Methodologies of systematic structure-based coarse-graining

Alexander Lyubartsev

Division of Physical Chemistry
Department of Material and Environmental Chemistry
Stockholm University

E-CAM workshop:
State of the art in mesoscale and multiscale modeling

Dublin 29-31 May 2017

Outline

1. Multiscale and coarse-grained simulations

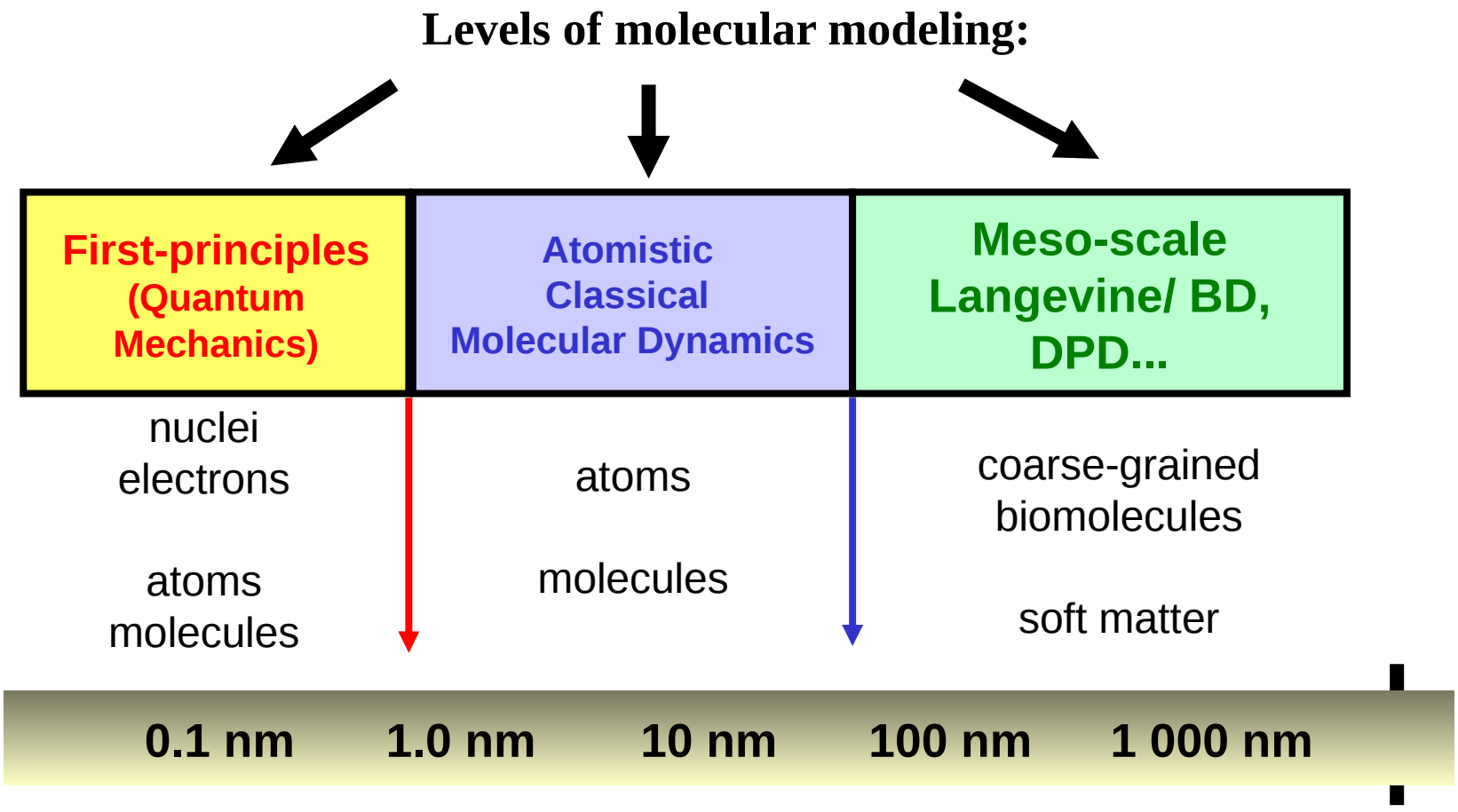
2. Systematic structure-based coarse-graining by Inverse Monte Carlo

3. Examples
   - water and ions
   - lipid bilayers and lipid assemblies

4. Software - MagiC

Example of systematic coarse-graining:
DNA in Chromatin: presentation 30 May

Computer modeling: from 1st principles to mesoscale



Larger scale ⇔ more approximations

Mesoscale Simulations

Length scale: > 10 nm (nanoscale: 10 - 1000 nm)

Atomistic modeling is generally not possible

box 10 nm - more than 10^5 atoms

even if doable for 10^5 - 10^6 particles - do not forget about time scale!

larger size : longer time for equilibration and reliable sampling

10^5 atoms - time scale should be above 1 μs = 1000 ns

Need approximations - coarse-graining

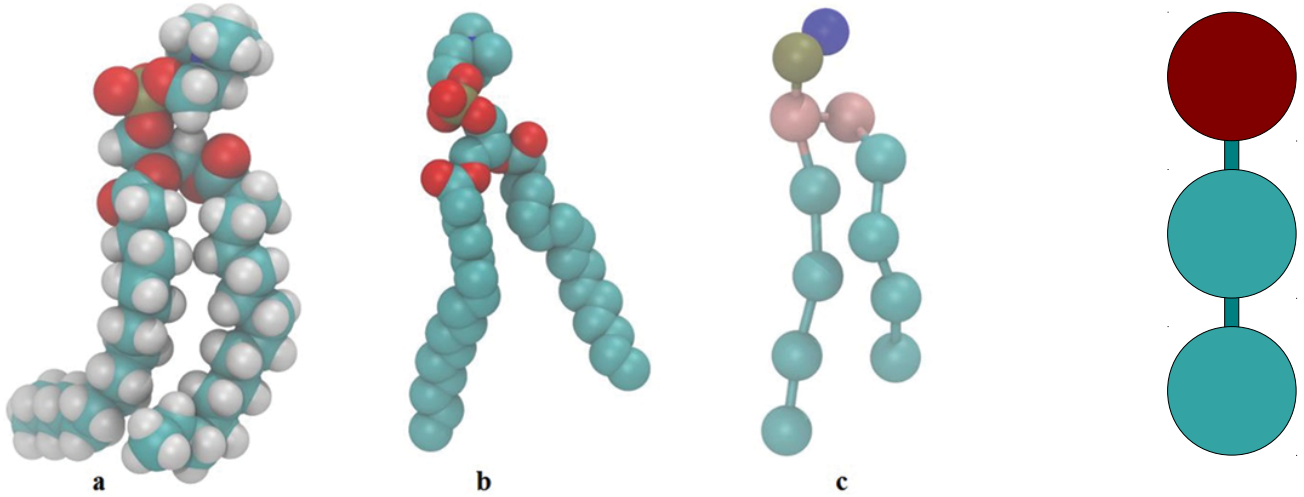
Coarse-graining – an example

Original size – 2.4Mb

Compressed to 24 Kb

Levels of coarse- graining

Level of coarse-graining can be different.
For example, for a DMPC lipid:



All-atom model
118 atoms

United-atom
model:
46 united atoms

Coarse-grained
10 sites:
Martini model

Coarse-grained
3 sites:
Cooke model

more details - chemical specificity

faster computations; larger systems

Coarse-graining of solvent:

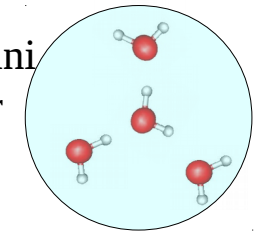
1) Explicit solvent:

One or several solvent molecules are united in a single site. E.g. Martini water model: one LJ particle representing 4 water molecules.

\+ : better description of hydrodynamic behaviour

\- : more computationally expensive
some artifacts remain
(polarization / dielecric effects)

Martini water



2) Implicit solvent:

No solvent particles but their effect is included into solvent-mediated effective potentials.
E.g.: primitive electrolyte model

\+ : Computationally very efficient especially for ”dilute” systems

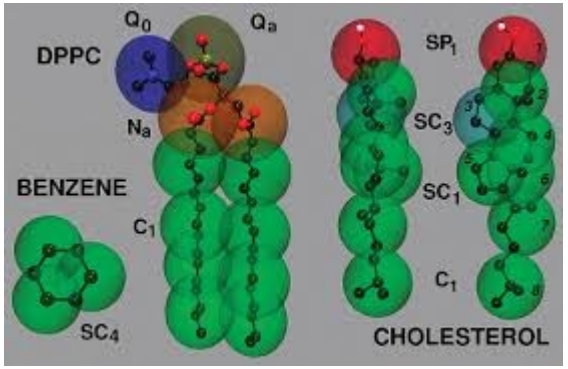
\- : more approximate dynamics
may be problem for inhomogeneous environment
(two different separated solvents, solvent - air interface etc)

Interactions for coarse-grained models

We determined the elements of the coarse-grained model.

