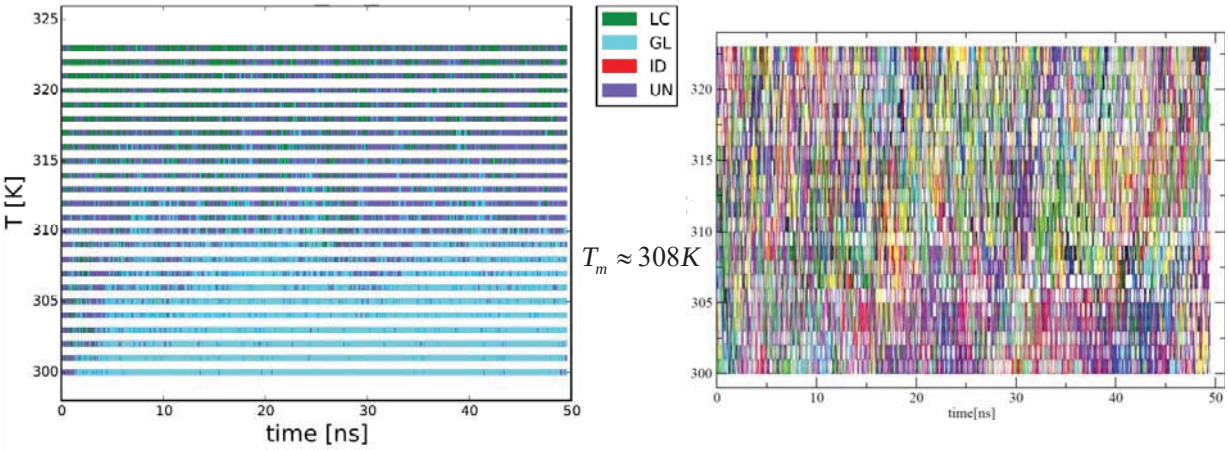
Thanks Pavel !

- Example: parallel tempering MD-simulation a GMP bilayer patch



$N = 24$

$\tau_{exc} = 10\,ps$



$T_m \approx 308K$

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods	*MD-based schemes for enhanced searching*	
	Altered parameters	***Altered potential energy***
Systematic search	**Altered masses**	**Use of soft-core atoms**
Random search	**Adiabatic decoupling**	**Diffusion-equation search**
Stepwise build-up	**Temperature annealing**	**Local-elevation search**
Genetic algorithm	**High-temperature sampling**	**Biasing (US)**
Multicopy "sampling"	**Parallel tempering**	**incl. LEUS (+λ,FB,B&S), metadyn or EDS**
		Hamiltonian replica exchange
Distance geometry	***Altered dimensionality*** *(and potential energy)*	***Altered prescription of motion***
Homology modelling	**Four-dimensional MD**	**PEACS**
	Coarse-grained MD	**SPEED**
...	**Multigraining**	**Monte Carlo sampling**
	Essential dynamics	**Markov-state modeling**
		SWARM MD

Use of soft-core atoms

Soft-core interaction :

Modify Coulomb and van der Waals interactions so that they are finite when atoms are on top of each other
→ atoms may go through each other with finite barriers

Use of soft-core atoms

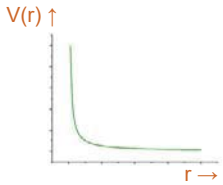
Use of soft-core non-bonded interactions

Thomas Beutler et al. Chem. Phys. Letters 222 (1994) 529-539

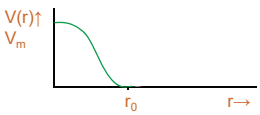
Use of non-physical potential energy terms

Physical non-bonded term: van der Waals $\frac{C_{12}}{r_{ij}^{12}}$

Coulomb $\frac{q_i q_j}{r_{ij}}$



Non-physical softer non-bonded term that allows atoms to pass through each other:



Conditions: $V(0) = V_m \quad V'(0) = 0$

$V(r_0) = 0 \quad V'(r_0) = 0$

1. V(r) is a function of r:

$$f(r) = a + br + cr^2 + dr^3$$

$$= V_m \left[1 - 3\left(\frac{r}{r_0}\right)^2 + 2\left(\frac{r}{r_0}\right)^3 \right]$$

$$f'(r) = -6V_m \frac{r}{r_0^2}\left[1 - \left(\frac{r}{r_0}\right) \right]$$

Use of soft-core atoms

2. V(r) is a function of r^2:

$$g(r) = a + br^2 + cr^4$$

$$= V_m\left[1 - \left(\frac{r}{r_o}\right)^2\right]^2$$

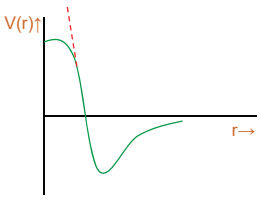
$$g'(r) = -4V_m \frac{r}{r_0^2}\left[1 - \left(\frac{r}{r_0}\right)^2\right]$$

3. V(r) is general van der Waals plus Coulomb form:

$$V(r) = \frac{C_{12}}{\left[\alpha + r_{ij}^6\right]^2} - \frac{C_6}{\left[\alpha + r_{ij}^6\right]} + \frac{1}{4\pi\varepsilon_0\varepsilon_r}\frac{q_i q_j}{\left[\alpha + r_{ij}^6\right]^{1/6}}$$

$\lim_{\alpha \to 0}$ = standard form

$\alpha \neq 0$: V(0) = finite V'(0) = 0



Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods	*MD-based schemes for enhanced searching*	
	Altered parameters	***Altered potential energy***
Systematic search	**Altered masses**	**Use of soft-core atoms**
Random search	**Adiabatic decoupling**	**⇨ Diffusion-equation search**
Stepwise build-up	**Temperature annealing**	**Local-elevation search**
Genetic algorithm	**High-temperature sampling**	**Biasing (US)**
Multicopy "sampling"	**Parallel tempering**	**incl. LEUS (+λ,FB,B&S), metadyn or EDS**
		Hamiltonian replica exchange
Distance geometry	***Altered dimensionality*** *(and potential energy)*	***Altered prescription of motion***
Homology modelling	**Four-dimensional MD**	**PEACS**
	Coarse-grained MD	**SPEED**
...	**Multigraining**	**Monte Carlo sampling**
	Essential dynamics	**Markov-state modeling**
		SWARM MD

Diffusion equation search

Piela et al. JPC 93 (1989) 3339

Diffusion equation: (1D PDE)

$$\frac{\partial^2}{\partial x^2} f(x,t) = \frac{\partial}{\partial t} f(x,t)$$

(constant D: omitted here)

Solutions depend on boundary conditions:

Initial values: $f(x,0) = f_0(x)$

Boundary values: $f(\pm\infty, t) = 0$

or $f(x,t) = f(x+L,t) \Rightarrow$ periodic

Solution ($f(\pm\infty,t)=0$):

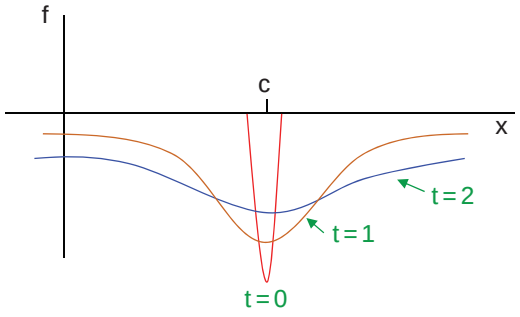
$$f(x,t) = \frac{A}{\sqrt{1+4Bt}} e^{-\frac{B[x-c]^2}{[1+4Bt]}}$$

c: center
A: amplitude
B: width at t=0

Diffusion equation search

Example:

$f_0(x) = -\delta(x-c)$



Note:

- parameter t (time) controls modification of f(x,t)

→ In the following, this equation is used to deform the potential-energy surface and the parameter *t* has no relationship to a time; so, I will write it as *q*

- Change of f(x,q) as a function of q is proportional to the local curvature (in x), so **barriers melt, minima fill up as q increases**

Diffusion equation search

Piela et al. J. Phys. Chem. 93 (1989) 3339:

Use functions that are solutions of the diffusion equation as means to find the global minimum

1. Choose for the potential energy

$V(\vec{r}, q) =$ Gaussian (or solution D.E.)

2. Deform V(r,q) by letting $q=0 \rightarrow q_{max}$

3. Minimize V(r,q) while reforming surface: $q=q_{max} \rightarrow 0$

Kind of potential energy simulated annealing

Diffusion equation search

Modification (Huber et al.): *J. Phys. Chem.* ***101*** *(1997) 5926-5930*

1. Deform *individual terms* in the potential energy function:

- dihedral angle
- van der Waals, use soft-core
- Coulomb

(if each term is a solution of the diffusion equation, then the sum will be too [linearity])

2. Such that *positions of minima* of each term are *not changed*

3. Thereby possibly relaxing the condition that V(r,q) must be a solution of the diffusion equation

(which means that the minima of the overall potential energy may still move a bit – but hopefully not too much!)

Diffusion equation search

Deformation of the dihedral angle term

GROMOS87:

$$V(\varphi) = K_{\varphi}\left[1 + \cos\left(n\varphi - \delta\right)\right]$$

Solution to diffusion equation (periodic):

$$f(x, t) = e^{-\omega^2 t}\cos(\omega x - \delta)$$

Deformable dihedral angle term:

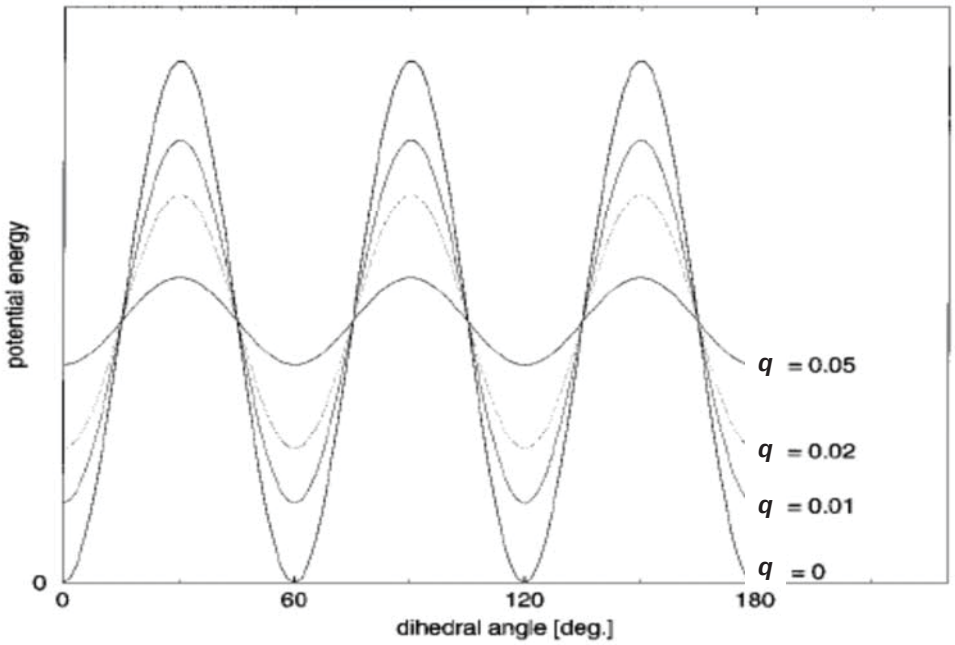
$$V(\varphi, q_{\varphi}) = K_{\varphi}\left[1 + e^{-n^2 q_{\varphi}}\cos(n\varphi - \delta)\right]$$

Note:

- For $q_{\varphi} = 0 \rightarrow$ standard GROMOS87
- For $q_{\varphi} = \infty \rightarrow$ constant K_{φ}
- **Positions of minima are independent of q_{φ}**

Diffusion equation search

Dihedral-angle term in the interaction function



For larger values of the parameter q the surface is smoother, leaving maxima and minima where they are

Diffusion equation search

Deformation of the van der Waals term

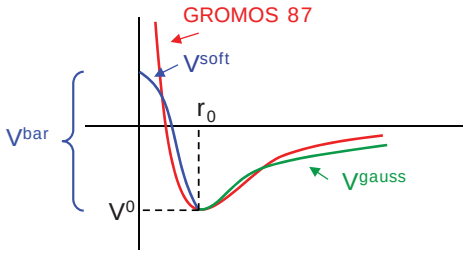
GROMOS87: $V(r) = \frac{C_{12}}{r^{12}} - \frac{C_6}{r^6}$ minimum V^0 at r_0

Partial solutions diffusion equation (+ soft-core):

$$V(r) = \begin{cases} V^{soft}(r) & 0 \le r \le r_0 \\ V^{gauss}(r) & r > r_0 \end{cases}$$

$$V^{soft}(r) = V^0 + V^{bar}\left[1 - 2\left(\frac{r}{r_0}\right)^2 + \left(\frac{r}{r_0}\right)^4\right]$$

$$V^{gauss}(r) = V^0 e^{-B[r-r_0]^2} \quad \text{with } B = \frac{2}{r_0^2}$$



Note: $V(0) = \text{finite}$ $\quad V'(0) = 0$ $\quad V''(r) = -\frac{4V^0}{r_0^2}$ for $r \downarrow r_0$

$V(r)$ and $V'(r)$ continuous at $r = r_0$

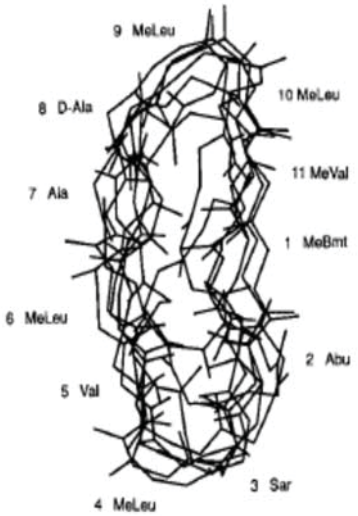
Deformable van der Waals term: $V^{soft}(r, q_{nb}) = V^{soft}(r)\left[1 + 4Bq_{nb}\right]^{-1/2}$

$$V^{gauss}(r, q_{nb}) = V^0 e^{\frac{B[r-r_0]^2}{[1+4Bq_{nb}]}}\left[1 + 4Bq_{nb}\right]^{-1/2}$$

Note: $V(r, q = 0) = V^{vdW}(r)$ ***(but: this implies a discontinuity in V at q=0)***

Diffusion equation search

A tough test case for searching: Cyclosporin A



11 residues
49 torsional angles
57 NOE distance restraints

1. **distance geometry**
   27 structures
   ↓
   9 classes
   very difficult structures
   high energy barriers between them

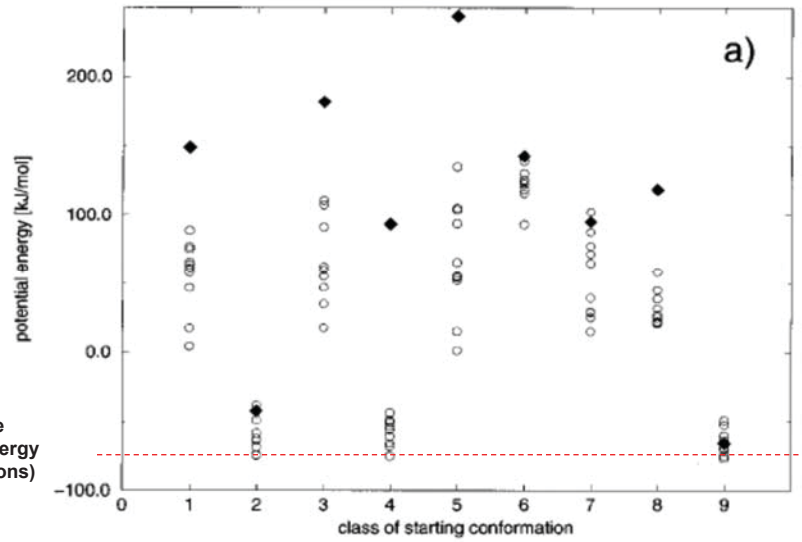
2. **structure refinement**
   standard 3D-MD
   1-2 correct

Challenge: how to get *all 9* different starting structures *converged* to the lowest-energy one

Diffusion equation search

Simulated annealing technique

Ten trials per high-energy starting structure



♦ Start

○ Individual runs

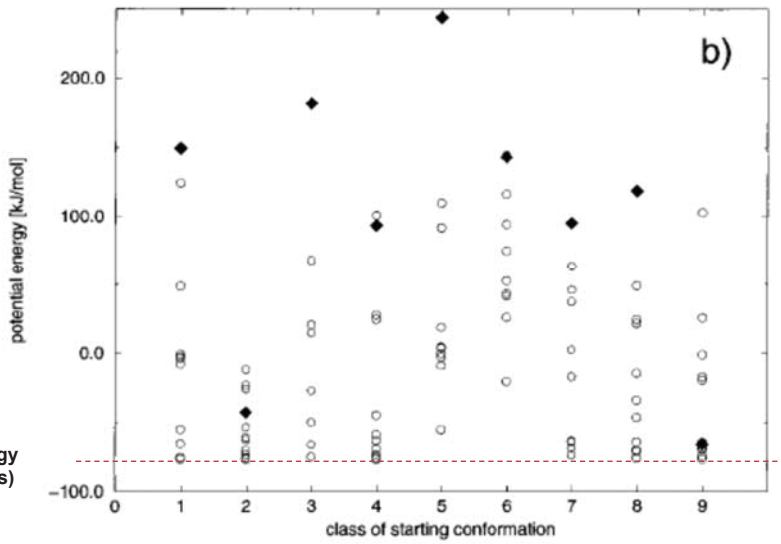
Correct structure (low potential energy and NOE violations)

Three reach lowest energy

Diffusion equation search

Diffusive soft-core technique

Ten trials per high-energy starting structure



♦ Start

○ Individual runs

Correct structure (low potential energy and NOE violations)

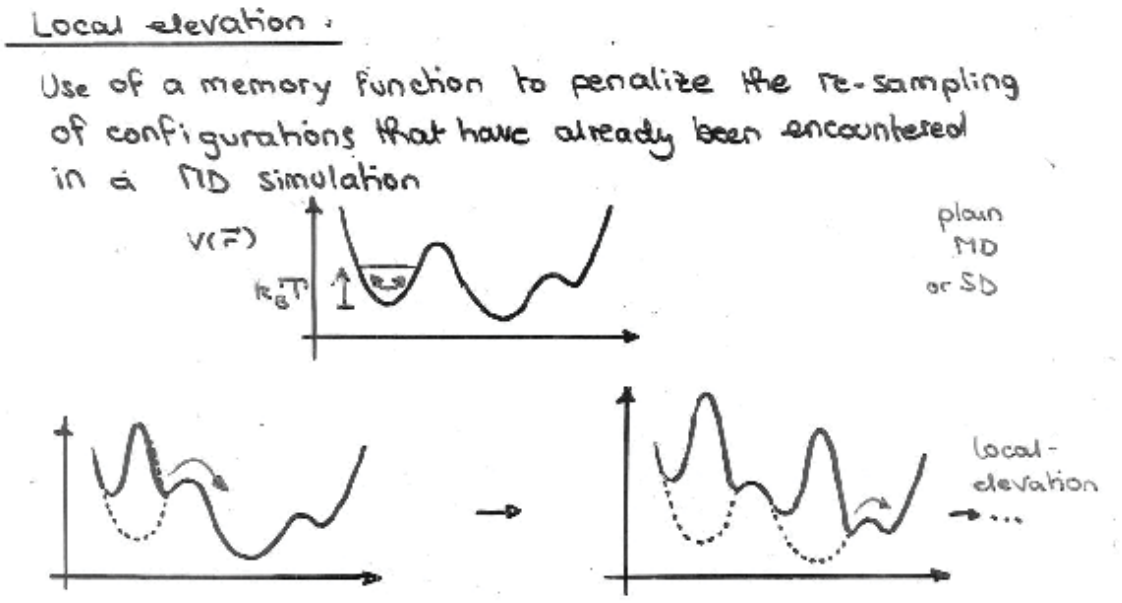
Seven reach lowest energy

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods	*MD-based schemes for enhanced searching*	
	Altered parameters	***Altered potential energy***
Systematic search	**Altered masses**	**Use of soft-core atoms**
Random search	**Adiabatic decoupling**	**Diffusion-equation search**
Stepwise build-up	**Temperature annealing**	**⇨ Local-elevation search**
Genetic algorithm	**High-temperature sampling**	**Biasing (US)**
Multicopy "sampling"	**Parallel tempering**	**incl. LEUS (+λ,FB,B&S), metadyn or EDS**
		Hamiltonian replica exchange
Distance geometry	***Altered dimensionality*** *(and potential energy)*	***Altered prescription of motion***
Homology modelling	**Four-dimensional MD**	**PEACS**
...	**Coarse-grained MD**	**SPEED**
	Multigraining	**Monte Carlo sampling**
	Essential dynamics	**Markov-state modeling**
		SWARM MD

Local-elevation search



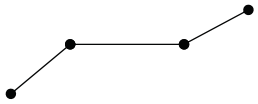
Local-elevation search

Methods to search conformational space

Idea: Include information obtained so far during the simulation into the search scheme: **memory function**

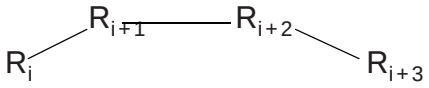
A. Characterize molecular conformations using:

- cartesian coordinates *too many*



- *torsional angles* φ, ψ, χ

- *dihedral angles* spanning residues:



Review searching: M. Christen & W.F. van Gunsteren J. Comput. Chem. ***29*** *(2007) 157 - 166*

Local-elevation search

B. Penalize the visited conformations by changing the energy function V as function of time

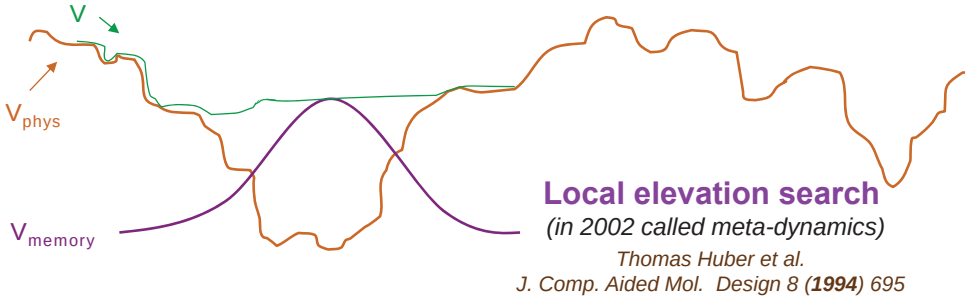
$$V\left(\{\varphi_i\}\right) = V_{phys}\left(\{\vec{r}_i\}\right) + V_{memory}\left(\{\varphi_i\}\right)$$

- potential energy term that pushes molecule out of the current conformation $\{\varphi_i^0\}$

$$V_{memory} = c \cdot \text{Number}\left(\varphi_i \text{ at } \varphi_i^0\right) \cdot e^{-\sum_i \left(\varphi_i - \varphi_i^0\right)^2 / 2\omega^2}$$

$$\downarrow$$

of conformations for which $\varphi_i^0 - \Delta\varphi_i < \varphi_i < \varphi_i^0 + \Delta\varphi_i$



Local elevation search
(in 2002 called meta-dynamics)
Thomas Huber et al.
*J. Comp. Aided Mol. Design 8 (**1994**) 695*

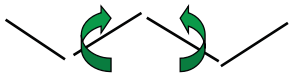
Local-elevation search

Implementation

1. Use torsion angles, ϕ_i
2. Each conformation $\phi_1, \phi_2, \phi_3, \ldots, \phi_n = \phi^n$
3. Discretise to M parts → M^n grid points $\phi^n{}_0$
4. Gaussian function at grid points: $V_{mem}(\phi^n) = k_{mem} N_{\phi^n_0} e^{-\frac{(\phi^n - \phi^n_0)^2}{2w^2}}$
5. $V_{total} = V_{phys} + V_{mem}$

A toy application

Pentane (two torsional angles)



Complete space can be mapped out

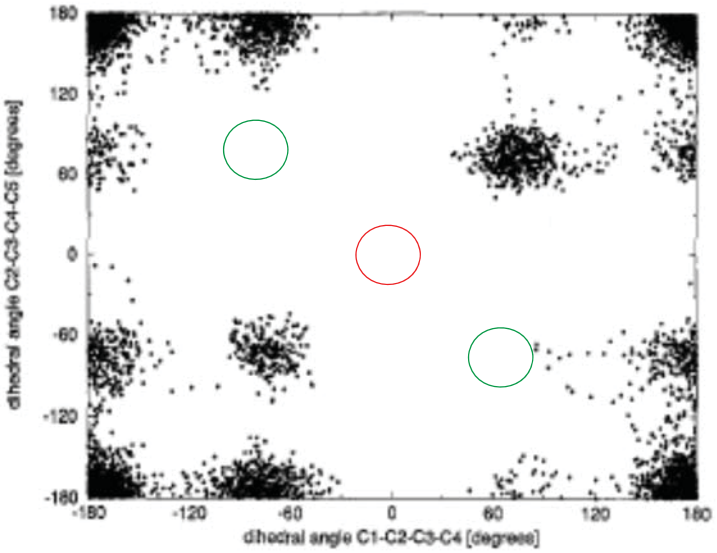
Local-elevation search

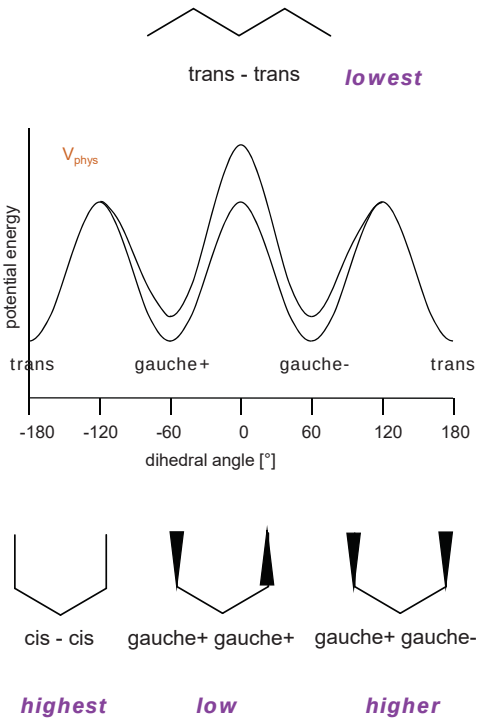
Test case: pentane

Thomas Huber et al.
J. Comp. Aided Mol. Design 8 (1994) 695

2 dihedral angles (3 minima each) → 9 low V_{phys} conformers

free SD-simulation (united atoms)
simulation time 100 ps, T=300 K, GROMOS force field





