Modelling of Solvation Thermodynamics by Using a Combination of Reference Interaction Site Model Theory and Multi-grid Numerical Methods

University of Strathclyde
Glasgow

Volodymyr Sergiievskyi

Department of Physics; Scottish Universities Physics Alliance

A thesis presented in fulfilment of the requirements for the degree of

Doctor of Philosophy

2012

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to award of a degree.

Signed:

Date:

To my wolverine

Abstract

Solvation Free Energy (SFE) is a fundamental quantity in chemical physics. It describes solvation behavior of substances in liquid media and has many important applications in solution chemistry, biophysics, pharmaceutics, medicine, and environmental sciences. In many applications (for example, screening of drug-candidate databases in a drug-discovery process) it is important to have a fast and accurate method for solvation free energy calculation. In this thesis two new methods for fast and accurate SFE calculations are proposed. The methods combine the theoretical basis of the integral equation theory of liquids with advanced computational techniques. The theoretical part of the methods is based on the Reference Interaction Site Model (RISM) and the three-dimensional RISM (3DRISM) molecular theories and semi-empirical models for SFE calculations. The computational part of the methods is based on the multi-grid scheme which drastically increases the computational performance. Additional investigations of speed and accuracy of calculations are performed to determine the optimal parameters of the methods which allow one to calculate the SFE with a required accuracy and minimal computational expenses. The methods are benchmarked on extended sets of small organic and drug-like compounds. It is shown that both (RISM and 3DRISM-based) methods can be successfully used for SFE calculations. It is shown that the parameters of the methods are transferable between different classes of compounds. The average computation time per typical drug-like compound of about 20 atoms is 17 seconds on a single CPU core for the RISM-based method and about 3.5 minutes for more accurate 3DRISM-based method.

Acknowledgements

I would like to acknowledge my supervisor, Prof. Maxim Fedorov for organizing of my PhD research. I am thankful to Prof. Gennady N. Chuev for organizing the course of lectures in Statistical Mechanics in the Max Planck Institute for Mathematics in the Sciences which I attended during my PhD study. These lectures helped my to understand the basics of the Integral Equation Theory of Liquids which is the theoretical base of my current research. I am deeply grateful to Dr. Andrey I. Frolov for fruitful discussions which helped me a lot to improve my understanding of physical chemistry and statistical physics. I would like to thank Dr. Ekaterina Ratkova for useful advices and consultations in Quantum Mechanical calculations. I am grateful to Wolfgang Hackbusch for insightful discussions on different aspects of multi-grid mathematical background. I would like to acknowledge Dr. David S. Palmer for productive collaboration.

I would like to acknowledge financial support of Max Planck Society, University of Strathclyde, Deutsche Forschungsgemeinschaft (DFG) - German Research Foundation, Research Grant FE 1156/2-1 and REA Research Executive Agency, Grant No. 247500 ”BioSol”, Programme: FP7-PEOPLE-2009-IRSES.

Contents

List of Figures

List of Tables

List of abbreviations

1D	one dimensional
3D	three dimensional
6D	six dimensional
B3LYP	Becke, three-parameter, Lee-Yang-Parr (method)
CHELPG	CHarges from Electrical Potential Grid-based method
DIIS	Direct Inverse in the itarative subspace (method)
DME	DiMethy Ether
FFT	Fast Fourier Transform
GB	Generalized-Born (equation)
GF	Gaussian Fluctuations (expression)
GMRES	Generalized Minimal RESidual (method)
HFE	Hydration Free Energy
HNC	Hyper Netted Chain (closure)
IETL	Integral Equation Theroy of Liquids
KH	Kovalenko-Hirata (closure)
MDIIS	Modified DIIS (method)
MD	Molecular Dynamics
MG	Multi-grid (method)
MOZ	Molecular Ornstein Zernike (equation)
MSPCE, MSPC/E	Modified SPC/E (water model)
NR	Newton-Raphson (method)
OPLS	Optimized Potentials for Liquid Simulations
OZ	Ornstein-Zernike (euqation)
PB	Poisson-Boltzmann (equation)
PMV	Partial Molar Volume
PW	Partial Wave (expression)
RISM	Reference Interaction Site Model
RMSD	Root Mean Square Deviation
SCF	Self-Consisten Field
SFE	Solvation Free Energy
SPCE, SPC/E	Simple Point Charge, Extended (water model)
UC	Universal Correction (method)

Chapter 1

Introduction

Eo quod in multa sapientia multa sit indignatio
et qui addit scientiam addat et laborem
Ecclesiastes 1:18

Accurate calculation of hydration free energies of organic molecules is a long-standing challenge in chemical physics and computational physical chemistry and is important in many aspects of research in the pharmaceutical and agrochemical industries. For example, many of the pharmacokinetic properties of potential drug molecules are defined by their *in vivo* solvation and acid-base behavior which can be estimated from their hydration free energies. [1–7]

Commonly used methods for hydration free energy calculation may be categorized as either explicit or implicit. In the first approach each solvent molecule is treated explicitly and molecular simulation methods are used to sample the conformational phase space [1,2,8–13]. In the second approach solvent is described as a continuum medium and the implicit effect of the solvent on a solute is determined by solving either the Poisson-Boltzmann (PB) or Generalized-Born (GB) equation [14–18]. While explicit solvent methods are scientifically more rigorous, the implicit models are often preferred because they are computationally less expensive. Still, explicit models can be successfully used in scientific investigations. However, in many cases they are too computationally expensive to be used in practical industrial applications. One of the applications where the speed of calculations is critical is the drug-discovery industry where one needs to filter large databases of drug-like candidates which can contain thousands and millions of compounds. For such kind of applications only implicit methods can be used. However, not all of the implicit methods are accurate enough. In some cases the errors of Solvation Free Energy (SFE) predictions can be of 2.5-3.5 kcal/mol which equates to a $\sim$ 2 log unit error in the related pharmacokinetic property (estimated from $\Delta G_{solv} = -RT \ln K$). Therefore, these methods are not accurate enough for many chemical applications [19–21]. Integral Equation Theory of Liquids (IETL) is an alternative framework for calculation of hydration free energies [22–24]. In terms of computational expenses IETL offers a compromise between computationally expensive fully-atomistic simulations [25–27] and rather approximate continuum electrostatic models [14, 17, 28]. Unlike PB or GB methods IETL retains information about the solvent structure (in terms of density correlation functions). However, unlike explicit methods IETL-based methods estimate SFE without long Molecular Dynamics (MD) or Monte Carlo (MC) simulations. At present there are several approaches based on the integral equations. The six-dimensional Molecular Ornstein–Zernike (MOZ) theory is used to calculate the three-dimensional (3D) hydration structure in molecular liquids [29,30]. The site–site Ornstein–Zernike (SSOZ) integral equation is used to calculate the properties of complex solute-solvent systems in the Reference Interaction Site Model (RISM) formalism developed by Chandler, Anderson and others [31–33]. The theory has been applied successfully to calculation of the structural and thermodynamical properties of various chemical and biological systems [34–49].

To solve the OZ equation one should complete it by a *closure* relation which makes the system OZ equation plus closure solvable. However, the closure relation incorporates a so-called *bridge* function which is practically incomputable due to the infinite number of terms in the exact representation of this function [50,51]. Therefore, in practice one uses approximate closure relations [24, 52]. Nowadays only few efforts were done to solve the six-dimensional molecular OZ equation numerically. This can be explained by a high computational complexity of the problem. Using the modern computers straightforward calculations on the six-dimensional grid with a moderate resolution are already feasible [53]. However, such computations are still extremely computationally demanding. Other class of methods uses low-rank decomposition of the correlation functions. In that case translational and rotational components of the functions are separated and some basis functions are used for representation of the rotational degrees of freedom. Typically, for the rotational degrees of freedom the basis set of rotational invariants is used [54, 55]. In some cases some advanced techniques, like hierarchical matrix decomposition, can be used to reduce the computational complexity of the method [56]. For an additional discussion of the computational complexity of operations in different low-rank formats one may refer to Appendix A of this thesis.

Despite the recent progress in the six-dimensional MOZ theory currently it was tested only for small and simple molecules. This is explained by a high computational cost of numerical solution of the six-dimensional problem with reasonable accuracy. In practical applications one usually uses simplified models, and the Reference Interaction Sites Model (RISM) is one of the most popular among them [31–33]. The main assumption of the RISM is that the molecular correlation functions can be represented as a sum of spherically symmetric functions. The operations with the spherically symmetric functions can be reduced to the operations with only their radial parts and that makes the RISM integral equations quasi one-dimensional.

The six-dimensional MOZ equation is replaced by a set of one-dimensional non-linear integral equations. From a computational point of view it is relatively inexpensive to solve the RISM equations numerically for small molecular solutes ($< 10^2$ atoms). Typically solutions of the RISM equations give a qualitatively correct description of the solvent structure around the solute. RISM-based methods have many important applications. The RISM equations can be self-consistently used to introduce an implicit solvent model in quantum mechanical calculations (RISM-SCF method) [57–61]. RISM theory gives end-point expressions for solvation free energy calculation [52, 62]. We note though that the original formulae for SFE calculations [52, 62] provide only qualitative predictions of trends in the differences of SFEs for different compounds [63]. Recently there were proposed several semi-empirical methods for parameterization of the RISM SFE expressions, such as Structural Descriptor Correction(SDC) method [64, 65], Atom Type Correction (ATC) method [66, 67], and others [41, 63]. The best of these methods are able to predict SFEs of different polyfragment organic molecules with the accuracy of around 1 kcal/mol [64, 65]. However, RISM-based semi-empirical methods need to use a considerable number of empirical parameters to achieve a good accuracy of SFE calculations.

Another approximation of the Ornstein-Zernike equation is the three-dimensional RISM (3DRISM) [68, 69] where a solute molecule is represented as a three dimensional object. This model provides better spatial description of solute-solvent correlations than the RISM. The 3DRISM-based methods are currently widely used in biochemical applications for description of solvation properties of biomolecules [46, 70–72]. As it has been recently shown, the 3DRISM-based Universal Correction (UC) method accurately predicts thermodynamic parameters of hydrated organic molecules including drug-like molecules using only two empirical parameters [73, 74]. However, for small molecules, a numerical solution of the multidimensional 3DRISM equations requires significantly more computational time than a solution of the RISM equations [74]. High computational expenses of the 3DRISM calculations are a real bottleneck of this method that inhibit wider applications of this technique.

Coming back to the history of the IETL the first algorithm used for solving OZ-like integral equations was presumably the Picard iteration method [22]. This method is easy to implement. However, it has a comparably low convergence rate. One may use faster convergent schemes, such as the Newton-Raphson (NR) iteration [75], the NR-GMRES algorithm [76], the method of direct inversion in iterative subspace (DIIS) [77], the combination of the modified NR and the DIIS iteration [78] or the vector extrapolation technique [79] . For the 3DRISM equations it was recently proposed to use the Modified DIIS (MDIIS) method [77]. Recently an efficient 3DRISM equations solver which uses the MDIIS algorithm was implemented in the Amber molecular modeling software [71]. However, for the grids with a large number of points and/or molecular systems with a large number of interacting sites these methods are computationally expensive. An alternative way to increase efficiency is to use a multi-scale approach. The commonly used approach is the combined two-level NR-Picard scheme, so-called Gillan method [80]. Similar two-level NR-Picard schemes are used also in the Labík-Malijevský-Voňka method [81,82] or in the wavelet-based methods [56,83–85]. Although two-level methods give an essential improvement with respect to the Picard iteration, two level approaches have limitations because the resolutions on the fine-grid and on the coarse-grid cannot differ too much from each other. Multi-level methods can be used to overcome these limitations. Recently an effective multilevel NR-GMRES algorithm has been applied to the one-dimensional problem [86]. This algorithm does not have such restrictions as the two-level methods; NR-GMRES algorithm is a fast convergent method and it needs much less iteration steps to converge than the Picard iteration method. However, faster convergence in terms of number of iteration steps does not necessarily mean better performance in terms of computer time. The Newton-Raphson method requires computation and inversion of the Jacobian matrix of size $N \times N$, where N is a number of discretization points on the coarse grid (typically $N \approx 10^1 - 10^2$). Although the implementation described in Ref. [86] does not require inversion of a Jacobian matrix it nevertheless requires operations with matrices of size $N \times N$ which makes each iteration step computationally expensive. For the RISM, where the number of coarse-grid points N is not very large these additional computational expenses for each iteration step can be compensated with much smaller number of iteration steps. However, it is not so for the three dimensional problem where the number of grid points grows cubically with respect to the one-dimensional problem (the matrix size could be as large as $10^6 \times 10^6$.

Despite the fact that many methods use sophisticated algorithms to enhance computational performance often these methods do not fully exploit the advantages of the multi-scale approach. The *multi-grid* method is a multi-scale technique in a sense that the iteration makes use of different grids (different discretization levels). The multi-grid approach uses an advanced iteration scheme which makes it more efficient than the simple multi-scale iteration methods [87]. The multi-grid scheme is not restricted to a specific type of iterations and can be applied to any kind of iteration process. Multi-grid methods are actively used in different applications in computational chemistry [88–92]. In our work we use the multi-grid approach for speeding-up the calculations. We propose new methods for solving the RISM and the 3DRISM problems based on the multi-grid scheme. We are focused on the practical application of the methods to the solvation free energy calculations. We perform additional investigations to determine some guidelines for choosing algorithms' parameters which are optimal for the solvation free energy calculations. We test the numerical performance of the proposed RISM multi-grid method and compare this method to the one-grid Picard iteration and to the nested Picard iteration methods. In our work we investigate the performance of two modifications of the 3DRISM multi-grid algorithm where the multi-grid is combined with (i) the Picard iteration method (MG-Picard); and (ii) with the MDIIS method (MG-MDIIS) respectively. By benchmarking of these methods on a set of model compounds we determine the optimal grid parameters for solvation (hydration) free energy calculations. We test the numerical performance of the proposed 3DRISM methods and compare these methods to the standard Picard iteration method and to the MDIIS method. To test the effectiveness of the proposed methods we benchmark the speed and accuracy of the methods on the extended sets of organic compounds. To test the effectiveness of the RISM multi-grid method we use the set of 63 bioactive drug-like compounds from Ref. [93]. In our work we perform the RISM calculations for 63 compounds from Ref. [93], discuss efficiency of different RISM-SFE expressions and perform a parameterization which allows one to improve computational results. The set of 99 organic compounds from the paper [74] was chosen for testing of the 3DRISM multi-gird methods. The 3DRISM calculations for these molecules were performed with the MG-MDIIS algorithm. The SFEs of 99 molecules from the set were calculated with the Universal Correction (UC) method and compared to the experimental results.

1.1 Goals

The main goal of this thesis is development of fast and accurate methods for Solvation Free Energy (SFE) calculations and make them reliable for molecular biophysics and medicine. This goal requires the solution of several computational and theoretical problems. On the one hand, in many cases it is critical to have fast methods for SFE calculation. On the other hand, the method should provide a reasonable accuracy of the SFE calculations to be useful for practical applications. To achieve a high accuracy of the calculations the RISM and the 3DRISM molecular theories in combination with the semi-empirical SFE calculation methods are used in this work. The multi-grid technique is used to speed-up the RISM and the 3DRISM calculations. The tasks of the current work are:

1. Development of fast multi-grid methods for solving of the RISM and the 3DRISM numerical problems.

2. Investigation of the accuracy of different semi-empirical SFE calculation methods for chemical compounds from different chemical classes, including polyfragment drug-like molecules.

3. Investigation of the computational errors and computational performance of the methods as well as determination of the optimal parameters for fast and accurate calculations.

1.2 Structure of the thesis

The thesis consists of 7 chapters and one appendix.

The first chapter is introduction.

In Chapter 2 some basic concepts of the statistical mechanics and thermodynamics are described.

Chapter 3 contains description of the theoretical background of the integral equation theory of liquids, Reference Interaction Site Model(RISM) and three-dimensional RISM (3DRISM).

In Chapter 4 the ways to calculate the solvation free energy in the RISM and the 3DRISM approximations are discussed.

In Chapters 5 and 6 the RISM and the 3DRISM multi-grid methods for calculation of the solvation free energy are described. Descriptions of the both methods have the same structure which includes three parts: (i) description of the numerical method; (ii) determination of the optimal parameters of the method; (iii) benchmarking of the speed and accuracy of the method on a set of organic compounds.

In Chapter 7 the perspectives of the theory are discussed and some preliminary results of the ongoing research are described.

In Appendix A the low-rank format for efficient operations with multi-dimensional functions is described.

Chapter 2

Statistical Mechanical background

In this chapter the basic concepts in statistical mechanics, such as ensemble average, partition function, free energy etc are described. The chapter is mostly based on Refs. [94] and [95].

2.1 Systems under investigation

One of the main tasks of the statistical mechanics is description of common laws of the many particle systems. Let there be N particles in the system, and each of the particles have m degrees of freedom. Then the total number of degrees of freedom of the system is $s = m \cdot N$ According to the Hamilton's equation the system which has s degrees of freedom can be described by the s generalized coordinates $(q_1(t), \ldots, q_s(t))$ and s generalized momenta components $(p_1(t), \ldots, p_s(t))$. For such a system the Hamiltonian equations hold [96]:

$$\dot{q}_i = \frac{\partial \mathcal{H}}{\partial p_i} \qquad \dot{p}_i = -\frac{\partial \mathcal{H}}{\partial q_i} \qquad i = 1 \ldots s \tag{2.1}$$

where $\dot{p}_i = \partial p_i / \partial t$, $\dot{q}_i = \partial q_i / \partial t$, $\mathcal{H} = \mathcal{H}(p_1, \ldots, p_s, q_1, \ldots, q_s)$ is the Hamiltonian of the system. If the Hamiltonian $\mathcal{H}$ is known the equations (2.1) can be solved numerically for any initial conditions $p_i(t_0) = p_i^0, q_i(t_0) = q_i^0, i = 1 \ldots s$ which allows to predict the state of the system at any moment t. Using this information it is possible to calculate the quantities of interest: temperature, pressure, density, residence time etc. This approach is the base for Molecular Dynamics (MD) simulations. Despite of simplicity and universality of this method MD simulations require comparably large computational resources. Today the typical size of simulated systems is only about 1000-10000 molecules and typical simulation time is of 100 ns. The limit which can be achieved with the modern computational resources lies at 10^9-10^{11} molecules simulated for several milliseconds [97–99].