Next question: ***How to set up the force field?***

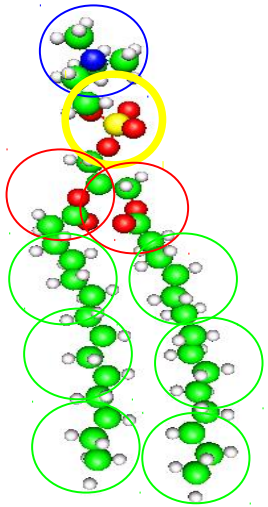
A) Empirically, to reproduce
available experimental data
example: Martini FF

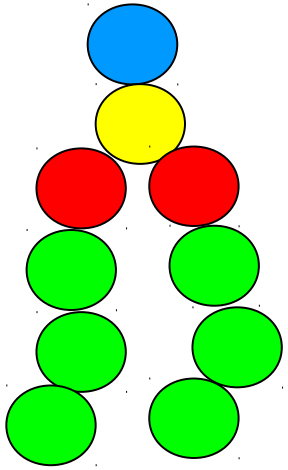


B) Bottom-up (systematic multiscale) approach:
Derive coarse-grained force field from the atomistic force field
(as atomistic force field can be derived from ab-initio computations)

Bottom-up approaches

Idea to derive parameters for CG models from results of atomistic simulations





$R_j = \theta(r_{1,} \ldots r_n)$

(centers of masses)

$U^{FG}(r_{1,} r_{2,} \ldots, r_n)$

n=118

$U^{CG}(R_{1,} R_{2,} \ldots, R_m) = ?$

m=10

- Force matching - fit forces
  (Izvekov & Voth, JPC-B,109,2469(2005))

- Structure based - fit RDF

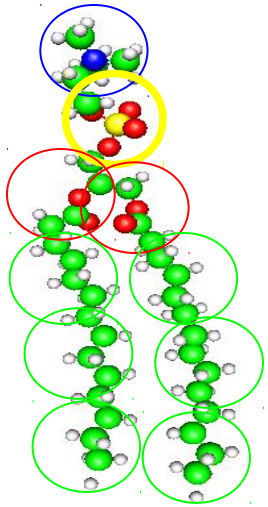
Henderson theorem:

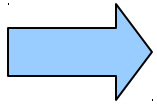
"For each set of RDF there is an unique set of pair potentials reproducing this set of RDFs"
(R. L. Henderson, Phys. Lett. A 49, 197 (1974))

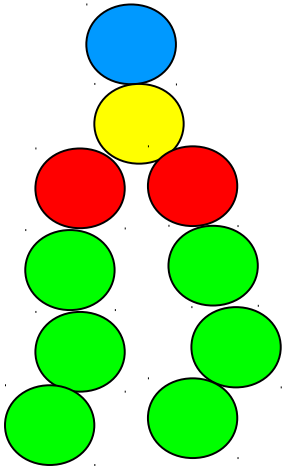
- Relative entropy (information loss) minimization
  (M.S.Shell, JCP, 129, 144108(2008))
  equivalent to fitting RDFs

Formal solution:
N-body mean force potential

Original (FG = fine grained system)







$H^{FG}(r_{1,}r_{2,}\dots,r_{n})$

$n = 118$

$R_j = \theta(r_{1,}\dots r_{n})$

$j = 1,\dots,10$

Usually, centers of mass of selected molecular fragments

Partition function :

$$Z = \int \prod_{i=1}^{n} dr_i \exp\left(-\beta H_{FG}(r_{1,}\dots,r_n)\right) =$$

$$= \int \prod_{i_1}^{n} dr_i \prod_{j=1}^{N} dR_j \,\delta\left(R_j - \theta_j(r_{1,}\dots r_n)\right) \exp\left(-\beta H_{FG}(r_{1,}\dots,r_n)\right) =$$

$$= \int \prod_{j=1}^{N} dR_j \exp\left(-\beta H_{CG}(R_{1,}\dots,R_N)\right)$$

where $\beta = \dfrac{1}{k_B T}$

$$H_{CG}(R_{1,}\dots,R_N) = -\frac{1}{\beta} \ln \int \prod_{i=1}^{n} dr_i \prod_{j=1}^{N} \delta\left(R_j - \theta_j(r_{1,}\dots,r_n)\right) \exp\left(-\beta H_{FG}(r_{1,}\dots r_n)\right)$$

is the effective N-body coarse-grained potential = potential of mean force = free energy of the non-important degrees of freedom

N-body potential of mean force (CG Hamiltonian) $H_{CG}(R_1,...R_N)$:

Structure: all structural properties are the same

for any $A(R_1,..R_N)$: $\langle A \rangle_{FG} = \langle A \rangle_{CG}$

Thermodynamics: in principle yes (same partition function)

but: $H_{CG} = H_{CG}(R_1,...R_N , \beta, V)$

already N-body mean force potential is state point dependent
T-V dependence should be in principle taken into account in computing thermodynamic properties

Dynamics: can be approximated within Mori-Zvanzig formalism

Problem with $H_{CG}(R_1,...R_N)$: simulation with N-body potential is unrealistic. ***We need something usable – e.g. pair potentials!***
(or any other preferably one-dimensional functions)

$$H_{CG}\left(R_1, R_2, \ldots, R_N\right) \approx \sum_{i>j} U_{ij}\left(R_{ij}\right) \qquad R_{ij} = |R_i - R_j|$$

How to approximate ?

- minimize the difference (Boltzmann averaged) : Energy matching
- minimize the difference of gradient (force) Force matching
- minimize the relative entropy
- ***provide the same canonical averages, e.g RDF-s***
  - Inverse MC
  - Iterative Boltzmann inversion

.. and may be some other properties "Newton inversion"

These approaches are in fact interconnected ….

Inverse Monte Carlo

Model ⟶ (direct) Properties

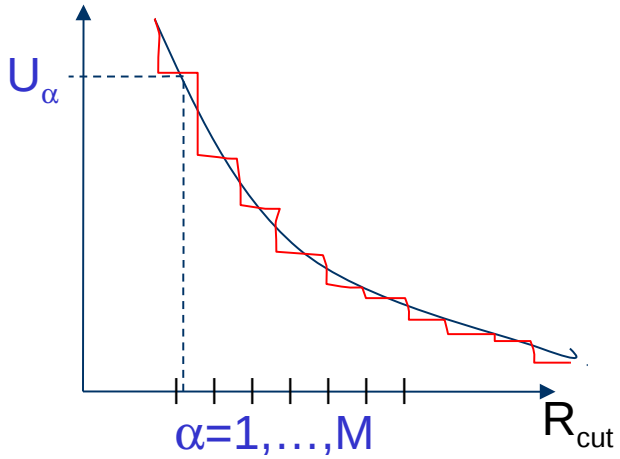
Model ⟵ (inverse) Properties

Interaction potential

E.g., radial distribution functions

Consider pair potentials: $H = \sum_{i,j} U(r_{ij})$

In tabulated (grid) approximation: $H = \sum_{\alpha} U_\alpha S_\alpha(r_i)$



$U_\alpha = U(R_{cut}\alpha/M)$ - potential within α-interval

S_α - number of particle's pairs with distance between them within α-interval

- defines RDF:

$$g(r_\alpha) = \frac{1}{4\pi r_\alpha^2 \Delta r} \frac{V}{N^2/2} \langle S_\alpha \rangle$$

Inversion algorithm

$$\{U_\alpha\} \xrightarrow{\text{direct}} \{\langle S\rangle_\alpha\}$$
$$\{U_\alpha\} \xleftarrow{\text{inverse}} \{\langle S\rangle_\alpha\}$$

Newton-Raphson solution of multidimensional non-linear problem

Jacobian of "RDF ↔ Potential" transformation:

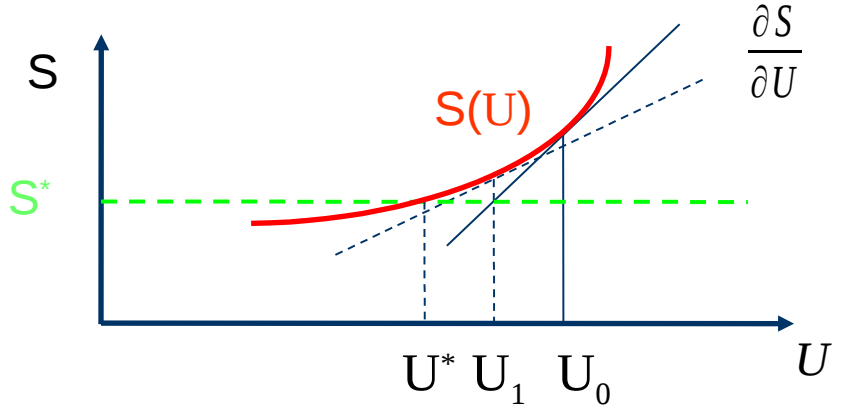
$$\boldsymbol{J} = \frac{\partial \langle S_\alpha \rangle}{\partial U_\gamma} = -\beta\left(\langle S_\alpha S_\gamma \rangle - \langle S_\alpha \rangle \langle S_\gamma \rangle\right)$$

$$\Delta U_\alpha = \boldsymbol{J}^{-1} \Delta S_\gamma$$

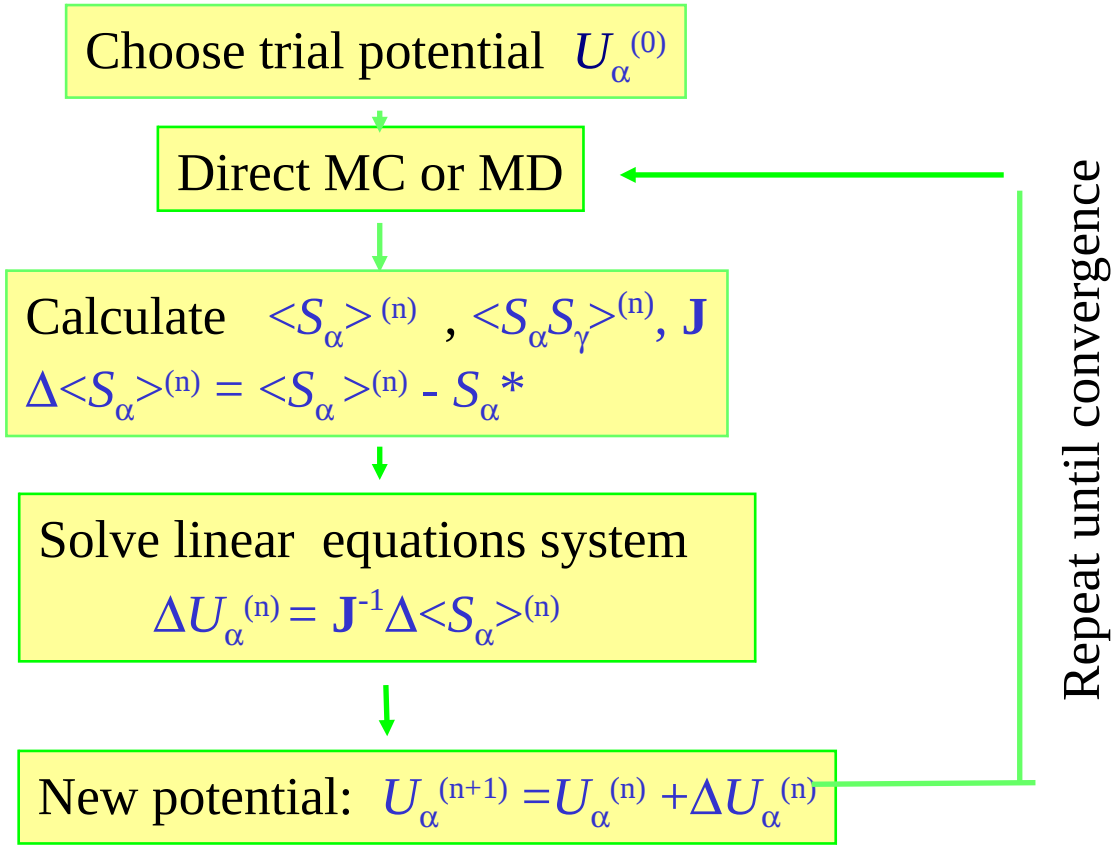
correction to the potential

deviation RDF from target

Can be found iteratively



Algorithm:



If no convergence, a regularization procedure can be applied:

$\Delta\langle S_\alpha \rangle^{(n)} = a(\langle S_\alpha \rangle^{(n)} - S_\alpha^*)$ with $a<1$

Additional comments:

- Relationsip " pair potential $\Leftrightarrow$ RDF " is unique
(Hendeson theorem: R. L. Henderson, Phys. Lett. A 49, 197 (1974)).