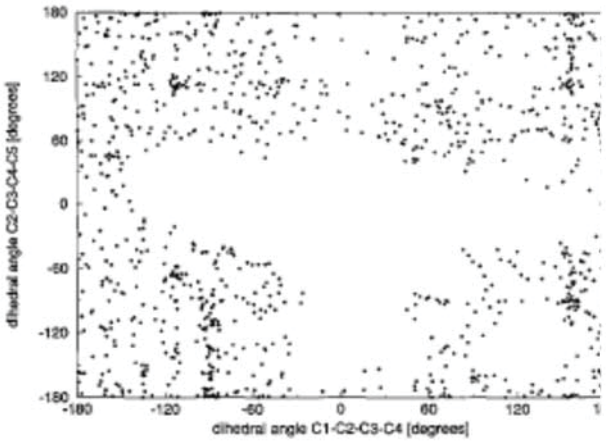
Higher-energy conformers are not (yet) sampled in 100 ps *normal* MD(SD) simulation

Local-elevation search

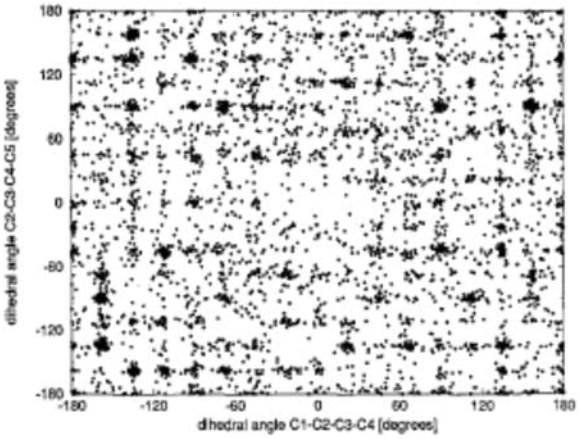
Local elevation search: pentane

Local-elevation simulation of pentane (united atoms) T=300 K,
Gaussian local-elevation function with k=5kJ/mol per MD step

simulation time **20ps**



simulation time **100ps**



Higher-energy conformations *are* sampled in 20 ps *local-elevation MD simulation*

Almost all conformations are sampled in 100 ps *LE*-MD simulation

Local-elevation search

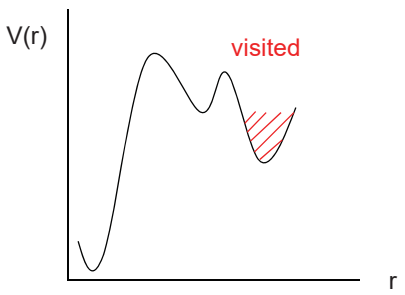
The local elevation simulation method

Normal simulation: relevant properties

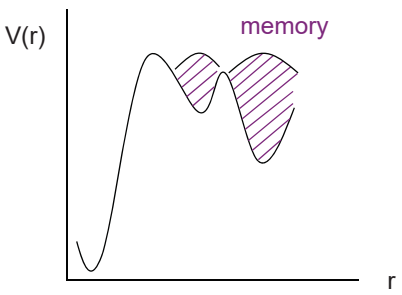
- Many conformers
  few visited
- Compact representation should be possible

Local-elevation simulation:

- run simulation
- store visited conformations (using compact representation)
- push system away when old conformation is seen



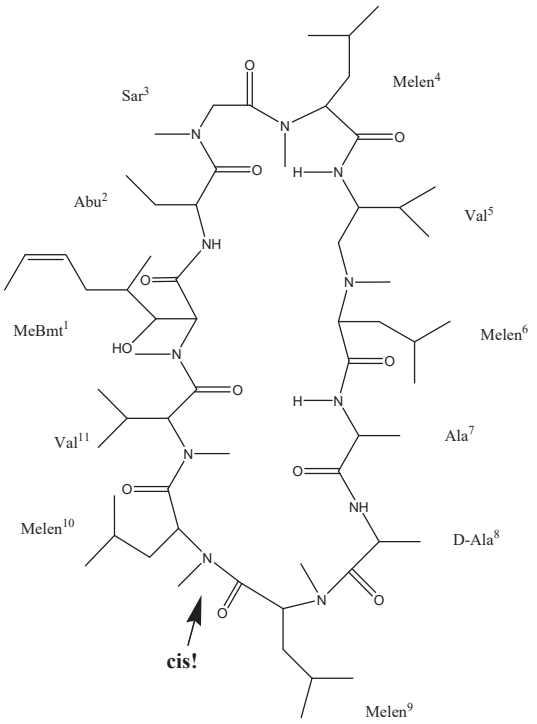
normal (MD)



Local elevation (MD)

Local-elevation search

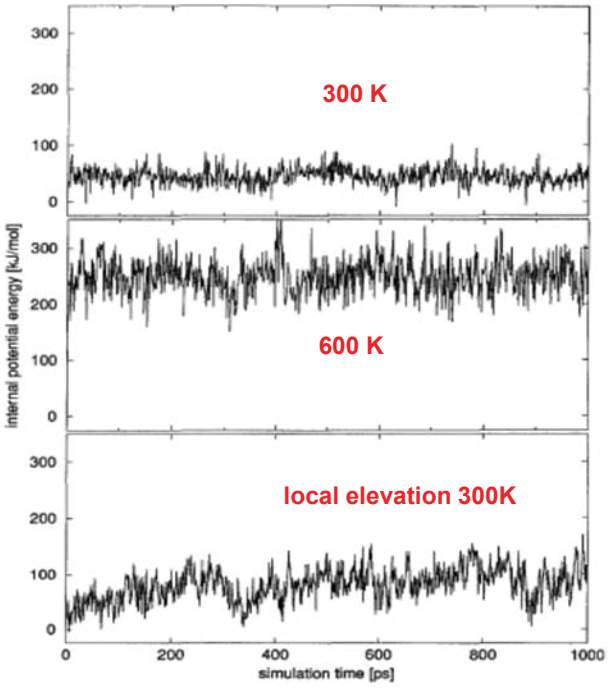
Cyclosporin A



- Amide bond (fixed to trans) ω-dihedral
- central bond of φ-dihedral
- central bond of ψ-dihedral

Local-elevation search

Cyclosporin A: potential energy



MD

MD

LE-MD

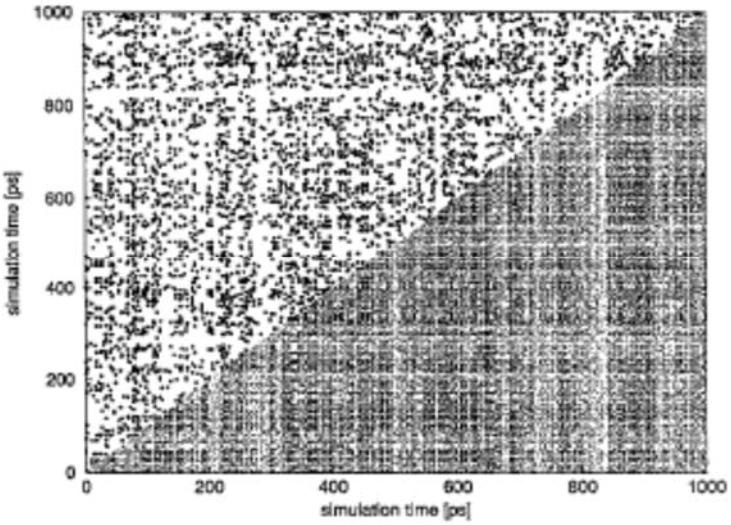
Local-elevation search

Cyclosporin A: similarity of conformations

Criterion: $\Delta\varphi_i$ ≤ 30° (upper), ≤ 45° (lower) } for each of the 11 φ-angles

164 = average number of visits of same conformer

SD simulation at 300 K



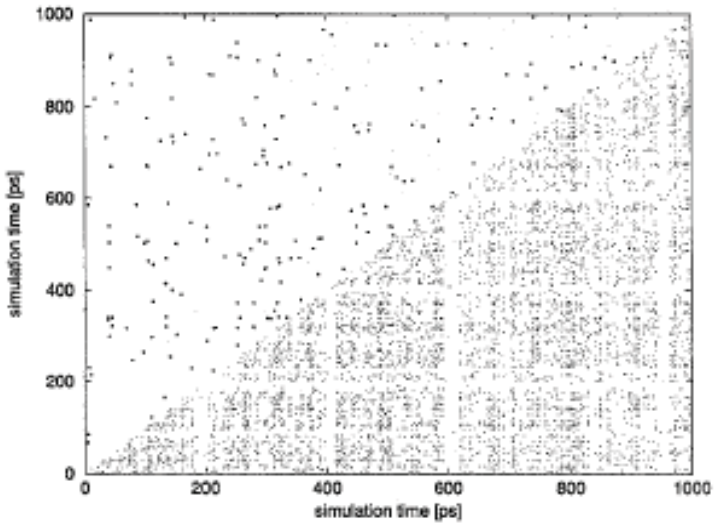
Local-elevation search

Cyclosporin A: similarity of conformations

Criterion: $\Delta\varphi_i$ ≤ 30° (upper), ≤ 45° (lower) } for each of the 11 φ-angles

26 visits on average

SD simulation at *600* K



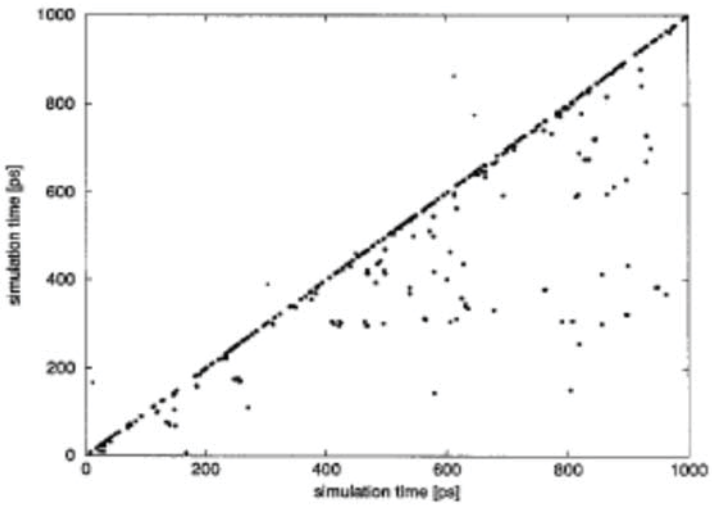
Local-elevation search

Cyclosporin A: similarity of conformations

Criterion: $\Delta\varphi_i$ ≤ *45°* (upper), ≤ *60°* (lower) } for each of the 11 φ-angles

1.6 visit on average

Local elevation simulation at *300* K



Local-elevation search

Ribonuclease A: RMS fluctuations in the loop region

Loop search

in vacuo

dashed: MD
solid: LE-MD

in solution

solid: LE-MD

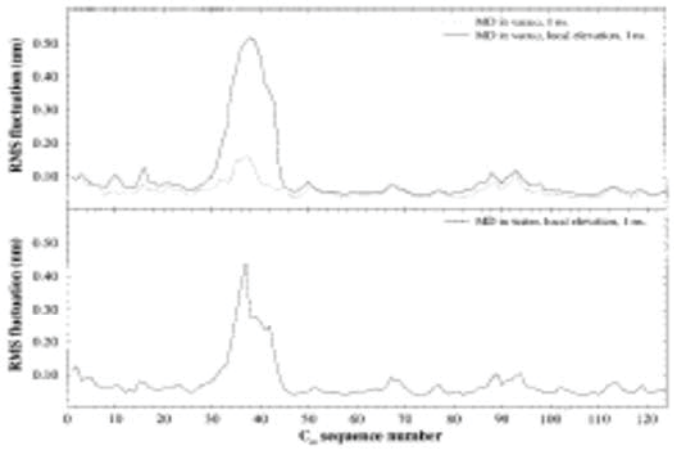


Figure 2. Root mean square fluctuations of atomic positions in ribonuclease A. Upper graph: simulations in vacuo with and without local elevation search. The simulation using local elevation search produces larger positional fluctuations than without, which is indicative of the larger conformational space searched. Lower graph: simulation in solvent with local elevation search. The fluctuations are smaller than those encountered in the vacuum simulations.

Local-elevation MD searches a much larger conformational space

Local-elevation search

Ribonuclease A: loop conformations

Standard MD **Local-elevation MD**

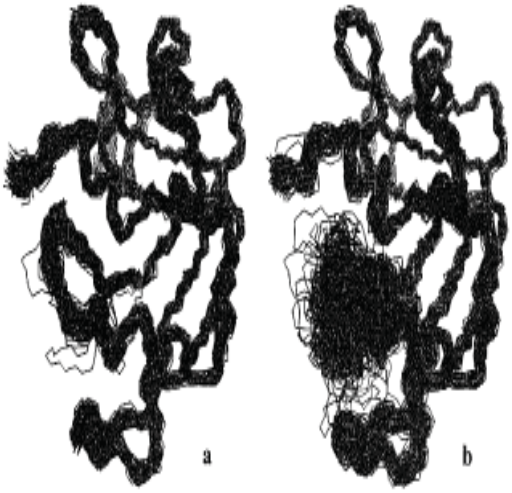


Figure 3. Superposition of 200 conformations of ribonuclease A taken at 5 ps intervals (total of 1 ns simulation time) in vacuo: (a) simulation without local elevation; (b) with local elevation search in the loop consisting of residues 33–43. The larger conformational space searched is apparent.

Local-elevation MD searches a much larger conformational space

Scott et al., J. Phys. Chem. ***A103*** *(1999) 3596-3607*

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods

Systematic search

Random search

Stepwise build-up

Genetic algorithm

Multicopy "sampling"

Distance geometry

Homology modelling

...

MD-based schemes for enhanced searching

Altered parameters

Altered masses
Adiabatic decoupling
Temperature annealing
High-temperature sampling
Parallel tempering

Altered potential energy

Use of soft-core atoms
Diffusion-equation search
Local-elevation search
Biasing (US)
incl. LEUS (+λ,FB,B&S), metadyn or EDS
Hamiltonian replica exchange

Altered dimensionality *(and potential energy)*

Four-dimensional MD
Coarse-grained MD
Multigraining
Essential dynamics

Altered prescription of motion

PEACS
SPEED
Monte Carlo sampling
Markov-state modeling
SWARM MD

Umbrella sampling (US)

- The idea of **umbrella sampling** (US) is to perform the MD simulation with a biased Hamiltonian

$$\mathcal{H}_b(\boldsymbol{r},\boldsymbol{p}) = \mathcal{H}(\boldsymbol{r},\boldsymbol{p}) + \mathcal{U}_b(\boldsymbol{r}) \qquad \mathcal{U}_b(\boldsymbol{r})$$ ***biasing potential***

→ The biasing potential should be designed in such a way that the biased trajectories sample **all the relevant states** with a **sufficient number of interconversion transitions**

Umbrella sampling (US)

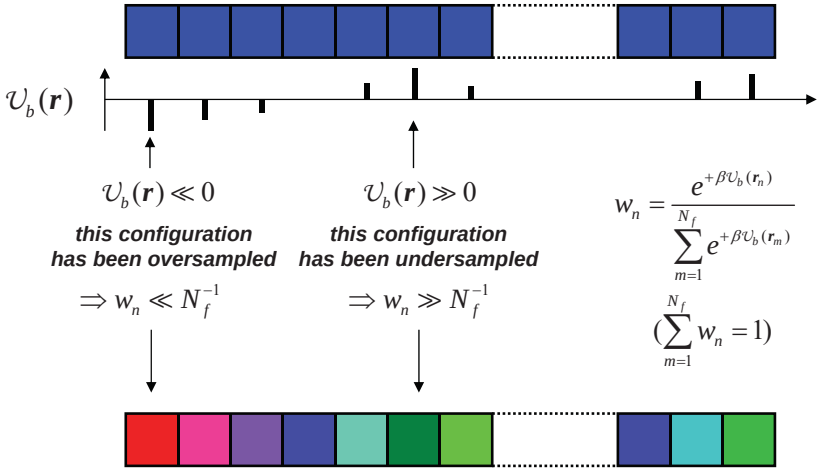
- Interpretation of the reweighting

→ Trajectory in the **original ensemble** (physical, unbiased) ***Boltzmann sampling: e.g. MD+thermostat***

all frames have equal weight in the ensemble average

$$w_n = N_f^{-1}$$

→ Trajectory in the **biased ensemble** (unphysical, biased)



sampling has been biased, ensemble averages are incorrect for the physical ensemble

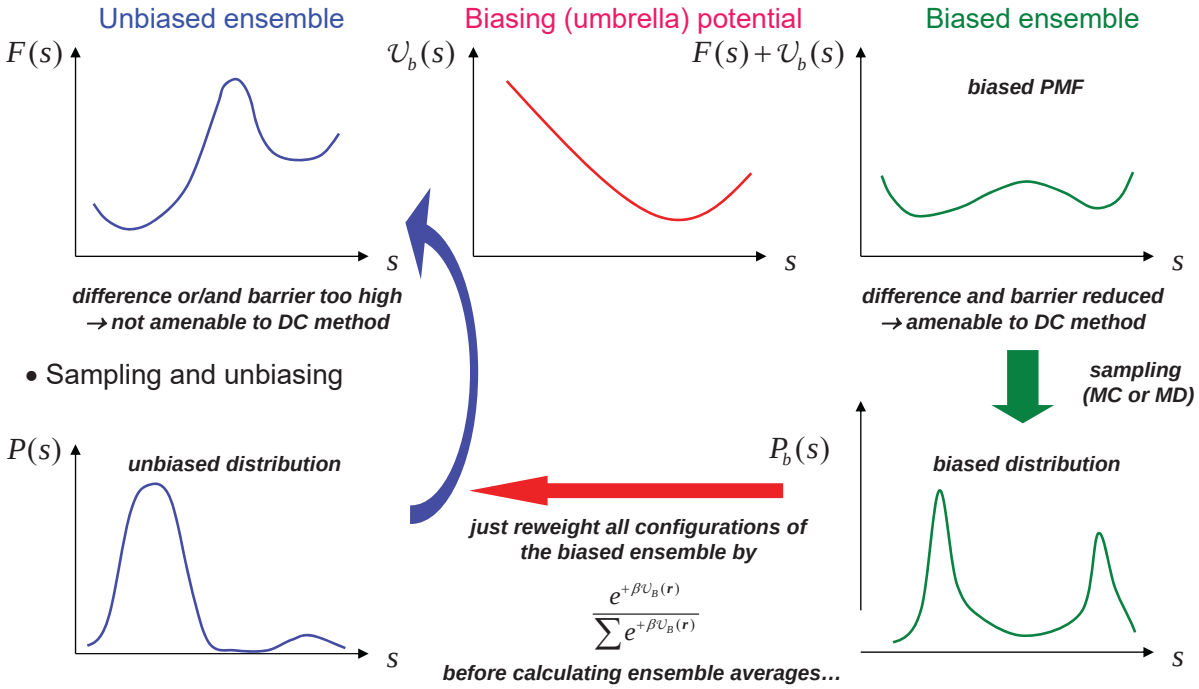
$$w_n = \frac{e^{+\beta\mathcal{U}_b(\boldsymbol{r}_n)}}{\sum_{m=1}^{N_f} e^{+\beta\mathcal{U}_b(\boldsymbol{r}_m)}} \qquad \left(\sum_{m=1}^{N_f} w_n = 1\right)$$

Reweighting

bias has been removed !

Umbrella sampling (US)

- With reaction coordinate



- Sampling and unbiasing

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods	***MD-based schemes for enhanced searching***	
	Altered parameters	***Altered potential energy***
Systematic search	**Altered masses**	**Use of soft-core atoms**
Random search	**Adiabatic decoupling**	**Diffusion-equation search**
Stepwise build-up	**Temperature annealing**	**Local-elevation search**
Genetic algorithm	**High-temperature sampling**	**Biasing (US)**
Multicopy "sampling"	**Parallel tempering**	**incl. LEUS (+λ,FB,B&S), metadyn or EDS**
Distance geometry		**Hamiltonian replica exchange**
Homology modelling	***Altered dimensionality*** *(and potential energy)*	***Altered prescription of motion***
...	**Four-dimensional MD**	**PEACS**
	Coarse-grained MD	**SPEED**
	Multigraining	**Monte Carlo sampling**
	Essential dynamics	**Markov-state modeling**
		SWARM MD

Local elevation umbrella sampling (LEUS)

• The *basic idea* underlying memory-based US is known under many names

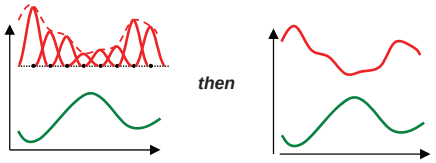
***deflation* (1969)**	***conformational flooding* (1995)**	***metadynamics* (2002)**	***gaussian-mixture US* (2009)**
***tunneling* (1985)**	***Engkvist-Karlström* (1996)**	***filling potential* (2003)**	***basin paving* (2010)**
***tabu search* (1989)**	***Wang-Landau* (2001)**	***adaptive reaction coordinate force* (2009)**	***LEUS* (2010)**
***local elevation* (1994)**	***adaptive biasing force* (2001)**		

→ but: implementation choices may affect a lot the *applicability* and *accuracy* in practice !

• Local elevation umbrella sampling (LEUS)

(our favorite flavor of this principle)
Hansen & Hünenberger
J. Comput. Chem., 31, 1 (2010).

Two-steps implementation



then

duration t_{LE}

LE BUILD-UP PHASE
non-equilibrium
→ rough biasing potential

$$G(Q) \approx -\mathcal{U}_{bias}(Q)$$

→ not very accurate or requires slow build-up

duration t_{US}

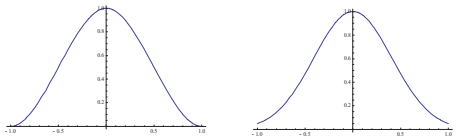
US SAMPLING PHASE
frozen biasing potential

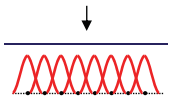
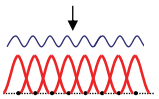
reweighting
→ "irons out" the roughness of the biasing potential

→ results less sensitive to build-up protocol

→ systematic error reduction upon increasing t_{US}

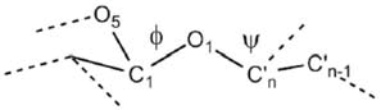
Truncated polynomial basis functions



	TRUNCATED POLYNOMIAL	***A COMPARABLE GAUSSIAN***
	$f(x) = (1-3x^2+2\lvert x^3\rvert)h(\lvert x\rvert-1)$	$f(x) = \exp(-3x^2)$
computation	***cheap***	***expensive***
range	***finite (next grid point)***	***formally infinite***
continuity	***yes (+derivative)***	***formally no (if cutoff)***
"ringing"	***no***	***yes***
even better: spline of order 2		 ***increases with build-up magnitude !***

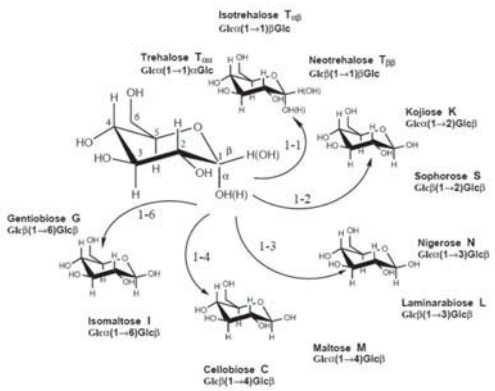
LEUS: Glucose-based disaccharides in water

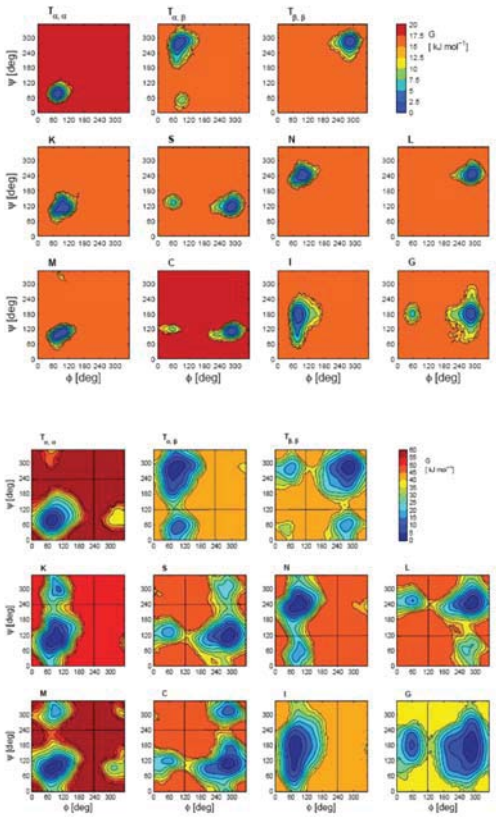
50 ns Plain MD
(initiated from X-ray structure)



2D subspace

(rotation timescales ~ 10 ns – 1 μs)





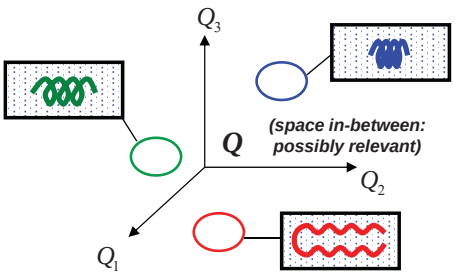
50+50 ns LEUS

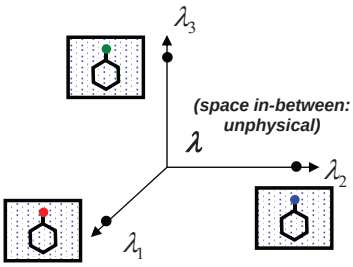
Perić-Hassler, Hansen, Baron & Hünenberger
Carbohydr. Res. 345, 1781 (2010)

LEUS calculations in multi-dimensional spaces

- LEUS (and analogous methods) can also be applied for calculations involving multi-dimensional relevant conformational or alchemical subspaces

Conformational ***Alchemical***





coupling scheme

$$\mathcal{H}_{cpl}(\boldsymbol{x};\lambda_1,\lambda_2,\lambda_3)$$

$$\begin{cases} \mathcal{H}_{cpl}(\boldsymbol{x};1,0,0)=\mathcal{H}_A(\boldsymbol{x}) \\ \mathcal{H}_{cpl}(\boldsymbol{x};0,1,0)=\mathcal{H}_B(\boldsymbol{x}) \\ \mathcal{H}_{cpl}(\boldsymbol{x};0,0,1)=\mathcal{H}_C(\boldsymbol{x}) \end{cases}$$

A Holy Graal of drug design !!!