2.2 Phase space. Ensemble. Micro-canonical ensemble

In contrast to the simulation methods statistical physics gives the possibility to find the physical-chemical quantities of interest without considering the movement of a single particle. Typically in reality the movement of the particles in the systems is quasi-random. This means, that the probability to find the particle in some point does not depend on the initial conditions and during the large period of time the system will reach any of the possible states. These assumptions allow us to assume that all the initial states are equivalently probable. The movement of N particles with generalized coordinates $q_1, \dots, q_s$ and generalized momenta $p_1, \dots, p_s$ can be equivalently described by the movement of the point with coordinates $(p_1, \dots, p_s, q_1, \dots, q_s)$ in $2s$-dimensional space. However, the system cannot reach all the points in $2s$ dimensional space due to some restrictions. Such restrictions can be for example constant volume of the system, constant pressure, temperature etc. The set of imaginary copies of the system which represent all possible states of the system under certain restrictions is called *ensemble*. One of the simplest examples of restrictions which can exist is the energy conservation law. The ensemble which contains the fixed number of particles where the energy conservation law holds is called the *microcanonical ensemble*. Each state of the system corresponds to some point in a $2s$ dimensional *phase space* of the system. We introduce the *distribution function* $f(p_1, \dots, p_s, q_1, \dots, q_s, t)$ such, that $f(p_1, \dots, p_s, q_1, \dots, q_s, t)dp_1 \dots dp_s dq_1 \dots dq_s dt$ is the probability to find the system during the infinitesimal time interval $[t; t + dt]$ in the infinitesimal parallelepiped of size $dp_1 \times \dots \times dp_s dq_1 \times \dots \times dq_s$ in the vicinity of the point $(p_1, \dots, p_s, q_1, \dots, q_s)$. We note that because the distribution function represents a probability it satisfies the following *normalization condition*:

$$\int f(p_1, \dots, p_s, q_1, \dots, q_s, t)dp_1 \dots dp_s dq_1 \dots dq_s = 1 \tag{2.2}$$

where integral is taken over all the possible states at moment t. Note that although we consider the system of classical particles, due to the uncertainty principle there is a so small elements of phase space that we are not able to distinguish different points in it. For each degree of freedom it holds that $dp_i dq_i \geq 2\pi\hbar$ [95]. Thus for s degrees of freedom the elementary volume in the phase space is $\geq (2\pi\hbar)^s$. Also, the uncertainty principle implies that for any finite system there is only finite number of distinguishable states. For each element of volume $\Delta p_1 \times \dots \times \Delta p_s \times \Delta q_1 \times \dots \times \Delta q_s$ the maximum number of distinguishable states is $\Delta p_1 \dots \Delta p_s \Delta q_1 \dots \Delta q_s)/(2\pi\hbar)^s$.

In most of the cases in the physical chemistry the goal of the investigation of the system is to determine some average physical quantity which describe the system (e.g. temperature, mean energy, density etc). Depending on approach that we use the algorithm for the calculation of these averages is different. Let $X(p_1, \dots, p_s, q_1, \dots, q_s, t)$ be some physical value which depends on the phase coordinates of the system. If we perform the MD simulations the trajectory of the system in the phase space is known. That means we know dependencies $q_i(t), p_i(t), \quad i = 1, \dots, s$. In that case we are able to calculate *time average* $\bar{X}$ of the value X by the following formula [50]:

$$\bar{X} = \lim_{T\to\infty} \frac{1}{T} \int\limits_0^T X(p_1(t), \dots, p_s(t), q_1(t), \dots, q_s(t), t) dt \tag{2.3}$$

In case of the thermodynamic description of the system we do not have the information about the trajectories of the particles. Instead of this we consider the density distribution function $f(p_1, \dots, p_s, q_1, \dots, q_s, t)$ which define the probability to find the system in the state $(p_1, \dots, p_s, q_1, \dots, q_s)$ at the moment t. Then instead of the time average we use the *ensemble average*. To do this we find the expected value $< X >$ of the physical value X:

$$< X(t) >= \int X(p_1, \dots, p_s, q_1, \dots, q_s, t) f(p_1, \dots, p_s, q_1, \dots, q_s, t) dp_1 \dots dp_s dq_1 \dots dq_s \tag{2.4}$$

where the integration is performed over all distinguishable states in the phase space. If both: the physical value X and the distribution function f are independent of time, the integration over time can be omitted and the ensemble average (2.4) can be rewritten as following:

$$< X >= \int X(p_1, \dots, p_s, q_1, \dots, q_s) f(p_1, \dots, p_s, q_1, \dots, q_s) dp_1 \dots dp_s dq_1 \dots dq_s \tag{2.5}$$

According to the basic assumptions of the statistical physics the ensemble average (2.5) is equivalent to the time average (2.3)

2.3 Continuity equation

Let's consider the motion of different points in the phase space. Although formally there is infinite number of points in the phase space it was discussed above that due to the uncertainty principle we should consider only finite number of them. Let $V = \int dp_1 \dots dp_s dq_1 \dots dq_s$ be the volume of the phase space. Then the maximum number of distinguishable points is $M = V/(2\pi\hbar)^s$. So, let us consider M points in the phase space which at the initial moment t_0 are distributed according to the distribution function f, which means that in the phase volume element of size $\Delta V = \Delta p_1 \times \dots \times \Delta p_s \times \Delta q_1 \times \dots \times \Delta q_s$ near the phase point $(p_1, \dots, p_s, q_1, \dots, q_s)$ there are (approximately) $M \cdot f(p_1, \dots, p_s, q_1, \dots, q_s, t_0)\Delta V$ points. Equation (2.1) uniquely defines trajectories of these points. The points cannot disappear, the new points cannot appear in time, so the total number of points is all the time constant. For the sake of uniformity we introduce the new coordinates $(x_1, \dots, x_{2s})$ in the following way:

$$x_i = q_i \qquad x_{i+s} = p_i \qquad i = 1, \dots, s \tag{2.6}$$

Then the density distribution function f can be written as a function of $(x_1, \ldots, x_{2s})$:

$$f(x_1, \ldots, x_{2s}, t) \equiv f(p_1, \ldots, p_s, q_1, \ldots, q_s, t) \tag{2.7}$$

According to the Hamiltonian's equations (2.1) the time derivatives (velocities) of the coordinates $(x_1, \ldots, x_{2s})$ are known at each point of the phase space at each moment of time:

$$\begin{aligned} \dot{x}_i(x_1, \ldots, x_{2s}, t) \equiv \dot{p}_i &= -\frac{\partial \mathcal{H}(x_1, \ldots, x_{2s})}{\partial x_{i+s}} \\ \dot{x}_{i+s}(x_1, \ldots, x_{2s}, t) \equiv \dot{q}_i &= \frac{\partial \mathcal{H}(x_1, \ldots, x_{2s})}{\partial x_i} \end{aligned} \tag{2.8}$$

where $\dot{x}_j \equiv \partial x_j / \partial t$, $j = 1 \ldots 2s$.

Let us consider a small parallelepiped in the $2s$-dimensional space with the center at $(x_1^0, \ldots, x_{2s}^0)$ and the volume of $\Delta x_1 \times \cdots \times \Delta x_{2s}$ with the edges parallel to the coordinate axes. Because the total number of points in the system is constant, the number of points inside the parallelepiped changes only due to the particles' flow through the faces of the parallelepiped. The parallelepiped in $2s$ dimensional space has $4s$ faces (two faces in each direction). Each face of the parallelepiped is a $2s-1$ dimensional set of points which is obtained by fixing one of the coordinates. Let us define by F_i^+ and F_i^- two faces in the direction x_i, namely:

$$\begin{aligned} F_i^- &= \{(x_1, \ldots, x_{i-1}, x_i^0 - \frac{\Delta x_i}{2}, x_{i+1}, \ldots, x_{2s}) : x_j^0 - \frac{\Delta x_j}{2} \leq x_j \leq x_j^0 + \frac{\Delta x_j}{2} \text{where} j \neq i\} \\ F_i^+ &= \{(x_1, \ldots, x_{i-1}, x_i^0 + \frac{\Delta x_i}{2}, x_{i+1}, \ldots, x_{2s}) : x_j^0 - \frac{\Delta x_j}{2} \leq x_j \leq x_j^0 + \frac{\Delta x_j}{2} \text{where} j \neq i\} \end{aligned} \tag{2.9}$$

Let us consider motion of the system at moment t during so small period of time Δt that all the velocities of phase points $\dot{x}_i$ and the distribution function $f(x_1, \ldots, x_{2s}, t)$ do not change much. We define by $n(F_i^\pm, t)$ the number of particles which flows through the face $F_i^\pm$ during the time interval $[t, t + \Delta t]$. The density of the phase points near the face is defined by the distribution function f, the velocity of the particles in the direction x_i is $\dot{x}_i$. Thus the number of particles which flows through the face $F_i^\pm$ can be calculated by integrating over the face the product of density multiplied by velocity:

$$\begin{aligned} &n(F_i^\pm, t) = \\ &\Delta t \int_{F_i^\pm} f(x_1, \ldots, x_i^0 \pm \tfrac{\Delta x_i}{2}, \ldots x_{2s}, t) \cdot \dot{x}_i(x_1, \ldots, x_i^0 \pm \tfrac{\Delta x_i}{2}, \ldots x_{2s}, t) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_{2s} \end{aligned} \tag{2.10}$$

If the parallelepiped is small enough the integral in (2.10) can be approximated by the product:

$$\begin{aligned} &n(F_i^\pm, t) \approx \\ &f(x_1^0, \ldots, x_i^0 \pm \tfrac{\Delta x_i}{2}, \ldots x_{2s}^0, t) \dot{x}_i(x_1^0, \ldots, x_i^0 \pm \tfrac{\Delta x_i}{2}, \ldots x_{2s}^0, t) \Delta x_1 \ldots \Delta x_{i-1} \Delta x_{i+1} \ldots \Delta x_{2s} \Delta t \end{aligned} \tag{2.11}$$

To make notations shorter we use the following definition:

$$w_i(x_1,\dots,x_{2s},t) \equiv f(x_1,\dots,x_{2s},t)\dot{x}_i(x_1,\dots,x_{2s},t) \tag{2.12}$$

Also, let us define $\Delta x^{2s} \equiv \Delta x_1 \dots \Delta x_{2s}$. Then $\Delta x_1 \dots \Delta x_{i-1}\Delta x_{i+1}\dots\Delta x_{2s} \equiv \Delta x^{2s}/\Delta x_i$ In this notations the expression (2.11) is written as following:

$$n(F_i^{\pm},t) \approx w_i(x_1^0,\dots,x_i^0 \pm \frac{\Delta x_i}{2},\dots x_{2s}^0)\frac{\Delta x^{2s}}{\Delta x_i}\Delta t \tag{2.13}$$

To find the number of particles $\Delta n_i(t)$ which flows in the direction x_i and left in the parallelepiped we need to subtract from the number of particles which comes through the face F_i^- the number of particles which flows out through the face F_i^+. Thus we have:

$$\begin{aligned}&\Delta n_i(t) = \\ &n(F_i^-,t) - n(F_i^+,t) \approx \\ &\approx \frac{\left(w_i(x_1^0,\dots,x_i^0 - \frac{\Delta x_i}{2},\dots x_{2s}^0) - w_i(x_1^0,\dots,x_i^0 + \frac{\Delta x_i}{2},\dots x_{2s}^0)\right)\Delta t \Delta x^{2s}}{\Delta x_i}\end{aligned} \tag{2.14}$$

The total change of the number of particles Δn during the time Δt in all directions is the sum of changes in each of directions. We can write this in the following way:

$$\begin{aligned}&\Delta n = \\ &\Delta t \sum_{i=1}^{2s} \frac{\Delta x^N}{\Delta x_i}\left(w_i(x_1^0,\dots,x_i^0 - \Delta x_i/2,\dots,x_{2s},t - w_i(x_1^0,\dots,x_i^0 + \Delta x_i/2,\dots,x_{2s},t)\right)\end{aligned} \tag{2.15}$$

On the other hand, the change of the number of particles is the change of mean density multiplied by the volume of the parallelepiped. If the parallelepiped is small enough we can approximate the mean density with the density in the center of the parallelepiped. Thus we have:

$$\Delta n = \Delta f \cdot \Delta V = \left(f(x_1^0,\dots,x_{2s}^0,t+\Delta t) - f(x_1^0,\dots,x_{2s}^0,t)\right)\Delta x_1 \dots \Delta x_{2s} \tag{2.16}$$

We divide (2.15) and (2.16) by $\Delta t \Delta x_1 \dots \Delta x_{2s}$ and equate them. We obtain the following relation:

$$\begin{aligned}&\frac{f(x_1^0,\dots,x_{2s}^0,t+\Delta t) - f(x_1^0,\dots,x_{2s}^0,t)}{\Delta t} = \\ &\sum_{i=1}^{2s} \frac{w_i(x_1^0,\dots,x_i^0 - \Delta x_i/2,\dots,x_{2s}^0) - w_i(x_1^0,\dots,x_i^0 + \Delta x_i/2,\dots,x_{2s}^0)}{\Delta x_i}\end{aligned} \tag{2.17}$$

Taking the limit $\Delta x_i \to 0, \Delta t \to 0$ we have the definitions of the derivatives in both sides:

$$\frac{\partial f}{\partial t} = -\sum_{i=1}^{2s} \frac{\partial w_i}{\partial x_i} \tag{2.18}$$

Because $w \equiv f \cdot \dot{x}_i$, we may write the following relation:

$$\frac{\partial f}{\partial t} + \sum_{i=0}^{2s} \frac{\partial (f \cdot \dot{x}_i)}{\partial x_i} = 0 \tag{2.19}$$

This equation is called the *continuity equation*.

2.4 Liouville equation

Using the formulae (2.6) we can write the continuity equation (2.19) in the phase coordinates p_i, q_i. In that case, the continuity equation (2.19) is written as

$$\frac{\partial f}{\partial t} + \sum_{i=1}^{s} \left(\frac{\partial (f \cdot \dot{q}_i)}{\partial q_i} + \frac{\partial (f \cdot \dot{p}_i)}{\partial p_i} \right) = 0 \tag{2.20}$$

Opening the brackets in derivatives, we have

$$\frac{\partial f}{\partial t} + \sum_{i=1}^{s} \left(\frac{\partial f}{\partial q_i} \dot{q}_i + \frac{\partial f}{\partial p_i} \dot{p}_i \right) + \sum_{i=1}^{s} f \cdot \left(\frac{\partial \dot{q}_i}{\partial q_i} + \frac{\partial \dot{p}_i}{\partial p_i} \right) = 0 \tag{2.21}$$

From the Hamiltonian equations (2.1) we have:

$$\frac{\partial \dot{p}_i}{\partial p_i} = -\frac{\partial^2 \mathcal{H}}{\partial q_i \partial p_i}$$

$$\frac{\partial \dot{q}_i}{\partial q_i} = \frac{\partial^2 \mathcal{H}}{\partial p_i \partial q_i}$$

This means, that second sum in (2.21) is canceled, and equation reads as:

$$\frac{\partial f}{\partial t} + \sum_{i=1}^{s} \left(\frac{\partial f}{\partial q_i} \dot{q}_i + \frac{\partial f}{\partial p_i} \dot{p}_i \right) = 0 \tag{2.22}$$

One can also notice that the left hand side of equation (2.22) is the full derivative of the phase density f. Thus, the equation can be rewritten in a more compact way:

$$\frac{df}{dt} = 0 \tag{2.23}$$

There is an important corollary of the Liouville equation (2.23). Let us consider a closed system in the equilibrium state. This means that the distribution function f of this system does not explicitly depend on time:

$$f(p_1, \ldots, p_s, q_1, \ldots, q_s, t) \equiv f(p_1, \ldots, p_s, q_1, \ldots, q_s) \tag{2.24}$$

By integrating the Liouville equation (2.23) we have:

$$\int\limits_{t_0}^{t_1} \frac{df}{dt}dt = f(p_1(t_1), \ldots, p_s(t_1), q_1(t_1), \ldots, q_s(t_1)) - f(p_1(t_0), \ldots, p_s(t_0), q_1(t_0), \ldots, q_s(t_0)) = 0 \tag{2.25}$$

Under considerations of the statistical mechanics the probability to find the system in certain state does not depend on the initial configuration. Thus relation (2.25) means that distribution function is a constant of motion of the system:

$$f(p_1, \ldots, p_s, q_1, \ldots, q_s) = f_0 = \text{const} \tag{2.26}$$

This in turn means that all the accessible microstates of the system are equiprobable. Although the above conclusions were obtained for the microcanonical ensemble where the energy conservation and Hamiltonian equations (2.1) hold the same in principle could be true for some other systems as well. Let in the system itself the energy conservation law does not hold. As it was mentioned above due to the uncertainty principle the finite system can have only finite number of distinguishable states and thus only the finite number of different energies. Let $\epsilon_1, \epsilon_2, ..., \epsilon_m$ be all the possible energies of the system. Let A_i be the set of all possible states with the energy ϵ_i. We will call these sets *energy levels* of the system. One of the basic assumptions of the statistical physics is that the probability to find the particle in certain state does not depend on the initial state and on the particular trajectories of the particles. Thus, we can assume that the system stays at each energy level for relatively long time. In that time the energy of the system is constant and thus there exists such Hamiltonian $\mathcal{H}$ that satisfies Hamiltonian equations (2.1), and thus all the accessible states with the same energy are equiprobable and the probability to find the system in a certain state depends only on the energy of this state.