- In practice the inverse problem is often ill-defined, that is noticeable different potentials yield RDFs not differing by eye on a graph

- Another scheme to correct the potential (Iterative Boltzmann inversion):
$U^{(n+1)}(r) = U^{(n)}(r) + kT \ln(g^{(n)}(r)/g^{ref}(r))$

  - may yield different result from IMC though RDFs are similar
  - may not completely converge in multicomponent case
  - may be reasonable to use in the beginning of the IMC iteration process

- Analogy with Renormalization group Monte Carlo
(R.H.Svendsen, PRL, 42, 859 (1979))

Molecular (multiple-site) systems

1) a set of different site-site potentials
2) intramolecular potentials: bonds, angles, torsions

We use the same expression: $H = \sum_{\alpha} U_{\alpha} S_{\alpha}$

But now α runs over all types of potentials: non-bonded and Intramolecular. If α corresponds to a intramolecular potential, then $\langle S_{\alpha} \rangle$ is the corresponding bond, angle or torsion distribution

R D F $g_{ij}(r)$	$\Leftrightarrow$	pair CG potentials $U_{ij}(r)$
bond length distribution		bond potentials
angle distribution		angle potentials
torsion distribution		torsion potentials

Treatment of electrostatics:

If sites are charged, we separate electrostatic part of the potential as:

$$U_{tot}\left(r_{ij}\right) = U_{short}\left(r_{ij}\right) + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}}$$

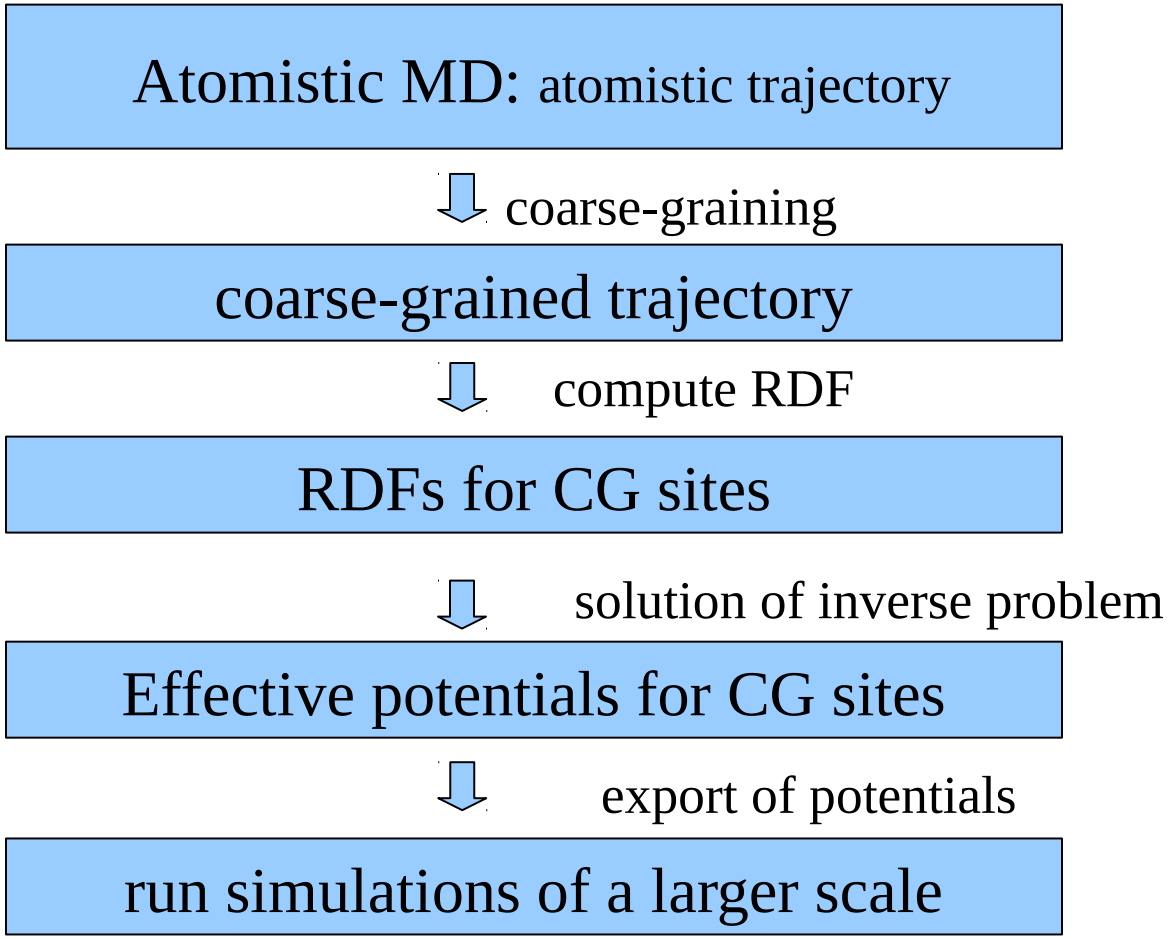
q_i - sum of charges from atomistic model

ε - dielectric permittivity: either experimental; or extracted from fitting of asymptotic behaviour of effective potentials)

Electrostatic part: computed by Ewald summation (or PME)

U_{short} - updated in the inverse procedure

Systematic Multiscale Approach:
Linking atomistic and coarse-grained simulations



Coarse-graines simulations:

- Molecular dynamics with local stochastic thermostat adding friction and random forces: Langevine, Brownian, V-rescale in Gromacs,...

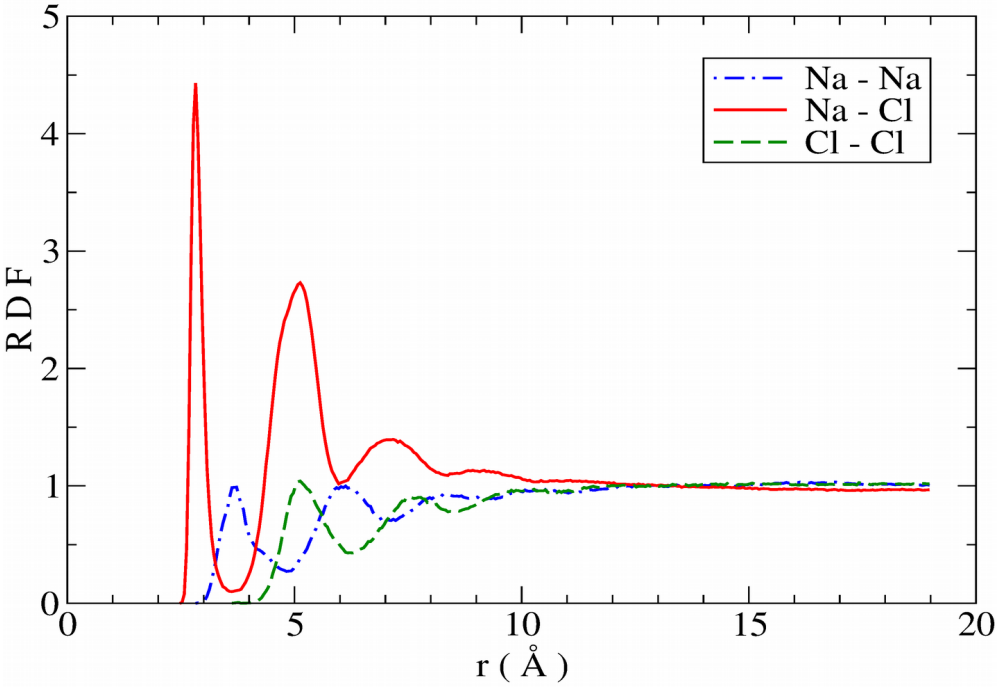
- DPD (pairwise random and friction forces)