→ but: the *memory costs* only allows for a limited number of dimensions

N_{LF} *number of local functions (grid points) required* ⇒ $N_{LF} = f \times M^N$

- M *number of grid points per dimension*
- N *dimensionality of the subspace*
- f *fraction of subspace to be mapped out*

→ and: the build-up time represents a further limitation on the number of dimensions

t_{LE} *required build-up time* ⇒ $t_{LE} = N_{LF} \times t_{LF}$ t_{LF} *required (average) visiting time per local function (grid point)*

$d\boldsymbol{Q}$

generally maps many orthogonal dimensions, with barriers and local minima

→ possible up to *N=3* or so (if only solvent and "fast" dof are averaged out)

e.g. $M = 30$ $N = 3$ $f = 0.5$ $t_{LF} = 5\,ps$ ⇒ $N_{LF} = 13500$ $t_{LE} = 67.5\,ns$

Fragment-based LEUS (FB-LEUS)

- LEUS (and analogous methods) can also be applied for calculations involving multi-dimensional relevant conformational or alchemical subspaces

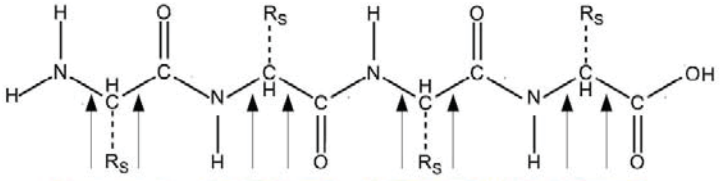
→ but what to do above *N=3* or so ???

e.g. decapeptide ϕ and ψ $M = 30$ $N = 20$ $f = 0.5$ $t_{LF} = 5\,ps$ ⇒ $N_{LF} = 1.7 \cdot 10^{29}$ $t_{LE} = 8.7 \cdot 10^{26}\,ns$ *(27 billion years !)*

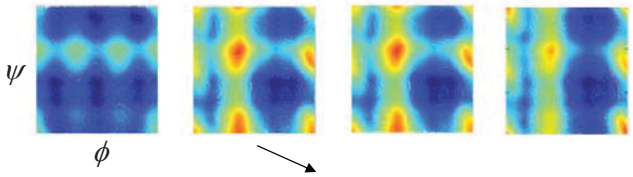
- A possible approach: fragment-based LEUS (FB-LEUS)

Hansen, Daura, & Hünenberger J. Chem. Theory Comput. 6, 2596 (2010).

e.g. Ala tetrapeptide in water



LEUS free-energy maps, blocked "monopeptide" fragments (LEUS 15+50 ns)



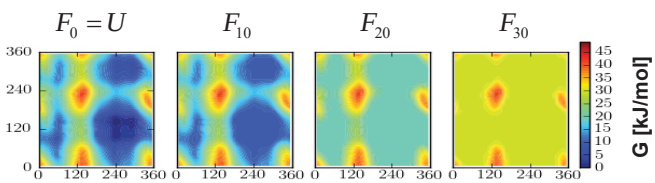
Assume:

$$G(\boldsymbol{Q}^{(8)}) \approx \sum G_n(\boldsymbol{Q}^{(2)})$$

Use: ↓

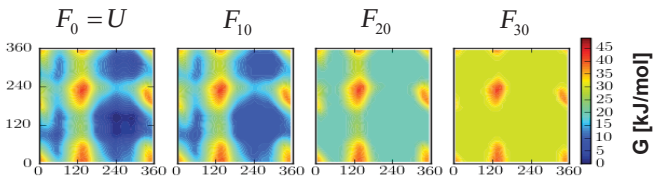
$$\mathcal{U}_{bias}(\boldsymbol{Q}^{(8)}) = \sum \mathcal{U}_{bias,n}(\boldsymbol{Q}^{(2)})$$

biased free-energy maps for various fragment-based biasing potentials (2D → acceptable N_{LF} & t_{LE})



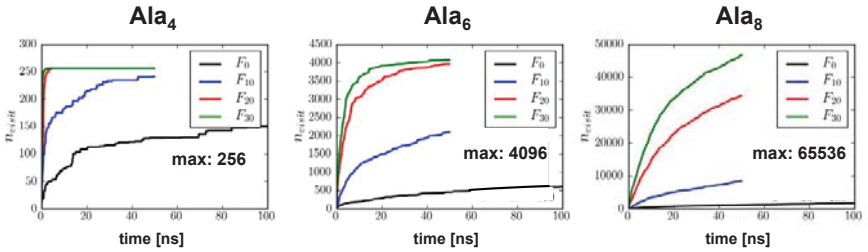
FB-LEUS: Ala oligopeptides in water

• Results



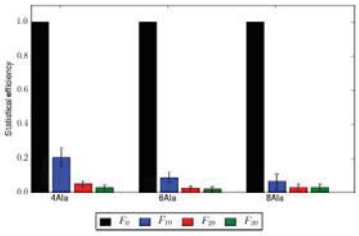
→ Searching efficiency (WOUAOUW !!!)

sampling phase unique conformations visited (four conformations per linkage)



→ Statistical efficiency (HEM...)

e.g. for F_{30}, the contribution of the sampled frames to the statistics relevant for the physical ensemble are 0.4, 0.04 and 0.002 % for Ala_4, Aal_6 and Ala_8



$$F = N_F^{-1} \exp[-\sum_{k=1}^{N_F} p_k \ln p_k]$$

$$p_k = \left\{ \sum_{l=1}^{N_F} \exp[\beta \mathcal{U}_{bias,l}] \right\}^{-1} \exp[\beta \mathcal{U}_{bias,k}]$$

$$\rightarrow F = \left\langle \exp[-\beta \mathcal{U}_{bias}] \right\rangle_{phys}^{-1} \exp[-\left\langle \beta \mathcal{U}_{bias} \right\rangle_{phys}]$$

• A complicated way of doing "random scanning" → much (much) less efficient then plain MD !!!

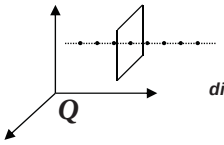
Basic requirements for a "good" memory-based biasing potential

(learning from the failure)

• **Low internal dimensionality**

→ acceptable memory and build-up duration costs

internal: referring to the memory grid, not to the relevant subspace itself

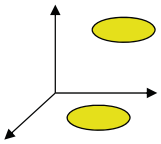


e.g. 1D bias in high-dimensional space

• **Minimal irrelevant volume**

→ acceptable sampling duration, high statistical efficiency

irrelevant: referring regions we are not interested in

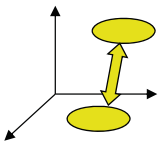


states or transition regions of interest

→ no need to sample the rest (except for transitions !)

• **Sufficient number of transitions** between relevant regions

→ accurate relative free energies



we want many transitions here

→ "narrow high-flow path" rather than "broad boulevard" (minimal irrelevant volume !)

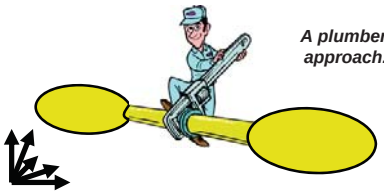
⇒ The biasing potential needs to have a **problem-adapted geometry** → *i.e. the proposed method is neither explorative nor extrapolative*

→ decide in advance what are the relevant and irrelevant regions (to be avoided, except narrow paths for transitions)

purely extrapolative schemes (single ensemble for predicting change to any state [not specified in advance]) are seldom reliable...

Ball-and-stick LEUS (B&S-LEUS)

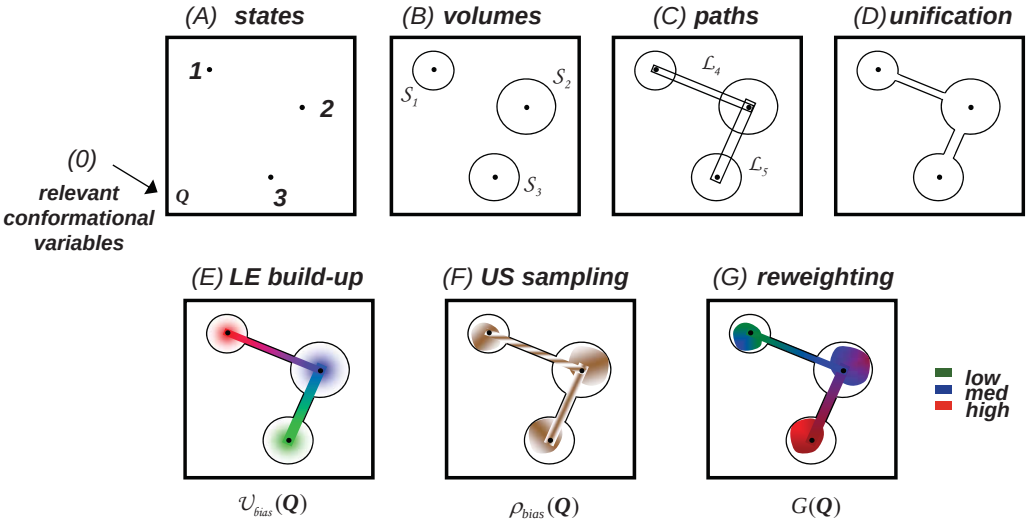
Hansen & Hünenberger J. Chem. Theory Comput. 6, 2622 (2010).



A plumber's approach...

Halvor Hansen (Now doing railway planning in Norway – only a pseudo-1D [actually 2D] problem... a pity !)

- Our attempt to fullfil all these requirements simultaneously
- Seven steps of the procedure



Simple 2D test system

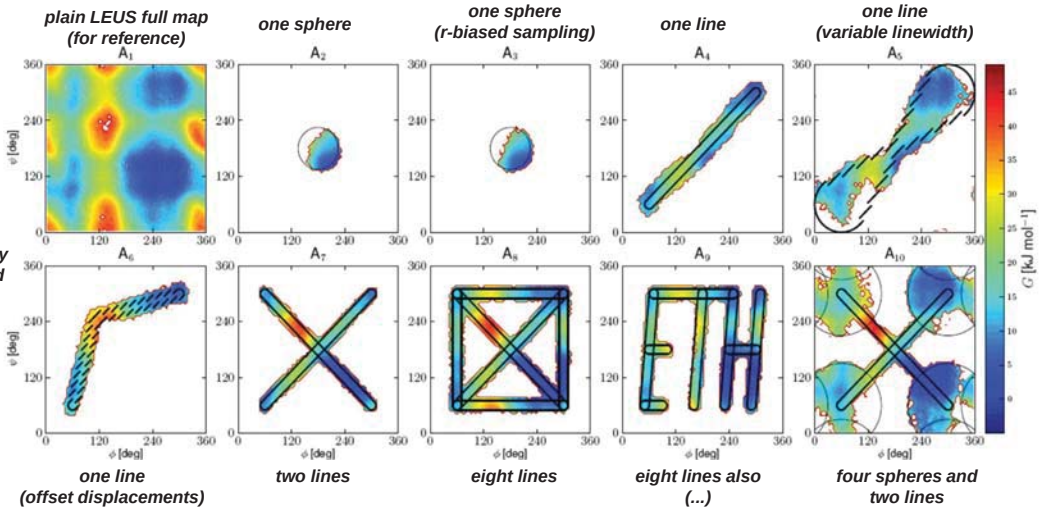
- Blocked alanine monopeptide in explicit (1300) water, GROMOS 53A6
- Conformational subspace $\boldsymbol{Q} = \sigma^{-1}(\phi, \psi)$ $\quad \sigma = 1^o$
- Various combinations of spheres and lines
- 10 grid points per sphere, 20 grid points per line
- 5-20 ns LE + 1-5 ns US (plain LEUS: 15+50)

peptide:

CH_3-CO-NH-CH(CH_3)-CO-NH-CO-CH_3

ϕ $\quad$ ψ

grid cells of 5° edge, maps optimally superimposed



→ Highly versatile in terms of active subspace (here surface) definition

→ Maps are identical within the sampled region

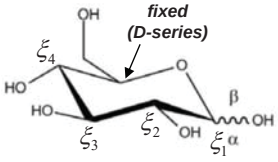
The "mother" of all hexopyranoses (7D)

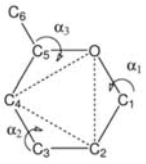
- Artificial hexopyranose in explicit (1200) water, GROMOS 56A$_{CARBO}$
  (we called it the "mother" of all hexopyranoses)

mixed conformational and alchemical

- Conformational subspace $\boldsymbol{Q} = \sigma^{-1}(\xi_1, \xi_2, \xi_3, \xi_4, \alpha_1, \alpha_2, \alpha_3) \quad \sigma = 1^o$

Epimery/anomery at ring carbons (improper dihedral)



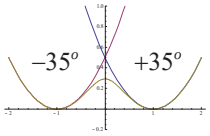


Pickett & Strauss ring puckering coordinates

$^4C_1 \rightarrow -35^o(1,1,1)$

$^1C_4 \rightarrow 35^o(1,1,1)$

- Double-well improper-dihedral potential (EDS)



Barrier of about 50 kJ mol^{-1}

→ the effect of this additional bias is also removed at the reweighting stage

- 32 spheres and 80 lines

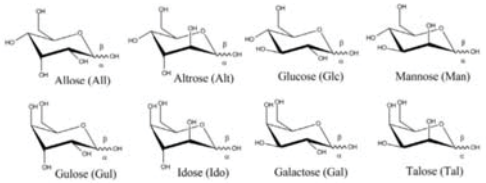
Spheres: $\boldsymbol{Q}_i = 35\left[2(b_{i,1}, b_{i,2}, b_{i,3}, b_{i,4}, b_{i,5}, b_{i,5}, b_{i,5}) - \boldsymbol{1}\right]$ *with* $\boldsymbol{b}_i = \boldsymbol{BIN}(i) \quad i = 1..32$

Lines: *+ Manhattan (Taxicab) tree of 80 lines [5 neighbors differing in 1 bit]*

We also tried a minimal-spanning tree of 31 lines (poorer convergence)

- 15 grid points per sphere, 21 grid points per line

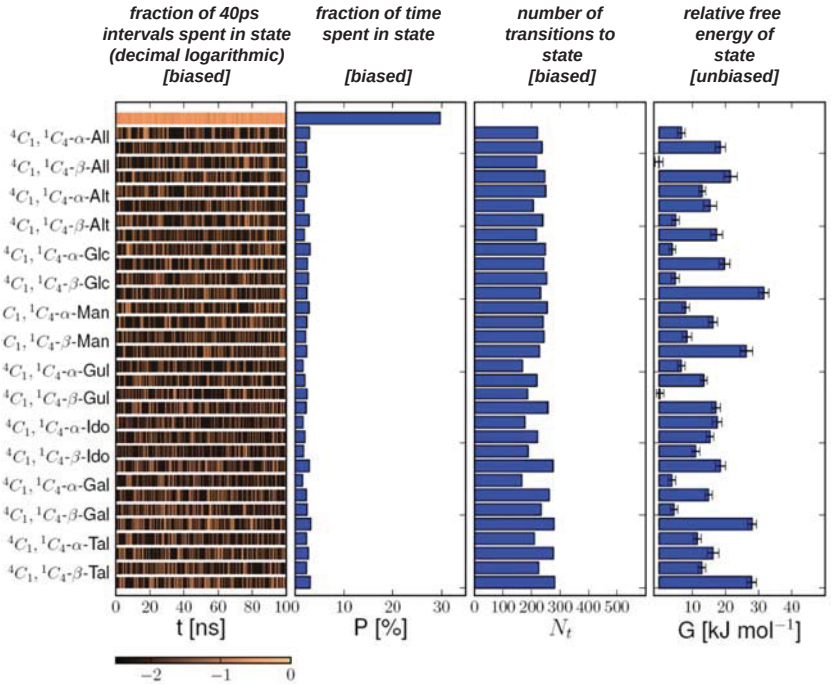
- 100 ns LE + 100 ns US



$^4C_1$ *epimers and anomers*

i	b_i	isomer	i	b_i	isomer
0	00000	$^4C_1$-α-idose	16	10000	$^4C_1$-β-idose
1	00001	$^1C_4$-α-idose	17	10001	$^1C_4$-β-idose
2	00010	$^4C_1$-α-altrose	18	10010	$^4C_1$-β-altrose
3	00011	$^1C_4$-α-altrose	19	10011	$^1C_4$-β-altrose
4	00100	$^4C_1$-α-talose	20	10100	$^4C_1$-β-talose
5	00101	$^1C_4$-α-talose	21	10101	$^1C_4$-β-talose
6	00110	$^4C_1$-α-mannose	22	10110	$^4C_1$-β-mannose
7	00111	$^1C_4$-α-mannose	23	10111	$^1C_4$-β-mannose
8	01000	$^4C_1$-α-gulose	24	11000	$^4C_1$-β-gulose
9	01001	$^1C_4$-α-gulose	25	11001	$^1C_4$-β-gulose
10	01010	$^4C_1$-α-allose	26	11010	$^4C_1$-β-allose
11	01011	$^1C_4$-α-allose	27	11011	$^1C_4$-β-allose
12	01100	$^4C_1$-α-galactose	28	11100	$^4C_1$-β-galactose
13	01101	$^1C_4$-α-galactose	29	11101	$^1C_4$-β-galactose
14	01110	$^4C_1$-α-glucose	30	11110	$^4C_1$-β-glucose
15	01111	$^1C_4$-α-glucose	31	11111	$^1C_4$-β-glucose

The "mother" of all hexopyranoses (7D)



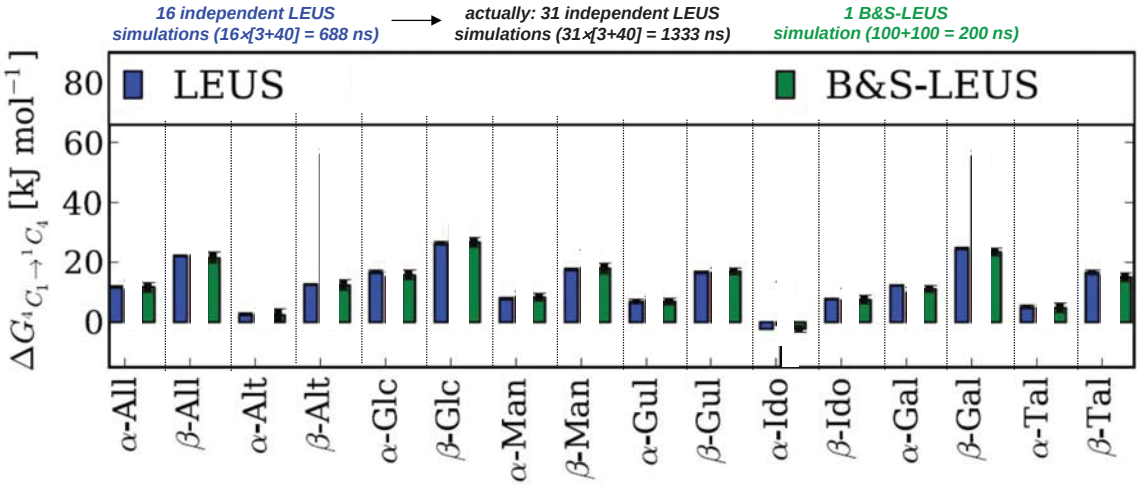
US phase 32 states grouped by chair pairs

→ all states are visited with almost equal probabilities

→ only 30% of the configurations are unassigned (small irrelevant volume, mainly the lines)

→ as many as 200-300 transitions to each state

The "mother" of all hexopyranoses (7D)



→ agreement (rmsd 0.7 kJ mol^{-1}) is within statistical error

→ for comparable statistical errors, and for the 31 calculated free-energy changes (chair-chair, but also epimerization and anomerization), there is a reduction of a factor 31 in the number of simulations and of a factor 6.7 in CPU time

→ with window-based US (rather than LEUS or B&S-LEUS), this calculation would represent a tremendous amount of human effort and CPU time !!!

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods	***MD-based schemes for enhanced searching***	
	Altered parameters	***Altered potential energy***
Systematic search	**Altered masses**	**Use of soft-core atoms**
Random search	**Adiabatic decoupling**	**Diffusion-equation search**
Stepwise build-up	**Temperature annealing**	**Local-elevation search**
Genetic algorithm	**High-temperature sampling**	**Biasing (US)**
Multicopy "sampling"	**Parallel tempering**	**incl. LEUS (+λ,FB,B&S), metadyn or EDS**
		⇨ **Hamiltonian replica exchange**
Distance geometry	***Altered dimensionality*** *(and potential energy)*	***Altered prescription of motion***
Homology modelling	**Four-dimensional MD**	**PEACS**
...	**Coarse-grained MD**	**SPEED**
	Multigraining	**Monte Carlo sampling**
	Essential dynamics	**Markov-state modeling**
		SWARM MD

Hamiltonian replica exchange

- Simulate *N* **independent replicas** of the system in parallel, each at a **different Hamiltonian**
- At **regular intervals** τ_{exc}, we attempt to **swap configurations** (coordinates and velocities) between two adjacent systems

Probability of the initial two-system state

$$p_i \sim \exp\left(-\beta(\mathcal{H}(\boldsymbol{x})+\mathcal{H}'(\boldsymbol{x}'))\right)$$

➡

Probability of the swapped two-system state

$$p_f \sim \exp\left(-\beta(\mathcal{H}(\boldsymbol{x}')-\mathcal{H}'(\boldsymbol{x}))\right)$$

$$\boldsymbol{x} = (\boldsymbol{r}, \boldsymbol{p})$$

- To ensure that the system pair remains Boltzmann-distributed, we **accept or reject** the swap according to a **Monte Carlo criterion**

$$P_{acc}(i \to f) = \begin{cases} 1 & if \quad p_f > p_i \\ p_f / p_i & otherwise \end{cases} \quad \textit{or} \quad = \begin{cases} 1 & if \quad \Delta \le 0 \\ \exp(-\Delta) & otherwise \end{cases}$$

with $p_f / p_i = \exp\left(-\beta(\mathcal{H}(\boldsymbol{x}')-\mathcal{H}(\boldsymbol{x})+\mathcal{H}'(\boldsymbol{x})-\mathcal{H}'(\boldsymbol{x}'))\right)$ *with* $\Delta = \beta(\mathcal{H}(\boldsymbol{x}')-\mathcal{H}(\boldsymbol{x})+\mathcal{H}'(\boldsymbol{x})-\mathcal{H}'(\boldsymbol{x}'))$

- It is easily seen that this procedure results in a set of *N* **canonical ensembles,** each at one of the selected Hamiltonians

This means in particular that the dynamics is nonsense...

  → if you look at **one Hamiltonian**, the trajectory is **discontinuous** (systems come and go)

  → if you look at **one system**, the trajectory **hops across Hamiltonians** (enabling enhanced sampling when there are orthogonal barriers!)

  → this is analogous to parallel tempering

Hamiltonian replica exchange

- Two common applications

  → Replica exchange in the **coupling parameter** λ in an alchemical perturbation

  → Replica exchange in the **biasing strength** in an US calculation (ground system = no bias)

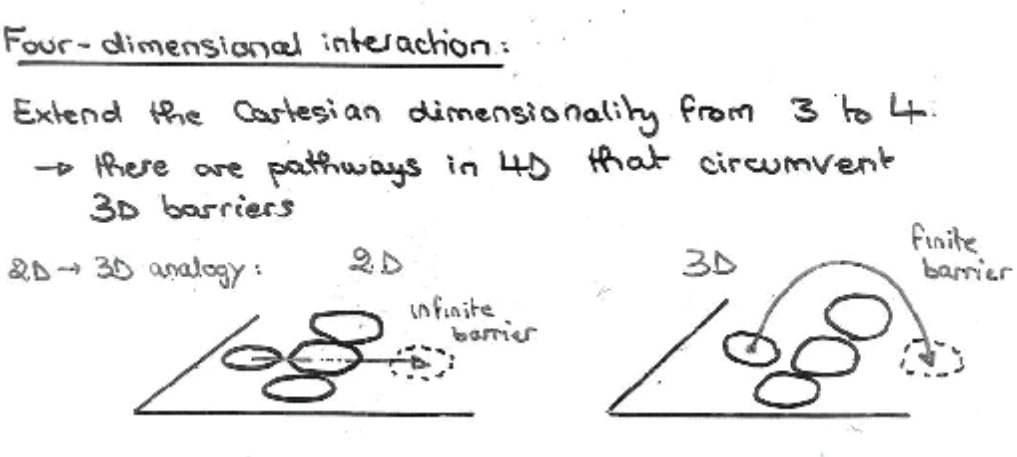
e.g. FB-LEUS + H-REX is a powerful combo !