2.5 Gibbs distribution

Let us consider a closed system which is composed of a body in environment. Let the body has certain finite volume and contain a fixed number of particles (the body does not exchange the particles with the environment) and is in dynamical energetic equilibrium with the environment. Let $(p_1^1, \ldots, p_{s_1}^1, q_1^1, \ldots, q_{s_1}^1)$ be the phase coordinates of the body and $(p_1^2, \ldots, p_{s_2}^2, q_1^2, \ldots, q_{s_2}^2)$ be the phase coordinates of the environment. Possible states of the body form the *canonical ensemble*. For the sake of clarity below we use the following notations:

$$\begin{array}{ll} p^{[1]} = (p_1^1, \ldots, p_{s_1}^1) & q^{[1]} = (q_1^1, \ldots, q_{s_1}^1) \\ p^{[2]} = (p_1^2, \ldots, p_{s_2}^2) & q^{[2]} = (q_1^2, \ldots, q_{s_2}^2) \\ dp^{[1]} = dp_1^1 \ldots dp_{s_1}^1 & dq^{[1]} = dq_1^1 \ldots dq_{s_1}^1 \\ dp^{[2]} = dp_1^2 \ldots dp_{s_2}^2 & dq^{[1]} = dq_1^2 \ldots dq_{s_2}^2 \end{array} \tag{2.27}$$

Let $\mathcal{H}(p^{[1]}, q^{[1]}, p^{[2]}, q^{[2]})$ be the Hamiltonian of the system and the probability to find the system in some state is described by the distribution function $f_{12}(p^{[1]}, p^{[2]}, q^{[1]}, q^{[2]})$. Let us assume that the distribution function of the body $f_1(p^{[1]}, q^{[1]})$ is independent from distribution function of the environment $f_2(p^{[2]}, q^{[2]})$. In that case the following relation holds:

$$f_{12}(p^{[1]}, p^{[2]}, q^{[1]}, q^{[2]}) = f_1(p^{[1]}, q^{[1]}) \cdot f_2(p^{[2]}, q^{[2]}) \tag{2.28}$$

This assumption is reasonable for big systems. Although the total energy of the system "the body + environment" is constant the energy of its components could vary. We introduce the function $E(p^{[1]}, q^{[1]})$ which describes the average energy of the system when the body is in the state $(p^{[1]}, q^{[1]})$. The function $E(p^{[1]}, q^{[1]})$ can be calculated by averaging of the Hamiltonian of the system over the phase coordinates of the environment:

$$E(p^{[1]}, q^{[1]}) = \int \mathcal{H}(p^{[1]}, p^{[2]}, q^{[1]}, q^{[2]}) f_2(p^{[2]}, q^{[2]}) dp^{[2]} dq^{[2]} \tag{2.29}$$

Multiplying (2.29) by $f_1(p^{[1]}, q^{[1]})$ and integrating over phase coordinates of the body we have the following:

$$\int E(p^{[1]}, q^{[1]}) f_1(p^{[1]}, q^{[1]}) dp^{[1]} dq^{[1]} = \int \mathcal{H}(p^{[1]}, p^{[2]}, q^{[1]}, q^{[2]}) f_1(p^{[1]}, q^{[1]}) f_2(p^{[2]}, q^{[2]}) dp^{[2]} dq^{[2]} dp^{[1]} dq^{[1]} \tag{2.30}$$

Let $E_{tot} = \mathcal{H}(p^{[1]}, p^{[2]}, q^{[1]}, q^{[2]}) = \text{const}$ be the total energy of the system. Assuming that $f_{12} = f_1 \cdot f_2$ we have the following:

$$\int E(p^{[1]}, q^{[1]}) f_1(p^{[1]}, q^{[1]}) dp^{[1]} dq^{[1]} = E_{tot} \tag{2.31}$$

As it was discussed above due to the uncertainty principle there is only a finite number of different indistinguishable points in the phase space of a finite system. The phase space of the body has s_1 degrees of freedom thus the minimal element of the phase space is $(2\pi\hbar)^{s_1}$. Then the integral in (2.31) can be rewritten in a form of a sum over different states:

$$\sum_k E(p_k^{[1]}, q_k^{[1]}) f(p_k^{[1]}, q_k^{[1]}) (2\pi\hbar)^{s_1} = E_{tot} \tag{2.32}$$

where k runs over all distinguishable points in the phase space of the body.

Because there is only a finite number of the distinguishable states of the body there is also only a finite number of the mean energies of the body. Let $\epsilon_1 < \epsilon_2 < \cdots < \epsilon_m$ be all the possible values of the average energy of the body. Let g_i be the number of states with the mean energy ϵ_i:

$$g_i = |\{(p^{[1]}, q^{[1]}) : E(p^{[1]}, q^{[1]}) = \epsilon_i\}| \tag{2.33}$$

where the symbol $|\cdot|$ denotes the cardinality of a set (number of elements in a set for finite sets). If we assume that all the states of the system with certain energy ϵ_i are equiprobable (which is reasonable to assume, considering the corollary of the Liouville equation) we can define the function $f_E(\epsilon_i)$ which is equal to the distribution function f_1 in the points where the mean energy of the system is ϵ_i:

$$f^E(\epsilon_i) = f_1(p^{[1]}, q^{[1]}) \iff E(p^{[1]}, q^{[1]}) = \epsilon_i \tag{2.34}$$

Then we can re-group the summands in the sum (2.32) and rewrite it in the following way:

$$\sum_{i=1}^{m} g_i f^E(\epsilon_i)\epsilon_i \cdot (2\pi\hbar)^{s_1} = E_{tot} \tag{2.35}$$

Let us consider the system at some moments $t_1, t_2, \ldots, t_N$ such that the distribution functions at these moments are independent of each other. Let n_i be the number of moments when the mean energy of the system is ϵ_i:

$$n_i = |\{t_k : E(p^{[1]}(t_k), q^{[1]}(t_k)) = \epsilon_i\}| \tag{2.36}$$

Obviously the sum of n_i is N:

$$\sum_{i=1}^{m} n_i = N \tag{2.37}$$

If the number of moments N is large enough, then n_i is proportional to the probability $P(\epsilon_i)$ to find the body in the state with the mean energy ϵ_i:

$$P(\epsilon_i) \approx \frac{n_i}{N} \tag{2.38}$$

The distribution function f^E for the states with certain energy ϵ_i can be found using the following formula:

$$f^E(\epsilon_i) \cdot (2\pi\hbar)^{s_1} = \frac{P(\epsilon_i)}{g_i} \tag{2.39}$$

where g_i is defined by (2.33) and $(2\pi\hbar)^{s_1}$ is an elementary element of the phase space of the body which corresponds to one state. Putting (2.39) to the (2.35) we have the following relation:

$$\sum_{i=1}^{m} P(\epsilon_i)\epsilon_i = E_{tot} \tag{2.40}$$

Using (2.38) we have the following constraint for n_i:

$$\sum_{i=1}^{m} n_i\epsilon_i = N \cdot E_{tot} \tag{2.41}$$

Let us consider that $n_1, \ldots, n_m$ are given. Let us now count the number W of ways to choose different states of the system at the moments $t_1, \ldots, t_N$ such that there will be n_1 states with the energy ϵ_1, n_2 states with the energy ϵ_2 etc. To do it let us count the number of ways to choose n_i states at each energy level. As it was defined above, there are g_i states in the system which give the energy ϵ_i. We need to choose n_i of them. Because the system can in principle in two moments come to the same state, there are in total $g_i^{n_i}$ variants to do such choice. However, if we are interested only in the number of the states but not distinguish the order, we need to divide the total number of ways by the number of possible permutations of n_i objects, which is $n_i!$. Thus the total number W_i of ways to choose n_i states from g_i possibilities not regarding the order of chosen states is $g_i^{n_i}/n_i!$:

$$W_i = \frac{g_i^{n_i}}{n_i!} \tag{2.42}$$

The total number of the ways W to choose the states of the system is the product of the numbers of the ways to choose the states at each energy level ϵ_i. Thus we get the following formula:

$$W(n_1, \ldots, n_m) = \prod_{i=1}^{m} \frac{g_i^{n_i}}{n_i!} \tag{2.43}$$

The more time passed the more states the system can reach. Thus the number of available states grows with time. The value W also grows with time until reaches its maximum value because it is connected to the total number of available states. That means that at the equilibrium state which formally corresponds to the infinite time function $W(n_1, \ldots, n_m)$ has maximum under constraints (2.37) and (2.41) [100]. Because $W > 0$, we can find the maximum of the function $\ln W$ instead of finding the maximum of the function W, because these two functions have extrema in the same points. From (2.43) we have:

$$\ln W = \sum_{i=1}^{m} (n_i \ln g_i - \ln n_i!) \tag{2.44}$$

To calculate $\ln n_i!$ we can use the Stirling's approximation [101]. Because typically $n_i \gg 1$ we can use the simplest integral approximation of the $\ln n_i!$, namely:

$$\ln n_i! = \sum_{k=1}^{n_i} \ln k \approx \int_1^{n_i} \ln x dx \tag{2.45}$$

The integral can be taken by parts. Using that $d \ln x = dx/x$ we have:

$$\ln n_i! \approx n_i \ln n_i - \int_1^{n_i} x/x dx = n_i \ln n_i - n_i \tag{2.46}$$

And thus (2.44) can be approximated by the following relation:

$$\ln W(n_1, \ldots, n_m) \approx \sum_{i=1}^{m} (n_i \ln g_i - n_i \ln n_i + n_i) \tag{2.47}$$

To find the maximum of the function under certain constraints we can use the method of Lagrange multipliers [102]. The Lagrange function will contain the logarithm $\ln W$ and constraints (2.37), (2.41) multiplied by the Lagrange multipliers α and β respectively:

$$L(n_1, \ldots, n_m, \alpha, \beta) = \ln W(n_1, \ldots, n_m) - \alpha \left(\sum_{i=1}^{m} n_i - N \right) - \beta \left(\sum_{i=1}^{m} n_i \epsilon_i - N \cdot E_{tot} \right) \tag{2.48}$$

The necessary maximum condition is that all the partial derivatives $\partial L / \partial n_i$ be zero. Thus, we have:

$$\frac{\partial L}{\partial n_i} = \frac{\partial \ln W}{\partial n_i} - \alpha - \beta \epsilon_i = 0 \tag{2.49}$$

Using approximation (2.47) we get the following relation:

$$\ln g_i - \ln n_i - 1 + 1 - \alpha - \beta \epsilon_i = 0 \tag{2.50}$$

Now we can find the number of states n_i:

$$n_i = g_i e^{-\alpha - \beta \epsilon_i} \tag{2.51}$$

Using the definitions of (2.38), (2.39) we can find the distribution function $f^E(\epsilon_i)$:

$$f^E(\epsilon_i) = \frac{P(\epsilon_i)}{g_i (2\pi\hbar)^{s_1}} = \frac{n_i}{N g_i (2\pi\hbar)^{s_1}} = A e^{-\beta \epsilon_i} \tag{2.52}$$

where $A \equiv e^{-\alpha} / (N (2\pi\hbar)^{s_1})$. It is necessary to note, that constants A and β are the properties of the whole system and do not depend on the current state of the system. Also, in the constraint (2.41) E_{tot} is the total energy of the system "body + environment". However, we always can choose the zero level for the energy in such a way, that mean energy of the environment is zero. Because the interactions between the body and environment are weak and we assume that the distribution functions of the body and environment are independent from each other, we may neglect the contribution to the total energy from the interactions between the particles of body and environment. In that case the relation (2.52) signifies that the probability to find the body in the state $(p^{[1]}, q^{[1]})$ is exponentially proportional to the mean energy of this state. The probability distribution of kind (2.52) is called *Gibbs distribution.*

2.6 Entropy

From the above considerations we see that the number of accessible states of the system plays an important role in statistical mechanics. Let us consider the microcanonical ensemble with the constant energy. Let Q^S be the number of states in the system. Q^S is a multiplicative value. Indeed, let we have a complex system which is composed of two independent subsystems. Let Q^S_1 and Q^S_2 be the numbers of accessible states in each of subsystems respectively. Then for each of Q^S_1 states of the first system one can choose any of Q^S_2 states of the second systems. The total number of states in the system is the product of numbers of states in its subsystems:

$$Q^S = Q^S_1 \cdot Q^S_2 \tag{2.53}$$

However, in practice it is more convenient to have some additive quantity which is directly connected to Q^S. Such a quantity is $\ln Q^S$. Thus, for the microcanonical ensemble we define the *entropy* of the system in the following way:

$$S = k_B \ln Q^S \tag{2.54}$$

where Q^S is the number of accessible states in the microcanonical ensemble, $k_B \approx 1.3806503 \cdot 10^{-23} J/K$ is the Boltzmann constant which is introduced due to the historical reasons and is a coefficient in proportionality between energy and temperature. In the microcanonical ensemble the Liouville equation (2.23) holds and all the states are equiprobable. Let $f(p_1, \ldots, p_s, q_1, \ldots, q_s) \equiv f^E(E_0)$ be the distribution function of the system, where E_0 is the energy of the system. Let us write the normalization condition for the function f:

$$\int f(p_1, \ldots, p_s, q_1, \ldots, q_s) dp_1 \ldots dp_s dq_1 \ldots dq_s = 1 \tag{2.55}$$

Using the discreteness of the phase space and considering that the distribution function is constant we have the following:

$$\sum_{j=1}^{Q^S} (2\pi\hbar)^s f^E(E_0) = Q^S (2\pi\hbar)^s f^E(E_0) = 1 \tag{2.56}$$

And so we have the relation for the number of states in the system:

$$Q^S = \frac{1}{(2\pi\hbar)^s f^E(E_0)} \tag{2.57}$$

Putting this to (2.54) we have the following relation for the entropy:

$$S = -k_B \ln \left((2\pi\hbar)^s f^E(E_0) \right) \tag{2.58}$$

For the systems with non-constant total energy the entropy is defined by statistical averaging the entropies of different energy levels. Thus in the canonical ensemble we have the following definition for the entropy

$$S = -k_B \sum_{i=1}^{m} P(\epsilon_i) \ln \left((2\pi\hbar)^s f^E(\epsilon_i) \right) \tag{2.59}$$

where $\epsilon_1 < \cdots < \epsilon_m$ are the energy levels of the body, $P(\epsilon_i)$ is the probability to find the body at the energy level ϵ_i, $f^E(\epsilon_i) = P(\epsilon_i)/g_i$ where g_i is the number of states on the energy level ϵ_i. Considering that in canonical ensemble the distribution function $f^E(\epsilon_i)$ follows the Gibbs distribution (2.52) we have the following expression:

$$S = -k_B \sum_{i=1}^{m} P(\epsilon_i) \left(\ln \left((2\pi\hbar)^s A \right) - \beta\epsilon_i \right) \tag{2.60}$$

Using the relation (2.41) and the normalizing rule $\sum_i P(\epsilon_i) = 1$ we get the following:

$$S = -k_B \left(\ln \left((2\pi\hbar)^s A \right) - \beta\bar{E} \right) \tag{2.61}$$

where $\bar{E}$ is the mean energy of the body. Applying the Gibbs distribution (2.52) to the expression in brackets we obtain the final relation for the entropy of canonical ensemble:

$$S = -k_B \ln(2\pi\hbar)^s f^E(\bar{E}) \tag{2.62}$$

Note, that in the definition (2.62) entropy is still an additive quantity. Indeed, let we have two independent subsystems with s_1 and s_2 degrees of freedom respectively. Let $\bar{E}_1$, $\bar{E}_2$, f_1^E, f_2^E be the average energies and distribution functions of these systems respectively. The entropies of these systems S_1 and S_2 are calculated using the equation (2.62):

$$S_1 = -k_B \ln(2\pi\hbar)^{s_1} f_1^E(\bar{E}_1) \qquad S_2 = -k_B \ln(2\pi\hbar)^{s_2} f_2^E(\bar{E}_2) \tag{2.63}$$

The system which combine these two subsystems will have s_1+s_2 degrees of freedom, an average energy $\bar{E}_1 + \bar{E}_2$ and distribution function f_{12}^E. Then the entropy S_{12} of the whole system is

$$S_{12} = -k_B \ln(2\pi\hbar)^{s_1+s_2} f_{12}^E(\bar{E}_1 + \bar{E}_2) \tag{2.64}$$

The fact that the subsystems are independent means that$f_{12}^E(\bar{E}_1 + \bar{E}_2) = f_1^E(\bar{E}_1) \cdot f_2^E(\bar{E}_2)$. Putting this to (2.64) we have

$$S_{12} = -k_B \ln(2\pi\hbar)^{s_1} f_1^E(\bar{E}_1)(2\pi\hbar)^{s_2} f_2^E(\bar{E}_2) = S_1 + S_2 \tag{2.65}$$

2.7 Temperature

Let us consider a closed system. The more time passed the more states can reach the system. Thus the number of states available to the system does not decay. The entropy of the closed system (2.54) is proportional to the number of available states, so the entropy does not decrease with time, and reaches its maximum at the equilibrium state. Let the closed system is composed of K independent canonical subsystems. Let $S_1, \dots, S_K$ and $\bar{E}_1, \dots, \bar{E}_K$ be the entropies and the average energies of the subsystems respectively. The total energy of the system E_0 is the sum of the average energies of the subsystems:

$$E_0 = \sum_{i=1}^{K} \bar{E}_i \tag{2.66}$$

The total entropy of the system S_0 is the sum of the entropies of the subsystems:

$$S_0 = \sum_{i=1}^{K} S_i \tag{2.67}$$

Let us consider the total entropy of the system S_0 as a function of energies of the subsystems:

$$S_0 = S_0(\bar{E}_1, \dots, \bar{E}_K) = \sum_{i=1}^{K} S_i(\bar{E}_i) \tag{2.68}$$

At the equilibrium state the entropy of the system $S_0(\bar{E}_1, \dots, \bar{E}_K)$ has a local maximum provided that the constraint (2.66) holds. We use the Lagrange's multipliers method and write the Lagrange function for the entropy S_0 [95]:

$$L(\bar{E}_i, \dots, \bar{E}_K, \gamma) = \sum_{i=1}^{K} S_i(\bar{E}_i) - \gamma \left(\sum_{i=1}^{K} \bar{E}_i - E_0 \right) \tag{2.69}$$

The necessary condition of extremum is equality of all the partial derivatives $\partial L / \partial \bar{E}_i$ to zero. This leads us to the following equations:

$$\frac{\partial L}{\partial \bar{E}_i} = \frac{\partial S_i}{\partial \bar{E}_i} - \gamma = 0 \qquad i = 1, \dots, K \tag{2.70}$$

Because the entropy $S_i(\bar{E}_i)$ depends only on one energy $\bar{E}_i$ the partial derivatives $\partial S_i / \partial \bar{E}_i$ can be replaced with the full derivatives $dS_i / d\bar{E}_i$. From (2.70) it follows that for any $1 \le i, j \le K$ it holds the following:

$$\frac{dS_i}{d\bar{E}_i} = \frac{dS_j}{d\bar{E}_j} = \gamma = \text{const} \tag{2.71}$$

Let us introduce the value $T = 1/\gamma$ which we call *the temperature* of the system. Note, that for the system in the equilibrium state the temperature of all its subsystems is constant. So, in the canonical ensemble in the equilibrium three values remain constant: number of particles N, volume V and temperature T. That's why the canonical ensembles are often referenced as *NVT ensembles*. Now we can determine the unknown coefficients A, β in the Gibbs distribution (2.52). Using the entropy representation (2.61) for the entropies S_i we have the following:

$$S_i = -k_B(\ln((2\pi\hbar)^{s_i} A_i) - \beta_i \bar{E}_i) \tag{2.72}$$

where A_i, β_i are unknown Lagrange multipliers. Putting (2.72) to (2.71) we have:

$$\frac{dS_i}{d\bar{E}_i} = k_B \beta_i = \gamma = \frac{1}{T} = \text{const} \tag{2.73}$$

So, for any $1 < i, j < K$ it holds $\beta_i = 1/(k_B T) = \beta_j \equiv \beta$.