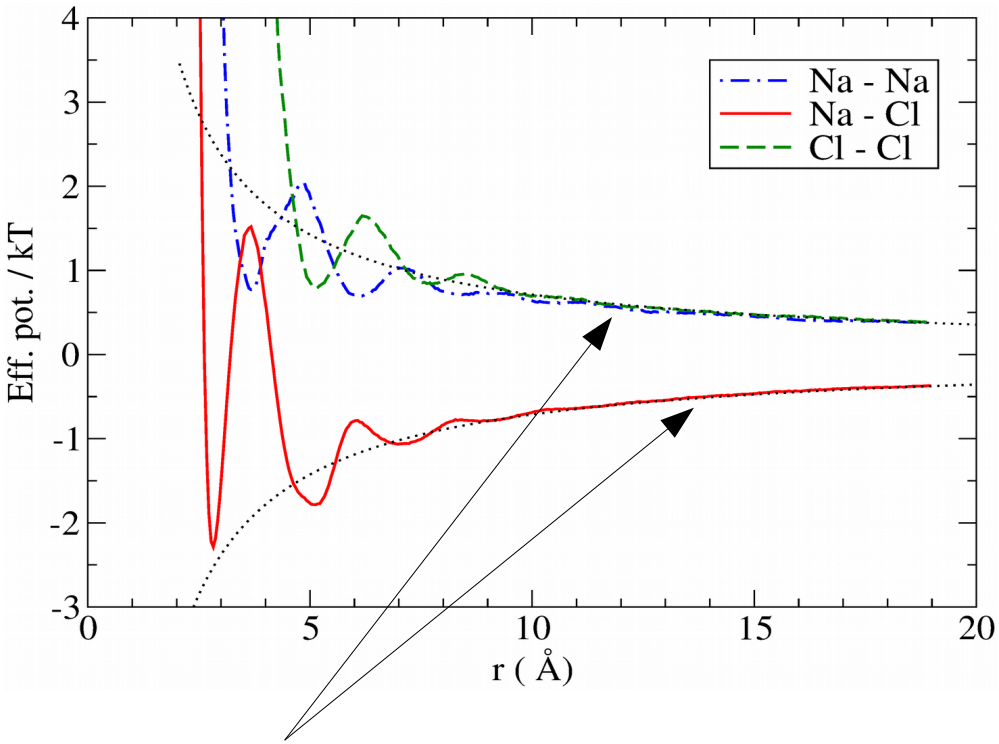
- Monte Carlo (including kinetic MC)

Simple example: NaCl ion solution

Ion-ion RDFs
(from atomistic MD)

Ion-ion effective
potentials



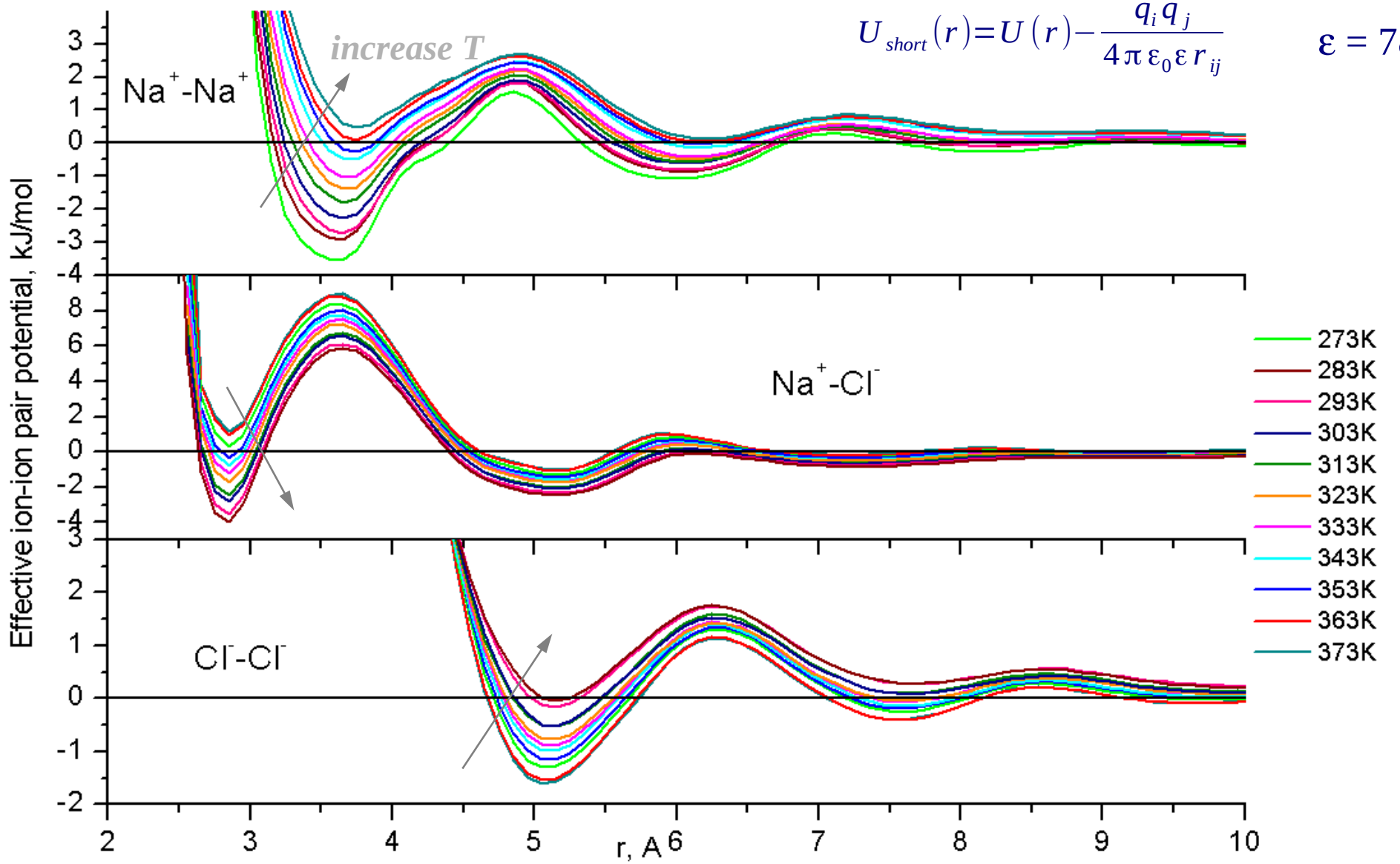


MD: 20 Na^+ and Cl^- ions,
2000 water molecules

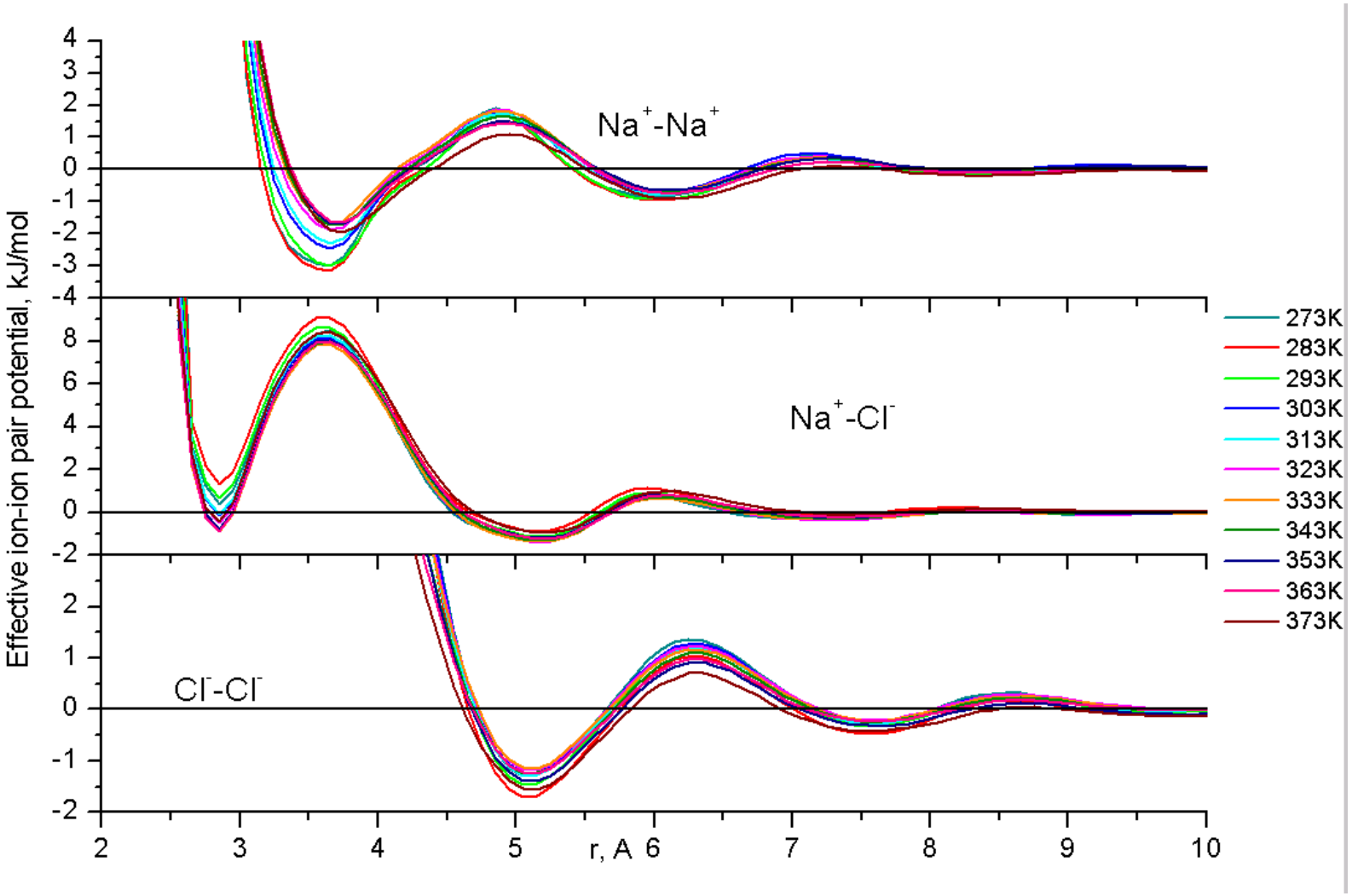
Coulombic asymptotic:

$$U(r) \underset{r\to\infty}{\rightarrow} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon r_{ij}}$$