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods	*MD-based schemes for enhanced searching*	
	Altered parameters	***Altered potential energy***
Systematic search	**Altered masses**	**Use of soft-core atoms**
Random search	**Adiabatic decoupling**	**Diffusion-equation search**
Stepwise build-up	**Temperature annealing**	**Local-elevation search**
Genetic algorithm	**High-temperature sampling**	**Biasing (US)**
Multicopy "sampling"	**Parallel tempering**	**incl. LEUS (+λ,FB,B&S), metadyn or EDS**
		Hamiltonian replica exchange
Distance geometry	***Altered dimensionality*** *(and potential energy)*	***Altered prescription of motion***
Homology modelling	**Four-dimensional MD**	**PEACS**
...	**Coarse-grained MD**	**SPEED**
	Multigraining	**Monte Carlo sampling**
	Essential dynamics	**Markov-state modeling**
		SWARM MD

Molecular dynamics in four dimensions



Molecular dynamics in four dimensions

Techniques to pass over barriers
MD in more than 3 dimensions

van Schaik et al., J. Mol. Biol. 234 (1993) 751

The 4D-space:

- 4 linearly independent orthonormal } basis vectors: $e_x\ e_y\ e_z\ e_\omega$ ($1^{st}\ 2^{nd}\ 3^{rd}\ 4^{th}$)
- arbitrary vector: $\vec{r} = xe_x + ye_y + ze_z + \omega e_\omega$
- scalar or dot product of $\vec{r}$ and $\vec{r}'$: $\vec{r} \cdot \vec{r}' = xx' + yy' + zz' + \omega\omega'$
- length or distance $r_{ij}^2 = x_{ij}^2 + y_{ij}^2 + z_{ij}^2 + \omega_{ij}^2$
- vector or cross product cannot be defined

Hamilton in 4D:

$$H_{4D} = \sum_{i=1}^{N} \frac{1}{2} m_i \left[v_{xi}^2 + v_{yi}^2 + v_{zi}^2 \right] + \sum_{i=1}^{N} \frac{1}{2} m_i v_{\omega i}^2 + V_{phys}^{scalar}\left(\{\vec{r}_i\}\right) + V_{phys}^{vector}\left(\{\vec{r}_i\}\right) + \left\{ \sum_{i=1}^{N} \frac{1}{2} k_\omega \omega_i^2 \right\}$$

V_{phys}^{scalar} = bonds, angles, torsional angles, non-bonded interactions, NOE restraints

V_{phys}^{vector} = improper torsions (chirality), X-ray restraints

Equations of motion in 4D:

- Newton for x, y, z and ω
- start from $\{\omega_i\} \neq 0$
- couple to separate temperature baths

Molecular dynamics in four dimensions

Backprojection from 4D to 3D:

1. - use penalty function $k_\omega \rightarrow \infty$
   - reduce T_ω to zero
   - perform rotation in 4D to minimize 3D projection
2. - decouple ω_i coordinates from $x_i\ y_i\ z_i$

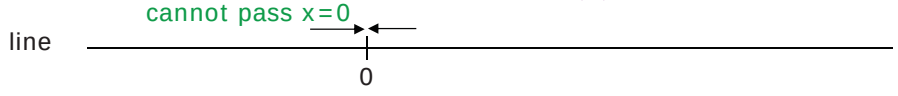
Extension of dimensionality:

Interaction function: $V(\vec{r}) = \frac{1}{r^2}$ $\qquad r^2 = x^2 + y^2 + z^2 + \ldots$

One dimension: $V(x) = \frac{1}{x^2}$

↑

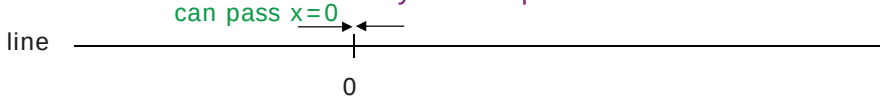
pole at $x=0 \Rightarrow V(0) = \infty$



Two dimensions: $V(x,y) = \frac{1}{x^2 + y^2}$

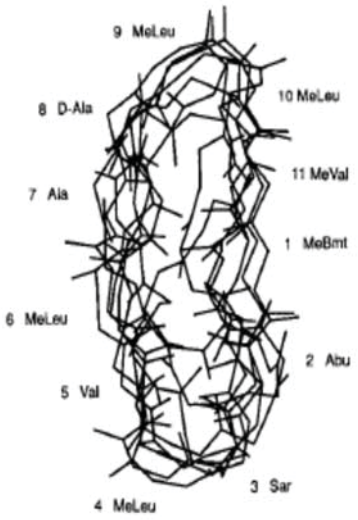
↑

for $y \neq 0$ no pole at $x=0$



Molecular dynamics in four dimensions

Cyclosporin A



11 residues
49 torsional angles
57 NOE distance restraints

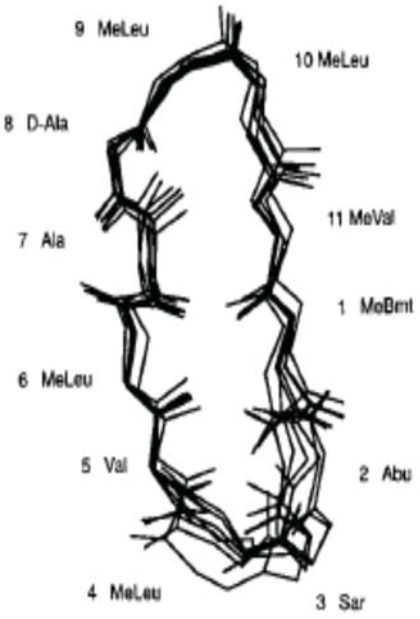
1. **distance geometry**
   27 structures
   ↓
   9 classes
   very difficult structures
   high energy barriers between them

2. **structure refinement**
   standard 3D-MD
   1-2 correct

Challenge: how to get *all 9* different starting structures *converged* to the lowest-energy one

Molecular dynamics in four dimensions

Cyclosporin A



seven 4D-MD refined structures

are well converged

Molecular dynamics in four dimensions

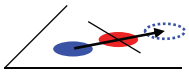
Results of the refinement of Cyclosporin A using DG, 3D-MD, PEACS and 4D-MD

	Distance geometry (DG)			3D-MD annealed from 1200 K			PEACS annealed from 1200 K			4D-MD annealed from 1200 K		
Struc-ture	E_{pot} kJ/ mol	Σ_{viol} Å	Δr_{MDS1} Å	E_{pot} kJ/ mol	Σ_{viol} Å	Δr_{MDS1} Å	E_{pot} kJ/ mol	Σ_{viol} Å	Δr_{MDS1} Å	E_{pot} kJ/ mol	Σ_{viol} Å	Δr_{MDS1} Å
xclass11	195	3.1	2.3	57	3.2	3.1	0	1.4	2.0	-39	0.5	0.4
xclass21	77	3.0	1.7	51	4.0	1.7	-65	0.9	0.3	-70	0.2	0.3
xclass31	214	2.9	3.0	35	3.9	2.9	-66	1.1	0.4	17	1.5	1.6
xclass41	133	4.0	1.6	-11	0.9	1.8	-26	1.2	1.9	-52	0.5	0.4
xclass51	235	5.7	1.3	73	3.6	2.8	-65	1.0	0.6	-25	1.8	0.4
xclass61	318	3.2	3.4	73	4.5	3.5	-8	2.0	1.9	-62	0.5	0.6
xclass71	142	3.9	1.4	-1	3.2	1.9	27	1.1	1.3	134	4.5	2.3
xclass81	160	2.7	2.0	31	2.5	2.5	150	1.2	2.1	-47	0.8	0.4
xclass91	144	3.0	1.2	-82	1.4	0.3	-50	2.5	0.5	-75	0.8	0.2
correct:		**0**			**1-2**			**3-4**			**6**	

deviation from correct structure

4D-MD ?

- Possibly the reason I decided to do my Ph.D. with Wilfred (1992) !



$$\mathcal{U}_{3D}(\mathbf{r}_{3D}) \rightarrow \mathcal{U}_{4D}(\mathbf{r}_{4D})$$

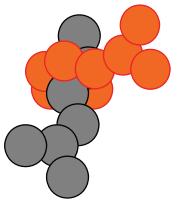
- I tried it during my post-doc in San Diego (1998)... but...

- The three most scary scenes of nature...

The fight of the snake and the mongoose

The Australian "choking" tree

A Phe trying to behead a Met (projected from 4D space)



Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

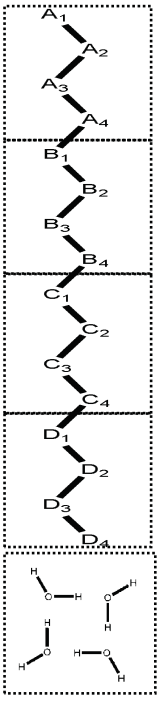
Systematic and heuristic search methods	*MD-based schemes for enhanced searching*	
	Altered parameters	*Altered potential energy*
Systematic search	Altered masses	Use of soft-core atoms
Random search	Adiabatic decoupling	Diffusion-equation search
Stepwise build-up	Temperature annealing	Local-elevation search
Genetic algorithm	High-temperature sampling	Biasing (US)
Multicopy "sampling"	Parallel tempering	incl. LEUS (+λ,FB,B&S), metadyn or EDS
		Hamiltonian replica exchange
Distance geometry	*Altered dimensionality* (*and potential energy*)	*Altered prescription of motion*
Homology modelling	Four-dimensional MD	PEACS
...	Coarse-grained MD	SPEED
	Multigraining	Monte Carlo sampling
	Essential dynamics	Markov-state modeling
		SWARM MD

Multigraining

Coarse-grained versus fine-grained models

liquid alkanes: hexadecane

AL (λ=0)
All-atom model
(non-hydrogen)
16 (CH_2 or CH_3) atoms



MAP
"mapped"
all-atom
configurations

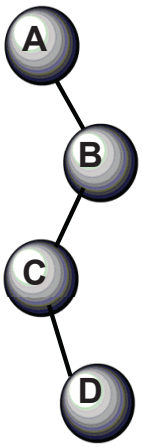
Centre of mass
$A_1 - A_4$

Centre of mass
$B_1 - B_4$

Centre of mass
$C_1 - C_4$

Centre of mass
$D_1 - D_4$

CG (λ=1)
Coarse-grained model
4 atoms



Compare: - structural characteristics
- energetic / entropic characteristics

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods

Systematic search

Random search

Stepwise build-up

Genetic algorithm

Multicopy "sampling"

Distance geometry

Homology modelling

...

MD-based schemes for enhanced searching

Altered parameters

Altered masses

Adiabatic decoupling

Temperature annealing

High-temperature sampling

Parallel tempering

Altered potential energy

Use of soft-core atoms

Diffusion-equation search

Local-elevation search

Biasing (US)

incl. LEUS (+λ,FB,B&S), metadyn or EDS

Hamiltonian replica exchange

Altered dimensionality
(and potential energy)

Four-dimensional MD

Coarse-grained MD

Multigraining

Essential dynamics

Altered prescription of motion

PEACS

SPEED

Monte Carlo sampling

Markov-state modeling

SWARM MD

Multigraining

Algorithm for mixed FG/CG simulation

Multigraining MD

fine-grained
topology,
configuration: $\mathbf{r}^{fg}$

coarse-grained
topology,
configuration: $\mathbf{r}^{cg}$

1. update coarse-grained configuration (using virtual-grain definition)

2a. calculate λ-dependent energies and forces (according to coupling parameter λ)

2b. calculate λ -dependent energies and forces (according to coupling parameter λ)

3. distribute forces from coarse-grained on fine-grained particles using virtual-grain definition $g(\mathbf{r}^{fg})$ and add potential energy terms

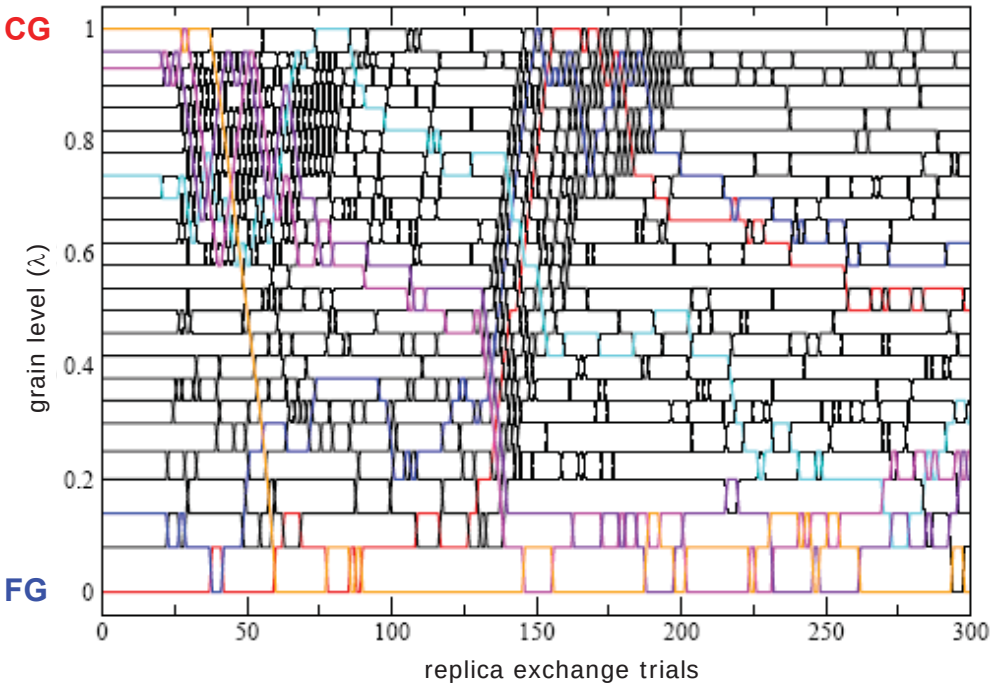
4a. propagate velocities and positions using the leap-frog scheme

4b. propagate velocities and positions of non-mapped solvent particles using the leap-frog scheme

Replica exchange

Multi-grained simulation of liquid octane

grain level of the 24 replicas during 300 replica exchange steps

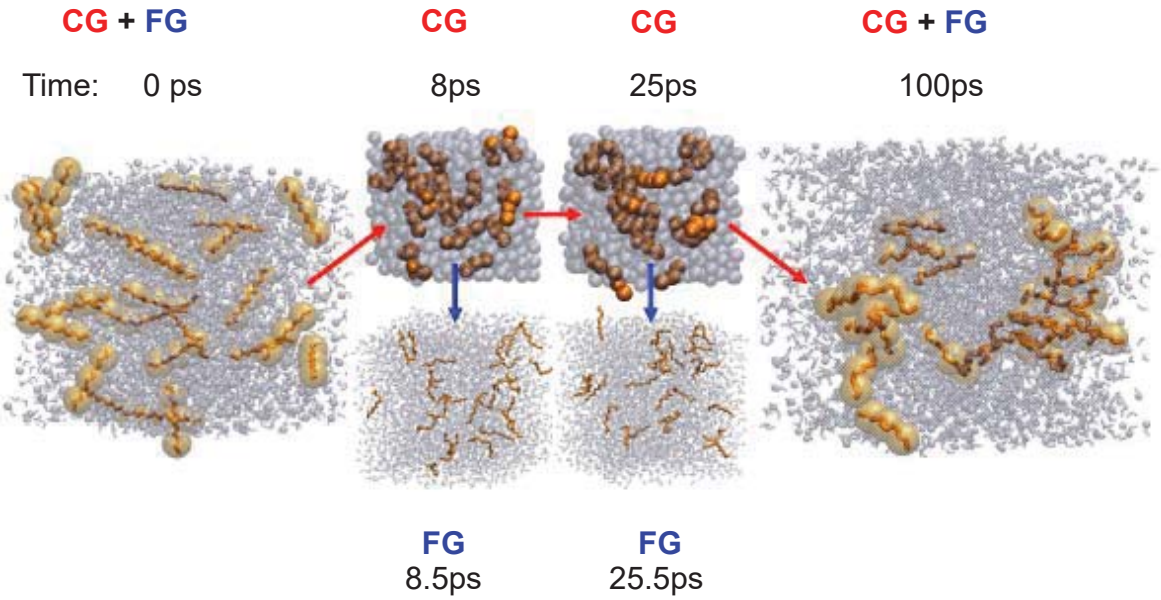


FG/CG replica-exchange simulation enhances sampling

Multigraining

Multi-grained simulation of 25 hexadecanes in water

M. Christen & W.F. van Gunsteren, J. Chem. Phys, ***124*** *(2006) DOI:10.1063/1.2187488*



CG level simulation with occasional switching to FG level enhances exploration of FG conformational space

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods

Systematic search

Random search

Stepwise build-up

Genetic algorithm

Multicopy "sampling"

Distance geometry

Homology modelling

...

MD-based schemes for enhanced searching

Altered parameters

Altered masses
Adiabatic decoupling
Temperature annealing
High-temperature sampling
Parallel tempering

Altered potential energy

Use of soft-core atoms
Diffusion-equation search
Local-elevation search
Biasing (US)
incl. LEUS (+λ,FB,B&S), metadyn or EDS
Hamiltonian replica exchange

Altered dimensionality *(and potential energy)*

Four-dimensional MD
Coarse-grained MD
Multigraining
Essential dynamics

Altered prescription of motion

PEACS
SPEED
Monte Carlo sampling
Markov-state modeling
SWARM MD

Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods

Systematic search

Random search

Stepwise build-up

Genetic algorithm

Multicopy "sampling"

Distance geometry

Homology modelling

...

MD-based schemes for enhanced searching

Altered parameters

Altered masses
Adiabatic decoupling
Temperature annealing
High-temperature sampling
Parallel tempering

Altered potential energy

Use of soft-core atoms
Diffusion-equation search
Local-elevation search
Biasing (US)
incl. LEUS (+λ,FB,B&S), metadyn or EDS
Hamiltonian replica exchange

Altered dimensionality *(and potential energy)*

Four-dimensional MD
Coarse-grained MD
Multigraining
Essential dynamics

Altered prescription of motion

PEACS
SPEED
Monte Carlo sampling
Markov-state modeling
SWARM MD

PEACS

Modified molecular dynamics (PEACS):

Potential Energy Annealing Conformational Search **Enhances barrier crossing**

MD plus $\frac{dV_{pot}(t)}{dt} = \frac{1}{\tau_v}\left[V_{reference} - V_{pot}(t)\right]$

slowly lowered

R.C. van Schaik et al., J. Comp.-Aided Mol. Des. ***6*** *(1992) 97-112*

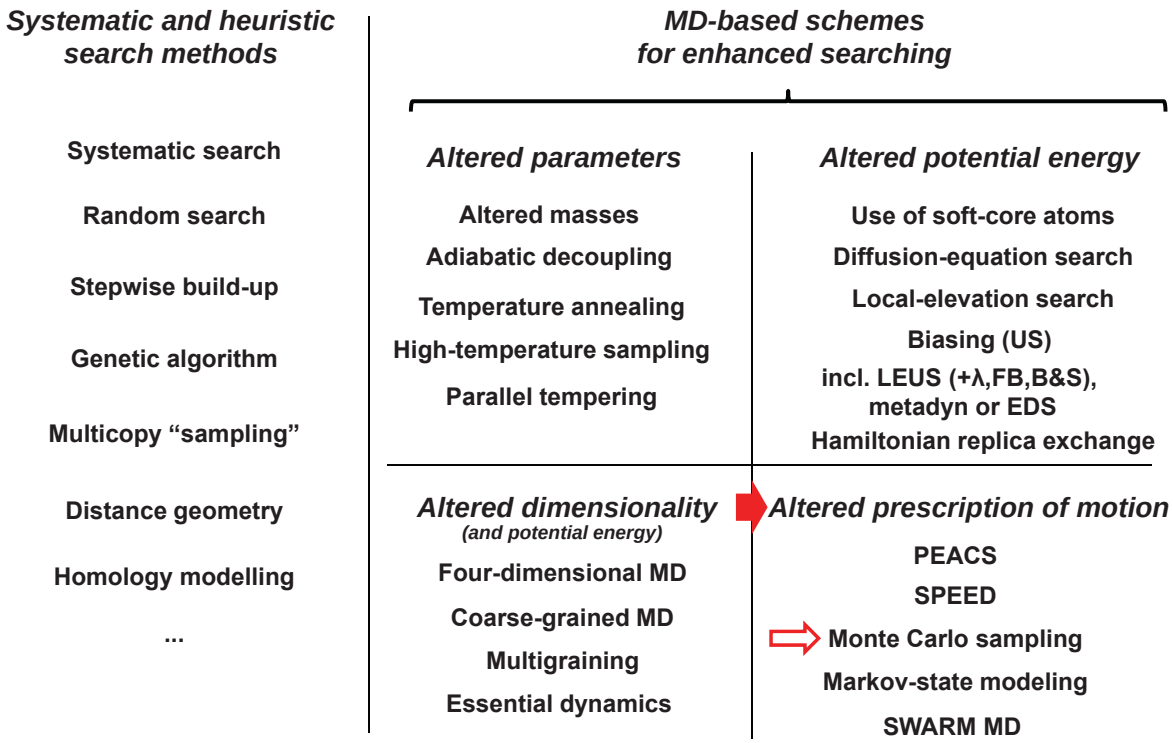
Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods	***MD-based schemes for enhanced searching***	
	Altered parameters	***Altered potential energy***
Systematic search	**Altered masses**	**Use of soft-core atoms**
Random search	**Adiabatic decoupling**	**Diffusion-equation search**
Stepwise build-up	**Temperature annealing**	**Local-elevation search**
Genetic algorithm	**High-temperature sampling**	**Biasing (US)**
Multicopy "sampling"	**Parallel tempering**	**incl. LEUS (+λ,FB,B&S), metadyn or EDS**
		Hamiltonian replica exchange
Distance geometry	***Altered dimensionality*** *(and potential energy)*	***Altered prescription of motion***
Homology modelling	**Four-dimensional MD**	**PEACS**
...	**Coarse-grained MD**	⇨ **SPEED**
	Multigraining	**Monte Carlo sampling**
	Essential dynamics	**Markov-state modeling**
		SWARM MD

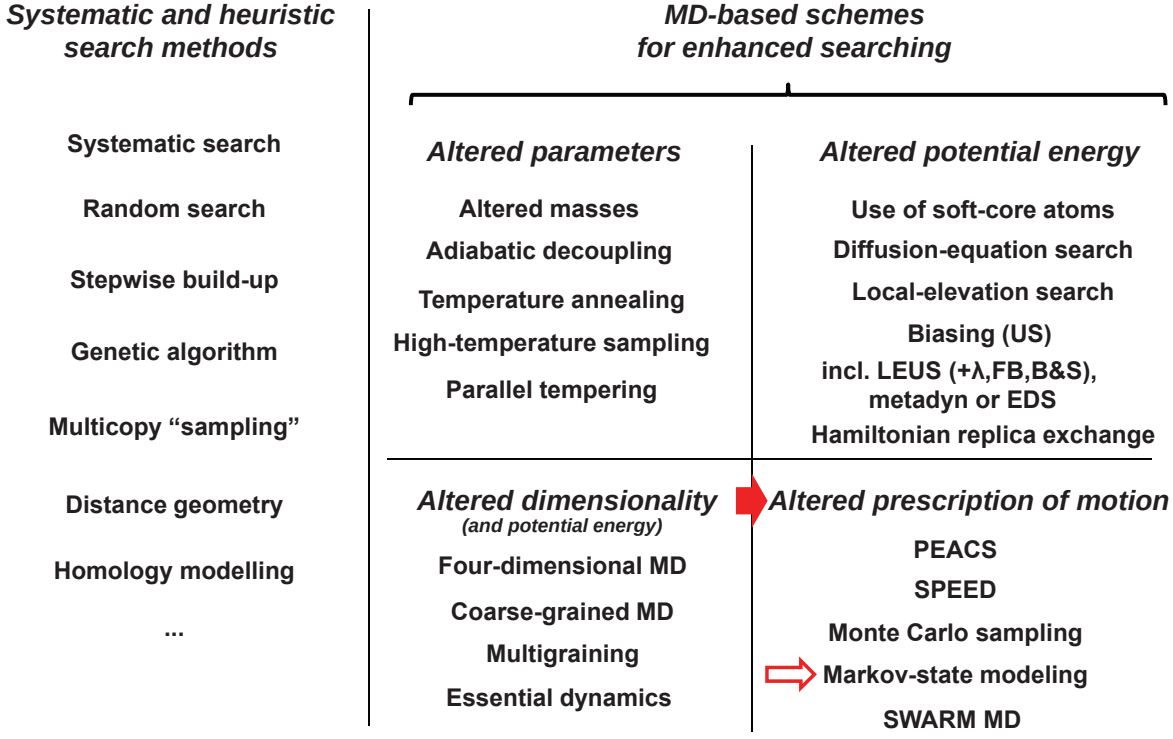
Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are



Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are



Overview of conformational search methods

- The (rough) categories of method to search conformational space, along with illustrative examples (out of many more!) are

Systematic and heuristic search methods	*MD-based schemes for enhanced searching*	
	Altered parameters	*Altered potential energy*
Systematic search	Altered masses	Use of soft-core atoms
Random search	Adiabatic decoupling	Diffusion-equation search
Stepwise build-up	Temperature annealing	Local-elevation search
Genetic algorithm	High-temperature sampling	Biasing (US)
Multicopy "sampling"	Parallel tempering	incl. LEUS (+λ,FB,B&S), metadyn or EDS
		Hamiltonian replica exchange
Distance geometry	*Altered dimensionality* (and potential energy)	*Altered prescription of motion*
Homology modelling	Four-dimensional MD	PEACS
...	Coarse-grained MD	SPEED
	Multigraining	Monte Carlo sampling
	Essential dynamics	Markov-state modeling
		⇨ SWARM MD

SWARM MD

Multi-copy search techniques: the SWARM method

Idea: combine a swarm of molecules with molecular trajectories into a cooperative system that searches conformational space (like a swarm of insects)

Implementation:

each molecule is, in addition to the physical forces,

subject to (artificial) forces that drive the trajectory of each molecule toward an average of the trajectories of the swarm of molecules