Now we can also calculate the coefficient A in the Gibbs distribution (2.52). To do this we can use the normalization condition for the density distribution. Putting the Gibbs distribution (2.52) to the normalization condition (2.2) we have the following relation:

$$A \int e^{-\beta \bar{E}(p_1,\ldots,p_s,q_1,\ldots,q_s)} dp_1 \ldots dp_s dq_1 \ldots dq_s = 1 \tag{2.74}$$

where $\beta = (k_B T)^{-1}$. Let us introduce the *partition function* of the system, which is defined in the following way:

$$Z_N = (2\pi\hbar)^{-s} \int e^{-\beta \bar{E}(p_1,\ldots,p_s,q_1,\ldots,q_s)} dp_1 \ldots dp_s dq_1 \ldots dq_s \tag{2.75}$$

Thus the Gibbs distribution (2.52) for the canonical ensemble read as follows:

$$f(p_1, \ldots, p_s, q_1, \ldots, q_s) = \frac{1}{(2\pi\hbar)^s Z_N} e^{-\beta \bar{E}(p_1,\ldots,p_s,q_1,\ldots,q_s)} \tag{2.76}$$

2.8 Free Energy

Let us try to determine the physical meaning of the partition function Z_N and coefficient A in the Gibbs distribution (2.52). Let us write the definition of entropy of the system (2.62) considering the Gibbs distribution (2.52). We obtain the following:

$$S = -k_B \ln\left(\frac{1}{Z_N} e^{-\beta \bar{E}}\right) = k_B \ln Z_N + \frac{k_B}{k_B T} \bar{E} \tag{2.77}$$

Multiplying both parts by the temperature T and rearranging the summand we come to the following relation:

$$-k_B T \ln Z_N = \bar{E} - TS \tag{2.78}$$

The quantity $-k_BT\ln Z_N$ is called *Helmholtz free energy* of the system. We will denote it by the symbol $\mathcal{F}$:

$$\mathcal{F} = -k_BT\ln Z_N = \bar{E} - TS \tag{2.79}$$

Let's determine a thermodynamical sense of the Helmholtz free energy. In the thermodynamics the energy which the body receives via the thermal interaction is called *heat* and is denoted by Q. The first principle of thermodynamics which follows from the energy conservation law states that the sum of the change of internal energy and the work performed by the body is equal to the heat transferred to the body:

$$dQ = d\bar{E} + dW \tag{2.80}$$

where W is the work performed by the body. Typically we assume that the work of the body is a mechanical work performed due to changing of the volume of the body. As it is known from mechanics, the work is a scalar product of the force by the displacement, and the force in turn is a product of the pressure by the surface area. In the equilibrium state the average pressure $\bar{P}$ is constant at each point and the pressure itself varies very little around the average value. Let the surface area of the reservoir which contain the body is A. Let after performing of the work dW each infinitesimal part of surface area dA_i moved by the distance dr_i in the direction perpendicular to the surface element dA_i. The force which acts on the surface element dA_i is $\bar{P}dA_i$ and the work to move this element is $dW_i = \bar{P}dA_idr_i$. The value dA_idr_i is an elementary change of volume dV_i. Thus $dW_i = \bar{P}dV_i$. The total work dW can be found as the sum of elementary works:

$$dW = \sum \bar{P}dV_i = \bar{P}dV \tag{2.81}$$

In the canonical ensemble the volume is fixed thus $dV = 0$ and mechanical work is not performed: $dW = 0$. Then from (2.80) we have $dQ = d\bar{E}$. Considering (2.73) we can conclude that the whole heat in canonical ensemble is transfered to the entropy:

$$dQ = d\bar{E} = TdS \tag{2.82}$$

Let us consider two phase process. At the first phase the body receives some heat dQ but the volume of body is constant. In that case the energy change $d\bar{E}_1$ increases the entropy of the body $dQ = d\bar{E}_1 = TdS$. At the second phase the body does not receive heat, but increase its volume by dV. According to (2.80) we have $0 = dE_2 + PdV$ where dE_2 is the energy change at second phase. The total energy change of the process dE is the sum of total energies of the pases:

$$dE = dE_1 + dE_2 = TdS - PdV = TdS - dW \tag{2.83}$$

Rearranging the summands and using the definition (2.79) we can conclude that the work performed by the body is equal to the decrease of the Helmholtz free energy $d\mathcal{F}$:

$$dW = -(dE - TdS) = -d\mathcal{F} \tag{2.84}$$

So Helmholtz free energy shows an ability of the body to perform a work in the isothermal process. It is necessary to notice that Helmholtz free energy should not be used for processes where the volume is not constant. Relation (2.84) only shows that amount of the energy received during the heat transfer in canonical ensemble in principle can be released in some other process. To describe a processes where the volume of the system changes we can use the *Gibbs free energy* which includes the term PV which reflects the mechanical work potential. The definition of the Gibbs free energy G is written in the following way:

$$G = \mathcal{F} + PV = H - TS = \bar{E} + PV - TS \tag{2.85}$$

where H is *enthalpy* of the system.

The change of the free energy shows whether the process is probable or not. If free energy decreases then the energy is released during the process, so this process is spontaneous. So, free energy can be used to determine the most probable state of the system.

2.9 Grand Canonical Ensemble

Before we considered only the systems which have a fixed number of particles. However, in the physical chemistry it is often necessary to describe the processes where the concentrations of substances may change. Such processes are for example diffusion and chemical reactions. Let us consider the system which contains m substances with changeable concentrations. Let $N_i, i = 1 \ldots, m$ be the quantity of the particles of i^{th} type introduced to the system. The average energy of the system $\bar{E}$ depends on the number of the introduced particles: $\bar{E} = \bar{E}(N_1, \ldots, N_m)$. Let $\mu_i, i = 1 \ldots m$ be the energy change after introducing of one particle of i^{th} type to the system. Formally we can write the following relation:

$$\mu_i = \frac{\partial E(N_1, \ldots, N_m)}{\partial N_i} \tag{2.86}$$

where $\bar{E}$ is the average energy of the system and partial derivative means the energy change after the minimal possible change of the number of particles. The value μ_i is called the *chemical potential* of the particles of i^{th} type in the given system. Strictly speaking chemical potential of the particle depends on the concentrations of the substances in the system, i.e $\mu_i = \mu_i(N_1, \ldots, N_m)$. However if the numbers of introduced particles $N_1, \ldots, N_m$ are small with respect to the total number of particles in the system the concentrations will not essentially change and thus the chemical potentials $\mu_1, \ldots, \mu_m$ can be assumed to be constant.

Let us consider the system at a fixed temperature T in a fixed volume V with fixed chemical potentials of particles $\mu_1, \ldots, \mu_m$. Considering that for the fixed number of particles $d\bar{E} = TdS$ the full differential of the energy can be written in the following way:

$$d\bar{E} = TdS + \sum_{i=1}^{m} \mu_i dN_i \tag{2.87}$$

By rearranging the summands and dividing by T we obtain the following relation for dS:

$$dS = \frac{d\bar{E}}{T} - \sum_{i=1}^{m} \frac{\mu_i}{T} dN_i \tag{2.88}$$

If $N_i, i = 1, \ldots, m$ are small, we can assume that the entropy is linear with respect to $N_1, \ldots, N_m$. From (2.77) it follows that the entropy is also linear with respect to the energy. So, considering the coefficients in (2.88) we can write the following relation for the entropy:

$$S(N_1, \ldots, N_m) = \frac{1}{T}\left(-\Omega + \bar{E} - \sum_{i=1}^{m} N_i \mu_i\right) \tag{2.89}$$

where Ω is some constant called *Grand potential.* From (2.89) immediately follows the definition of the grand potential:

$$\Omega = \bar{E} - TS - \sum_{i=1}^{m} \mu_i N_i \tag{2.90}$$

Let us find the probability density $f^G(E, N_1, \ldots, N_m)$ to find the system is in some state $(p_1, \ldots, p_s, q_1, \ldots, q_s)$ given that the average energy of this state is E and the numbers of introduced particles of types $1 \ldots m$ are $N_1, \ldots, N_m$. The definition of f^G given above can be written rigorously in the following way:

$$\begin{array}{c} f^G(E_0, N_1, \ldots, N_m) = f(p_1, \ldots, p_s, q_1, \ldots, q_s) \iff \\ \iff s = s_0 + \sum_{i=1}^{m} N_i s_i, E(p_1, \ldots, p_s, q_1, \ldots, q_s) = E_0 \end{array} \tag{2.91}$$

where s_0 is the number of degrees of freedom in the system without any introduced particles, s_i is the number of degrees of freedom of the particle of i^{th} type, $E(p_1, \ldots, p_s, q_1, \ldots, q_s)$ is the average energy of the system in certain state defined by (2.29). Using the relation (2.62) we can find the entropy for any given number of introduced particles $N_1, \ldots, N_m$:

$$S(N_1, \ldots, N_m) = -k_B \ln(2\pi\hbar)^{s_0 + \sum N_i s_i} f^G(E, N_1, \ldots, N_m) \tag{2.92}$$

Putting this to (2.89) we have the following:

$$-k_B \ln(2\pi\hbar)^{s_0 + \sum N_i s_i} f^G(E, N_1, \ldots, N_m) = \frac{1}{T}\left(-\Omega + E - \sum_{i=1}^{m} \mu_i N_i\right) \tag{2.93}$$

From this relation we can find f^G:

$$f^G(E, N_1, \ldots, N_m) = \frac{1}{(2\pi\hbar)^{s_0+\sum N_i s_i}} e^{\beta\Omega} e^{-\beta E+\sum \beta\mu_i N_i} \tag{2.94}$$

The grand partition function Ω can be found from the normalization rule for f^G. The total probability to find the system in some state with some number of particles $N_1, \ldots, N_m$ is unity. Thus, considering (2.94) we can write the following normalization:

$$e^{\beta\Omega} \cdot \sum_{N_1,\ldots,N_m} \frac{\lambda_1^{N_1} \ldots \lambda_m^{N_m}}{(2\pi\hbar)^{s_0+\sum N_i s_i}} \int e^{-\beta E(p_1,\ldots,p_s,q_1,\ldots,q_s)} dp_1 \ldots dp_s dq_1 \ldots dq_s = 1 \tag{2.95}$$

where sum is taken over all possible values of $N_1, \ldots, N_m$, $\lambda_i \equiv e^{\mu_i}$ is *activity* of the i^{th} type of particles. The grand potential can be written in the following form:

$$\Omega = k_B T \ln \frac{1}{\Xi} = -k_B T \ln \Xi \tag{2.96}$$

where $\Xi = \sum_{N_1,\ldots N_m} \lambda_1^{N_1} \ldots \lambda_m^{N_m} Z_{N_1,\ldots,N_m}$ is the *Grand partition function*, $Z_{N_1,\ldots,N_m}$ is the canonical partition function for particular values $N_1, \ldots, N_m$.

2.10 Solubility and Solvation Free energy

The solubility in water and other fluids in the human body is one of the important properties of drugs and drug-like compounds used in medicine because it shows the drug ability to be delivered to the place there it works. Free energy is the energetic characteristic of the chemical system which allows one to describe and predict many chemical processes including the dissolution process. The dissolution process by itself can be divided into two phases: 1) Destroying the crystal structure, transition from the crystal to the “free” state. 2) Solvatation of the molecule. The energy of the transition from the crystal to the “free” state can be with a reasonable accuracy estimated with quantum-mechanical methods [7]. The second phase (solvation of the molecule) is energetically described by the solvation free energy (SFE). SFE is the free energy change during the transfer of the compound from the free (gas) phase to a solution. Although SFE can be measured experimentally, these experiments are quite complicated from the technical point of view especially for compounds with low solubility and low volatility [64]. Computation of the SFE is also a challenging task, because solvation process involves many-body interactions of the solvent molecules. Typically calculation of the SFE with the molecular dynamics or Monte Carlo simulations may take several days or even weeks, and will not necessarily be accurate enough. In our work we use much faster method which is based on the classical density functional theory and integral equation theory of liquids.

Chapter 3

Integral Equation Theory of Liquids. RISM and 3DRISM

In this chapter the basics of the classical density functional theory and the reference interaction site model (RISM) are described. The chapter contains definitions of the distribution and correlation functions, free-energy functional methods, derivation of Ornstein-Zernike, RISM and 3DRISM equations and closure relations. Theoretical questions which are necessary for numerical solution of RISM equations are discussed. The chapter is mostly based on Refs. [50], [95] and [24]. It is necessary to note, that many derivations in the textbooks are given for systems of spherical particles. In this chapter we develop the derivations for the six-dimensional case and multi-particle systems.

3.1 System under investigation

In the previous chapter we did not make any assumptions about the particles in the systems. We only assume that there are many particles in the system, and that the motion of the particles is quasi-chaotic. We defined the macroscopic parameters of the system, such as entropy, temperature and free energy and relate them to each other. This allows us to predict the state of the system and the energy changes in the thermodynamic processes. Integral equation theory of liquids (IETL) allows one to predict some specific microscopic and macroscopic thermodynamic parameters of fluids, i.e. the local solvent structure around the solute, free energy of solvation etc. The simplest model of fluid is the system composed of the spherical particles which interact via the pairwise additive potential. This model allows one to describe noble gases, electrolyte solutions and other systems where the shape of the particles does not affect much the properties of the fluid. However, properties of many physical systems of interest essentially depend on the shape of the molecules and partial charges of the atoms of the molecules. For those systems the model of spherical particles is unable to give reliable results. That's why we will use the model which accounts the shape of the molecules. The molecules in our model are described as rigid objects. This means that the distances of the atoms in the molecule are fixed. Position of the rigid molecule can be described by three coordinates of the center of mass $\mathbf{r} = (x, y, z)$ and three Euler angles $\boldsymbol{\theta} = (\theta_x, \theta_y, \theta_z)$. Interactions between the molecules can be described by the potential $U_{12}(\mathbf{r_2} - \mathbf{r_1}, \boldsymbol{\theta_1} - \boldsymbol{\theta_2})$ where $(\mathbf{r_1}, \boldsymbol{\theta_1})$ and $(\mathbf{r_2}, \boldsymbol{\theta_2})$ are the positions and rotations of the first and second molecules correspondingly. This model gives much better description of the fluid than the spherical particles model. However, it also does not consider many effects related to the motion of the molecules, e.g. conformational changes of the molecules, vibrations of the atoms etc. The model gives relatively good description of the fluids composed of small or rigid molecules, e.g. water, carbon hydroxide, benzene etc. This model is not ideal by any means; however it can give insights to many important systems in physical chemistry. The reason why we use this model is that more detailed models with non-rigid molecules often appear to be too complicated to be simulated from the computational point of view.

3.2 Configuration integral

In many cases the general formula (2.79) is not suitable for practical free energy calculations due to unknown partition function (2.75). However there are some systems for which the partition function can be calculated easily. One of such systems is the ideal gas. For non-ideal gas systems one can separate the ideal and non-ideal contributions to the free energy which also can simplify the calculations. Let us now consider the general case of the system of N interacting rigid particles. Total energy of the system can be represented as a sum of kinetic energy term K and potential energy term U, where kinetic energy depends only on the generalized momenta of the system and the potential energy depends only on generalized coordinates of the system. Let us consider the canonical ensemble which contains N molecules. We assume that the molecules are rigid. The potential energy U depends on positions $\mathbf{r_1}, \ldots, \mathbf{r_N}$ and orientations $\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_N}$: $U = U(\mathbf{r_1}, \ldots, \mathbf{r_N}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_N})$. It is known that the motion of the rigid body can be described as the movement of its center of mass and rotation of the body around the perpendicular axes which come through the center of mass of the body [96]. Let $\mathbf{p_k} = (p_x^k, p_y^k, p_z^k)$ be the linear momentum of the k^{th} particle, $\mathbf{l_k} = (l_x^k, l_y^k, l_z^k)$ be the angular momentum of the k^{th} particle around the perpendicular axes which come through the center of mass of the body. As it is known from the course of mechanics, the kinetic energy of the body is a sum of translational and rotational kinetic energy. Let $\mathbf{v_0}(t)$ be the velocity of the center of mass of the molecule. At each moment of time t we can introduce the coordinate system which is located in the center of mass of the molecule and moving with the speed $\mathbf{v_0}(t)$. Let us choose at the moment t the coordinate axes of the system in some known "local" directions (e.g. which go through some certain atoms of the molecule). However, let the orientation of the axes be fixed with respect to the initial system. Let there are M different types of molecules in the system. Let $a_k \in \{1, \dots, M\}$ be the type of the k^{th} molecule. Let m_{a_k} be the mass of the molecule of type a_k. Thus, in a new coordinate system the center of mass of the molecule is immovable; however, the molecule by itself is rotating. The kinetic energy with respect to this system differs from the kinetic energy with respect to initial system by $m_{a_k} v_0(t)^2/2$. Let $J^x_{a_k}$, $J^y_{a_k}$, $J^z_{a_k}$ be the moments of inertia of the molecule of type a_k about the chosen axes. Then the rotational kinetic energy can be found with the following relation [1]:

$$K_{\text{rot}} = \frac{(l^k_x)^2}{2J^x_{a_k}} + \frac{(l^k_y)^2}{2J^y_{a_k}} + \frac{(l^k_z)^2}{2J^z_{a_k}} \tag{3.1}$$

The total kinetic energy of the molecule is a sum of the rotational kinetic energy and the translational kinetic energy of the center of mass of the molecule. It can be written in the following way:

$$K_k = \frac{(p^k_x)^2 + (p^k_y)^2 + (p^k_z)^2}{2m_{a_k}} + \frac{(l^k_x)^2}{2J^x_{a_k}} + \frac{(l^k_y)^2}{2J^y_{a_k}} + \frac{(l^k_z)^2}{2J^z_{a_k}} \tag{3.2}$$

For the sake of uniformity, let us use the following definitions:

$$\begin{array}{ll} p_{6k-5} \equiv p^k_x & p_{6k-2} \equiv l^k_x \\ p_{6k-4} \equiv p^k_y & p_{6k-1} \equiv l^k_y \\ p_{6k-3} \equiv p^k_z & p_{6k} \equiv l^k_z \end{array} \tag{3.3}$$

$$\begin{array}{l} B_{6k-5} \equiv B_{6k-4} \equiv B_{6k-3} \equiv m_{a_k} \\ B_{6k-2} \equiv J^x_{a_k} \\ B_{6k-1} \equiv J^y_{a_k} \\ B_{6k} \equiv J^z_{a_k} \end{array} \tag{3.4}$$

where $k = 1 \dots N$

Using the above definitions the kinetic energy of the system can be written in the following way:

$$K(p_1, \dots, p_{6N}) = \sum_{i=1}^{6N} \frac{p_i^2}{2B_i} \tag{3.5}$$

The total energy of the system is a sum of kinetic and potential energy components. This can be written in the following way:

$$E(p_1, \dots, p_{6N}, \mathbf{r_1}, \dots, \mathbf{r_N}, \boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_N}) = \sum_{i=1}^{6N} \frac{p_i^2}{2B_i} + U(\mathbf{r_1}, \dots, \mathbf{r_N}, \boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_N}) \tag{3.6}$$

where $U(\mathbf{r_1}, \dots, \mathbf{r_N}, \boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_N})$ is a potential energy of the system. Here and below saying "the total energy of the system" we mean the total average energy of the system, where averaging is done over the degrees of freedom of the environment as it was done in the section 2.5.