Temperature dependence



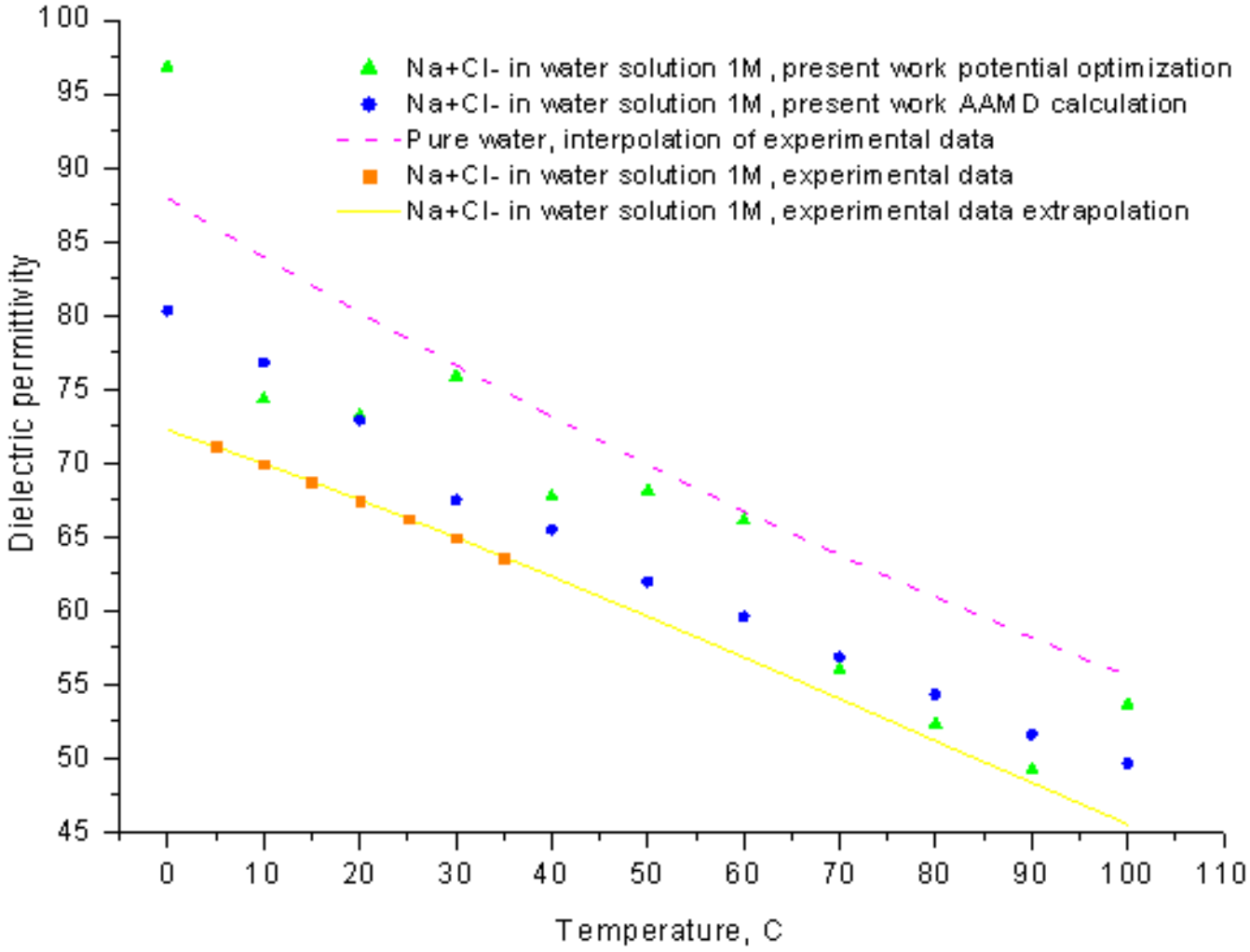
“Short range” effective potential: after optimizing ε



Temperature dependence can be effectively included in *ε*

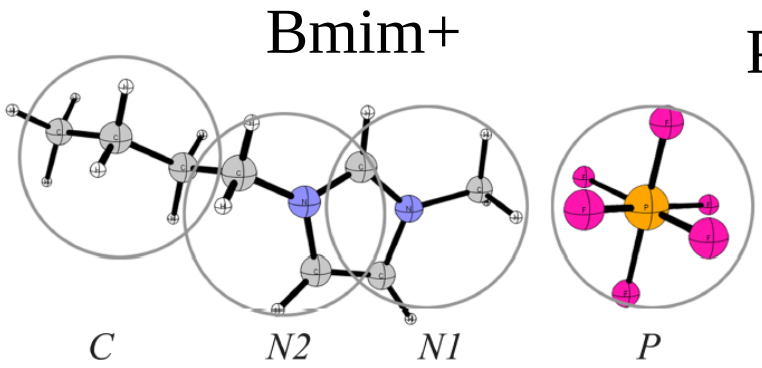
From tail asymptotic – extract “dielectric permitttivity”

Fit ε to minimize $U_{short}(r)=U(r)-\frac{q_i q_j}{4\pi\varepsilon_0\varepsilon r_{ij}}$ outside some cutoff

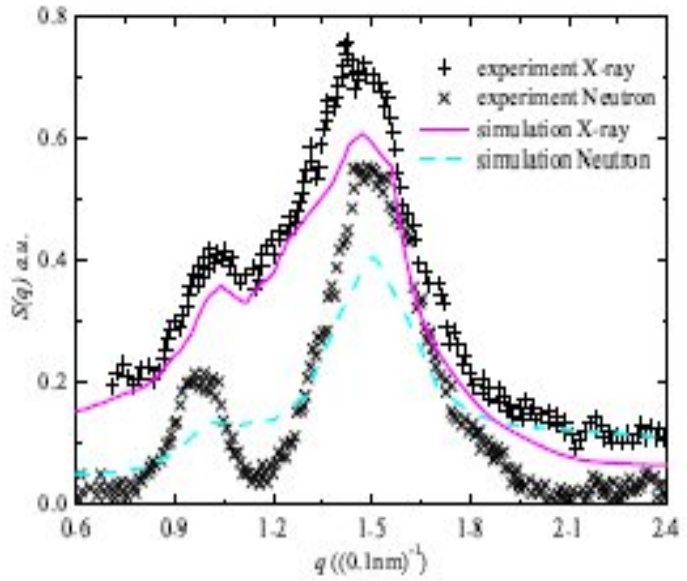


Reference:
A.A.Mirzoev, A.P.Lyubartsev
PCCP 13, 5722 (2011)

Ionic liquids



Long-range electrostatic interactions important:
CG models are needed to access long length- and time-scale



Atomistic MD: 128 ion pairs -> RDF -> effective CG potentials

Scattering factor obtained in CG-simulations with 4000 ion pairs

(Y.L Wang, A.P.Lyubartsev, Z.Y. Lu, A.Laaksonen, PCCP, 15, 7701 (2013)

Coarse-grained model for DMPC lipid

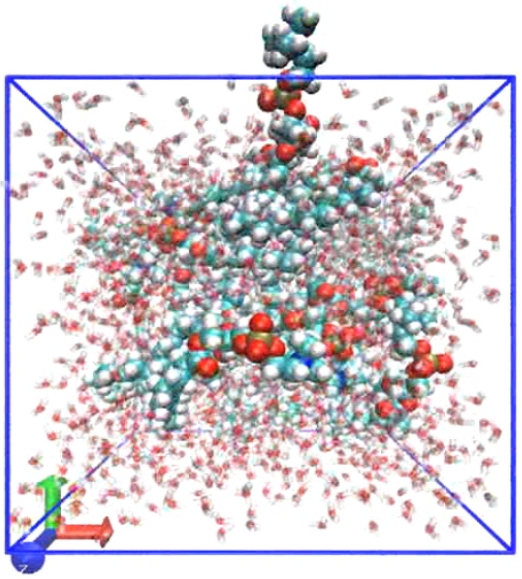
earlier work: Lyubartsev, Eur. Biophys. J.,35, 53 (2005)
recent: Mirzoev & Lyubartsev, J.Comp. Chem., 35, 1208 (2014)

Atomistic MD:

- 60 lipid molecules (DMPC)
  dissolved in 1800 waters
  + complementary simulations
  with 3 other lipid/water ratio

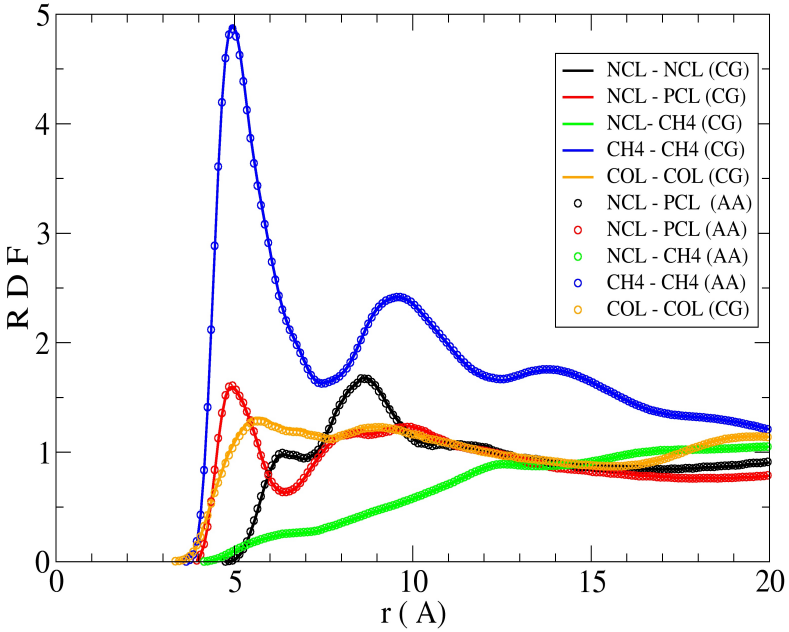
- Initial state - randomly dissolved

- RDFs calculated during 400 ns
  after 100 ns equilibration, T=303K

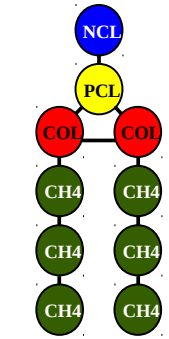


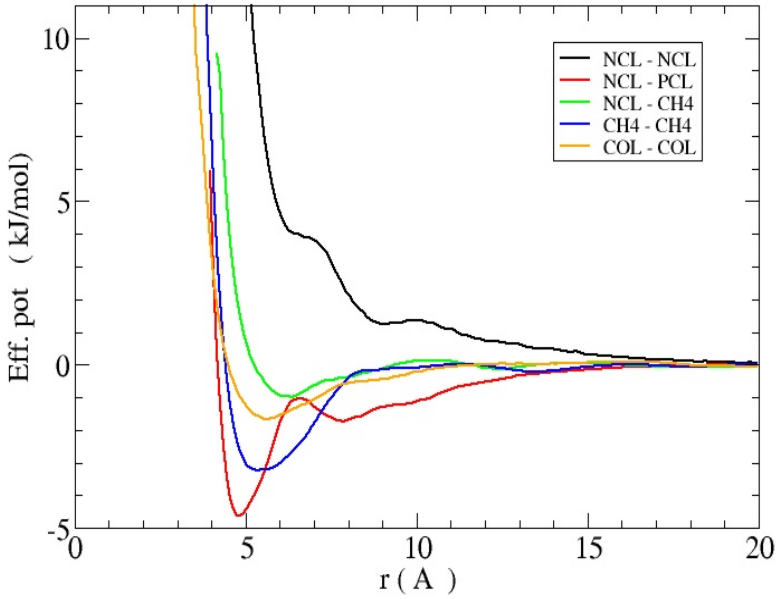
Computation of structure-based CG potentials:

Atomistic simulations (60 lipids + 1800 water / 400 ns) →
Structural Information (RDF) →
Coarse-Grained Potentials by the inverse MC



IMC

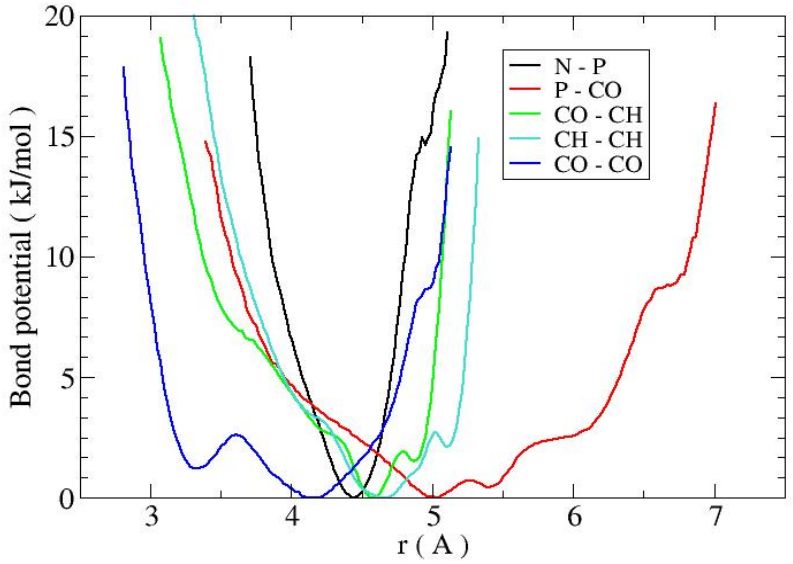




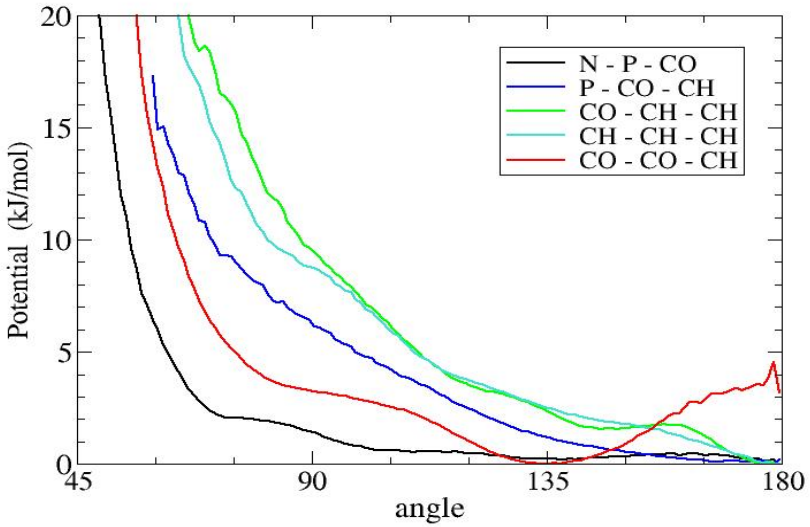
Effective potentials provide for coarse-grained models the same RDFs (structure) as atomistic simulations but require 2-4 order less CPU time for the same system

Intramolecular (bonded) potentials:

Bond potentials



Angle potentials

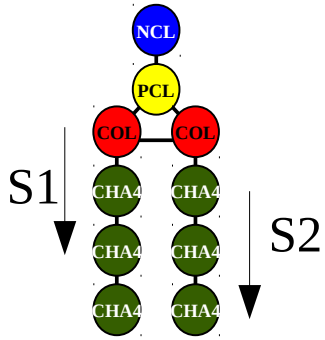


Multiple minima in bond potentials – reflect trans – gauche conformations of the atomistic representation

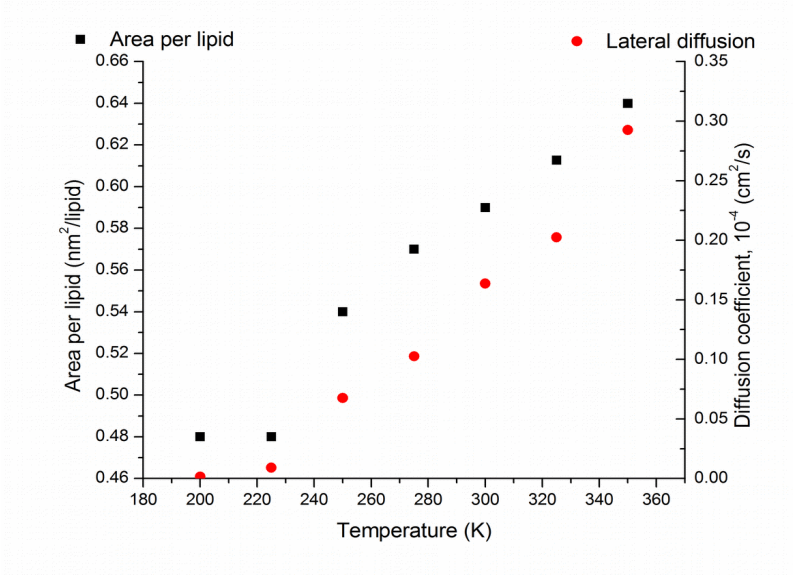
CG and atomistic DMPC bilayer

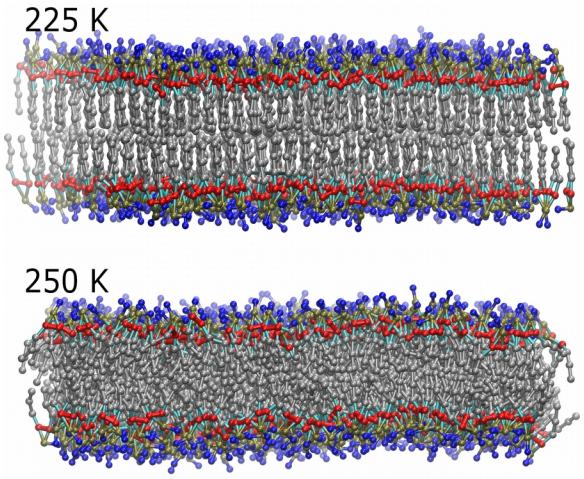
DMPC lipid bilayer properties (303K):

	Area per lipid	Compressibility 10^{-10} N/nm	Order parameter S1	S2
atomistic	60	1.9	0.57	0.52
CG	59	2.5	0.56	0.52



Phase transition on CG DMPC bilayer

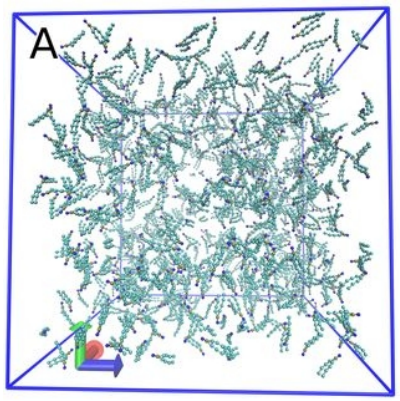


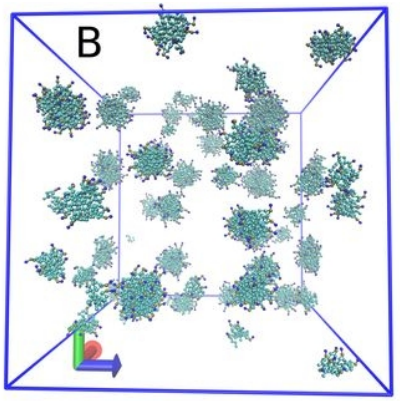


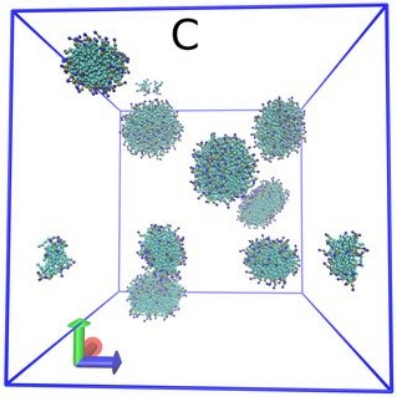
A.Mirzoev, A.P.Lyubartsev, J.Comp.Chem.,35, 1208-1215 (2014)

Lipid selfassembly into vesicle:

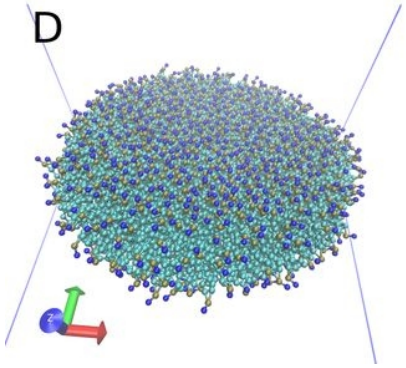
1000 lipids; start from random distribution