Huber and van Gunsteren: J.Phys.Chem. ***A102*** *(1998) 5937-5943*

SWARM-MD: Searching configurational space by cooperative molecular dynamics

P.H. Hünenberger

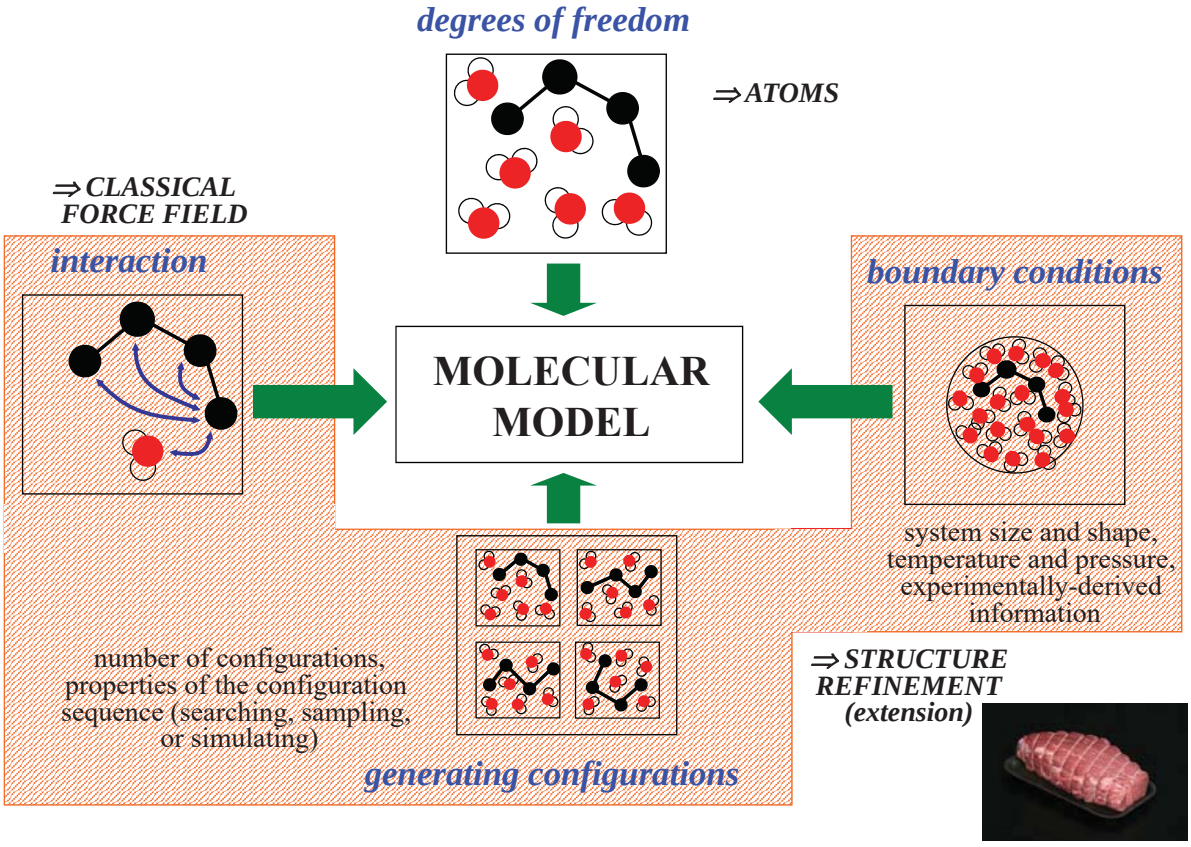
COMPUTER SIMULATION OF MOLECULAR SYSTEMS

Herbstsemester 2019
Tuesday 9:45-11:30 a.m.
HCI D2

Lecture 529-0004-00
www.csms.ethz.ch/education/CSCBP

LECTURE 10 (WEEK 11):
Structure refinement

★ **Four basic choices defining a molecular model**



Overview of structure refinement

- The **problem** of **structure refinement**
  - → A set of **experimental observables** provides **average** information (over molecules and over time) on the sample present in a test tube

$\{Q^{exp}\}$ *quantities measured experimentally*

  - → The proper way to account for this information relies on **statistical mechanics**

the ensemble of all conformations $\{\boldsymbol{r}\}$

obeys a Boltzmann distribution $P(\boldsymbol{r}) \sim \exp[-\beta \mathcal{U}(\boldsymbol{r})]$

and each conformation leads to a signal $Q(\boldsymbol{r})$

so, the experimental signal is obtained by ensemble averaging

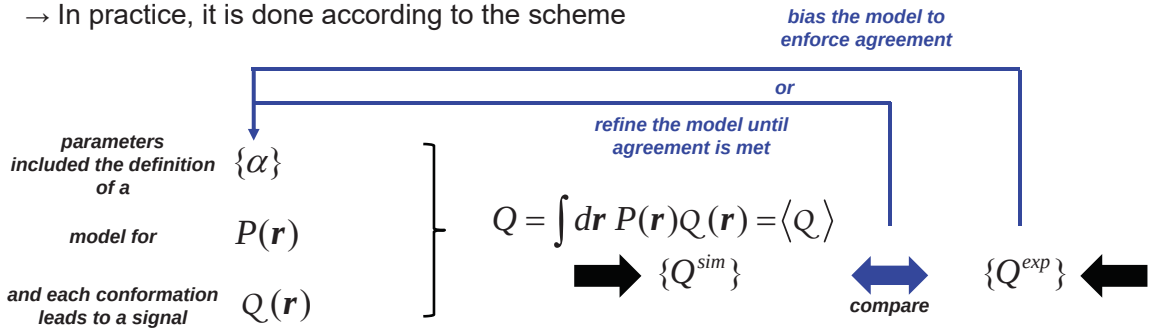
$$Q = \int d\boldsymbol{r}\, P(\boldsymbol{r}) Q(\boldsymbol{r}) = \langle Q \rangle \Rightarrow \{Q^{exp}\}$$

- Unfortunately, this alone **does not help** *even if we know* $Q(\boldsymbol{r})$
  - → It is easy to go in the **forward** direction $P(\boldsymbol{r})$ $Q(\boldsymbol{r})$ ➡ $\{Q^{exp}\}$
  - → But impossible to go in the **backward** direction $P(\boldsymbol{r})$ ⬅ $Q(\boldsymbol{r})$ $\{Q^{exp}\}$
- The reason is that
  - → There is **never enough data** to fully determine a $P(\boldsymbol{r})$ *A continuous function cannot be obtained from discrete data*
  - → Even if we postulate a single populated configuration $\boldsymbol{r}$ *(i.e. P(r)~δ(r))* there is also **seldom enough data** to specify it completely (*e.g. 3N-3* atomic coordinates)

A third issue is that $Q(\boldsymbol{r})$ *may be a non-invertible function (and may also be only approximate)*

Overview of structure refinement

- It follows that structure refinement always involves a **modeling component**
  - → In practice, it is done according to the scheme



- The function *P(r)* may be of **varying complexity**
  - → Single structure $\{\alpha\} \rightarrow \boldsymbol{r}$
  - → Single structure with alternative fragments *e.g. alternative sidechain positions in X-ray structures*
  - → Bundle of structures $\{\alpha\} \rightarrow \{\boldsymbol{r}\}$ *e.g. NMR bundle*
  - → Full conformational ensemble $P(\boldsymbol{r})$ *e.g. from MD trajectory*
- In **conventional** structure refinement
  - → One will rather refine a model iteratively until we have $\{Q^{sim}\} \approx \{Q^{exp}\}$

often: single structure or bundle normally in a "least-squares" sense → minimize a residual

- In **simulation-based** structure refinement
  - → One will rather bias a simulation using the values $\{Q^{exp}\}$ so that the generated trajectory directly satisfies $\{Q^{sim}\} \approx \{Q^{exp}\}$

Conventional structure refinement

- In **conventional** structure refinement one typically
  - → Refines *one model structure* (or a small number)
  - → *Does not care* too much about *probability distributions*
  - → Commonly uses *simple approximations* to *eliminate degrees of freedom*
  - → Usually refines the structure(s) iteratively (*minimization* of a residual)
- Example: conventional X-ray or NMR structure refinement ***e.g. DG / XPLOR***
- Typically

→ **Single structure** (Dirac delta function for the probability) ***"conventional" single structure refinement***

fewer model parameters to refine ↓

- Entirely free atom coordinates
- Standard covalent geometries + closest packing (excluded volume)
- Potential energy minimum of a simplistic force field

→ Problem 1: observable to parameter ratio is often << 1

→ Problem 2: observable information is not homogeneously distributed over the structure (regions over- versus underdetermined)

→ Problem 3: observables may be mutually incompatible in single structures (because needs an ensemble)

→ **Bundle of structures** ***"conventional" multiple structure refinement***

- Set of alternative structures obtained as above

→ Problem: is this really a proper Boltzmann-weighted ensemble? (i.e. what is the weight of each structure in the bundle? is the bundle "complete" or just a "sampler"?)

→ **Gaussian distribution** around reference structure
e.g. use of isotropic or anisotropic B-factors

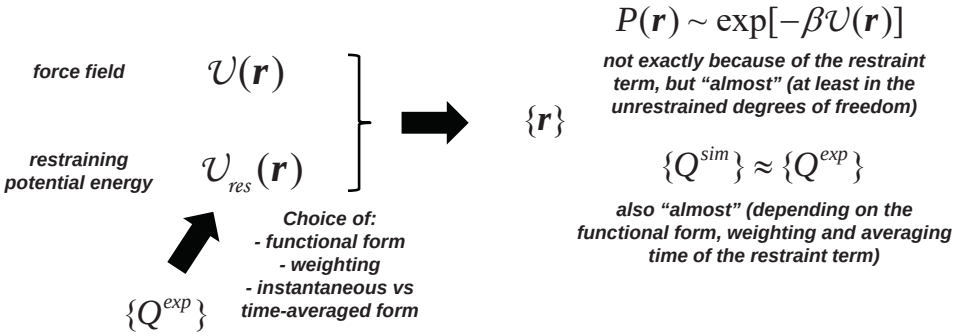
→ Not much better (biomolecules are not harmonic oscillators !)

Simulation-based structure refinement

- In **simulation-based** structure refinement one typically
  - → Refines an *ensemble* of structure in the form of an MD (+thermostat) trajectory
  - → Tries to enforce a (nearly) *Boltzmann distribution* of the structures
  - → Commonly uses a physics-based *force field*
  - → Usually refines the ensemble by biasing (inclusion of *restraints* in the dynamics)

nearly: Boltzmann up to the biasing restraints!

- Example: MD-based X-ray or NMR ensemble refinement ***e.g. GROMOS***
- Typically



→ The force field **complements** the missing information from the experiment in a (hopefully) physically meaningful way

Key issues in structure refinement

INSTANTANEOUS OBSERVABLE

$Q(\boldsymbol{r})$

Common observables
Approximations
Invertibility
Calculation cost

ENSEMBLE AVERAGING

$\langle Q \rangle$

Single structure vs ensemble
Convergence

FORCE-FIELD REPRESENTATION

$\mathcal{U}(\boldsymbol{r})$

Accuracy

EXPERIMENTAL DATA

Q^{exp}

Primary vs secondary
Quality and accuracy
Redundancy vs completeness

BIASING FUNCTION REPRESENTATION

$\mathcal{U}_{res}(\boldsymbol{r})$

Weighting
Instantaneous vs time-averaged

Illustrative example

Common errors, inefficiencies and misinterpretations

Observables that can be used in refinement

- Example of **instantaneous observables** $Q(\boldsymbol{r})$

– **NOE intensities** — *Interpretation: so-called model-free approach* — $\langle I_{ij} \rangle$
distances — $\langle r_{ij} \rangle$

– **J-coupling constants** — *Interpretation: Karplus equation* — $\langle {}^3J_{ij} \rangle$
dihedral angles — $\langle \varphi_{ixxj} \rangle$

– **Residual dipolar couplings** — $\langle D_{ij} \rangle$

– **Chemical shifts** — $\langle \sigma_i \rangle$

– **Structure factors (ampl.)** — $\langle F_{hkl} \rangle$
electron density — *Interpretation: phase* — $\langle \rho(\boldsymbol{r}) \rangle$

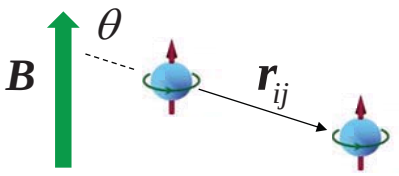
– **CD spectra** — $\langle I(\lambda) \rangle$

MOST COMMON → DISCUSSED IN MORE DETAILS IN THE NEXT SLIDES

+ NMR relaxation times, NMR order parameters, SAXS/WAXS structure factors, IR/Raman intensities, FRET efficiencies ...

NOE intensities

- The **Nuclear Overhauser Enhancement** (NOE) effect
  - → Transfer of nuclear spin polarization from one spin to another in space represents a cross-relaxation mechanism *via* dipole-dipole interactions
  - → The effect can be measured, *e.g.* as cross-peak intensities in a NOESY spectrum
  - → The NOE intensities are directly related to time-correlation functions



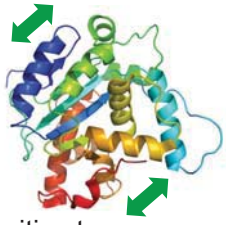
$$Q \sim C_t \left(\frac{3\cos^2\theta(t) - 1}{r_{ij}^3(t)} \right)$$

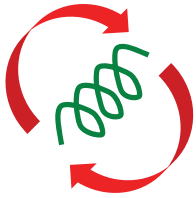
The cross-relaxation efficiency depends on how the connecting vector (length and orientation relative to magnetic field) varies in time

- If one assumes that the molecules in the sample are **randomly oriented**, the intensities map to (inverse-n^{th}-powed) **average distances**, depending on the **tumbling regime**

Slow tumbling limit (e.g. big protein)

$$I \sim \left\langle r_{ij}^{-3} \right\rangle$$





Fast tumbling limit (e.g. small peptide)

$$I \sim \left\langle r_{ij}^{-6} \right\rangle$$

  - → Going from intensities to average distances requires a tumbling model, the so-called "model-free" approach

this silly name is commonly twisted into «model-free model»

  - → Intermediate situations may be complicated!

you can also calculate intensities directly from MD via the time-correlation functions (complicated but has been done)

NOE intensities

- In general, the experimental data is formulated in terms of **distances between protons**

e,g. fast tumbling

$$I \sim \left\langle r_{ij}^{-6} \right\rangle \Rightarrow r_{ij}^{exp} \sim I^{1/6} \Rightarrow \left\{ r_{ij}^{exp} \right\}$$

experimentally inferred inter-proton distances (sometimes: just distance classes like short/medium/long)

- Things to keep in mind
  - → There is already a **tumbling model** involved in this "experimental" data
  - → These values are truly **averages**, *i.e.* they may not be fulfilled simultaneously in a single structure
  - → The averaging overweighs **short distances** (because negative exponent), *i.e.* intermittent contacts between otherwise remote pairs give a shorter distance than the plain average
  - → The **absence of a cross-peak** does not automatically imply a large distance
  - → The distances may also sometimes be **slightly underestimated**

other cross-relaxation mechanisms, e.g. via an intermediate nucleus

- In **conventional** refinement
  - → Use the distance information along with a primitive force-field to refine a single structure (or a bundle thereof) *via* distance geometry
- In **simulation-based** refinement
  - → **Pseudo-atom** sites / corrections – when equivalent or stereospecific protons not resolved
  - → **Virtual-atom** sites / corrections – when using united atoms
  - → Monitor disagreement by **violations**

$$\left\langle r_{ij}^{sim} \right\rangle - r_{ij}^{exp}$$

only positive deviations are considered a violation

only large deviations (>0.1nm) are considered important

Virtual and pseudo atoms

- **Virtual atoms** correspond to H atoms that are not in the force-field united-atom representation

  → distance is calculated based on virtual atom

  → no distance correction

- **Pseudo atoms** correspond to multiple H atoms which could not be distinguished in the experiment (common signal)

  → distance is calculated on pseudo site

  → distance bound must be increased appropriately (pseudo atom correction)

group	configuration	atom type	correction term Δr^{ps} on distance restraint r^0 (nm)	geometric code ICDR1 or ICDR2 *)
CH1 (aliphatic)		virtual	.00	1
CH1 (aromatic)		virtual	.00	2
CH2 (stereospecific)		virtual	.00	4
CH2 (non-stereospecific)		pseudo	.09	3
CH3		pseudo	.10	5
CH3 (non-stereospecific, Val, Leu)		pseudo	.22	6
CH3 (non-stereospecific, t-butyl)		pseudo	.23	7
COG (center of geometry)		pseudo	.00	-1
COM (center of mass)		pseudo	.00	-2

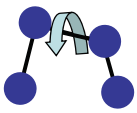
TABLE 8.1. Virtual and pseudo hydrogen atoms and distance restraint correction terms.

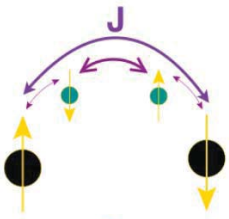
J-coupling constants

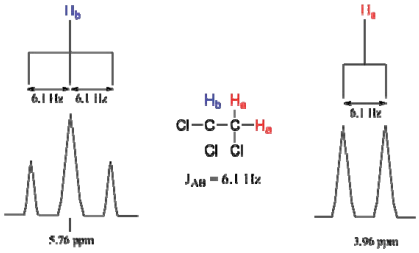
- The spin-spin **scalar coupling constants** (J-value)

  → Coupling between two nuclear spins *via* the electrons in the bonds

  → Most relevant are vicinal proton-proton coupling constants ($^3J^{HH}$)







- The vicinal J-values strongly depend on the **dihedral angle** between the coupled protons

  → Connecting equations are called a **Karplus equations**, and often take the form

$$^3J_{ij} \approx a\cos^2\varphi + b\cos\varphi + c$$

  → The parametrization is **empirical** (compare X-ray structures to J-values for many compounds, possibly complement with QM)

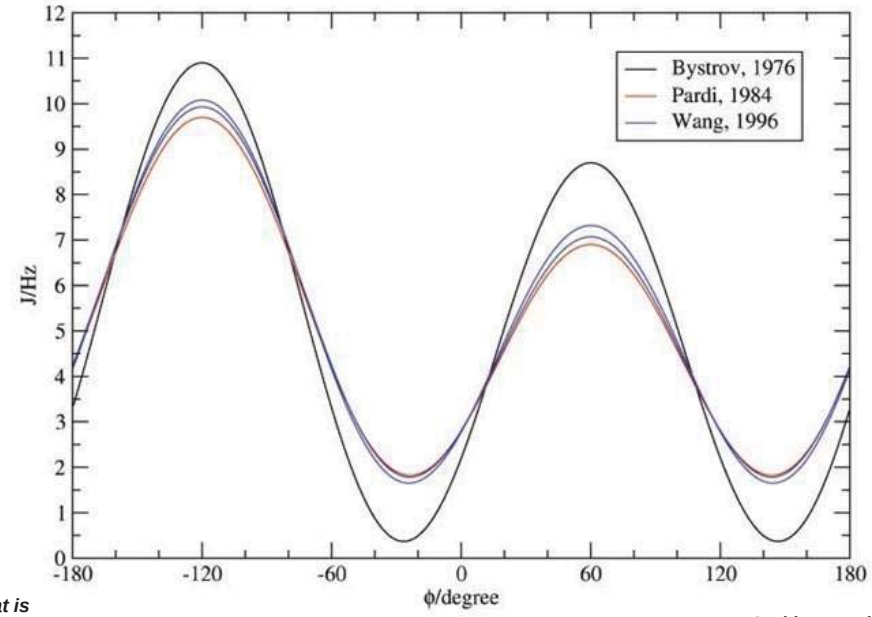
  → As a result, there are **many Karplus equations** around, usually designed for *specific classes of dihedral angles* in *specific compounds* based on a *specific training set* of molecules

J-coupling constants

Karplus curve for peptide ϕ angle (three parametrizations; based on fit of NMR J-value vs X-ray structures or/and QM calculations)

Karplus relation for $^3J(H_N\text{-}H_\alpha)$
as function of the backbone ϕ torsional angle

$$^3J(\phi) = a\cos^2(\phi+\delta) + b\cos(\phi+\delta) + c; \quad \delta = -60°$$



No clue what is the 4th curve... (two blue curves?)

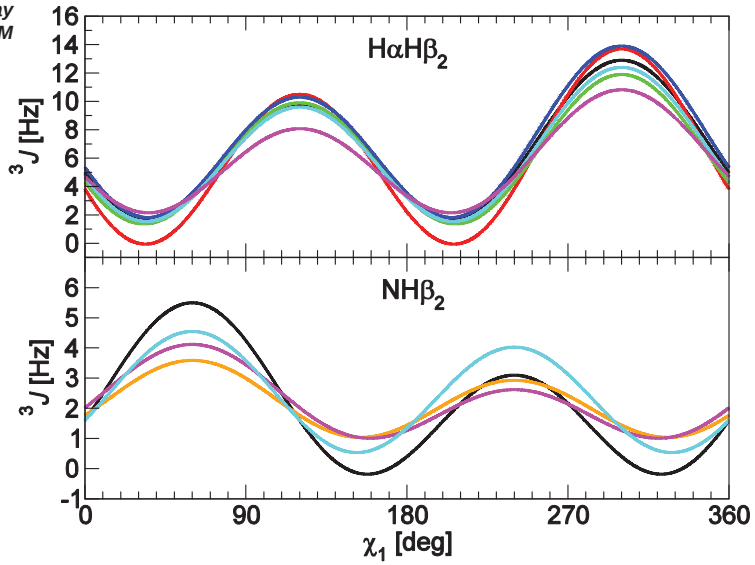
Accuracy of $^3J(\phi)$ is about 1 Hz

And it sometimes occurs that a measured J is up to 1 Hz below the min or above the max!

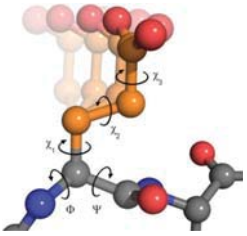
J-coupling constants

Karplus curve for peptide sidechain χ_1 angle (seven parametrizations; based on fit of NMR J-value vs X-ray structures or/and QM calculations)

Karplus relations $^3J(H_\alpha\text{-}H_\beta)$ and $^3J(N\text{-}H_\beta)$
as function of the side-chain χ_1 torsional angle



Abraham, 1962
Deber, 1971
DeMarco, 1978
Cung, 1978
Fischman, 1980
Perez(NMR), 2001
Perez(Xray), 2001



$$^3J(\chi_1) = a\cos^2(\chi_1 + \delta) + b\cos(\chi_1 + \delta) + c;$$

$\delta = -120°$ ($H_\alpha H_{\beta 2}$) or $\delta = +120°$ ($NH_{\beta 2}$)

Accuracy of $^3J(\chi_1)$ is definitely not better than 1 Hz

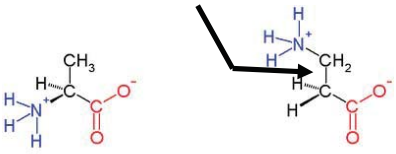
J-coupling constants

Accuracy of some parametrisations of the Karplus relation

Karplus relation : $^3J_{HH} = A \cos^2\theta + B \cos\theta + C$

A, B and C from
(a) Wüthrich (1978, 1984)
(b) Case (1994)
(c) Bax (1996)
(d) Rüterjans (1999)

Karplus curve for β-peptide ω angle (seven parametrizations; based on fit of NMR J-value vs X-ray structures or/and QM calculations)



L-α-alanine β-alanine

Table 3. Comparison of simulated and experimental ^3J coupling constants [Hz] for the Val-Phe β-dipeptides in methanol. Different stereoisomers (*R,S*) at Phe and different simulation temperatures are considered.[24]

		Simulated (298 K)				(340 K)	Exptl (298 K)
Protons	Isomer	(a)	(b)	(c)	(d)	(a)	
$^3J_{H^{\alpha}H^{\beta}}$ (Val)	*SR*	5.4	5.4	5.5	4.9	5.7	6.1
	SS	6.5	6.9	6.7	6.1	6.1	7.2
$^3J_{H^{\alpha}H^{\beta}}$ (Val)	*SR*	6.1	6.4	6.3	5.8	5.8	5.8
	SS	5.4	5.5	5.5	5.0	5.7	4.8
$^3J_{H^{\beta}H^{\gamma}}$ (Val)	*SR*	8.2				8.8	9.2
	SS	11.0				9.8	9.9
$^3J_{H^{\beta}H^{\gamma}}$ (Val)	*SR*	4.5				4.2	4.5
	SS	3.8				4.3	4.2
$^3J_{H^{\alpha}H^{\beta}}$ (Val)	*SR*	7.2				6.8	8.5
	SS	6.2				6.1	8.4
$^3J_{H^{\alpha}H^{\beta}}$ (Phe)	*SR*	7.0	7.4	7.2	6.6	6.1	8.4
	SS	7.1	7.5	7.3	6.7	6.2	8.4
$^3J_{H^{\beta}H^{\alpha}}$ (Phe)	*SR*	4.0				4.1	5.8
	SS	3.5				4.2	5.3
$^3J_{H^{\beta}H^{\alpha}}$ (Phe)	*SR*	10.8				10.6	8.2
	SS	11.3				10.3	8.3
$^3J_{H^{\beta}H^{\alpha}}$ (Phe)	*SS*	3.9				4.2	5.0
$^3J_{H^{\beta}H^{\alpha}}$ (Phe)	*SS*	11.0				10.6	9.2

Variation 0.5 – 0.8 Hz

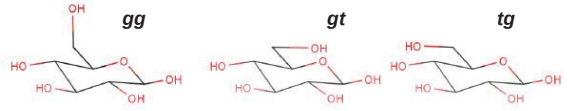
Chem. Eur. J., 9 (2003), 5838-5849

Simulations predict which of the isomers has the largest $^3J_{HH}$ value
Differences between SR and SS are small

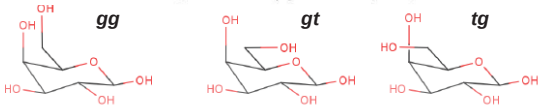
J-coupling constants

Lonardi, Oborský & Hünenberger Helv. Chim. Acta 100 (2017) e1600158

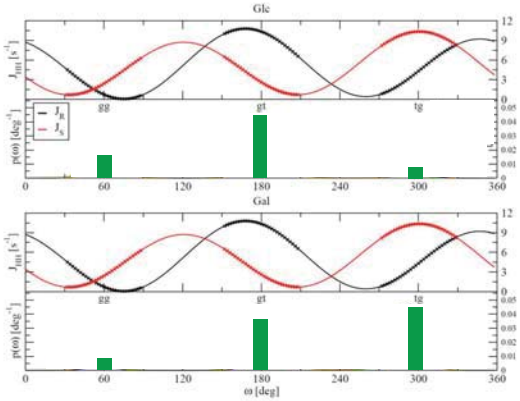
- In **conventional** refinement
  → Pick a Karplus equation
  → Postulate a (simplistic!) φ-population model
  → Fit the populations to the J-values



β-D-Glucopyranose (Glc)



β-D-Galactopyranose (Gal)

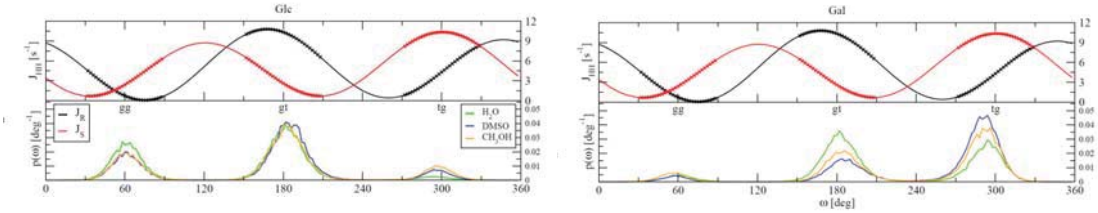


2 J-values (R and S), 3 populations, one constraint (sum of populations)

Population estimates now depend on the choice of a Karplus equation AND of a simple population model!