[1] In principle, the same formula can be derived using the diagonalization of the moment of inertia tensor

Let us find the partition function of the system $Z_{N_1\dots N_M}$ where $N_1\dots N_M$ are the numbers of particles of different kind in the system. In the general definition of partition function (2.75) the integration is performed over all distinguishable states of the system. We suppose that the particles of each kind are indistinguishable. Then there are $N_k!$ ways to select N_k particles of kind k. The total number of ways to select particles of all kinds is $N_1!\cdot\dots\cdot N_M!$. We need to normalize the integral in the definition of the partition function considering this number. Then the partition function $Z_{N_1\dots N_M}$ is defined in the following way:

$$Z_{N_1\dots N_M} = \\ \frac{1}{(2\pi\hbar)^{6N}N_1!\dots N_M!}\int e^{-\beta\sum_{i=1}^{6N}\frac{p_i^2}{2B_i}-\beta U(\mathbf{r_1},\dots,\mathbf{r_N},\boldsymbol{\theta_1},\dots,\boldsymbol{\theta_N})}dp_1\dots dp_{6N}d\mathbf{r_1}\dots d\mathbf{r_N}d\boldsymbol{\theta_1}\dots d\boldsymbol{\theta_N} \tag{3.7}$$

One can notice that the integration over the momenta components can be done separately:

$$\int e^{-\beta\sum_{i=1}^{6N}\frac{p_i^2}{2B_i}}dp_1\dots dp_{6N} = \prod_{i=1}^{6N}\int_{-\infty}^{\infty} e^{-\beta\frac{p_i^2}{2B_i}}dp_i \tag{3.8}$$

To calculate contribution of each component we can use a known formula for Gaussian integral [103]:

$$\int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi} \tag{3.9}$$

To prove this relation we define $I = \int_{-\infty}^{\infty} e^{-x^2}dx$. Then $I^2 = \int e^{-x^2}dx\int e^{-y^2}dy = \int\int e^{-(x^2+y^2)}dxdy$. Switching to the polar coordinates we have $x = r\cos\phi$, $y = r\sin\phi$, $dxdy = rdrd\phi$, $I^2 = \int_0^{\infty}\int_0^{2\pi} e^{-r^2}rdrd\phi = -\pi\int_0^{\infty} e^{-r^2}(-2r)dr$. Considering that $-2rdr = d(-r^2)$ we have $I^2 = -\pi(e^{\infty} - e^0) = \pi$, thus $I = \sqrt{\pi}$. Using this equality we can write the following:

$$\prod_{i=1}^{6N}\int_{-\infty}^{\infty} e^{-\beta\frac{p_i^2}{2B_i}}dp_i = \prod_{i=1}^{6N}\sqrt{2B_ik_BT}\int e^{-\left(\frac{p_i}{\sqrt{2B_ik_BT}}\right)^2}\frac{dp_i}{\sqrt{2B_ik_BT}} = \prod_{i=1}^{6N}\sqrt{2\pi B_ik_BT} \tag{3.10}$$

Considering the definitions of B_i (3.4) we can return to the original masses and momenta of inertia of the particles:

$$\prod_{i=1}^{6N}\sqrt{2\pi B_ik_BT} = \prod_{i=1}^{N}(2\pi k_BT)^{6/2}(m_{a_i})^{3/2}(J^x_{a_i}J^y_{a_i}J^z_{a_i})^{1/2} \tag{3.11}$$

Putting this to the definition of the partition function (3.7) and considering that $(2\pi\hbar)^{6N} = \prod_a(2\pi\hbar)^{6N_a}$ we have the following:

$$Z_{N_1\dots N_M} = \\ \prod_{a=1}^{M}\left(\frac{(2\pi\hbar k_BT)^3(m_a^3J_a^xJ_a^yJ_a^z)^{1/2}}{(2\pi\hbar)^6}\right)^{N_a}\frac{1}{N_a!}\int e^{-\beta U(\mathbf{r_1},\dots,\mathbf{r_N},\boldsymbol{\theta_1},\dots,\boldsymbol{\theta_N})}d\mathbf{r_1}\dots d\mathbf{r_N}d\boldsymbol{\theta_1}\dots d\boldsymbol{\theta_N} \tag{3.12}$$

We define the *configuration integral* $Q_{N_1 \ldots N_M}$ in the following way:

$$Q_{N_1 \ldots N_M} = \int e^{-\beta U(\mathbf{r_1}, \ldots, \mathbf{r_N}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_N})} d\mathbf{r_1} \ldots d\mathbf{r_N} d\boldsymbol{\theta_1} \ldots d\boldsymbol{\theta_N} \tag{3.13}$$

Then the partition function can be written in the following form:

$$Z_{N_1 \ldots N_M} = \prod_{a=1}^{M} \frac{D_a^{N_a}}{N_a!} Q_{N_1 \ldots N_M} \tag{3.14}$$

where $D_a \sqrt{m_a^3 J_a^x J_a^y J_a^z} (k_B T)^3 (2\pi)^{-3} \hbar^{-6}$.

We will use the functions $Z_{N_1 \ldots N_M}$ and $Q_{N_1 \ldots N_M}$ in our work. However, we find it also necessary to write the exact expression for the functions Z_N and Q_N for the liquid of N identical spherical particles. In that case there are no integration over the angular degrees of freedom in (3.7), and the partition function Z_N is written in the following way:

$$Z_N = \frac{1}{\Lambda^{3N} N!} Q_N \tag{3.15}$$

where $Q_N = \int \exp(-\beta U(\mathbf{r_1}, \ldots, \mathbf{r_N})) d\mathbf{r_1} \ldots d\mathbf{r_N}$, $\Lambda = \sqrt{2\pi\hbar^2/(m k_B T)}$ is the *thermal de Broglie wavelength*, m is a mass of the particle.

3.3 Density distribution functions

Putting the expression for the average energy of the system (3.6) to the Gibbs distribution (2.52) we obtain the following:

$$\begin{aligned} f(p_1, \ldots, p_{6N}, \mathbf{r_1}, \ldots, \mathbf{r_N}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_N}) &= \prod_{i=1}^{6N} \left(\frac{e^{-\beta \frac{p_i^2}{2B_i}}}{\hbar \sqrt{2\pi B_i k_B T}} \right) \cdot \frac{N_1! \ldots N_M!}{Q_{N_1 \ldots N_M}} e^{-\beta U(\mathbf{r_1}, \ldots, \mathbf{r_N}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_N})} \\ &\equiv f_p(p_1, \ldots, p_{6N}) \cdot f_q(\mathbf{r_1}, \ldots, \mathbf{r_N}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_N}) \end{aligned} \tag{3.16}$$

where kinetic and potential components f_p, f_q of the distribution function f are defined in the following way:

$$f_p(p_1, \ldots, p_{6N}) = \prod_{i=1}^{6N} \frac{1}{\hbar \sqrt{2\pi B_i k_B T}} e^{-\beta \frac{p_i^2}{2B_i}} \tag{3.17}$$

$$f_q(\mathbf{r_1}, \ldots, \mathbf{r_N}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_N}) = \frac{N_1! \ldots N_M!}{Q_{N_1 \ldots N_M}} e^{-\beta U(\mathbf{r_1}, \ldots, \mathbf{r_N}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_N})} \tag{3.18}$$

We can see that kinetic component f_p is a product of the probability distributions of momenta components. In such a way we can find the probability that the momenta degree of freedom has a certain value p_i:

$$f_p^i(p_i) = \frac{1}{2\pi\hbar \sqrt{B_i k_B T}} e^{-\beta \frac{p_i^2}{2B_i}} \tag{3.19}$$

For the distribution f_q in a general case such a representation is not possible, because the potential energy U connects the coordinates of all particles. Speaking about the potential component, we also rarely need to know positions of all particles. It is more useful to know the mean number of particles at some selected positions, where number of such positions is much smaller than number of particles. For the one-component system we define the n-particle density distribution function $\rho^{(n)}(\mathbf{r_1},\ldots,\mathbf{r_n},\boldsymbol{\theta_1},\ldots,\boldsymbol{\theta_n}))$ in the following way:

$$\rho^{(n)}(\mathbf{r_1},\ldots,\mathbf{r_n},\boldsymbol{\theta_1},\ldots,\boldsymbol{\theta_n}) = \frac{N!}{(N-n)!}\int \frac{e^{-\beta U(\mathbf{r_1},\ldots,\mathbf{r_N},\boldsymbol{\theta_1},\ldots,\boldsymbol{\theta_N})}}{Q_N} d\mathbf{r_{n+1}}\ldots d\mathbf{r_N} d\boldsymbol{\theta_{n+1}}\ldots d\boldsymbol{\theta_N} \tag{3.20}$$

The similar definition can be given for the multi-component systems as well. However, in a general case this definition is too cumbersome to be written. For the practical applications the most interesting are one-particle and two-particle density distribution functions. We define the one-particle distribution function ρ^a of particles of type a in the following way:

$$\rho^a(\mathbf{r_1},\boldsymbol{\theta_1}) = N_a \int \frac{e^{-\beta U(\mathbf{r_1},\ldots,\mathbf{r_N},\boldsymbol{\theta_1},\ldots,\boldsymbol{\theta_N})}}{Q_{N_1\ldots N_M}} d\mathbf{r_2}\ldots d\mathbf{r_N} d\boldsymbol{\theta_2}\ldots d\boldsymbol{\theta_N} \tag{3.21}$$

where it is assumed that the particle with the coordinates $(\mathbf{r_1},\boldsymbol{\theta_1})$ has type a. Two-particle distribution function ρ^{ab} for the particles of types a, b is defined in the following way:

$$\rho^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2}) = N_a(N_b-\delta_{ab})\int \frac{e^{-\beta U(\mathbf{r_1},\ldots,\mathbf{r_N},\boldsymbol{\theta_1},\ldots,\boldsymbol{\theta_N})}}{Q_{N_1\ldots N_M}} d\mathbf{r_3}\ldots d\mathbf{r_N} d\boldsymbol{\theta_3}\ldots d\boldsymbol{\theta_N} \tag{3.22}$$

where δ_{ab} is a Kronecker's delta and it is assumed that the particles with the coordinates $(\mathbf{r_1},\boldsymbol{\theta_1})$, $(\mathbf{r_2},\boldsymbol{\theta_2})$ have types a and b correspondingly.

Also, we define the dimensionless *pair density correlation function* g^{ab} in the following way:

$$g^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2}) = \frac{\rho^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})}{\rho^a(\mathbf{r_1},\boldsymbol{\theta_1})\rho^b(\mathbf{r_2},\boldsymbol{\theta_2})} \tag{3.23}$$

We note that considering the properties of the Dirac delta-function the definitions of the density distribution functions (3.21), (3.22) can be written in the following form [50]:

$$\rho^a(\mathbf{r_1},\boldsymbol{\theta_1}) = \int \sum_{i=1}^{N_a} \delta(\mathbf{r}_i^a-\mathbf{r_1})\delta(\boldsymbol{\theta}_i^a-\boldsymbol{\theta_1}) \frac{e^{-\beta U(\mathbf{r_1},\ldots,\mathbf{r_N},\boldsymbol{\theta_1},\ldots,\boldsymbol{\theta_N})}}{Q_{N_1\ldots N_M}} d\mathbf{r_1}\ldots d\mathbf{r_N} d\boldsymbol{\theta_1}\ldots d\boldsymbol{\theta_N} \tag{3.24}$$

$$= \left\langle \sum_{i=1}^{N_a} \delta(\mathbf{r}_i^a-\mathbf{r_1})\delta(\boldsymbol{\theta}_i^a-\boldsymbol{\theta_1}) \right\rangle$$

$$\rho^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2}) = \left\langle \sum_{i=1}^{N_a}\sum_{j=1}^{N_b} (1-\delta_{ab}\delta_{ij})\delta(\mathbf{r}_i^a-\mathbf{r_1})\delta(\boldsymbol{\theta}_i^a-\boldsymbol{\theta_1})\delta(\mathbf{r}_j^b-\mathbf{r_2})\delta(\boldsymbol{\theta}_j^b-\boldsymbol{\theta_2}) \right\rangle \tag{3.25}$$

3.4 Free Energy functional

We considered systems with the uniform density. However, if some external field acts on the particles of the system the density can change. Also, we can consider not necessarily the equilibrium state of the system. In such cases it can be necessary to consider dependency of the free energy of the system on the external factors such as external field or density distribution. Let the system contains $N_1, \ldots, N_M$ particles of types $1, \ldots, M$ correspondingly. We denote as $(\mathbf{r}_i^a, \boldsymbol{\theta}_i^a)$ the coordinates of the i^{th} particle of type a. Then the external field V_N can be expressed as follows:

$$V_N(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) = \sum_{a=1}^{M} \sum_{i=1}^{N_a} \phi^a(\mathbf{r}_i^a, \boldsymbol{\theta}_i^a) \tag{3.26}$$

where $N = \sum_a N_a$, $\mathbf{r}^{[N]} \equiv (\mathbf{r}_1^1, \ldots, \mathbf{r}_{N_1}^1, \ldots, \mathbf{r}_1^M, \ldots, \mathbf{r}_{N_M}^M)$,
$\boldsymbol{\theta}^{[N]} \equiv (\boldsymbol{\theta}_1^1, \ldots, \boldsymbol{\theta}_{N_1}^1, \ldots, \boldsymbol{\theta}_1^M, \ldots, \boldsymbol{\theta}_{N_M}^M)$, $\phi^a(\mathbf{r}, \boldsymbol{\theta})$ is an external field which acts on the particle of type a at position $(\mathbf{r}, \boldsymbol{\theta})$.

Using the definition of the Gibbs distribution (2.76) together with the relation (3.16) we can write the following:

$$f_p(p_1, \ldots, p_{6N}) f_q(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) = \frac{1}{(2\pi\hbar)^{6N} Z_{N_1 \ldots N_M}} e^{-\beta \sum \frac{p_i^2}{2B_i} - \beta U(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) - \beta V(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]})} \tag{3.27}$$

Integrating both parts over the momenta components $p_1, \ldots, p_{6N}$ we get the following:

$$f_q(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) = \prod_{a=1}^{M} \frac{D_a^{N_a}}{N_a!} \frac{1}{Z_{N_1 \ldots N_M}} e^{-\beta U(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) - \beta V(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]})} \tag{3.28}$$

Form this relation we can express $Z_{N_1 \ldots N_M}$ as a function of f_q in a certain point $(\mathbf{r_1}, \ldots, \mathbf{r_N}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_N})$:

$$Z_{N_1 \ldots N_M}(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) = \prod_{a=1}^{M} \frac{D_a^{N_a}}{N_a!} \frac{e^{-\beta U(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) - \beta V(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]})}}{f_q(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]})} \tag{3.29}$$

Similarly to the general definition of the free energy of the system (2.79) we can define the free energy as a function of the density in certain point:

$$\begin{aligned} &\mathcal{F}(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) = -k_B T \ln Z_{N_1 \ldots N_M} = \\ &k_B T \ln \prod_{a=1}^{M} \frac{N_a!}{D_a^{N_a}} + k_B T \ln f_q(\mathbf{r_1}, \ldots, \mathbf{r_N}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_N}) + U(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) + V(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) \end{aligned} \tag{3.30}$$

We introduce the *Free energy functional* as the ensemble average of the expression (3.30) over all distinguishable points of the phase space, namely:

$$\begin{aligned} &\mathcal{F}[\mathbf{f_q}] = \left\langle \mathcal{F}(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) \right\rangle = \\ &k_B T \sum_{a=1}^{M} \ln \frac{N_a!}{D_a^{N_a}} + \int f_q(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) \left(k_B T \ln f_q(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) + U(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) + V(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) \right) d\mathbf{r}^{[N]} d\boldsymbol{\theta}^{[N]} \end{aligned} \tag{3.31}$$

3.5 Minimization property of the free energy functional

One of the important properties of the free energy functional is that for the equilibrium density $f_q^0 = A/Z_{N_1 \dots N_M} \exp(-\beta U - \beta V)$ where $A = \prod_a D_a^{N_a}/N_a!$ it has the minimum[2].