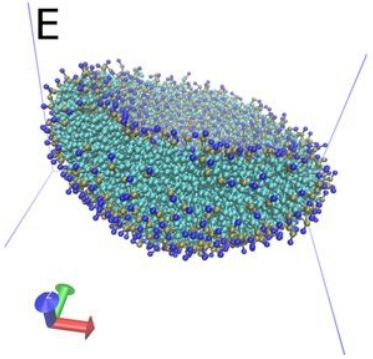


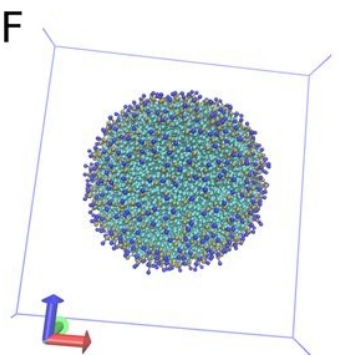


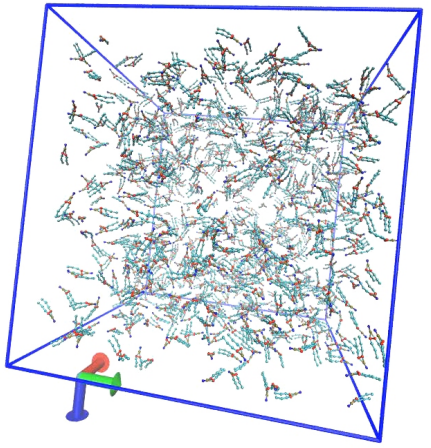


movie



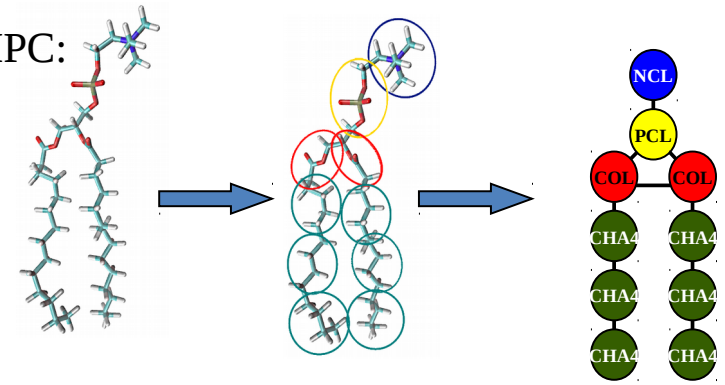




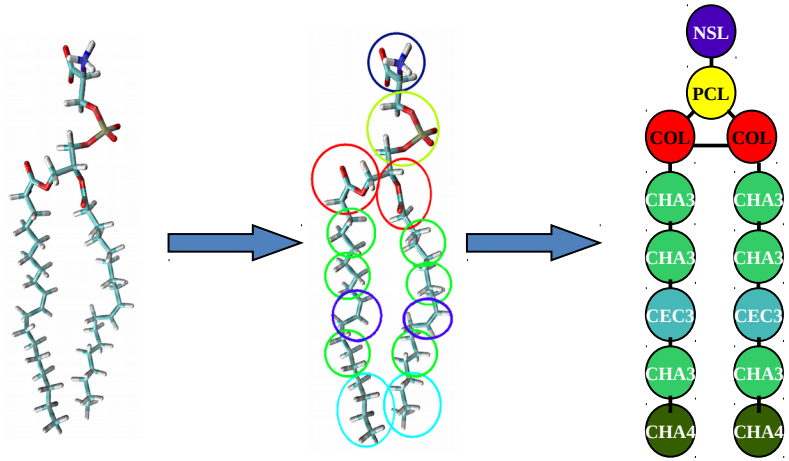


Other coarse-grained lipids

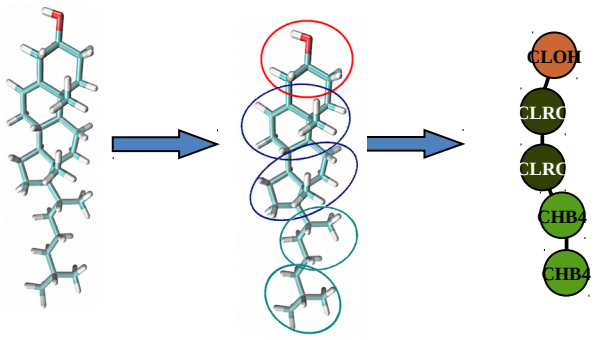
DMPC:



DOPS



Chol:



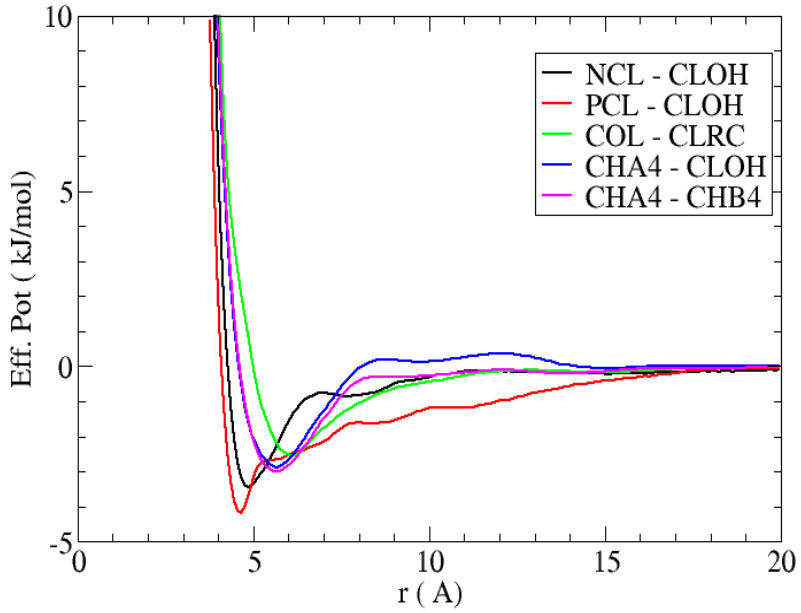
CG potentials for:

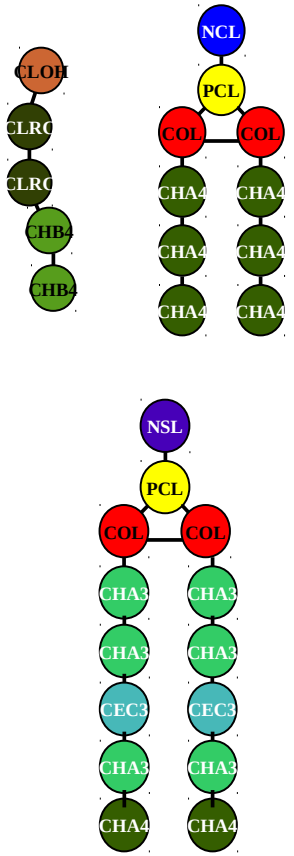
- PC and PS lipid head groups
- saturated and single-unsaturated tails
- cholesterol

CG potentials for other lipids

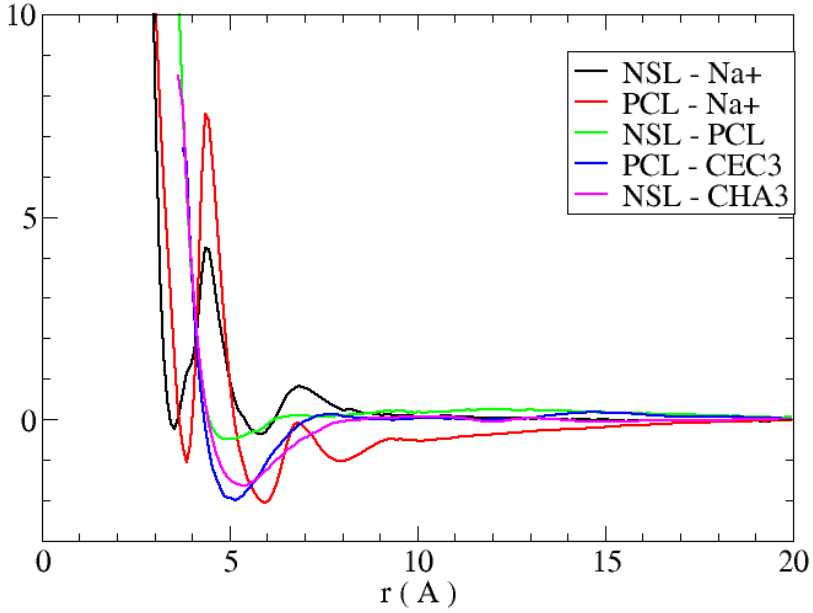
Atomistic MD: 30 lip1 + 30 lip2 + 1800 water mixture

DMPC - Cholesterol mixture





DOPS / DMPC mixture with ions



(H.Lopez et al, Advances in Experimental Medicine and Biology, vol 947, 173-206 (2017)).

MagiC Software

Any method can be used only if software is available...

MagiC: Software package implementing Inverse Monte Carlo and Iteraive Boltzmann inversion for calculation of effective potentials for coarse-grained models of arbitrary molecular systems

v 1.0 Released: 4 March 2013
A.Mirzoev, A.P.Lyubartsev, JCTC, 9, 1512 (2013)

Current stable version: 2.2 (July 2016)
Web site with download & documentation:
http://www.fos.su.se/~sasha/magic/

Developers web site with developer versions (2.3) and version archive:
https://bitbucket.org/magic-su/magic-2

The MagiC software

- Open source
- Freely available
- Documentation and examples/tutorials
- Easy to compile and install
- Can be used for arbitrary molecular system
- Python-based flexible
  pre-processing and post-processing
- Fortran-based kernel runs in parallel (MPI)

Other features:

Interactions:

- tabulated bond and angle potentials
- non-bonded tabulated potentials between different site types
- electrostatic outside cutoff by Ewald

Input (atomistic) trajectories:

- from Gromacs (pdb), NAMD, MDynaMix, xmol(xyz) format

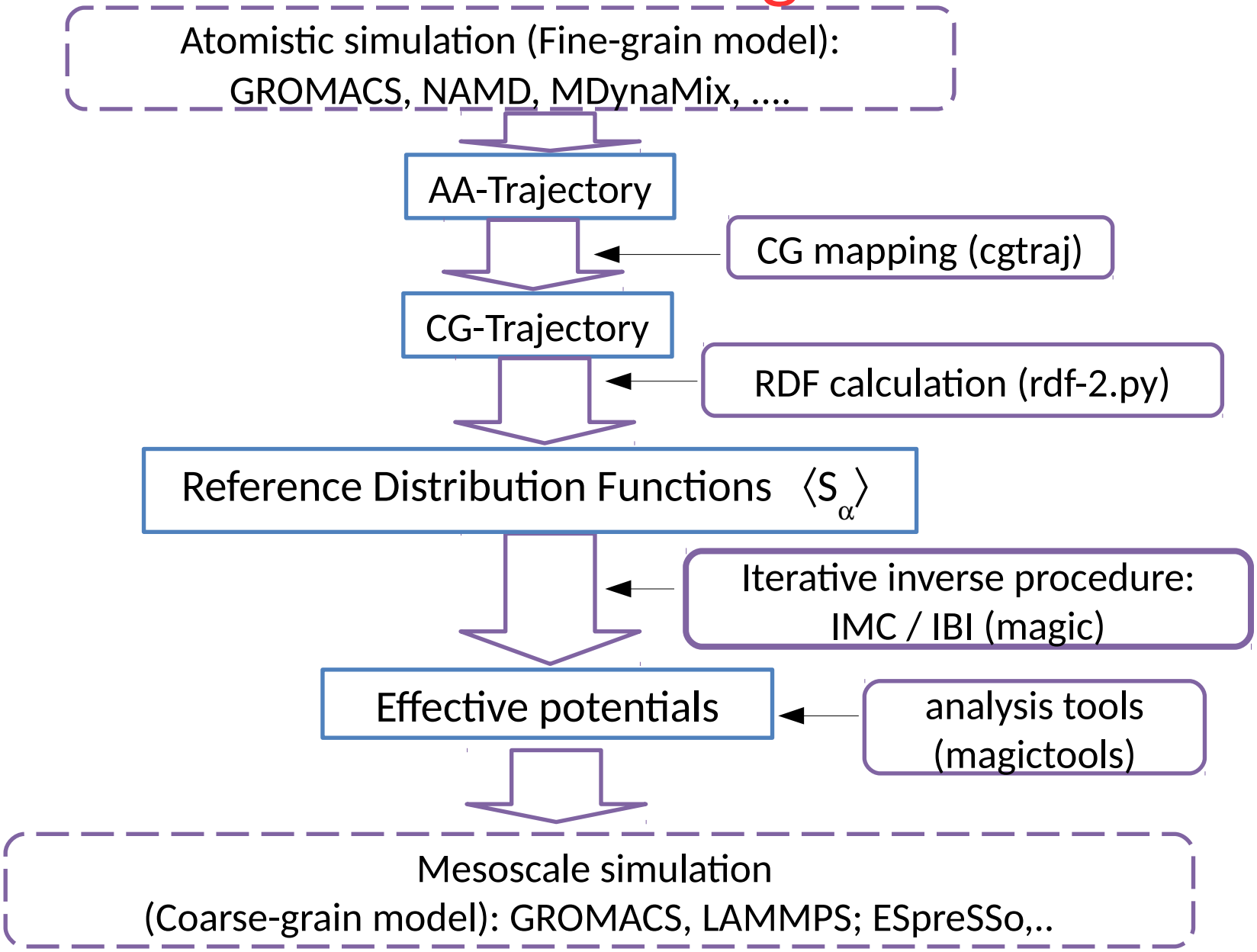
Methods:

- Iterative Boltzmann inversion and inverse Monte Carlo

Output: (tabulated potentials and CG topologies)

- Gromacs
- LAMMPS

Outline of the MagiC



Bead Mapping: cgtraj

Fortran-based utility inherited from MdynaMix (MD program)

Input: Atomistic trajectory (can be multiple files)

Current formats: MdynaMix; Gromacs – pdb; NAMD (.dcd); xmol (xyzz)

Mapping scheme: define which atoms in the atomistic model define each CG site

Output: CG trajectory (xmol format),
molecular description in terms of CG beads (.mmol file)

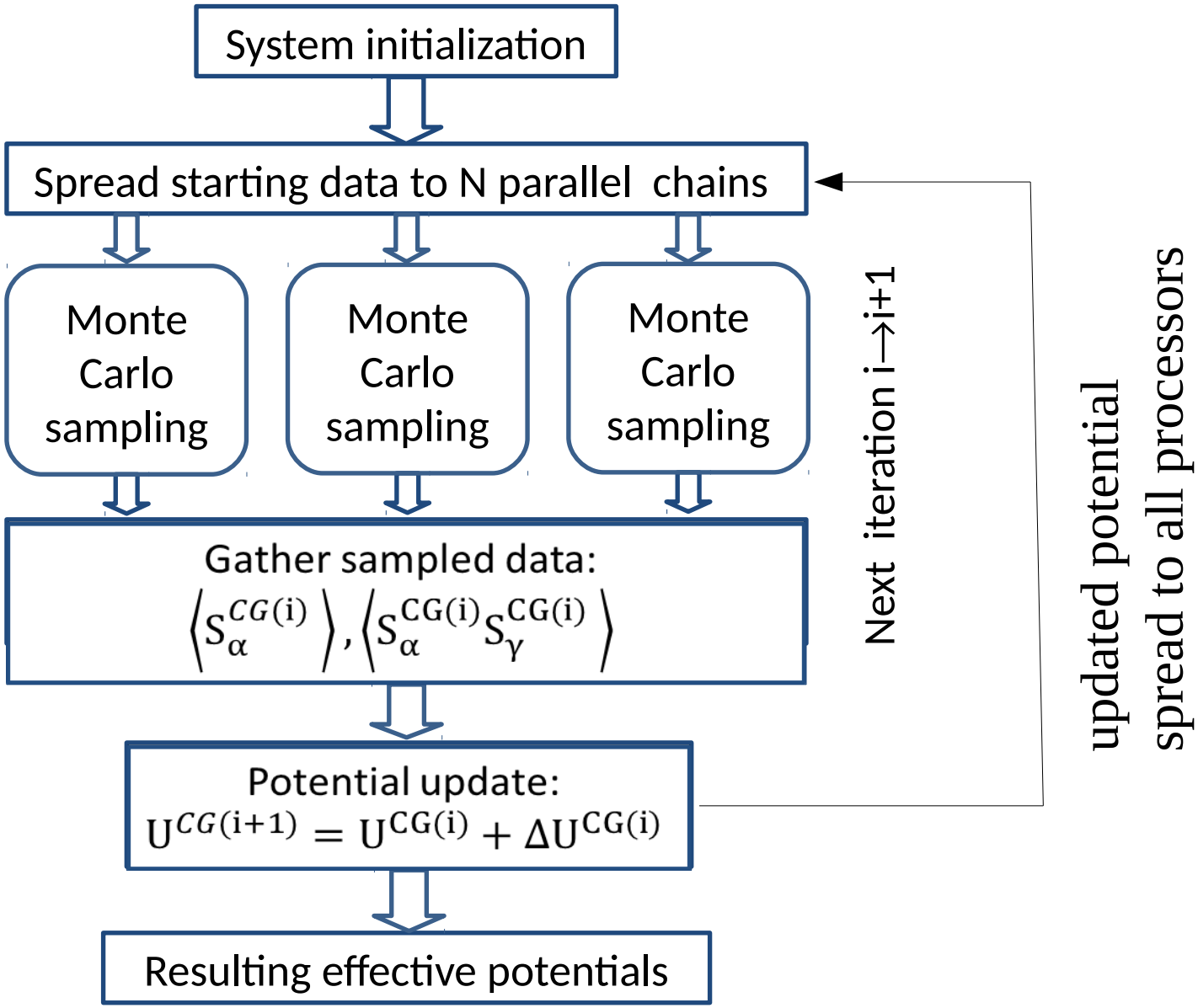
Step 2: Reference distribution functions

Input:
CG trajectory (from previous step)
Interaction definitions:
which CG sites have the same type?
Sites of same type have same interaction potential
Bonds: which CG sites are bound?
which bonds are of the same type?
Angles: angular interactions between CG sites. Can be generated automatically, but can be set (or adjusted) manualy

Output: RDF (including intramolecular bond and angle distributions),
CG molecular description files (.mcm)

Python-based script (rdf.py)

Step 3: Iterative inverse procedure (MagiC core)



Magic core: Inverse problem solver

Input: Parameter file given as

Keyword = value (s)

Defines:

- Simulated system (GC molecules, their number, box size,...)
- MC run parameters (number of steps, max step size, Ewald,…)
- Inverse solver parameters (IBI/IMC; regularization, num of iterations)
- output (what and how much you will see in the output)

Output: Effective Potentials (.pot), MC CG trajectory (.xmol), log output

MagicTools

- Load MagicTools

  > ipython

  > import MagicTools

- Inverse procedure convergence analysis:
  >MagicTools.Deviation('01.dmpc16-100a.new.out')
- RDF convergence plots:

  > MagicTools.AnalyzeIMCOuput('01.dmpc16-100a.new.out‘, iters=(1,5,7,9,10))

- Export potentials and topology to GROMACS

  > MagicTools.GromacsTopology('dmpc_NM.CG.mcm','dmpc.cg.top')

  > pots=MagicTools.GetPotsFromFile_pot('01.dmpc16-100a.i010.pot')

  > MagicTools.PotsExport2Gromacs(pots)

- Effective dielectric permittivity calculation
- Pressure correction

Useful hints:

- Start from zero non-bonded and Boltzmann-inverted bonded potentials
- In the beginning: small regularization parameter (< 0.1, sometimes even 0.02!), which can be increased to ~0.5 upon convergency
- Smaller number of MC steps in the beginning; larger number upon convergence
- Parallelization (many replicas) improves convergence greatly!
  Use as many processors as you can
- Do not run unattended! Control how the things are going

Messages to take home:

- Structure-based coarse-graining provides a working method to build interaction potentials for coarse-grained models
- Be aware on transferability problem: CG potentials are state point dependent (temperature, concentration..)
- Inverse MC allows to deal with complex CG systems consisting of multiple molecular types, bonds and angular interactions: mixed lipid systems, DNA – ions, etc

Acknowledgements

- Alexander Mirzoev
- Joakim Jämbeck
- Erik Brandt
- Dmitri Zhurkin
- Matus Rebic
- Aatto Laaksonen
- Yonglei Wang
- Lars Nordenskiöld
- Nikolai Korolev

- $$: Swedish Research Council (Vetenskapsrådet);

EU FP7 Programme ”MembreneNanoPart”, H2020 ”SmartNanoTox”

- CPU: Swedish National Infrastructure for Computing (SNIC)