- In **simulation-based** refinement
  → Compare directly to J-values

The peak widths and locations also play a role, mixed with the integrated populations

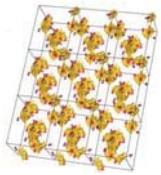


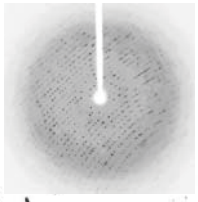
Structure factors

- In X-ray crystallography

Electron density in unit cell

$\rho(\boldsymbol{r})$





Reflection intensities

$I(\boldsymbol{h})$

- Alternative formulation of structure factors

$$F(\vec{h}) = \iiint_{\text{unit cell}} d^3\vec{s}\ \rho(\vec{s})\ e^{i 2\pi \vec{h}\cdot\vec{s}}$$

$\rho(\vec{s})$: electron density

RECIPROCAL SPACE: structure factors

- Fourier series

$$\rho(\vec{s}) = \sum_{\vec{h}} F(\vec{h})\ e^{-i 2\pi \vec{h}\cdot\vec{s}}$$

phase: special tricks, e.g. isomorphous replacement

REAL SPACE: electron density

- Observable: intensity

$$I(\vec{h}) = |F(\vec{h})|^2$$

AND phase information is not available directly!

in practice, corrections required (Lorentz, polarization, absorption corrections)

- Issues in **conventional** refinement
  - → Reflection intensities to structure factors (phase?)
  - → Electron density map to atomic coordinates (density of data [resolution]? static and dynamic averaging? site occupancies?)

Key issues in structure refinement

INSTANTANEOUS OBSERVABLE

$Q(\boldsymbol{r})$

Common observables
Approximations
Invertibility
Calculation cost

ENSEMBLE AVERAGING

$\langle Q \rangle$

Single structure vs ensemble
Convergence

FORCE-FIELD REPRESENTATION

$\mathcal{U}(\boldsymbol{r})$

Accuracy

EXPERIMENTAL DATA

Q^{exp}

Primary vs secondary
Quality and accuracy
Redundancy vs completeness

BIASING FUNCTION REPRESENTATION

$\mathcal{U}_{res}(\boldsymbol{r})$

Weighting
Instantaneous vs time-averaged

Illustrative example

Common errors, inefficiencies and misinterpretations

Effects of ensemble (motional) averaging

- Effect of **averaging**

$$\left\langle Q(\vec{r})\right\rangle_{\vec{r}} \equiv \int Q(\vec{r})\, \mathrm{P}(\vec{r})\, d\vec{r} \equiv \frac{\int Q(r) e^{-V^{phys}(\vec{r})/k_B T}\, d\vec{r}}{\int e^{-V^{phys}(\vec{r})/k_B T}\, d\vec{r}}$$

$$\left\langle Q(\vec{r})\right\rangle_{\vec{r}} \neq Q(\vec{r})$$

averaged Q — single structure Q

→ in general, the ensemble cannot be summarized by a single structure

$$\neq Q\left(\left\langle \vec{r}\right\rangle_{\vec{r}}\right)$$

mean structure Q

→ the signal of the average structure is not the average signal over all structures (average structure: essentially meaningless!)

- Consequences

→ If a **single structure** reproduces the experimental data, this does not mean that the conformational ensemble **solely** consists of **this structure** (nearly always, it is not the case)

→ The observable/averaging may be **dominated** by specific structures in the ensemble, which do not need to have the **highest populations**, *i.e.* it may be that a small fractions of the conformations actually determines the experimental signal

→ **Different experimental signals** may be determined by **different conformations**, no single conformation being able to reproduce the entire signal

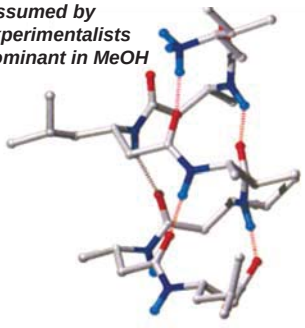
→ **Experimental data** is never sufficient to determine an entire conformational ensemble

Effects of ensemble (motional) averaging

The average structure $\langle \mathbf{r} \rangle_r$ is highly strained 6-ß-peptide in methanol: 34 NOE's

Angew. Chemie ***38*** *(1999) 236-240*

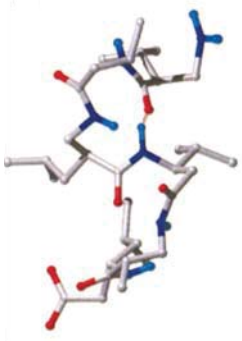
Classical XPLOR refinement: Assumed by the experimentalists to be dominant in MeOH



left-handed 3_{14}-helix
H-bonds: NH(i) – O(i+2)

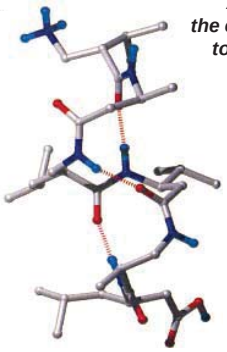
1.3 % of the conformations

3 NOE's can only be fulfilled by these types conformations



average structure in MeOH : distorted right-handed helix – only one H-bond: NH(4) – O(1)

Average structure



Assumed by the experimentalists to be dominant in pyridine

right-handed helix
H-bonds: NH(i) – O(i+1, i-3)

35 % of the conformations

MD in MeOH

→ The MD simulation in MeOH satisfies all NOE bounds in terms of $\langle Q(\mathbf{r}) \rangle$

→ Neither the 3_{14}-helix, nor the right-handed helix, nor the average structure can satisfy all the NOE bounds simultaneously

Conclusion: - average structure may be meaningless
- use ***primary*** exp. data (NOE's), ***not secondary*** experimental data (structures) ***to compare with***

Here, the structure claimed by the experimentalists in MeOH was incorrect or, at least, not representative (the MD ensemble is different and equally acceptable)

Key issues in structure refinement

INSTANTANEOUS OBSERVABLE

$Q(\boldsymbol{r})$

Common observables
Approximations
Invertibility
Calculation cost

ENSEMBLE AVERAGING

$\langle Q \rangle$

Single structure vs ensemble
Convergence

FORCE-FIELD REPRESENTATION

$\mathcal{U}(\boldsymbol{r})$

Accuracy

EXPERIMENTAL DATA

Q^{exp}

Primary vs secondary
Quality and accuracy
Redundancy vs completeness

BIASING FUNCTION REPRESENTATION

$\mathcal{U}_{res}(\boldsymbol{r})$

Weighting
Instantaneous vs time-averaged

Illustrative example

Common errors, inefficiencies and misinterpretations

Force-field representation

- Choices in the **force-field definition**

  → degrees of freedom (elementary particles)

  electrons
  atoms
  united atoms
  implicit-solvent
  residue beads

  → *chemical shifts*
  → *virtual H's*
  → *crude!*

  ➡ ***should be sufficient to represent/calculate appropriately*** $Q(\boldsymbol{r})$

  → interaction function (force field)

  «common» choice: (united-)atom, no solvent

  e.g. GROMOS ➡ $\mathcal{U}(\boldsymbol{r})$ ***quality of the functional form? quality of the parametrization?***

  → configuration generation (search, sample, simulate)

  database scan (e.g. PDB)
  distance geometry
  + simplified force field
  explicit-solvent
  MD+thermostat
  ...

  ➡ ***ideally generate a Boltzmann ensemble*** $P(\boldsymbol{r}) \sim \exp[-\beta \mathcal{U}(\boldsymbol{r})]$

  «common» choices: do we really have Boltzmann statistics?

- Choices in the **observable calculation**

  → definition of the relevant instantaneous observable ➡ $Q(\boldsymbol{r})$ ***Nearly always involves assumptions and approximations!!!***

- Then, we just have to **average** and **compare** with experiment

  the experimental signal is obtained by ensemble (simulation) averaging

  $Q = \int d\boldsymbol{r}\, P(\boldsymbol{r}) Q(\boldsymbol{r}) = \langle Q \rangle$

  compare with

  $\{Q^{sim}\} \Leftrightarrow \{Q^{exp}\}$

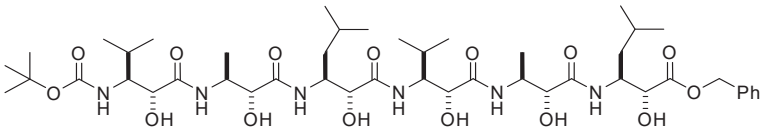
Force-field representation

- The **"dream" situation**

$\mathcal{V}(\boldsymbol{r})$ and $\mathcal{Q}(\boldsymbol{r})$ ***correct*** + ***infinite sampling*** } ***problem solved !***

→ Only happens in exceptional cases (then it may also be coincidental: experimental data may actually provide little relevant information and be compatible with many models, even the crudest)

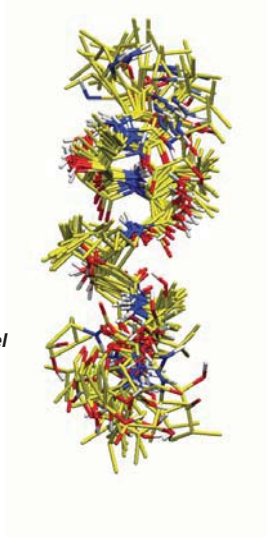
- Example: **β-hexapeptide in methanol**



→ **experiment** delivers 41 NOE's & 12 ^{3}J-couplings

→ **NMR bundle** (conventional XPLOR refinement) suggests 2_8-helix

Bundle of 20 NMR model structures (protection groups not shown)



Gademann et al., Angew. Chem. Int. Ed. Engl. 42 (2003) 1534

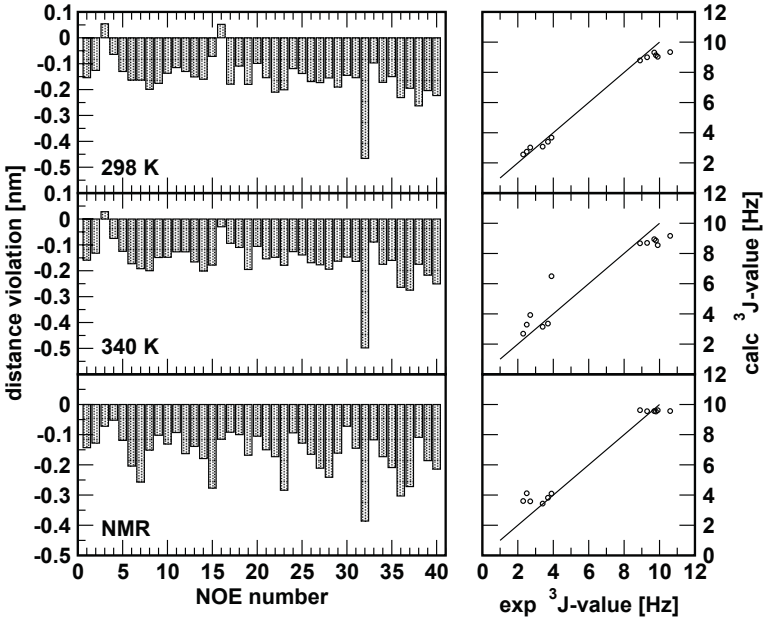
→ 100 ns **unbiased** MD simulations in explicit methanol with GROMOS 45A3 at 298 or 340 K (started from an extended fonformation)

at 298K:
2 violations of about 0.05 nm
J-value rms deviation of 0.44 Hz

Glaettli & van Gunsteren Angew. Chem. Int. Ed. Engl. 43 (2004) 6312

Force-field representation

- NOE distance violations & backbone ^{3}J-values



- at 298 K
  2 violations (~0.05 nm)

  average deviation from exp. J-values: 0.44 Hz

- at 340 K
  1 violation (~ 0.03 nm)

  average deviation from exp. J-values: 0.91 Hz

- NMR bundle
  no violation

  average deviation from exp. J-values: 0.57 Hz

- GROMOS force field reproduces experimental data without any restraining

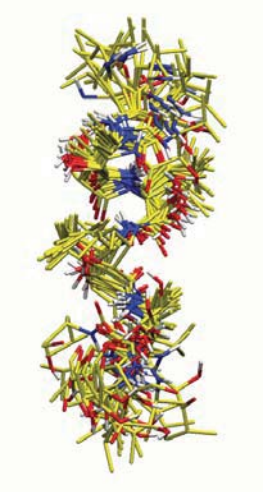
This is a case where the force-field information is sufficient to generate an ensemble compatible with the experimental data (nice – but maybe also a bit lucky ?)

Force-field representation

- Example: **β-hexapeptide in methanol**

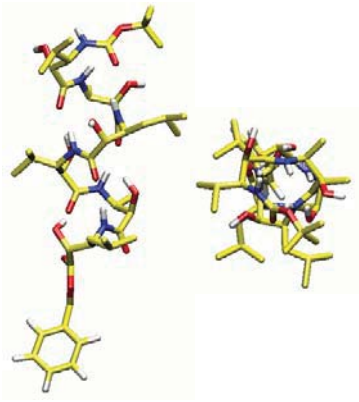
  → Experiment delivers 41 NOE's & 12 ^{3}J-couplings

  → Two conformational ensembles reproduce about equally well the experimental data!



→ **NMR bundle** (conventional refinement) suggests 2_8-helix

Gademann et al., Angew. Chem. Int. Ed. Engl. 42 (2003) 1534



→ **unbiased MD** (GROMOS 45A3) suggests dominant 2.5_{12}-helix (34% population)

Glaettli & van Gunsteren, Angew. Chem. Int. Ed. Engl. 43 (2004) 6312

→ ... more on this case later ...

Key issues in structure refinement

INSTANTANEOUS OBSERVABLE

$Q(\boldsymbol{r})$

Common observables
Approximations
Invertibility
Calculation cost

ENSEMBLE AVERAGING

$\langle Q \rangle$

Single structure vs ensemble
Convergence

FORCE-FIELD REPRESENTATION

$\mathcal{V}(\boldsymbol{r})$

Accuracy

EXPERIMENTAL DATA

Q^{exp}

Primary vs secondary
Quality and accuracy
Redundancy vs completeness

BIASING FUNCTION REPRESENTATION

$\mathcal{V}_{res}(\boldsymbol{r})$

Weighting
Instantaneous vs time-averaged

Illustrative example

Common errors, inefficiencies and misinterpretations

Experimental data

Which type of experimental data or quantities Q

$\langle Q \rangle_{sim}$ ↔ $\langle Q \rangle_{exp}$

W.F. van Gunsteren, J. Dolenc, & A.E. Mark, Curr. Opin. Struct. Biol., 18 (2008) 149-153

Distinguish between:

1. primary experimental data Q^0: observable quantities Q that are ***directly measured***

   Examples: peak location and intensity from X-ray diffraction or NMR spectroscopic measurements, ^3J–values

2. secondary (***derived***) experimental data Q^d:
   quantities Q for which (non-observed) values are ***derived from*** (observed) values of primary experimental data Q^0 by applying a given procedure f: $Q^d = f(Q^0)$
   which involves assumptions and approximations

   Examples: textbook structures, refined molecular structures, torsional angle values, NOE-derived proton-proton distances, NMR order parameters

You don't want to compare simulations against another (often cruder) model, but really against experiment !

Comparison of	may reflect the quality of:
a. $\langle Q^0 \rangle_{sim}$ with $\langle Q^0 \rangle_{exp}$	→ the molecular model
b. $\langle Q^d \rangle_{sim}$ with $\langle Q^d \rangle_{exp} = f(\langle Q^0 \rangle_{exp})$	→ the model and the procedure f

In reality $\langle Q^d \rangle_{exp}$ may carry little experimental information

Experimental data

Quality of the experimental data Q^{exp}

Test of force field and NMR data
for Hen Egg White Lysozyme

Experimental data
(Smith et. al., 1991, 1993; Buck et. al., 1995; Schwalbe et. al., 2001)

1158 NOE's **derived inter-proton distances (set1 1993)**

1525 NOE's **derived inter-proton distances (set2 2001)**

95 $^3J_{HN\alpha}$-coupling constants

100 $^3J_{\alpha\beta}$-coupling constants

124 backbone and 28 side-chain order parameters

X-ray coordinates (PDB 1aki, 1.5 A)

NMR coordinates (PDB 1e8l, ensemble of 50 structures)

Experimental data

NOE distance bound violations in HEWL

NOE bound violations computed from MD trajectories (43A1(1996)/45A3(2001)) **against** ***two sets of experimental NOE distance bounds*** from Smith *et. al.* (set1, 1993) and from Schwalbe *et. al.* (set2, 2001)

Averaging period (ns)	Number of violations (set1) out of 1158 NOE's			Mean violation $\langle R_E - R_O \rangle$	
	>0.1 nm	*>0.2 nm*	*> 0.3 nm*		
0.0-0.5	25/44	9/15	2/6	0.017/0.024	
0.5-1.5	31/44	11/15	3/3	0.020/0.024	**1993 set**
1.5-3.5	41/56	11/27	5/17	0.023/0.034	
0.0-3.5	23/43	9/17	3/6	0.019/0.026	
	Number of violations (set2) out of 1525 NOE's (30% more)				
	>0.1 nm	*>0.2 nm*	*> 0.3 nm*		
0.0-0.5	21/43	4/9	0/0	0.015/0.021	
0.5-1.5	22/47	2/14	0/2	0.017/0.021	**2001 set**
1.5-3.5	27/60	6/12	0/6	0.017/0.026	
0.0-3.5	20/40	2/7	0/1	0.014/0.020	

Over time (1993 →2001) the experimental data converged towards simulated ones

But: this may be misleading if the extra NOEs are short-ranged !

But: the force-field did not seem to improve ??? ;-)

Key issues in structure refinement

INSTANTANEOUS OBSERVABLE

$Q(\boldsymbol{r})$

Common observables
Approximations
Invertibility
Calculation cost

ENSEMBLE AVERAGING

$\langle Q \rangle$

Single structure vs ensemble
Convergence

FORCE-FIELD REPRESENTATION

$\mathcal{U}(\boldsymbol{r})$

Accuracy

EXPERIMENTAL DATA

Q^{exp}

Primary vs secondary
Quality and accuracy
Redundancy vs completeness

BIASING FUNCTION REPRESENTATION

$\mathcal{U}_{res}(\boldsymbol{r})$

Weighting
Instantaneous vs time-averaged

Illustrative example

Common errors, inefficiencies and misinterpretations

Accounting for motional averaging

Choice of biasing function $V^{Q,restr}\left(\langle Q(\vec{r})\rangle ; Q^{exp}\right)$ to bias $\langle Q(\vec{r})\rangle$ towards Q^{exp}

- **Form:**
  - Half or full harmonic at short range, i.e. for $\langle Q\rangle \approx Q^{exp}$
  - Bounded gradient (slope = force) at long range, i.e. for $\langle Q\rangle \neq Q^{exp}$
  - Continuous, continuous derivative

  Example: $V^{Q,restr}(\vec{r}) = \frac{1}{2}K^{Qr}\left[\langle Q(\vec{r})\rangle - Q^{exp}\right]^2$
  + linear beyond $Q^{exp} + \Delta Q$

- **Averaging:**

  Time: $\langle Q(\vec{r})\rangle \equiv \frac{1}{\tau\left[1-e^{-t/\tau}\right]}\int_0^t e^{-\left[t-t'\right]/\tau} Q(\vec{r}(t'))dt'$

  Molecules: $\langle Q(\vec{r})\rangle \equiv \sum_{n=1}^{N_M} p_n Q(\vec{r}_n)$ $\quad p_n = \begin{cases} 1 & \rightarrow \text{non-Boltzmann} \\ \dfrac{e^{-V(\vec{r}_n)/k_BT}}{\sum_{n=1}^{N_M} e^{-V(\vec{r}_{n'})/k_BT}} & \rightarrow \text{Boltzmann} \end{cases}$

  Examples: 15 Tyr in Tendamistat: averaging <…> essential
  CI-2: <distance> is o.k.
  $(NPNA)_3$: $<^3$J-value> is a major problem

- **Parameters:** calculate RMS-fluctuations and deviations

Accounting for motional averaging

Atom-atom distance restraining and multiple conformations

Half harmonic (attractive only):
A too long distance is considered to be a disagreement with experiment – but a too short distance is not, because other experimental effects might have attenuated the NOE signal (e.g. three-center spin diffusion)

Linearized:
we want to avoid too strong forces on large violations

Distance restraint:

d_0 involving atoms *i* and *j*

actual distance:

$$d(t) = \sqrt{\left[\vec{r}_i(t) - \vec{r}_j(t)\right]^2}$$

Instantaneous restraint

(half harmonic):

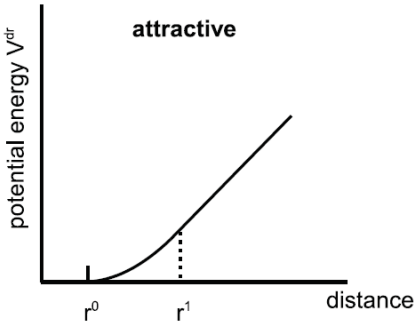
$V^{dr} = 0$ if $d(t) < d_0$

$V^{dr} = \frac{1}{2}K_{dr}\left[d(t) - d_0\right]^2$ if $d(t) > d_0$

$f_{ix} = 0$ if $d(t) < d_0$

$f_{ix} = -K_{dr}\left[d(t) - d_0\right]\frac{x_{ij}(t)}{r_{ij}(t)}$ if $d(t) > d_0$

With linearization:



Virtual and pseudo atoms

- **Virtual atoms** correspond to H atoms that are not in the force-field united-atom representation
  → distance-restraint force calculated on virtual atom, and redistributed to neighbor atoms
  → no distance correction

- **Pseudo atoms** correspond to multiple H atoms which could not be distinguished in the experiment (common signal)
  → distance-restraint force calculated on pseudo site, and redistributed to neighbor atoms
  → distance bound must be increased appropriately (pseudo atom correction)

group	configuration	atom type	correction term Δr^{ps} on distance restraint r^0(nm)	geometric code ICDR1 or ICDR2 *)
CH1 (aliphatic)		virtual	.00	1
CH1 (aromatic)		virtual	.00	2
CH2 (stereospecific)		virtual	.00	4
CH2 (non-stereospecific)		pseudo	.09	3
CH3		pseudo	.10	5
CH3 (non-stereospecific, Val, Leu)		pseudo	.22	6
CH3 (non-stereospecific, t-butyl)		pseudo	.23	7
COG (center of geometry)		pseudo	.00	-1
COM (center of mass)		pseudo	.00	-2

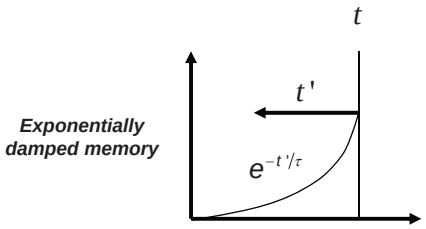
TABLE 8.1. Virtual and pseudo hydrogen atoms and distance restraint correction terms.

Accounting for motional averaging

Time average:

$$\bar{d}(t) = \left[t^{-1} \int_0^t d(t')^{-3} dt' \right]^{-1/3}$$

$$\text{discrete:} = \left[N_t^{-1} \sum_{n=1}^{N_t} d(r_{ij}(t_n))^{-3} \right]^{-1/3}$$



Exponentially damped memory

Time-average restraint:

$$V^{dr} = 0 \qquad \text{if } \bar{d}(t) < d_0$$

$$V^{dr} = \tfrac{1}{2} K_{dr} \left[\bar{d}(t) - d_0 \right]^2 \qquad \text{if } \bar{d}(t) > d_0$$

$$f_{ix} = 0 \qquad \text{if } \bar{d}(t) < d_0$$

$$f_{ix} = -K_{dr} \left[\bar{d}(t) - d_0 \right] \left[-\frac{1}{3} \bar{d}(t)^4 \right] N_t^{-1} * \left[-3 d(t)^{-4} \right] \frac{x_{ij}(t)}{r_{ij}(t)} \qquad \text{if } \bar{d}(t) > d_0$$

Force becomes smaller with growing N_t, so damp the memory

Time-average with damped memory:

Put factor $e^{-t'/\tau}$ in time average:

$$\bar{d}(t) = \left[\frac{1}{\tau \left[1 - e^{-t/\tau} \right]} \int_0^t e^{-t'/\tau} d(t - t')^{-3} dt' \right]^{-1/3}$$

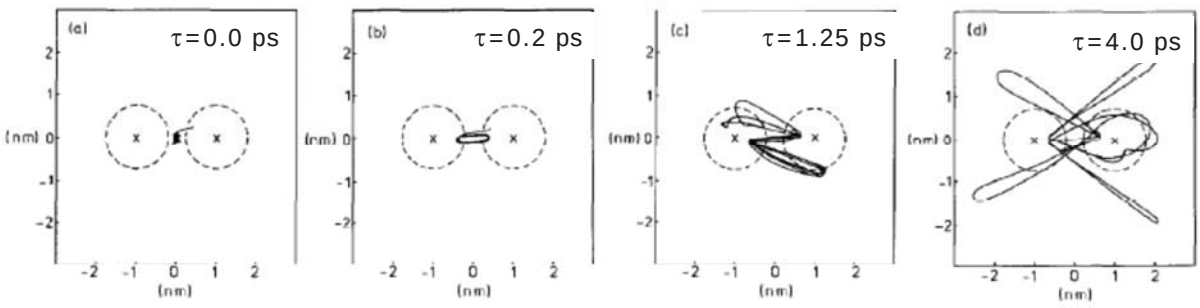
Normalization (d = cst → average = d)

Accounting for motional averaging

Non-linear averaging: a simple 2-dimensional 3-atom example

A.E. Torda et al., Chem. Phys. Lett, ***157*** *(1989) 289-294*

3 particles, only distance restraint forces, τ = memory relaxation time



3 particle system: 2 particles fixed 2 nm distance from each other
1 particle freely moving with **2 distance restraints**
of 0.8 nm length to the fixed particles

Refinement:

$$\bar{d}(t) = \left[\frac{1}{\tau\left[1 - e^{-t/\tau}\right]} \int_0^t e^{-t'/\tau}\, d(t-t')^{-3} dt' \right]^{-1/3}$$

$$V^{dr} = \tfrac{1}{2} K_{dr} \left[\bar{d}(t) - d_0\right]^2$$

Remark:
time-dependent Hamiltonian
→ no energy conservation !
→ we pump energy into the system
→ thermostatize well !