For the f_q^0 it holds the following:

$$k_B T \ln f_q^0(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) = k_B T \ln A/Z_{N_1 \dots N_M} - U - V \tag{3.32}$$

Putting this to the definition of the free energy (3.31) we have the following:

$$\mathcal{F}[\mathbf{f_q^0}] = \int f_q^0(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) \left(k_B T \ln A/Z_{N_1 \dots N_M} - U - V + U + V\right) d\mathbf{r}^{[N]} d\boldsymbol{\theta}^{[N]} = k_B T \ln A/Z_{N_1 \dots N_M} \tag{3.33}$$

Using this formula we can find the difference $\Delta\mathcal{F} = \mathcal{F}[\mathbf{f_q}] - \mathcal{F}[\mathbf{f_q^0}]$. We obtain the following:

$$\Delta\mathcal{F} = \int f_q \left(k_B T \ln f_q + k_B T \ln \frac{A}{Z_{N_1 \dots N_M}} - k_B T \ln f_q^0\right) d\mathbf{r}^{[N]} d\boldsymbol{\theta}^{[N]} - k_B T \ln \frac{A}{Z_{N_1 \dots N_M}} \tag{3.34}$$

where $U + V$ was changed to $kT \ln A/Z - kT \ln f_q^0$ as a result of (3.32). Considering, that due to normalization $\int f_q = \int f_q^0 = 1$, we have the following

$$\Delta\mathcal{F} = k_B T \int \left(f_q \ln \frac{f_q}{f_q^0} + f_q - f_q^0\right) d\mathbf{r}^{[N]} d\boldsymbol{\theta}^{[N]} = k_B T \int f_q^0 \left(\frac{f_q}{f_q^0} \ln \frac{f_q}{f_q^0} + \frac{f_q}{f_q^0} - 1\right) d\mathbf{r}^{[N]} d\boldsymbol{\theta}^{[N]} \tag{3.35}$$

The integrand is always non-negative. To prove it we denote $f_q/f_q^0 \equiv 1/B$. We have $\frac{f_q}{f_q^0} \ln \frac{f_q}{f_q^0} + \frac{f_q}{f_q^0} - 1 = 1/B \ln 1/B + 1/B - 1 = 1/B(-\ln B + 1 - B)$. It is known, that for any $B > 0$ it holds that $\ln B \leq B - 1$, thus considering that $f_q, f_q^0 \geq 0$ we conclude that the integrand is always non-negative and $\Delta\mathcal{F} \geq 0$.

3.6 Free energy functional of the ideal gas in an external field

One of the basic assumptions of the density functional theory is that the free energy functional can be expressed as a functional of one-particle densities $\rho^1(\mathbf{r}^1, \boldsymbol{\theta}^1), \dots, \rho^M(\mathbf{r}^M, \boldsymbol{\theta}^M)$ of the particles of types $1, \dots, M$ correspondingly. In the ideal gas the particles do not interact to each other, so the probability to find the system in a certain configuration is a product of probabilities to find a particle at certain position. Considering the normalization of $\rho^a(\mathbf{r}, \boldsymbol{\theta})$ we can write the following:

$$f_q(\mathbf{r}^N, \boldsymbol{\theta}^N) = \prod_{a=1}^{M} \prod_{i=1}^{N_a} \frac{\rho^a(\mathbf{r}_i^a, \boldsymbol{\theta}_i^a)}{N_a} \tag{3.36}$$

[2] In this section we omit the arguments $(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]})$ of f_q, f_q^0 for the sake of simplicity.

Putting this to the definition of the free energy functional (3.31) and considering that for ideal gas in external field $U = 0$ we have the following:

$$\mathcal{F}[\boldsymbol{\rho^1}, \ldots, \boldsymbol{\rho^M}] = S_1 + S_2 + S_3 \tag{3.37}$$

where summands S_1, S_2, S_3 are defined in the following way:

$$S_1 = \sum_{a=1}^{M} k_B T \ln \frac{N_a!}{D_a^{N_a}} \tag{3.38}$$

$$S_2 = k_B T \int \prod_{b=1}^{M} \prod_{j=1}^{N_b} \frac{\rho^b(\mathbf{r}_j^b, \boldsymbol{\theta}_j^b)}{N_b} \ln \prod_{a=1}^{M} \prod_{i=1}^{N_a} \frac{\rho^a(\mathbf{r}_i^a, \boldsymbol{\theta}_i^a)}{N_a} d\mathbf{r}^{[N]} d\boldsymbol{\theta}^{[N]} \tag{3.39}$$

$$S_3 = \int \prod_{b=1}^{M} \prod_{j=1}^{N_b} \frac{\rho^b(\mathbf{r}_j^b, \boldsymbol{\theta}_j^b)}{N_b} \sum_{a=1}^{M} \sum_{i=1}^{N_a} \phi^a(\mathbf{r}_i^a, \boldsymbol{\theta}_i^a) d\mathbf{r}^{[N]} d\boldsymbol{\theta}^{[N]} \tag{3.40}$$

Using the Stirling's approximation (2.46) we rewrite the first summand in the following way:

$$S_1 \approx k_B T \sum_{a=1}^{M} (N_a \ln N_a - N_a - N_a \ln D_a) = k_B T \sum_{a=1}^{M} N_a (\ln \frac{N_a}{D_a} - 1) \tag{3.41}$$

Considering the normalization $\int \rho^a = N_a$ the first summand can be re-written in the following way:

$$S_1 \approx k_B T \sum_{a=1}^{M} \int \rho^a(\mathbf{r}, \boldsymbol{\theta}) \left(\ln \frac{N_a}{D_a} - 1 \right) d\mathbf{r} d\boldsymbol{\theta} \tag{3.42}$$

Considering that the logarithm of the product is the sum of logarithms we can re-write the second summand in the following way:

$$S_2 = \sum_{a=1}^{M} \sum_{i=1}^{N_a} k_B T \int d\mathbf{r}_i^a d\boldsymbol{\theta}_i^a \frac{\rho^a(\mathbf{r}_i^a, \boldsymbol{\theta}_i^a)}{N_a} \ln \frac{\rho^a(\mathbf{r}_i^a, \boldsymbol{\theta}_i^a)}{N_a} \int \prod_{j \neq i} \frac{\rho^a(\mathbf{r}_j^a, \boldsymbol{\theta}_j^a)}{N_a} \prod_{b \neq a} \prod_{k=1}^{N_b} \frac{\rho^b(\mathbf{r}_k^b, \boldsymbol{\theta}_k^b)}{N_b} d\mathbf{r}^{[N]} / d\mathbf{r}_i^a d\boldsymbol{\theta}^{[N]} / d\boldsymbol{\theta}_i^a \tag{3.43}$$

where $d\mathbf{r}^{[N]}/d\mathbf{r}_i^a d\boldsymbol{\theta}^{[N]}/d\boldsymbol{\theta}_i^a$ means integration over all coordinates except $(\mathbf{r}_i^a, \boldsymbol{\theta}_i^a)$. From the normalization $\int \rho^a/N_a = 1$ we conclude that the integration over $d\mathbf{r}^{[N]}/d\mathbf{r}_i^a d\boldsymbol{\theta}^{[N]}/d\boldsymbol{\theta}_i^a$ gives unity. Also, considering that all particles of the same type are indistinguishable, we can substitute the sum over i by the multiplication by N_a. Thus we have the following expression for the second summand:

$$S_2 = \sum_{a=1}^{M} \int \rho^a(\mathbf{r}, \boldsymbol{\theta}) \ln \frac{\rho^a(\mathbf{r}, \boldsymbol{\theta})}{N_a} d\mathbf{r} d\boldsymbol{\theta} \tag{3.44}$$

Similarly, the third summand (3.40) can be expressed in the following way:

$$\begin{split} S_3 = \\ \sum_{a=1}^{M}\sum_{i=1}^{M}\int d\mathbf{r}_i^a d\boldsymbol{\theta}_i^a \frac{\rho^a(\mathbf{r}_i^a,\boldsymbol{\theta}_i^a)}{N_a}\phi^a(\mathbf{r}_i^a,\boldsymbol{\theta}_i^a)\int\prod_{j\neq i}\frac{\rho^a(\mathbf{r}_j^a,\boldsymbol{\theta}_j^a)}{N_a}\prod_{b\neq a}\prod_{k=1}^{N_b}\frac{\rho^b(\mathbf{r}_k^b,\boldsymbol{\theta}_k^b)}{N_b}d\mathbf{r}^{[N]}/d\mathbf{r}_i^a d\boldsymbol{\theta}^{[N]}/d\boldsymbol{\theta}_i^a \end{split} \tag{3.45}$$

Considering the normalization of the ρ^a and changing sum over N_a identical particles to multiplication we have the following expression:

$$S_3 = \sum_{a=1}^{N_a}\int \rho^a(\mathbf{r},\boldsymbol{\theta})\phi^a(\mathbf{r},\boldsymbol{\theta})d\mathbf{r}d\boldsymbol{\theta} \tag{3.46}$$

Combining the expressions for S_1, S_2, S_3 we have:

$$\begin{split} \mathcal{F}[\boldsymbol{\rho}^{\mathbf{1}},\ldots,\boldsymbol{\rho}^{\boldsymbol{M}}] = \\ \sum_{a=1}^{M} k_BT\int \rho^a(\mathbf{r},\boldsymbol{\theta})\left(\ln\frac{N_a}{D_a} - 1 + \ln\frac{\rho^a(\mathbf{r},\boldsymbol{\theta})}{N_a}\right)d\mathbf{r}d\boldsymbol{\theta} + \sum_{a=1}^{N_a}\int \rho^a(\mathbf{r},\boldsymbol{\theta})\phi^a(\mathbf{r},\boldsymbol{\theta})d\mathbf{r}d\boldsymbol{\theta} \end{split} \tag{3.47}$$

Considering that the sum of logarithms is a logarithm of product we can cancel N_a and write the final expression for the free energy functional of ideal gas:

$$\mathcal{F}^{\text{id}}[\boldsymbol{\rho}^{\mathbf{1}},\ldots,\boldsymbol{\rho}^{\boldsymbol{M}}] = \sum_{a=1}^{M} k_BT\int \rho^a(\mathbf{r},\boldsymbol{\theta})\left(\ln\frac{\rho^a(\mathbf{r},\boldsymbol{\theta})}{D_a} - 1\right)d\mathbf{r}d\boldsymbol{\theta} + \sum_{a=1}^{M}\int \rho^a(\mathbf{r},\boldsymbol{\theta})\phi^a(\mathbf{r},\boldsymbol{\theta})d\mathbf{r}d\boldsymbol{\theta} \tag{3.48}$$

In a general case of non-ideal gas the free energy can be expressed as a sum of ideal and exchange parts:

$$\mathcal{F}[\boldsymbol{\rho}^{\mathbf{1}},\ldots,\boldsymbol{\rho}^{\boldsymbol{M}}] = \mathcal{F}^{\text{id}}[\boldsymbol{\rho}^{\mathbf{1}},\ldots,\boldsymbol{\rho}^{\boldsymbol{M}}] + \mathcal{F}^{\text{ex}}[\boldsymbol{\rho}^{\mathbf{1}},\ldots,\boldsymbol{\rho}^{\boldsymbol{M}}] \tag{3.49}$$

3.7 Functional derivatives

Let us consider a linear functional $\mathcal{F}[\boldsymbol{\rho}]$ which maps the function $\boldsymbol{\rho} : \mathbb{R}^3 \to \mathbb{R}$ to the real space. Let us consider the case then the functional $\mathcal{F}$ can be represented in the following way:

$$\mathcal{F}[\boldsymbol{\rho}] = \int_{\mathbb{R}^3} A[\boldsymbol{\rho}](\mathbf{r})\rho(\mathbf{r})d\mathbf{r} \tag{3.50}$$

where $A[\boldsymbol{\rho}] : \mathbb{R}^3 \to \mathbb{R}$ is defined for each $\boldsymbol{\rho}$. In that case the functional is called *differentiable* and the expression $A[\boldsymbol{\rho}]$ is called *functional derivative* of the functional $\mathcal{F}$. The functional derivative of the functional $\mathcal{F}$ with respect to the function $\boldsymbol{\rho}$ in the point $\mathbf{r}$ is denoted in the following way:

$$A[\boldsymbol{\rho}](\mathbf{r}) \equiv \frac{\delta\mathcal{F}}{\delta\rho(\mathbf{r})} \tag{3.51}$$

For non-linear functionals the functional derivative can be introduced by linearization of the functionals in some neighborhood. Let now $\mathcal{F}$ be a non-linear functional and there exists such neighborhood of a function $\boldsymbol{\rho_0}$ that for any function $\boldsymbol{\rho}$ from this neighborhood the following relation holds:

$$\mathcal{F}[\boldsymbol{\rho}] = \mathcal{F}[\boldsymbol{\rho_0}] + \int_{\mathbb{R}^3} A[\boldsymbol{\rho_0}](\mathbf{r})\,(\rho(\mathbf{r}) - \rho_0(\mathbf{r}))\,d\mathbf{r} + o(||\boldsymbol{\rho} - \boldsymbol{\rho_0}||) \tag{3.52}$$

where $||\cdot||$ is some norm of the functions defined on $\mathbb{R}^3$. In that case the functional $\mathcal{F}$ is called *differentiable* in the point $\boldsymbol{\rho_0}$ and $A[\boldsymbol{\rho_0}]$ is called the *functional derivative*. We should note that more strict and general definition of the functional derivatives can be given using the concepts of Fréchet and Gâteaux derivatives. More information can be found in Ref. [104].

Many properties of functional derivatives can be determined using the properties of the partial derivatives. Let us consider the case when the functional $\mathcal{F}$ is defined only for the piecewise constant functions. Let additionally these piecewise constant functions have final support $V \subset \mathbb{R}^3$. Let the set V is divided to the non-intersecting subsets $V_1, \ldots, V_N$. We define the functional $\mathcal{F}$ for all functions $\rho(\mathbf{r})$ which have constant values on the sets V_i. We choose the points $\mathbf{r_1}, \ldots, \mathbf{r_N}$ in such a way that $\mathbf{r_i} \in V_i, \quad i = 1, \ldots, N$. In that case expression (3.52) can be written in the following way:

$$\mathcal{F}[\boldsymbol{\rho}] = \mathcal{F}[\boldsymbol{\rho_0}] + \sum_{i=1}^{N} \frac{\delta \mathcal{F}}{\delta \rho_0(\mathbf{r_i})} \mu(V_i)\left(\rho_i - \rho_i^0\right) + o(||\boldsymbol{\rho} - \boldsymbol{\rho_0}||) \tag{3.53}$$

where $\mu(V_i)$ is the volume of the set V_i, $\rho_i = \rho(\mathbf{r_i})$, $\rho_i^0 = \rho_0(\mathbf{r_i})$. Using the properties of the partial derivatives we come to the following relation:

$$\frac{\delta \mathcal{F}}{\delta \rho_0(\mathbf{r_i})} \mu(V_i) = \frac{\partial \mathcal{F}}{\partial \rho_i^0} \tag{3.54}$$

Using this relation when $N \to \infty$ it is possible to obtain the properties of the functional derivatives.

For example, let us consider one of the simplest functional $\mathcal{F}[\boldsymbol{\rho}] \equiv \rho(\mathbf{r_0})$ where $\mathbf{r_0}$ is some given point. Our aim is to find the functional derivative $\delta \mathcal{F}/\delta \rho(\mathbf{r}) \equiv \delta \rho(\mathbf{r_0})/\delta \rho(\mathbf{r})$. Let us consider a case of the piecewise constant functions. Let $\mathbf{r_0} \in V_1$, $\mathbf{r} \in V_i$. From (3.54) we have the following:

$$\frac{\delta \rho(\mathbf{r_0})}{\delta \rho(\mathbf{r})} \mu(V_i) = \frac{\partial \rho_i^0}{\partial \rho_i} = \delta_{1i} \tag{3.55}$$

where δ_{1i} is the Kronecker delta. Now we consider the case $N \to \infty$, $\mu(V_i) \to 0$. We have the following relation:

$$\frac{\delta \rho(\mathbf{r_0})}{\delta \rho(\mathbf{r})} = \lim_{\mu(V_i) \to 0} \frac{\delta_{1i}}{\mu(V_i)} = \delta(\mathbf{r} - \mathbf{r_0}) \tag{3.56}$$

where $\delta(\mathbf{r}-\mathbf{r_0})$ is the Dirac delta function.

Other properties of the functional derivatives can be obtained using the similar procedure. The most of the properties of the partial derivatives are also valid for the functional derivatives if one changes symbols ∂ to δ and summation to the integration over $\mathbb{R}^3$.

3.8 Direct correlation functions

In the density functional theory responses of the free energy functionals to the local change of the density play an important role. To characterize this response one needs to know the functional derivatives of the free energy functional with respect to the density. Let us find the functional derivative of the free energy functional (3.49) with respect to the density of particles of type a in the point $(\mathbf{r_1}, \boldsymbol{\theta_1})$. Using the properties of functional derivatives we write the following:

$$\frac{\delta \mathcal{F}}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} = \sum_{b=1}^{M} \int \left(k_B T \frac{\delta \rho^b(\mathbf{r}, \boldsymbol{\theta})}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} + \frac{\delta \rho^b(\mathbf{r}, \boldsymbol{\theta})}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} \phi^a(\mathbf{r}, \boldsymbol{\theta}) \right) d\mathbf{r} d\boldsymbol{\theta} + \frac{\delta \mathcal{F}^{\mathrm{ex}}}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} \tag{3.57}$$

From the properties of the functional derivatives it we conclude that the following relation holds:

$$\frac{\delta \rho^b(\mathbf{r}, \boldsymbol{\theta})}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} = \delta_{ab} \delta(\mathbf{r_1} - \mathbf{r}, \boldsymbol{\theta_1} - \boldsymbol{\theta}) \tag{3.58}$$

where $\delta(\mathbf{r_1}-\mathbf{r}, \boldsymbol{\theta_1}-\boldsymbol{\theta}) = \delta(\mathbf{r_1}-\mathbf{r})\delta(\boldsymbol{\theta_1}-\boldsymbol{\theta})$ Using this relation we come to the following relation:

$$\frac{\delta \mathcal{F}}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} = k_B T \ln \frac{\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})}{D_a} + \phi^a(\mathbf{r_1}, \boldsymbol{\theta_1}) + \frac{\delta \mathcal{F}^{\mathrm{ex}}}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} \tag{3.59}$$

We define the *direct correlation function* $c^a(\mathbf{r_1}, \boldsymbol{\theta})$ in the following way:

$$c^a(\mathbf{r_1}, \boldsymbol{\theta}) \equiv -\beta \frac{\delta \mathcal{F}^{\mathrm{ex}}}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} \tag{3.60}$$

As it was proven in the section 3.5 the free energy functional reaches its minimum at the equilibrium density under constrains of the density normalization which can be written as $\int \rho^a = N_a$, $a = 1, \ldots, M$. To find this minimum we use the Lagrange multipliers method for functionals [104]. The Lagrange functional of the system can be written in the following way:

$$L[\boldsymbol{\rho^1}, \ldots, \boldsymbol{\rho^M}, \mu_1, \ldots, \mu_M] = \mathcal{F}[\boldsymbol{\rho^1}, \ldots, \boldsymbol{\rho^M}] - \sum_{a=1}^{M} \mu_a \left(\int \rho^a(\mathbf{r}, \boldsymbol{\theta}) d\mathbf{r} d\boldsymbol{\theta} - N_a \right) \tag{3.61}$$

The necessary condition of the minimum is $\delta L / \delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1}) = 0$. So we have the following:

$$\frac{\delta L}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} = \frac{\delta \mathcal{F}}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} - \mu_a = 0 \tag{3.62}$$

Using (3.59) and (3.60) we obtain the following expression for the direct correlation function:

$$c^a(\mathbf{r_1}, \boldsymbol{\theta_1}) = -\beta\mu_a + \beta\phi^a(\mathbf{r_1}, \boldsymbol{\theta_1}) + \ln \frac{\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})}{D_a} \tag{3.63}$$

Expressing ρ^a from the above relation we have the following:

$$\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1}) = D_a e^{\beta\mu_a} e^{-\beta\phi^a(\mathbf{r_1}, \boldsymbol{\theta_1}) + c^a(\mathbf{r_1}, \boldsymbol{\theta_1})} = \rho_0 e^{-\beta\phi^a(\mathbf{r_1}, \boldsymbol{\theta_1}) + c^a(\mathbf{r_1}, \boldsymbol{\theta_1})} \tag{3.64}$$

where $\rho_0 \equiv D_a e^{\beta\mu_a}$ is the uniform bulk density when there are no external force acting on the system.