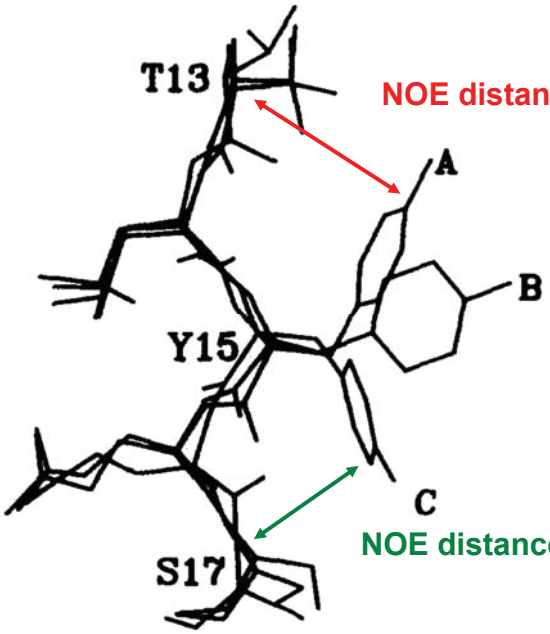
Average violations:

$$\bar{d}(t) = \left[\frac{1}{\tau\left[1 - e^{-t/\tau}\right]} \int_0^t d(t-t')^{-3} dt' \right]^{-1/3}$$

Accounting for motional averaging

Signal is a time average:
example Tendamistat

- small, 74 residue protein, 842 NOE distances from NMR
- conflicting NOE distances from the experiment
  - no single structure found that had no violations



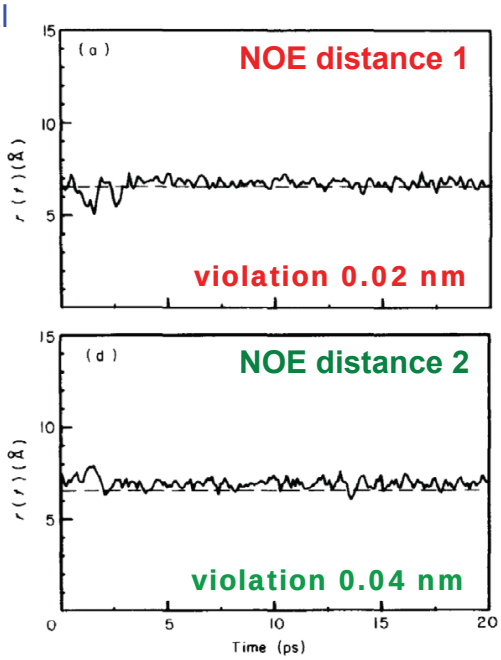
NOE distance 1: H_ε 15 Tyr - H_β 13 Thr

NOE distance 2: H_ε 15 Tyr - H_β 17 Ser

Accounting for motional averaging

Simulations using instantaneous distance restraints

- Applying ***instantaneous*** distance restraints
- puts an extra force on the atoms which pulls them to the experimental NOE distance



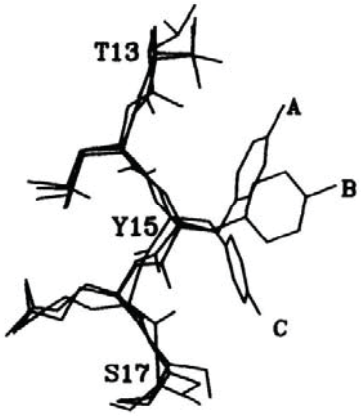
- Atoms do get close, but quite a lot of strain is present in the system

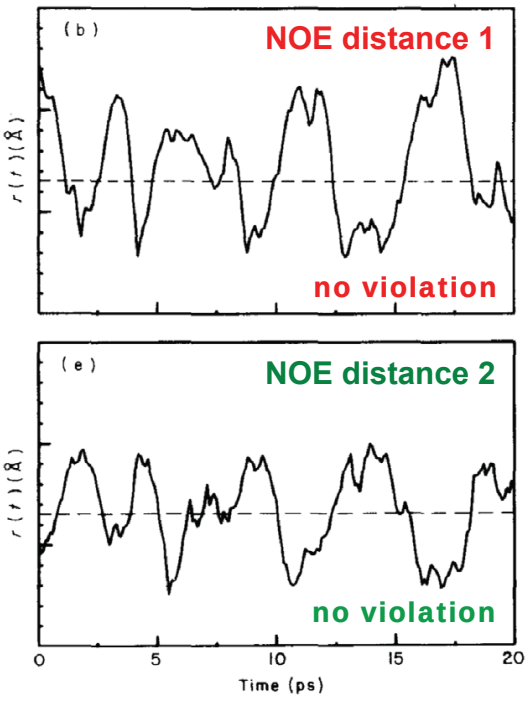
- **Small distance fluctuations**
- **Yet, *small* bound violations remain present**

Accounting for motional averaging

Simulations using time-averaged distance restraints

- ***Time-averaged*** distance restraints
- Extra forces on the atoms to enforce that the NOE distance is fulfilled on average
- Tyr13 is flipping back and forth





- **Large distance fluctuations**
- **Yet, *no* bound violations are present**

Accounting for motional averaging

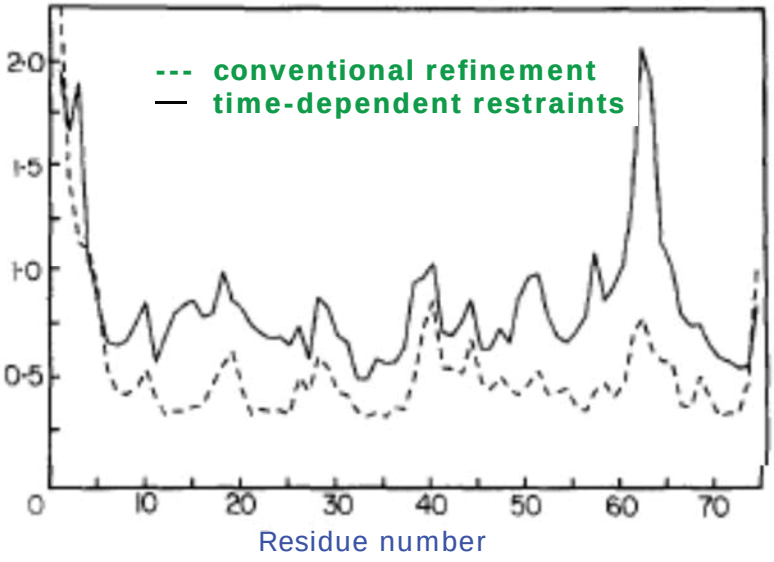
Tendamistat **20 ps MD simulation**

RMS fluctuation of C_α atoms

*J. Mol. Biol., **214** (1990) 223-235*

Small, 74 residue protein, 842 experimental NOE's

Root-mean-square atom-positional fluctuation(Å) of C_α atoms

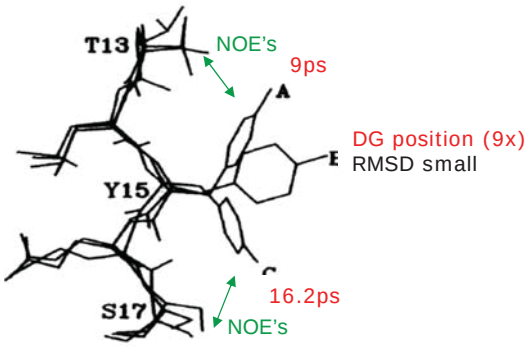


Conclusion: conventional refinement restricts atomic motion too much (instantaneous restraints)

Accounting for motional averaging

Conventional versus time-averaged refinement: Tendamistat

DG: one position (violations) in all 9 DG (distance geometry) structures
MD: many positions (no violations)
X-ray: no electron density



*J. Mol. Biol. **214** (1990) 223-235*

Conclusion:

- Convergence to **one** structure does **not** indicate that only one structure fits the experimental data!
- The experimental data are compatible with more mobility than is suggested by **static** modelling

Accounting for motional averaging

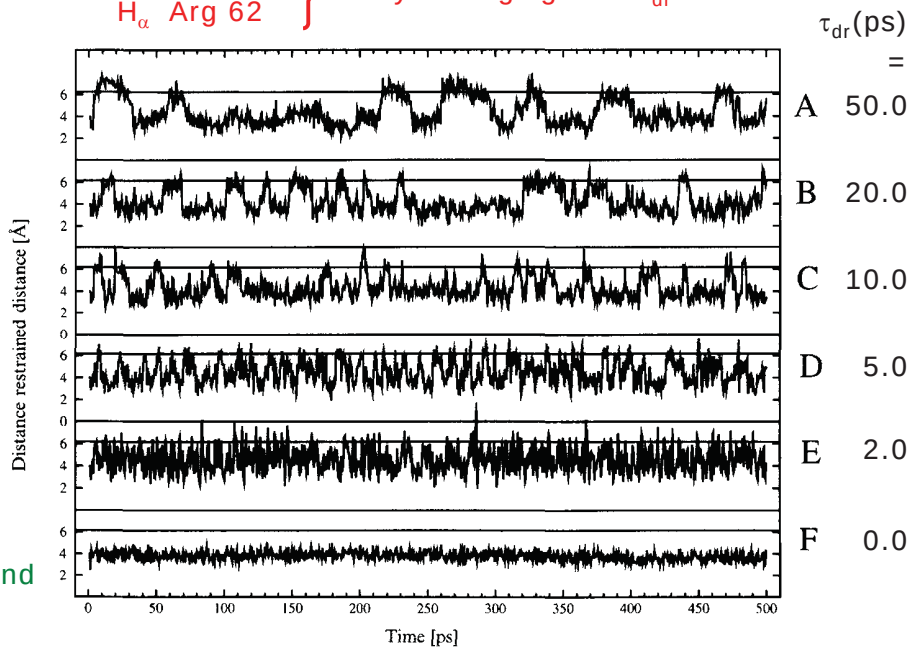
Artefacts due to time-averaged restraining: Chymotrypsin Inhibitor 2

*Alain Nanzer et al., J. Biomol. NMR, **6** (1995) 313-320*

MD in vacuo: 500ps

NOE distance restraining: H_N Trp 5, H_α Arg 62 } vary averaging time τ_{dr}

K_{dr} = 3000 kJmol^{-1}nm^{-2}



Restraining with short τ_{dr} may restrict

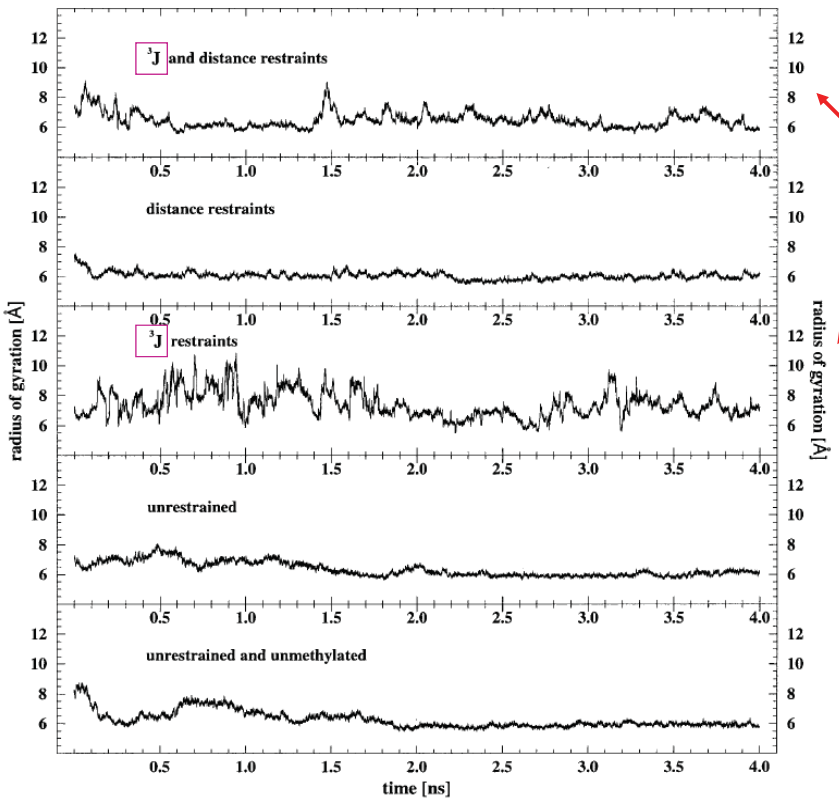
1. **fluctuations** and
2. **dynamics**

Accounting for motional averaging

NPNA-NPNA-NPNA peptide in water

*A. Nanzer et al., J. Mol. Biol., **267** (1997) 1012*

NPNA: Asn-Pro-Asn-Ala

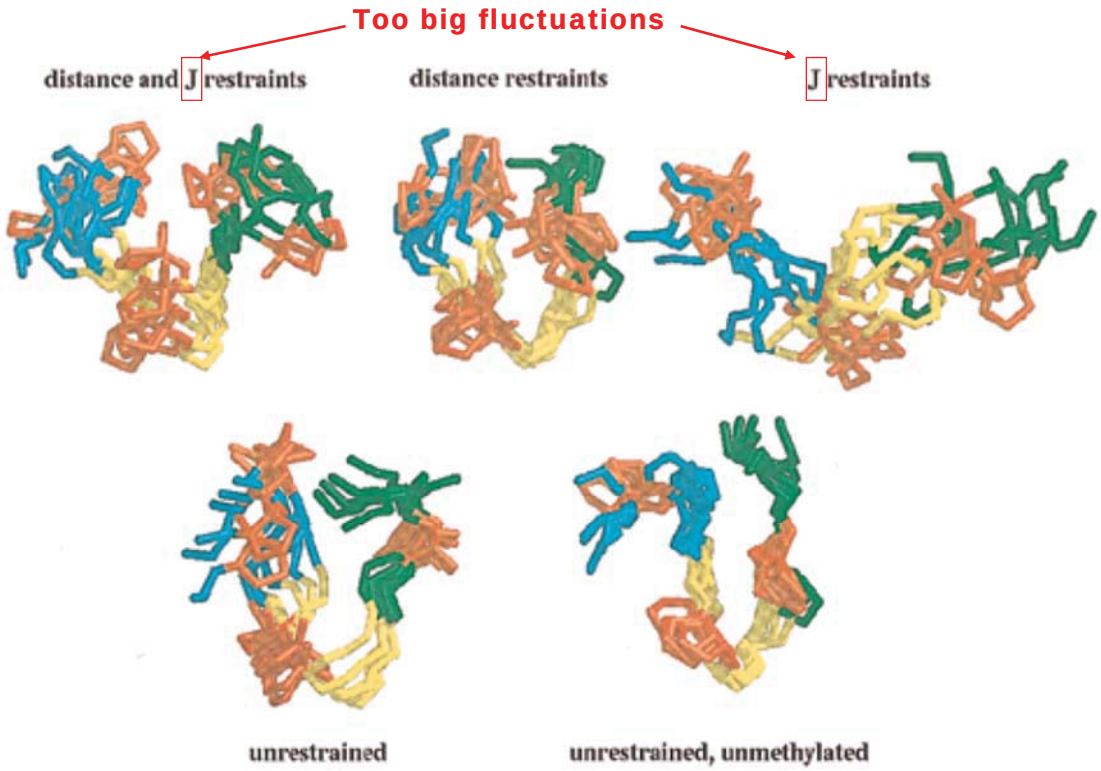


Too big fluctuations

if harmonic time-averaged ^{3}J-value restraints are applied

Accounting for motional averaging

NPNA peptide in water



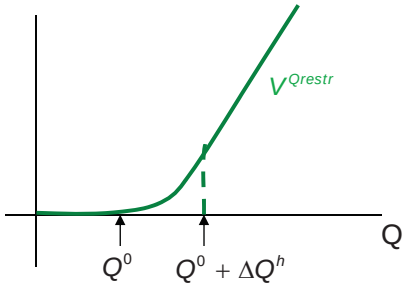
Accounting for motional averaging

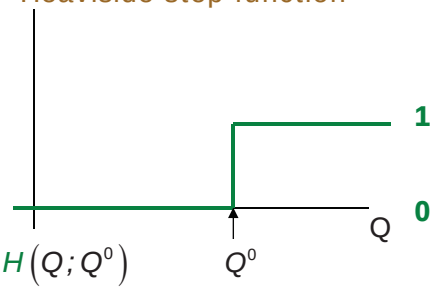
Problem of time-averaged restraining

when applied to ^{3}J-coupling constant data

$$V_n^{Qrestr}\left(\overline{Q_n};K_n^{Qr},Q_n^0\right)=\frac{1}{2}K_n^{Qr}\left[\overline{Q_n}-Q_n^0\right]^2 H\left(\overline{Q_n};Q_n^0\right)$$

Heaviside step function





Problem:

Restraining force keeps pushing $Q_n(t)$ to the left beyond Q_n^0 as long as $\overline{Q_n}(t) > Q_n^0$.

In distance restraining, this is a minor problem

- Van der Waals repulsion counteracts an attractive restraint
- r^{-3} averaging favors short distances
- generally only half-harmonic restraints (attractive) are used

In J-value restraining, this is a major problem

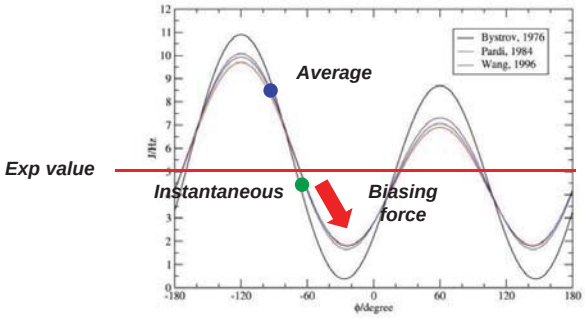
Solution ? Use both instantaneous and time-averaged ^{3}J-value for restraining (Scott et al., J. Biomol. NMR 12 (1998) 501-508)

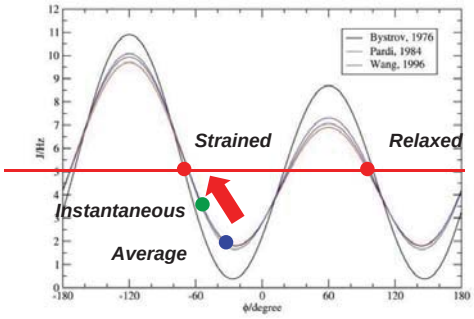
Accounting for motional averaging

- Time-averaged J-value restraining: two issues

  → Instantaneous vs average biasing

  → Multi-valued function





FIX:

For J-value restraining, consider the value of the instantaneous and of the average observable

Do not restrain to a measured value of an observable if the function connecting structure to observable is multiple-valued

Accounting for motional averaging

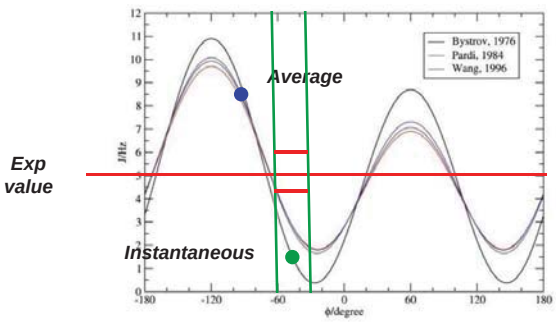
- Time-averaged J-value restraining: a possible solution...

  → ^{3}J-coupling constant biasing of the sampling by using **local-elevation** conformational sampling on torsional angles **instead of restraints**

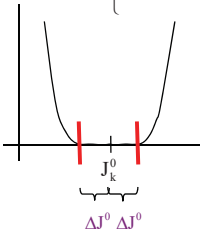
- Basic idea

  → If either the current of the average J-value is correct, do nothing

  → If both the current and the average J-value are incorrect correct, push system away from current dihedral value the local-elevation way



$$V^{Jrest}(J;J_k^0,\Delta J^0)=\begin{cases}\left(J(\phi_k(t))-J_k^0-\Delta J^0\right)^2 & \text{for } J(\phi_k(t))>J_k^0+\Delta J^0\\ \left(J(\phi_k(t))-J_k^0+\Delta J^0\right)^2 & \text{for } J(\phi_k(t))<J_k^0-\Delta J^0\\ 0 & \text{otherwise}\end{cases}$$



Flat-bottom harmonic restraint

Biasing potential (Gaussian)

$$V_{ki}^{le}\left(\phi_k(\mathbf{r}(t))\right)=K^{Jres}w_{\phi_{ki}}(t)\exp\left(-\left(\phi_k(t)-\phi_{ki}^0\right)^2/2\left(\Delta\phi^0\right)^2\right)$$

Memory-based wrighting

$$w_{\phi_{ki}}(t)=(\Delta t)^{-1}\int_0^t \delta_{\phi_k(\mathbf{r}(t'))\phi_{ki}^0}V^{Jrest}(J(\phi_k(\mathbf{r}(t')));J_k^0,\Delta J^0)V^{Jrest}(\bar{J}(\phi_k(\mathbf{r}(t')));J_k^0,\Delta J^0)dt'$$

One if dihedral within bin

Accounting for motional averaging

- Time-averaged J-value restraining

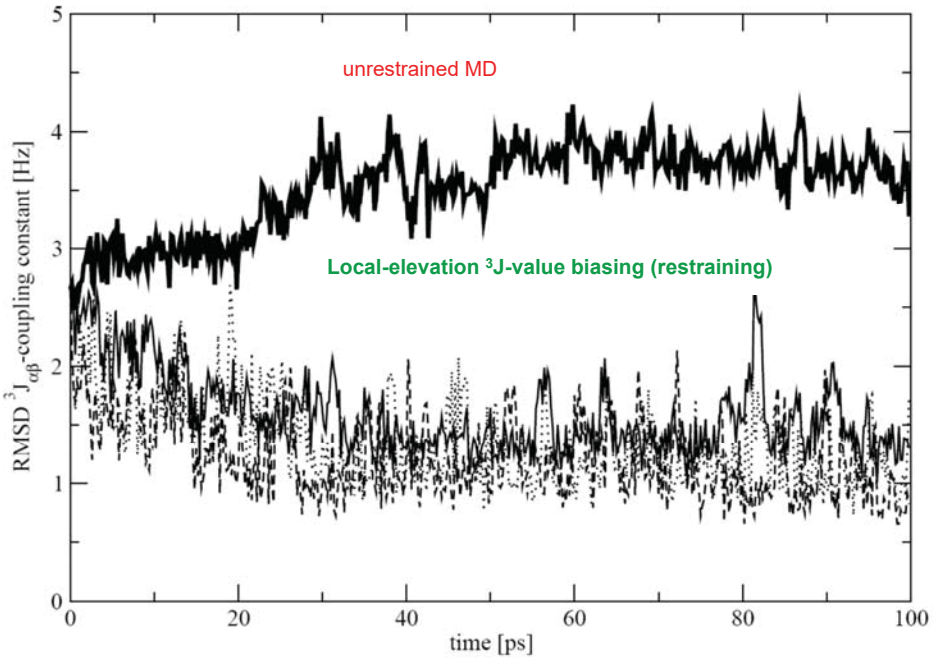
  → Application: Hen Egg White Lysozyme (37 $^3J_{\alpha\beta}$ -values)

$^3J_{\alpha\beta}$-coupling constants
(stereospecifically assigned)
χ_1 torsional angles

residue			$^3J^0$	residue			$^3J^0$
name	nr	proton		name	nr	proton	
Val	2	β	10.8	Phe	3	β_3	3.0
Cys	6	β_2	11.5	His	15	β_2	11.2
Asp	18	β_3	11.0	Tyr	20	β_3	11.7
Tyr	23	β_2	10.9	Asn	27	β_2	10.3
Val	29	β	11.1	Cys	30	β_2	5.3
Phe	34	β_3	5.0	Asn	39	β_2	4.5
Thr	40	β	4.5	Thr	43	β	3.7
Asn	46	β_3	4.7	Thr	47	β	2.6
Asp	48	β_2	2.6	Thr	51	β	9.3
Asp	52	β_2	11.6	Tyr	53	β_2	10.4
Asn	59	β_2	5.4	Arg	61	β_3	10.8
Asp	66	β_3	4.5	Thr	69	β	9.3
Leu	75	β_3	2.1	Asp	87	β_2	5.1
Ile	88	β	4.5	Thr	89	β	9.5
Val	92	β	10.1	Cys	94	β_2	4.0
Val	99	β	6.3	Val	109	β	8.0
Thr	118	β	4.2	Asp	119	β_2	4.9
Trp	123	β_2	10.6	Ile	124	β	4.6
Cys	127	β_2	11.6				