We can also introduce multi-particle direct correlation functions which are proportional to the higher derivatives of the free energy functional. The most interesting in practice is two-particle direct correlation function $c^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ which is defined in the following way:

$$c^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \equiv \frac{\delta c^a(\mathbf{r_1}, \boldsymbol{\theta_1})}{\delta \rho^b(\mathbf{r_2}, \boldsymbol{\theta_2})} = -\beta \frac{\delta^2 \mathcal{F}}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1}) \delta \rho^b(\mathbf{r_2}, \boldsymbol{\theta_2})} \tag{3.65}$$

Taking the functional derivative over $\rho^b(\mathbf{r_2}, \boldsymbol{\theta_2})$ from the expression (3.63) and using the properties of the functional derivatives we express the two particle function at the equilibrium state in the following way:

$$c^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \beta \frac{\delta \phi^a(\mathbf{r_1}, \boldsymbol{\theta_1})}{\delta \rho^b(\mathbf{r_2}, \boldsymbol{\theta_2})} + \frac{\delta_{ab}\delta(\mathbf{r_1} - \mathbf{r_2}, \boldsymbol{\theta_1} - \boldsymbol{\theta_2})}{\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} \tag{3.66}$$

From this expression we can find the derivative $\delta\phi^a(\mathbf{r_1}, \boldsymbol{\theta_1})/\delta\rho^b(\mathbf{r_2}, \boldsymbol{\theta_2})$:

$$\frac{\delta \phi^a(\mathbf{r_1}, \boldsymbol{\theta_1})}{\delta \rho^b(\mathbf{r_2}, \boldsymbol{\theta_2})} = k_B T \left(c^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) - \frac{\delta_{ab}\delta(\mathbf{r_1} - \mathbf{r_2}, \boldsymbol{\theta_1} - \boldsymbol{\theta_2})}{\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} \right) \tag{3.67}$$

3.9 Ornstein-Zernike equation

In this section I adapt the derivation of OZ equation from Ref. [105] to the case of multi-component solutions. The free energy functional can be regarded as a functional of the external field (3.26). In that case we can find a functional derivative of $\mathcal{F}$ with respect to ϕ^a. Using the definition of the Helmholtz free energy (2.79) we can write the following:

$$\frac{\delta \mathcal{F}}{\delta \phi^a(\mathbf{r}, \boldsymbol{\theta})} = -k_B T \frac{\delta \ln Z_{N_1 \ldots N_M}}{\delta \phi^a(\mathbf{r}, \boldsymbol{\theta})} = \frac{-k_B T}{Z_{N_1 \ldots N_M}} \frac{\delta Z_{N_1 \ldots N_M}}{\delta \phi^a(\mathbf{r}, \boldsymbol{\theta})} \tag{3.68}$$

Using the relation (3.14) between the partition function $Z_{N_1 \ldots N_M}$ and the configuration integral $Q_{N_1 \ldots N_M}$ we come to the following relation:

$$\frac{\delta \mathcal{F}}{\delta \phi^a(\mathbf{r}, \boldsymbol{\theta})} = \frac{-k_B T}{Q_{N_1 \ldots N_M}} \frac{\delta Q_{N_1 \ldots N_M}}{\delta \phi^a(\mathbf{r}, \boldsymbol{\theta})} \tag{3.69}$$

Using the definition of the configuration integral (3.13) and properties of functional derivatives we have the following:

$$\frac{\delta\mathcal{F}}{\delta\phi^a(\mathbf{r_1},\boldsymbol{\theta_1})}=\frac{-k_BT}{Q_{N_1\ldots N_M}}\int e^{-\beta U(\mathbf{r}^{[N]},\boldsymbol{\theta}^{[N]})-\beta\sum_{b=1}^{M}\sum_{i=1}^{N_b}\phi^b(\mathbf{r}_i^b,\boldsymbol{\theta}_i^b)}(-\beta)\sum_{b=1}^{M}\sum_{i=1}^{N_b}\frac{\delta\phi^b(\mathbf{r}_i^b,\boldsymbol{\theta}_i^b)}{\delta\phi^a(\mathbf{r_1},\boldsymbol{\theta_1})}d\mathbf{r}^{[N]}d\boldsymbol{\theta}^{[N]} \tag{3.70}$$

Using that $\delta\phi^b(\mathbf{r}_i^b,\boldsymbol{\theta}_i^b)/\delta\phi^a(\mathbf{r_1},\boldsymbol{\theta_1})=\delta_{ab}\delta(\mathbf{r}_i^b-\mathbf{r_1},\boldsymbol{\theta}_i^b-\boldsymbol{\theta_1})$ we have the following relation:

$$\frac{\delta\mathcal{F}}{\delta\phi^a(\mathbf{r_1},\boldsymbol{\theta_1})}=\int\frac{e^{-\beta U(\mathbf{r}^{[N]},\boldsymbol{\theta}^{[N]})-\beta\sum_{b=1}^{M}\sum_{i=1}^{N_b}\phi^b(\mathbf{r}_i^b,\boldsymbol{\theta}_i^b)}}{Q_{N_1\ldots N_M}}\sum_{i=1}^{N_a}\delta(\mathbf{r}_i^a-\mathbf{r_1},\boldsymbol{\theta}_i^a-\boldsymbol{\theta_1})d\mathbf{r}^{[N]}d\boldsymbol{\theta}^{[N]} \tag{3.71}$$

Using the representation (3.24) of ρ^a as an ensemble average of the δ-function we have:

$$\frac{\delta\mathcal{F}}{\delta\phi^a(\mathbf{r_1},\boldsymbol{\theta_1})}=\left\langle\sum_{i=1}^{N_a}\delta(\mathbf{r}_i^a-\mathbf{r_1},\boldsymbol{\theta}_i^a-\boldsymbol{\theta_1})\right\rangle=\rho^a(\mathbf{r_1},\boldsymbol{\theta_1}) \tag{3.72}$$

Now, let us find how does ρ^a changes with the change of the external field in a selected point. From (3.69), (3.72) we have the following:

$$\frac{\delta\rho^a(\mathbf{r_1},\boldsymbol{\theta_1})}{\delta\phi^b(\mathbf{r_2},\boldsymbol{\theta_2})}=\frac{k_BT}{Q^2_{N_1\ldots N_M}}\frac{\delta Q_{N_1\ldots N_M}}{\delta\phi^a(\mathbf{r_1},\boldsymbol{\theta_1})}\frac{\delta Q_{N_1\ldots N_M}}{\delta\phi^b(\mathbf{r_2},\boldsymbol{\theta_2})}+\frac{-k_BT}{Q_{N_1\ldots N_M}}\frac{\delta^2 Q_{N_1\ldots N_M}}{\delta\phi^a(\mathbf{r_1},\boldsymbol{\theta_1})\delta\phi^b(\mathbf{r_2},\boldsymbol{\theta_2})} \tag{3.73}$$

Using the relation (3.72) and (3.69) we can immediately write the expression for the first summand:

$$\frac{k_BT}{Q^2_{N_1\ldots N_M}}\frac{\delta Q_{N_1\ldots N_M}}{\delta\phi^a(\mathbf{r_1},\boldsymbol{\theta_1})}\frac{\delta Q_{N_1\ldots N_M}}{\delta\phi^b(\mathbf{r_2},\boldsymbol{\theta_2})}=\beta\rho^a(\mathbf{r_1},\boldsymbol{\theta_1})\rho^b(\mathbf{r_2},\boldsymbol{\theta_2}) \tag{3.74}$$

To find the expression for the second summand we can use (3.68), (3.71) to find the derivative of $\delta Q/\delta\phi^a(\mathbf{r_1},\boldsymbol{\theta_1})$. We have the following:

$$\begin{aligned}&\frac{-k_BT}{Q_{N_1\ldots N_M}}\frac{\delta^2 Q_{N_1\ldots N_M}}{\delta\phi^a(\mathbf{r_1},\boldsymbol{\theta_1})\delta\phi^b(\mathbf{r_2},\boldsymbol{\theta_2})}=\\&\frac{\delta}{\delta\phi^b(\mathbf{r_2},\boldsymbol{\theta_2})}\int\frac{\exp(-\beta U(\mathbf{r}^{[N]},\boldsymbol{\theta}^{[N]}-\beta V(\mathbf{r}^{[N]},\boldsymbol{\theta}^{[N]}))}{Q_{N_1\ldots N_M}}\sum_{i=1}^{N_a}\delta(\mathbf{r}_i^a-\mathbf{r_1},\boldsymbol{\theta}_i^a-\boldsymbol{\theta_1})d\mathbf{r}^{[N]}d\boldsymbol{\theta}^{[N]}\end{aligned} \tag{3.75}$$

Using the properties of the functional derivatives we obtain the following expression:

$$\begin{aligned}&\frac{-k_BT}{Q_{N_1\ldots N_M}}\frac{\delta^2 Q_{N_1\ldots N_M}}{\delta\phi^a(\mathbf{r_1},\boldsymbol{\theta_1})\delta\phi^b(\mathbf{r_2},\boldsymbol{\theta_2})}=\\&-\beta\int\frac{\exp(-\beta U(\mathbf{r}^{[N]},\boldsymbol{\theta}^{[N]}-\beta V(\mathbf{r}^{[N]},\boldsymbol{\theta}^{[N]}))}{Q_{N_1\ldots N_M}}\sum_{i=1}^{N_a}\sum_{j=1}^{N_b}\delta(\mathbf{r}_i^a-\mathbf{r_1},\boldsymbol{\theta}_i^a-\boldsymbol{\theta_1})\delta(\mathbf{r}_j^b-\mathbf{r_2},\boldsymbol{\theta}_j^b-\boldsymbol{\theta_2})d\mathbf{r}^{[N]}d\boldsymbol{\theta}^{[N]}\\&=-\beta\left\langle\sum_{i=1}^{N_a}\sum_{j=1}^{N_b}\delta(\mathbf{r}_i^a-\mathbf{r_1},\boldsymbol{\theta}_i^a-\boldsymbol{\theta_1})\delta(\mathbf{r}_j^b-\mathbf{r_2},\boldsymbol{\theta}_j^b-\boldsymbol{\theta_2})\right\rangle\end{aligned} \tag{3.76}$$

We can consider two cases: 1) $a \equiv b$, $i = j$ and 2) $a \neq b$ or $i \neq j$. We can formally write the following:

$$\begin{aligned}&\left\langle \sum_{i=1}^{N_a} \sum_{j=1}^{N_b} \delta(\mathbf{r}_i^a - \mathbf{r_1}, \boldsymbol{\theta}_i^a - \boldsymbol{\theta_1})\delta(\mathbf{r}_j^b - \mathbf{r_2}, \boldsymbol{\theta}_j^b - \boldsymbol{\theta_2}) \right\rangle \\ &= \delta_{ab} \left\langle \sum_{i=1}^{N_a} \delta(\mathbf{r}_i^a - \mathbf{r_1}, \boldsymbol{\theta}_i^a - \boldsymbol{\theta_1})\delta(\mathbf{r}_i^b - \mathbf{r_2}, \boldsymbol{\theta}_i^b - \boldsymbol{\theta_2}) \right\rangle \\ &+ \left\langle \sum_{i=1}^{N_a} \sum_{j=1}^{N_b} (1 - \delta_{ab}\delta_{ij})\delta(\mathbf{r}_i^a - \mathbf{r_1}, \boldsymbol{\theta}_i^a - \boldsymbol{\theta_1})\delta(\mathbf{r}_j^b - \mathbf{r_2}, \boldsymbol{\theta}_j^b - \boldsymbol{\theta_2}) \right\rangle\end{aligned} \tag{3.77}$$

Comparing this to (3.24), (3.25) we conclude that the first summand is equal to $\delta(\mathbf{r_1} - \mathbf{r_2}, \boldsymbol{\theta_1} - \boldsymbol{\theta_2})\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})$, the second summand is equal to $\rho^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2})$. So we have the following:

$$\frac{-k_B T}{Q_{N_1 \ldots N_M}} \frac{\delta^2 Q_{N_1 \ldots N_M}}{\delta\phi^a(\mathbf{r_1}, \boldsymbol{\theta_1})\delta\phi^b(\mathbf{r_2}, \boldsymbol{\theta_2})} = -\beta \left(\delta(\mathbf{r_1} - \mathbf{r_2}, \boldsymbol{\theta_1} - \boldsymbol{\theta_2})\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1}) + \rho^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \right) \tag{3.78}$$

This together with the relations (3.73), (3.74) gives the following expression for the functional derivative $\delta\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})/\delta\phi^b(\mathbf{r_2}, \boldsymbol{\theta_2})$:

$$\frac{\delta\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})}{\delta\phi^b(\mathbf{r_2}, \boldsymbol{\theta_2})} = \beta \left(\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})\rho^b(\mathbf{r_2}, \boldsymbol{\theta_2}) - \delta_{ab}\delta(\mathbf{r_1} - \mathbf{r_2}, \boldsymbol{\theta_1} - \boldsymbol{\theta_2})\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1}) - \rho^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \right) \tag{3.79}$$

We consider now the case with no external field: $\phi^a(\mathbf{r}, \boldsymbol{\theta}) \to 0$. In that case the density will tend to the equilibrium density $\rho^a(\mathbf{r}, \boldsymbol{\theta}) \to$ const and using the definition of the pair density correlation function (3.23) we can write $\rho^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \rho_0^a \rho_0^b g^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ We introduce the *total correlation function* $h^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ in the following way:

$$h^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = g^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) - 1 \tag{3.80}$$

Then the equation (3.79) can be written in a simpler form:

$$\left.\frac{\delta\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})}{\delta\phi^b(\mathbf{r_2}, \boldsymbol{\theta_2})}\right|_{V_N=0} = -\beta\rho^a \left(\rho^b h^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) + \delta_{ab}\delta(\mathbf{r_1} - \mathbf{r_2}, \boldsymbol{\theta_1} - \boldsymbol{\theta_2}) \right) \tag{3.81}$$

Now, let us look at the derivative $\delta\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})/\delta\rho^b(\mathbf{r_2}, \boldsymbol{\theta_2})$. Using the properties of the functional derivatives we can write $\delta\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})/\delta\rho^b(\mathbf{r_2}, \boldsymbol{\theta_2}) = \delta_{ab}\delta(\mathbf{r_1} - \mathbf{r_2}, \boldsymbol{\theta_1} - \boldsymbol{\theta_2})$. Alternatively, we can write the same relation using the chain rule for functional derivatives:

$$\left.\frac{\delta\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})}{\delta\rho^b(\mathbf{r_2}, \boldsymbol{\theta_2})}\right|_{V_N=0} = \sum_{c=1}^{M} \int \left.\frac{\delta\rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})}{\delta\phi^c(\mathbf{r_3}, \boldsymbol{\theta_3})}\right|_{V_N=0} \cdot \left.\frac{\delta\phi^c(\mathbf{r_3}, \boldsymbol{\theta_3})}{\delta\rho^b(\mathbf{r_2}, \boldsymbol{\theta_2})}\right|_{V_N=0} d\mathbf{r_3} d\boldsymbol{\theta_3} \tag{3.82}$$

Putting here expressions (3.67) (3.81) for the derivatives $\delta\phi/\delta\rho$ and $\delta\rho/\delta\phi$ we have the following:

$$\begin{aligned}&\delta_{ab}\delta(\mathbf{r_1}-\mathbf{r_2},\boldsymbol{\theta_1}-\boldsymbol{\theta_2}) = \\ &\sum_{c=1}^{M}\int(-\beta)\rho^a\left(\rho^c h^{ac}(\mathbf{r_1},\mathbf{r_3},\boldsymbol{\theta_1},\boldsymbol{\theta_3})+\delta_{ac}\delta(\mathbf{r_1}-\mathbf{r_3},\boldsymbol{\theta_1}-\boldsymbol{\theta_3})\right)\times \\ &\times k_BT\left(c^{cb}(\mathbf{r_3},\mathbf{r_2},\boldsymbol{\theta_3},\boldsymbol{\theta_2})-\frac{\delta_{cb}\delta(\mathbf{r_3}-\mathbf{r_2},\boldsymbol{\theta_3}-\boldsymbol{\theta_2})}{\rho^c}\right)d\mathbf{r_3}d\boldsymbol{\theta_3}\end{aligned} \tag{3.83}$$

Opening the brackets we come to the following relation:

$$\begin{aligned}&\delta_{ab}\delta(\mathbf{r_1}-\mathbf{r_2},\boldsymbol{\theta_1}-\boldsymbol{\theta_2}) = \\ &-\sum_{c=1}^{M}\rho^a\rho^c\int h^{ac}(\mathbf{r_1},\mathbf{r_3},\boldsymbol{\theta_1},\boldsymbol{\theta_3})c^{cb}(\mathbf{r_3},\mathbf{r_2},\boldsymbol{\theta_3},\boldsymbol{\theta_2})d\mathbf{r_3}d\boldsymbol{\theta_3}-\rho^a c^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2}) \\ &+\rho^a h^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})+\delta_{ab}\delta(\mathbf{r_1}-\mathbf{r_2},\boldsymbol{\theta_1}-\boldsymbol{\theta_2})\end{aligned} \tag{3.84}$$