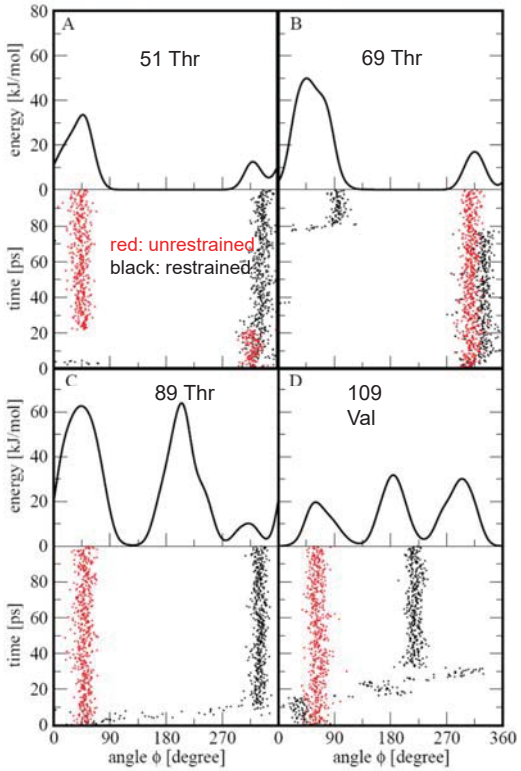
Accounting for motional averaging

- RMSD from experiment for 37 $^3J_{\alpha\beta}$-coupling constants



Accounting for motional averaging

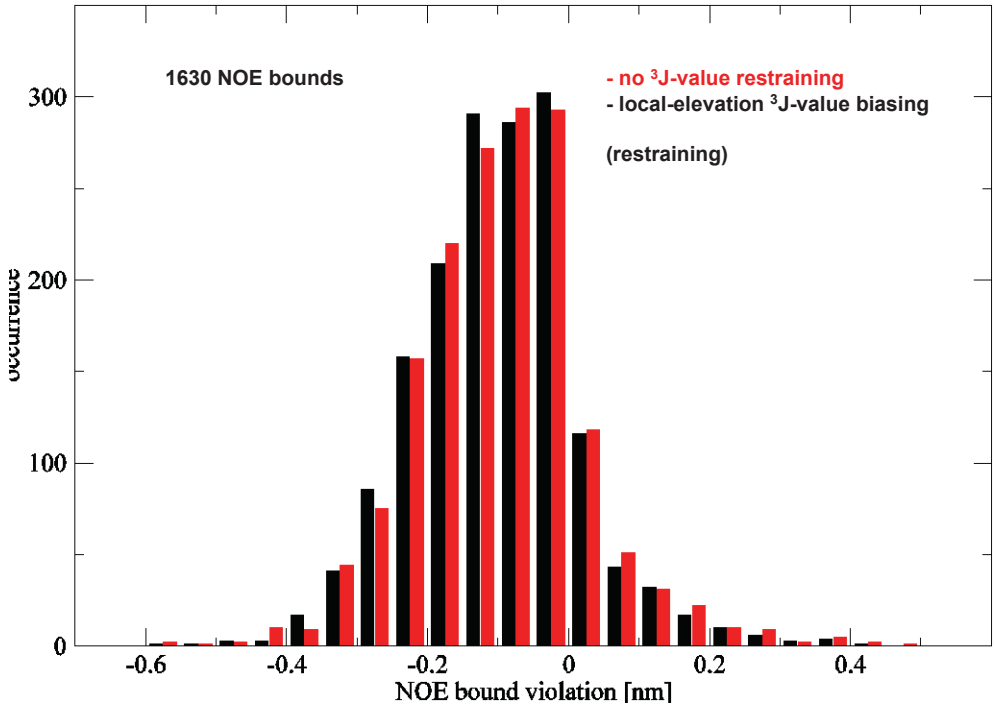
→ Local-elevation energies after 100ps torsional angle values each 0.2ps



residue		^3J-coupling constant		
name	nr	exp	unrestr	restr
Thr	51	9.3	4.3	9.7
Thr	69	9.3	12.5	9.8
Thr	89	9.5	2.5	9.9
Val	109	8.0	3.2	8.2

Accounting for motional averaging

- Hen Egg-white Lysozyme MD



Local-elevation ^{3}J-value restraining of 37 side-chain $^3J_{\alpha\beta}$-coupling constants in MD improves agreement of 1630 NOE atom-atom distances with experimental NOE-bounds

Accounting for motional averaging

Conclusions and Outlook

- A new application of the *local-elevation technique* to achieve *^3J-coupling constant restraining* to measured values is proposed
- Using this method it is possible to successfully *restrain ^3J-coupling constants without destabilising the overall molecular structure*. In the example Lysozyme an improvement on reproducing experimental NOE distance bounds is observed
- The method achieves *selectively enhanced sampling* by disfavouring conformations of dihedral angles with ^{3}J-coupling constants deviating from experiment. A *minimum interference by the restraints* compared to an unrestrained simulation is guaranteed
- The method is not very sensitive with respect to force constant and other parameters chosen, which makes it *suitable for inclusion of ^3J-value restraining in standard biomolecular NMR structure refinement*

M. Christen, B. Keller, & W.F. van Gunsteren, J. Biomol. NMR 39 (2007) 265-273

Accounting for motional averaging

Crystallographic refinement by MD with time-averaged structure factor restraints

Potential energy (target) function:

$$V(\vec{r}) = V^{phys}(\vec{r}) + V^{X-ray}(\vec{r})$$

$$V^{X-ray}(\vec{r}) = \frac{1}{2} k_r^{sf} \sum_{hkl} \left\{ \left| \left\langle F_{calc}\left(hkl;\vec{r}(t)\right)\right\rangle_t \right| - \left| F_{obs}\left(hkl\right)\right| \right\}^2$$

$$\left\langle F_{calc}\left(hkl\right)\right\rangle_t = \frac{1}{\tau\left[1-e^{-t/\tau}\right]} \int_0^t e^{-[t-t']/\tau} F_{calc}\left(hkl;\vec{r}(t')\right) dt'$$

τ = **memory relaxation time**

$$F(hkl) = F(\vec{k}) \propto \int \rho(\vec{r}) e^{i\vec{k}\cdot\vec{r}} d\vec{r} =$$ structure factor

$$\rho(\vec{r}) \propto \int F(\vec{k}) e^{-i\vec{k}\cdot\vec{r}} d\vec{k} =$$ electron density

X-ray diffraction measurement yields structure factor amplitudes: $\left|F(hkl)\right| = \left|F(\vec{k})\right|$

unknown

***not* the phases ϕ_{obs}:** $F_{obs}(real) = \left|F_{obs}\right| e^{i\varphi_{obs}}$

Key issues in structure refinement

INSTANTANEOUS OBSERVABLE

$Q(\boldsymbol{r})$

Common observables
Approximations
Invertibility
Calculation cost

ENSEMBLE AVERAGING

$\langle Q \rangle$

Single structure vs ensemble
Convergence

FORCE-FIELD REPRESENTATION

$\mathcal{V}(\boldsymbol{r})$

Accuracy

EXPERIMENTAL DATA

Q^{exp}

Primary vs secondary
Quality and accuracy
Redundancy vs completeness

BIASING FUNCTION REPRESENTATION

$\mathcal{V}_{res}(\boldsymbol{r})$

Weighting
Instantaneous vs time-averaged

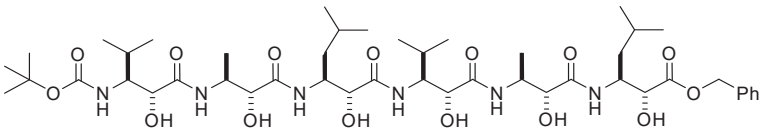
Illustrative example

Common errors, inefficiencies and misinterpretations

Illustrative examples

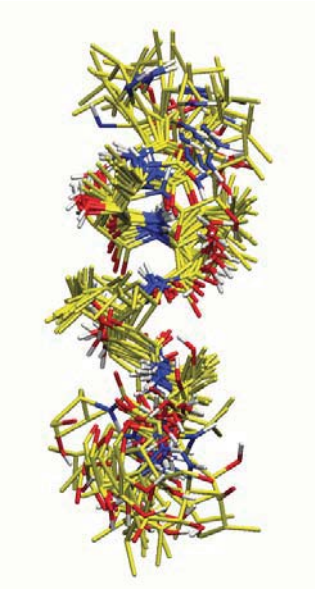
A β-hexapeptide

Alice Glättli



- β-hexapeptide with hydroxyl groups attached to the α-carbons
- NMR model structure suggests the formation of a 2_8-***P***-helix
- MD simulation from totally extended conformation at two different temperatures (298 K & 340K) using the GROMOS 45A3 force field
- No NOE-distance or J-value restraining in MD simulation

Glaettli & van Gunsteren, Angew. Chem. Int. Ed. Engl. 43 (2004) 6312



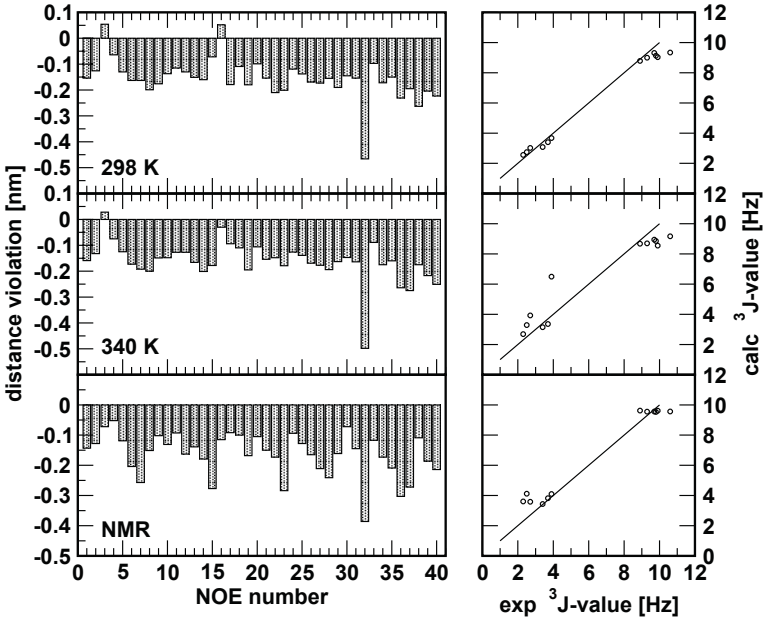
Bundle of 20 NMR model structures

(protection groups not shown)

Gademann et al., Angew. Chem. Int. Ed. Engl. 42 (2003) 1534

Illustrative examples

NOE Distance Violations & Backbone ^{3}J-values



- at 298 K
  2 violations (~0.05 nm)

  average deviation from exp. J-values: 0.44 Hz

- at 340 K
  1 violation (~ 0.03 nm)

  average deviation from exp. J-values: 0.91 Hz

- NMR bundle
  no violation

  average deviation from exp. J-values: 0.57 Hz

Illustrative examples

Occurrence of Hydrogen Bonds [%]

	MD simulation		NMR bundle
Donor-Acceptor	**298 K**	**340 K**	**X-PLOR**
NH(i)-O(i-2) [HB8]			
NH(3)-O(1)	0	1	20
NH(4)-O(2)	0	1	25
NH(5)-O(3)	2	4	10
NH(i)-O(i-3) [HB12]			
NH(3)-O(0)	0	30	0
NH(4)-O(1)	0	26	0
NH(5)-O(2)	0	35	0
NH(6)-O(3)	1	18	0
NH(i)-O(i+1) [HB10]			
NH(2)-O(3)	11	0	0
NH(5)-O(6)	11	1	0

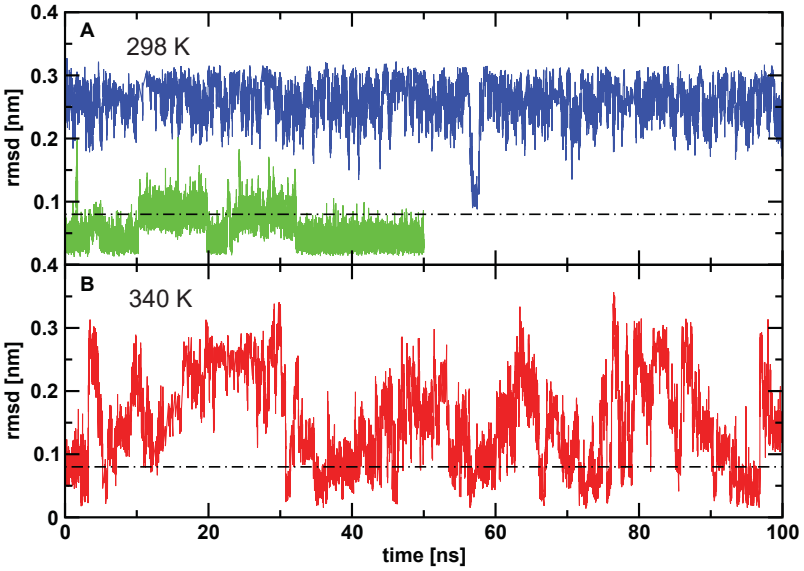
	MD simulation		NMR bundle
Donor-Acceptor	**298 K**	**340 K**	**X-PLOR**
OH(i)-O(i-1) [HB7]			
OH(6)-O(5)	0	14	0
OH(i)-O(i-2) [HB11]			
OH(4)-O(2)	0	8	10
OH(5)-O(3)	1	22	0
OH(6)-O(4)	1	10	0
OH(i)-O(i-3) [HB15]			
OH(4)-O(1)	1	26	0
OH(5)-O(2)	0	10	0
OH(i)-O(i+2) [HB13]			
OH(3)-O(5)	38	0	0

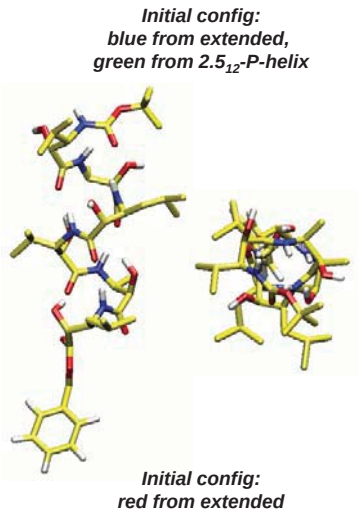
None of the H-bond patterns supporting the formation of a 2_8-***P***-helix were detected in the simulations.

Illustrative examples

Another possible Secondary Structure Element: 2.5_{12}-*P*-helix

Atom-positional root-mean-square deviation from 2.5_{12}-***P***-helix





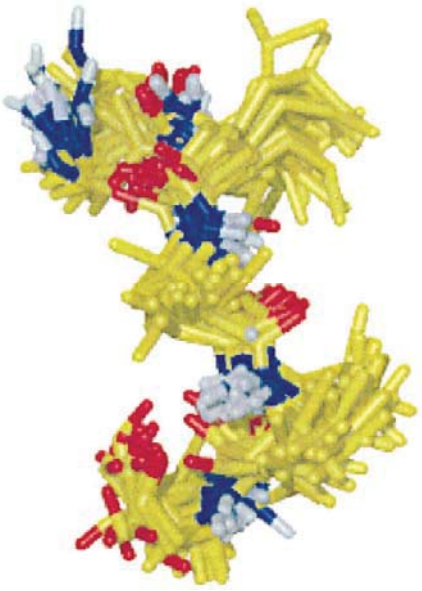
2_8-***P***-helix is ***virtually absent***, but

2.5_{12}-***P***-helix is ***35% populated*** at 340 K and ***stable*** at 298 K

Illustrative examples

Definition of a Conformer for a β-heptapeptide

20 structures forming one conformational cluster



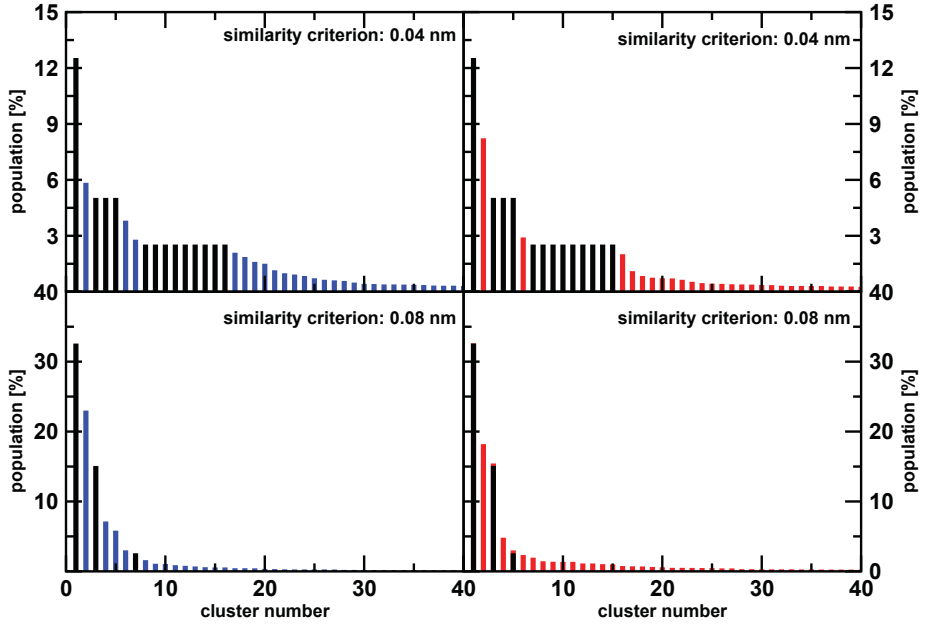
Backbone atom-positional RMSD (residues 2-6) ≤ 0.09 nm

Maximum RMSD between two members 0.16 nm

Illustrative examples

Conformational Analysis of the combined MD & NMR "Ensembles"

MD at 298 K + NMR bundle MD at 340 K + NMR bundle



No overlap between MD trajectory and NMR bundle of structures

Illustrative examples

Conclusions

1. MD simulation using a "thermodynamic" force field (GROMOS) (without NMR restraints) reproduces experimental NOE/J-value data equally good or better than a set of 20 NMR model structures derived by classical single structure refinement techniques (XPLOR)
   (aspect: force field problem)

2. Single structures may not be representative for the (Boltzmann) ensemble of structures in solution
   (aspect: ensemble problem)

3. Standard (NMR) structure refinement procedures should be revised in order to avoid the deposition of non-representative model structures in structure data banks
   (aspect: search problem)

4. Don't compare secondary (derived) data (structures, angles) but primary (measured) data (NOE's, ^{3}J-values) when comparing models with experimental data
   (aspect: experimental problem)

Angew. Chem. Int. Ed. ***43*** *(2004) 6312*

Key issues in structure refinement

INSTANTANEOUS OBSERVABLE

$Q(\boldsymbol{r})$

Common observables
Approximations
Invertibility
Calculation cost

ENSEMBLE AVERAGING

$\langle Q \rangle$

Single structure vs ensemble
Convergence

FORCE-FIELD REPRESENTATION

$\mathcal{U}(\boldsymbol{r})$

Accuracy

EXPERIMENTAL DATA

Q^{exp}

Primary vs secondary
Quality and accuracy
Redundancy vs completeness

BIASING FUNCTION REPRESENTATION

$\mathcal{U}_{res}(\boldsymbol{r})$

Weighting
Instantaneous vs time-averaged

Illustrative example

Common errors, inefficiencies and misinterpretations

Common errors, inefficiencies and misinterpretations

(arbitrary, non-Boltzmann sampled set)

1. **Conversion of a ^{3}J-value (other than extremes) to a ϕ-angle value with subsequent ϕ-angle restraining**
   (inverse Karplus relation is non-linear and multiple-valued)
2. **Instantaneous restraining**
   (neglect of averaging effects)
3. **Time-averaging restraining of ^{3}J-values using a restraining function only dependent on $\langle ^3J \rangle$ (too large fluctuations)**
4. **Use of non-Boltzmann weighting of conformers**
   (violates statistical mechanics)
5. **Use of equations of motion in internal, non-cartesian coordinates**
   Torsional dynamics is either inefficient or yields wrong dynamics
6. **Freezing of bond-angle degrees of freedom**
   (reduces motion and entropy, no gain in efficiency)
7. **Inadequate sampling**
   (many, but high-energy conformers)
8. **...**

Common errors, inefficiencies and misinterpretations

Refinement of protein structures using simulation

1. Use a ***thermodynamically calibrated force field***
2. Include essential degrees of freedom: ***solvent***

   Be certain that the solvent model is compatible with the protein one
3. Use the ***appropriate*** (experimental) ***thermodynamic state point***:
   a. Temperature
   b. Pressure
   c. pH
   d. Ionic strength (co-solvents)
4. ***Sample*** conformational space ***sufficiently*** and ***Boltzmann-weighted***
5. When using experimental data to bias the sampling:
   a. ***Use*** only ***primary*** (measured) not secondary (derived) experimental ***data***
   b. Do ***account for motional averaging***
   c. ***Do not restrain*** to a measured value of an ***observable if*** the function connecting structure to observable is ***multiple-valued***

Computer Simulation in Chemistry, Biology and Physics

P.H. Hünenberger

COMPUTER SIMULATION OF MOLECULAR SYSTEMS

Herbstsemester 2019
Tuesday 9:45-11:30 a.m.
HCI D2

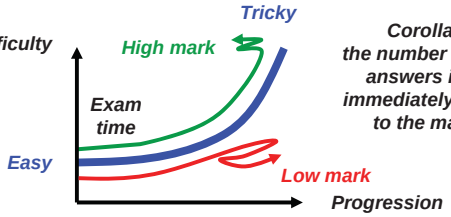
Lecture 529-0004-00
www.csms.ethz.ch/education/CSCBP

LECTURE 13 (WEEK 14):
Concluding remarks

About the examination

- February or August 2020
  - → **Oral** exam, **30'**, **no preparation**, Phil + Beisitzer
  - → **English or German** (or even French) as you prefer
  - → Atmosphere is **friendly**
  - → I normally **"improvise" the questions** on the flight

So that I can "tune" the level of difficulty, and we can "stop" a question where you don't know to move on to one where you know



Corollary: the number of good answers is not immediately related to the mark...

  - → Basic principle of an oral exam ➡
  - → Your **exercise grades** are used for "rounding" (*i.e.* component of ~0.5 [*i.e.* ±0.25])
  - → You don't get the mark before the Notenkonferenz

But when it is obvious, we usually say if "pass" or "fail" right after the exam

- Most important
  - → Be **relaxed**, **don't self-assess** during the exam, just **give your best** !
  - → Try to understand the **working principles** of the different methods
  - → It is a good idea to know some of the **key equations** "by heart" (there is no time to derive everything); but knowing "by heart" won't help at all if you don't know the working principles: you need both

If there is a question you cannot answer, just let it be – and focus on the next...

Example: if I ask you about TI and we spend 5' to just to get to the correct equation, these are 5' less to discuss more interesting aspects of TI (practical use, limitations, ...)

About the examination

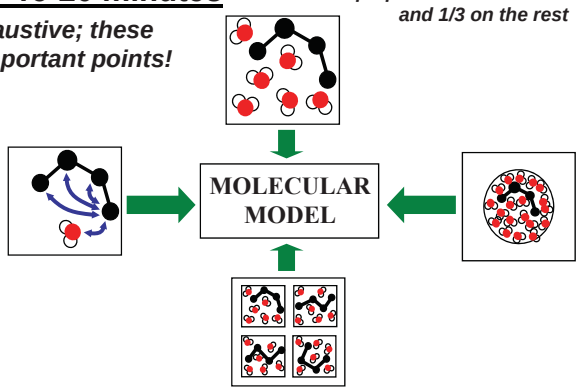
→ Examples of “starting” questions (non-exhaustive; just to give you an idea...)

- what are the main resolution levels in simulations (QM,MM,CG) and (broadly) their advantages/limitations/scalings
- what are the terms in a force-field ***(then we pick one to see what you know)***
- how do we integrate the equations of motion in MD ***(then we might ask the leap-frog equations; no derivation)***
- how do we sample something else than the microcanonical ensemble
- what are the main methods to calculate electrostatic interactions under PBC
- how would you calculate this or that property using MD ***(e.g. permittivity, diffusion constant, heat capacity, ...)***
- name free energy methods ***(then we pick one, e.g. DC, US, TI, FEP, EDS ... to see what you know; no derivation)***
- name enhanced-sampling methods ***(then we pick one, e.g. LE, DEM, 4DMD, to see what you know; no derivation)***
- ...

Basic principles: ~15-20 minutes

Advice: spend 2/3 of your preparation on these basics and 1/3 on the rest

Lists are non-exhaustive; these are just the most important points!



1) INTRODUCTION

- choice of the degrees of freedom
  - QM *vs* atomistic *vs* coarse-grained
  - implicit *vs* explicit solvent
  - computational cost/scaling/limitations
- ~~quantum chemistry~~

2) INTERACTION (FORCE FIELDS)

- basis of the classical description
- molecular topology
- covalent/nonbonded/unphysical force-field terms
  - know their forms (incl. eqs.) and usage
- calculating atomic forces (principle)
- force-field parameterization (main sources of data)

3) GENERATING CONFIGURATIONS

- EM, MC
  - know the basic principles (incl. eq. for MC)
- MD
  - know the basic equations of class. mech.
  - explain leap-frog integrator (incl. eqs. without derivation)
  - explain the use of constraints (SHAKE picture, not eqs.)
  - timestep choice, initial conditions
- SD
  - know the basic principles

4) BOUNDARY CONDITIONS

- main spatial boundary conditions
- main stat. mech. ensembles
- instantaneous temperature/pressure (incl. eqs.)
- thermostatting/barostatting (qualitative, no eqs.)

5) ELECTROSTATIC INTERACTIONS

- problems of finite-size and surface effects
- main approaches for implicit solvent
- main approaches for explicit solvent
  - working principle of cutoff, RF, LS (no eqs.)
  - atomic *vs* group-based cutoff
  - shortcomings and artifacts
- pairlisting algorithms (qualitative)

Advanvced topics: ~15 minutes

Lists are non-exhaustive; these are just the most important points!

→ Lectures 6: analysis

- ***Time series, average, fluctuation, distribution***
- ***Periodic gathering, rototranslational fitting***
- ***Calculation principle of main properties: internal coordinates, RMSD, RMSF, RDF, thermodynamic properties, diffusion constant, dielectric permittivity, ...***

→ Lectures 7&8: free energy

- ***Three types of free energy changes, use of thermodynamic cycles, use of dummy atoms, understand DC, US, TI, FEP, EDS (with equations)***
- ***Understand issues related to accuracy: hysteresis, cycle-closure, finite sampling, approximate quadrature, overcoming singularities (soft-core) [you can forget free-energy components and entropy calculations]***

→ Lectures 9: enhanced sampling

- ***Understand the principles (not the detailed equations) of the temperature annealing, parallel tempering, soft-core, diffusion equation, local elevation, 4DMD, parallel tempering and Hamiltonian replica exchange methods***

→ Lectures 10: refinement

- ***Understand the principles of structure refinement based on NOE-derived distances and J-values [you can forget the Xray refinement]***
- ***Understand the main approaches and difficulties in NMR-based refinement***

→ Exercises 1-6 plus Lecture 12-13 thinking questions

- ***I may also ask questions related to that (of course, I won't assume that you remember all the details of these exercises!)***