Canceling $\delta_{ab}\delta(\mathbf{r_1}-\mathbf{r_2},\boldsymbol{\theta_1}-\boldsymbol{\theta_2})$ and dividing both parts by ρ^a we come to the set of *Ornstein-Zernike* (OZ) equations:

$$\begin{aligned}&h^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2}) = c^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})+\sum_{c=1}^{M}\rho^c\int h^{ac}(\mathbf{r_1},\mathbf{r_3},\boldsymbol{\theta_1},\boldsymbol{\theta_3})c^{cb}(\mathbf{r_3},\mathbf{r_2},\boldsymbol{\theta_3},\boldsymbol{\theta_2})d\mathbf{r_3}d\boldsymbol{\theta_3} \\ &a,b=1,\ldots,M\end{aligned} \tag{3.85}$$

We note, that because the functions h^{ab}, c^{ab}, are symmetric with respect to their arguments, we can rewrite the OZ-equations in the following way:

$$h^{ba}(\mathbf{r_2},\mathbf{r_1},\boldsymbol{\theta_2},\boldsymbol{\theta_1}) = c^{ba}(\mathbf{r_2},\mathbf{r_1},\boldsymbol{\theta_2},\boldsymbol{\theta_1})+\sum_{c=1}^{M}\rho^c\int h^{ca}(\mathbf{r_3},\mathbf{r_1},\boldsymbol{\theta_3},\boldsymbol{\theta_1})c^{bc}(\mathbf{r_2},\mathbf{r_3},\boldsymbol{\theta_2},\boldsymbol{\theta_3})d\mathbf{r_3}d\boldsymbol{\theta_3} \tag{3.86}$$

Changing the labels $a\leftrightarrow b$, $\mathbf{r_1}\leftrightarrow\mathbf{r_2}$, $\boldsymbol{\theta_1}\leftrightarrow\boldsymbol{\theta_2}$ we come to the familiar form of the OZ equations:

$$h^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2}) = c^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})+\sum_{c=1}^{M}\rho^c\int c^{ac}(\mathbf{r_1},\mathbf{r_3},\boldsymbol{\theta_1},\boldsymbol{\theta_3})h^{cb}(\mathbf{r_3},\mathbf{r_2},\boldsymbol{\theta_3},\boldsymbol{\theta_2})d\mathbf{r_3}d\boldsymbol{\theta_3} \tag{3.87}$$

We also note, that in the uniform fluids the functions c^{ab}, h^{ab} depend only on relative shifts between the particles but not on positions of particles themselves. Thus, the OZ equation for uniform fluids can be written in the following form:

$$h^{ab}(\mathbf{r_2}-\mathbf{r_1},\boldsymbol{\theta_2}-\boldsymbol{\theta_1}) = c^{ab}(\mathbf{r_2}-\mathbf{r_1},\boldsymbol{\theta_2}-\boldsymbol{\theta_1})+\sum_{c=1}^{M}\rho^c\int c^{ac}(\mathbf{r_3}-\mathbf{r_1},\boldsymbol{\theta_3}-\boldsymbol{\theta_1})h^{cb}(\mathbf{r_2}-\mathbf{r_3},\boldsymbol{\theta_2}-\boldsymbol{\theta_3})d\mathbf{r_3}d\boldsymbol{\theta_3} \tag{3.88}$$

3.10 Closure relation

Although the Ornstein-Zernike equations (3.87) are fundamental equations of the integral equation theory of liquids they are not enough to calculate the correlation functions. Equations (3.87) give only M^2 relations for $2M^2$ unknown correlation functions h^{ab}, c^{ab}. To make the system of equations solvable we need to find yet M^2 independent relations between the functions h^{ab}, c^{ab}. Such relations can be found using the density functional theory. Let us consider the system with the origin connected to one of the particles. Let this particle has type a. We denote as $(\mathbf{r_0}, \boldsymbol{\theta_0}) \equiv (\mathbf{0}, \mathbf{0})$ the coordinates of this particle. Let the particles in the system interact via the pairwise-additive potential (4.4). In the coordinate system associated with one of the particles we can consider that other particles move in the external field generated by this particle. We denote this external field as $V(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]})$. It can be expressed in the following form:

$$V(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) = \sum_{b=1}^{M} \sum_{i=1}^{N_b} u^{ab}(\mathbf{r}_i^b, \boldsymbol{\theta}_j^b) \tag{3.89}$$

We note, that this expression becomes the same as the expression (3.26) if we substitute $\phi^b(\mathbf{r}, \boldsymbol{\theta})$ with $u^{ab}(\mathbf{r}, \boldsymbol{\theta})$. In the coordinate system associated with one of the particles the one-particle distribution functions $\rho^b(\mathbf{r}, \boldsymbol{\theta})$ can be written in the following way:

$$\rho^b(\mathbf{r}, \boldsymbol{\theta}) = \rho_0^b g^{ab}(\mathbf{r_1} - \mathbf{r_0}, \boldsymbol{\theta_1} - \boldsymbol{\theta_0}) \tag{3.90}$$

Using the relation (3.64) we have the following:

$$g^{ab}(\mathbf{r}, \boldsymbol{\theta}) = e^{-\beta u^{ab}(\mathbf{r}, \boldsymbol{\theta}) + c^b(\mathbf{r}, \boldsymbol{\theta})} \tag{3.91}$$

Let us consider the process of smooth transition from the system without inter-particle interaction to the given system [3]. Let the one-particle density during this transition changes by the following low:

$$\rho^b(\mathbf{r}, \boldsymbol{\theta}; \lambda) = \rho_0^b + \lambda \Delta\rho^b(\mathbf{r}, \boldsymbol{\theta}) \tag{3.92}$$

where $\Delta\rho^b(\mathbf{r}, \boldsymbol{\theta}) = \rho^b(\mathbf{r}, \boldsymbol{\theta}) - \rho_0^b$. The function $\rho^b(\mathbf{r}, \boldsymbol{\theta}; 0)$ corresponds to the initial uniform state and the function $\rho^b(\mathbf{r}, \boldsymbol{\theta}; 1)$ to the final state. Also, using the relation (3.90) and the definition of the total correlation function (3.80) we can write the following:

$$\Delta\rho^b(\mathbf{r_1}, \boldsymbol{\theta_1}) = \rho_0^b h^{ab}(\mathbf{r_0}, \mathbf{r_1}, \boldsymbol{\theta_0}, \boldsymbol{\theta_1}) \tag{3.93}$$

We denote as $c^b(\mathbf{r}, \boldsymbol{\theta}; \lambda)$ the direct correlation function which corresponds to the density $\rho^b(\mathbf{r}, \boldsymbol{\theta}; \lambda)$. According to the rules of differentiation and to the properties of the functional derivatives the

[3] Using the term *smooth transition* we suppose, that all the λ-dependent correlation and distribution functions are continuous with respect to λ. This in particular means that we can integrate the correlation functions over λ.

differential $dc^b(\mathbf{r},\boldsymbol{\theta};\lambda) = c(\mathbf{r},\boldsymbol{\theta};\lambda+d\lambda) - c(\mathbf{r},\boldsymbol{\theta};\lambda)$ can be written in the following way:

$$dc^b(\mathbf{r_1},\boldsymbol{\theta_1};\lambda) = \sum_{c=1}^{M}\int \frac{\delta c^b(\mathbf{r_1},\boldsymbol{\theta_1};\lambda)}{\delta\rho^c(\mathbf{r_3},\boldsymbol{\theta_3};\lambda)} d\rho^c(\mathbf{r_3},\boldsymbol{\theta_3};\lambda)d\mathbf{r_3}d\boldsymbol{\theta_3} \tag{3.94}$$

where $d\rho^c(\mathbf{r_3},\boldsymbol{\theta_3};\lambda) = \rho^c(\mathbf{r_3},\boldsymbol{\theta_3};\lambda+d\lambda) - \rho^c(\mathbf{r_3},\boldsymbol{\theta_3};\lambda)$. Using the relations (3.92), (3.65) and integrating over λ we have the following:

$$c^b(\mathbf{r_1},\boldsymbol{\theta_1}) = \sum_{c=1}^{M}\int_0^1 d\lambda \int c^{bc}(\mathbf{r_1},\mathbf{r_3},\boldsymbol{\theta_1},\boldsymbol{\theta_3};\lambda)\Delta\rho^c(\mathbf{r_3},\boldsymbol{\theta_3})d\mathbf{r_3}d\boldsymbol{\theta_3} \tag{3.95}$$

Here we use that $c^b(\mathbf{r},\boldsymbol{\theta};0) \equiv 0$, because the initial state corresponds to the ideal-gas. The relation (3.95) means that to calculate $c^b(\mathbf{r_1},\boldsymbol{\theta_1})$ for the final system one need to know functions c^{ab} for any λ. This in turn seriously complicates solution of the OZ equations. To simplify the closure relation different approximations are used. One of the widely used approximations is the *Hyper-Netted chain* (HNC) approximation. In the Hyper-netted chain approximation it is assumed that the functions c^{bc} do not depend on λ and can be substituted by the pair direct correlation function of the final system. In that case the integration over $d\lambda$ can be omitted. Using (3.93) we come to the following expression:

$$c^b(\mathbf{r_1},\boldsymbol{\theta_1}) = \sum_{c=1}^{M}\rho_0^c \int c^{bc}(\mathbf{r_1},\mathbf{r_3},\boldsymbol{\theta_1},\boldsymbol{\theta_3})h^{ca}(\mathbf{r_3},\mathbf{r_0},\boldsymbol{\theta_3},\boldsymbol{\theta_0})d\mathbf{r_3}d\boldsymbol{\theta_3} \tag{3.96}$$

Putting this expression to (3.91) and using the Ornstein-Zernike equation (3.87) we come to the HNC closure relation:

$$g^{ab}(\mathbf{r_{01}},\boldsymbol{\theta_{01}}) = e^{-\beta u(\mathbf{r_{01}},\boldsymbol{\theta_{01}})+h^{ab}(\mathbf{r_{01}},\boldsymbol{\theta_{01}})-c^{ab}(\mathbf{r_{01}},\boldsymbol{\theta_{01}})} \tag{3.97}$$

where we define $\mathbf{r_{01}} \equiv \mathbf{r_1} - \mathbf{r_0} \equiv \mathbf{r_1}$, $\boldsymbol{\theta_{01}} \equiv \boldsymbol{\theta_1} - \boldsymbol{\theta_0} \equiv \boldsymbol{\theta_1}$ to stress that the distances in the expression are relative to the selected particle. We note, that substituting (3.95) by (3.96) we neglect the dependency of the direct correlation function on λ. In a general case the closure relation is written in the following form:

$$g^{ab}(\mathbf{r_{01}},\boldsymbol{\theta_{01}}) = e^{-\beta u(\mathbf{r_{01}},\boldsymbol{\theta_{01}})+h^{ab}(\mathbf{r_{01}},\boldsymbol{\theta_{01}})-c^{ab}(\mathbf{r_{01}},\boldsymbol{\theta_{01}})+B^{ab}(\mathbf{r_{01}},\boldsymbol{\theta_{01}})} \tag{3.98}$$

where $B^{ab}(\mathbf{r_{01}},\boldsymbol{\theta_{01}})$ is a *Bridge* function defined in the following way:

$$B^{ab}(\mathbf{r_{01}},\boldsymbol{\theta_{01}}) = \sum_{c=1}^{M}\rho_0^c \int d\lambda \int h^{ac}(\mathbf{r_{03}},\boldsymbol{\theta_{03}})c^{bc}(\mathbf{r_{13}},\boldsymbol{\theta_{13}};\lambda)d\mathbf{r_3}d\boldsymbol{\theta_3} - h^{ab}(\mathbf{r_{01}},\boldsymbol{\theta_{01}}) + c^{ab}(\mathbf{r_{01}},\boldsymbol{\theta_{01}}) \tag{3.99}$$

3.11 Reference interaction site model

In a general case, solution of the six dimensional OZ equations (3.87) is a challenging problem from the computational point of view. Nowadays there are no algorithm implementations which solve the OZ equations in a general case for arbitrary molecules. Only few works exist which describe solution of the 6D OZ equations for simple small molecules, like water or ionic solutions. In practice simplification of the OZ equations are usually used. Probably the most popular of such simplifications is the Reference Interaction Site Model (RISM). The RISM was initially proposed by Chandler and Anderson [31] and was intensively investigated by the community until the more extended, 3DRISM model come to change it. The basic consideration of RISM theory is the assumption that the molecules in the system can be considered as the sets of sites (typically - atoms of the molecule). The main assumption of the RISM theory is that the molecular direct correlation function $c^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ can be expressed as the sum of spherically symmetric site-site correlation functions $c^{ab}_{\alpha'\beta'}$. Let us denote by $\mathbf{r_1}$, $\mathbf{r_2}$, $\ldots$ the centers of the molecules in the system, by $\mathbf{\hat{r}_1}$, $\mathbf{\hat{r}_2}$, $\ldots$ the coordinates of the sites of the molecules, by $\mathbf{d}^a_\alpha(\boldsymbol{\theta})$ shift of the site α of the molecule of type a with respect to the center of the molecule. This shift depends on the orientation of the molecule $\boldsymbol{\theta}$. If $\mathbf{r_i}$ is the absolute coordinate of the center of the i^{th} molecule of type a, $\mathbf{\hat{r}_i}$ is the coordinate of the site α of this molecule, then we can write $\mathbf{\hat{r}_i} = \mathbf{r_i} + d^a_\alpha(\boldsymbol{\theta_i})$, where $\boldsymbol{\theta_i}$ is the orientation of the molecule. We denote by $h^{ab}_{\alpha\beta}(\mathbf{\hat{r}_1}, \mathbf{\hat{r}_2})$ $c^{ab}_{\alpha'\beta'}(\mathbf{\hat{r}_1}, \mathbf{\hat{r}_2})$ the total and direct site-site correlation functions between the site α of molecule of type a and site β of molecule of type b. According to the assumptions of the RISM theory we can write the following relation for the direct correlation function $c^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2})$:

$$c^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \sum_{\alpha'\beta'} c^{ab}_{\alpha'\beta'}(\mathbf{\hat{r}'_1}, \mathbf{\hat{r}'_2}) \tag{3.100}$$

where the following relations hold:

$$\mathbf{\hat{r}'_1} = \mathbf{r_1} + \mathbf{d}^a_{\alpha'}(\boldsymbol{\theta_1}) \qquad \mathbf{\hat{r}'_2} = \mathbf{r_2} + \mathbf{d}^b_{\beta'}(\boldsymbol{\theta_2}) \tag{3.101}$$

The relation (3.100) can be re-written in the integral form where restrictions (3.101) are incorporated by introducing δ-functions into the integral representation. We have the following representation: [4]

$$c^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \sum_{\alpha'\beta'} \int \delta(\mathbf{r_1} + \mathbf{d}^a_{\alpha'}(\boldsymbol{\theta_1}) - \mathbf{\hat{r}'_1}) \delta(\mathbf{r_2} + \mathbf{d}^b_{\beta'}(\boldsymbol{\theta_2}) - \mathbf{\hat{r}'_2}) c^{ab}_{\alpha'\beta'}(\mathbf{\hat{r}'_1}, \mathbf{\hat{r}'_2}) d\mathbf{\hat{r}'_1} d\mathbf{\hat{r}'_2} \tag{3.102}$$

Total site-site correlation functions $h^{ab}_{\alpha\beta}(\mathbf{\hat{r}_1}, \mathbf{\hat{r}_2})$ are defined by averaging of the molecular total correlation function h^{ab} over the rotational degrees of freedom. Considering the restrictions

[4] Here and below to make the equations shorter we will sometimes omit multiple integral signs ($\int$). The number of integrations can be determined from the number of differentials under the integral.

(3.101) the definition can be written in the following way:

$$h^{ab}_{\alpha\beta}(\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2)=\frac{1}{\Omega^2}\int h(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})\delta(\mathbf{r_1}+\mathbf{d}^a_\alpha(\boldsymbol{\theta_1})-\hat{\mathbf{r}}_1)\delta(\mathbf{r_2}+\mathbf{d}^b_\beta(\boldsymbol{\theta_2})-\hat{\mathbf{r}}_2)d\mathbf{r_1}d\mathbf{r_2}d\boldsymbol{\theta_1}d\boldsymbol{\theta_2} \quad (3.103)$$

where $\Omega=\int d\boldsymbol{\theta}=4\pi^2$. Here we should note, that in principle we distinguish the density of the molecules ρ^a and the density of the molecule sites $\hat{\rho}^a$. The difference is due to the different normalization for these quantities. According to the definition (3.21) we have the following normalization for the molecular density:

$$\int \rho^a d\mathbf{r}d\boldsymbol{\theta}=\rho^a V\Omega=N_a \quad (3.104)$$

where V is the volume of the system. Thus we have $\rho^a=N_a/(V\Omega)$. However, for sites, which do not have angular degrees of freedom the normalization $\int\hat{\rho}^a d\mathbf{r}=N_a$ is usually used. This means that $\hat{\rho}^a=N_a/V=\rho^a\Omega$. Using these definitions we come to the following representation of the OZ equation (3.87):

$$h^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})=c^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})+\sum_{c=1}^{M}\frac{\hat{\rho}^c}{\Omega}\int c^{ac}(\mathbf{r_1},\mathbf{r_3},\boldsymbol{\theta_1},\boldsymbol{\theta_3})\cdot h^{cb}(\mathbf{r_3},\mathbf{r_2},\boldsymbol{\theta_3},\boldsymbol{\theta_2})d\mathbf{r_3}d\boldsymbol{\theta_3} \quad (3.105)$$

To make the expressions shorter, we will transform separately each summand in the OZ equation. Let us multiply both parts of the OZ equation (3.105) by $\Omega^{-2}\delta(\mathbf{r_1}+\mathbf{d}^a_\alpha(\boldsymbol{\theta_1})-\hat{\mathbf{r}}_1)\delta(\mathbf{r_2}+\mathbf{d}^b_\beta(\boldsymbol{\theta_2})-\hat{\mathbf{r}}_2)$ and integrate over $\mathbf{r_1}$, $\mathbf{r_2}$, $\boldsymbol{\theta_1}$, $\boldsymbol{\theta_2}$. The OZ equation will have the following form:

$$X=Y+Z \quad (3.106)$$

where X, Y, Z are defined with the following relations:

$$X=\frac{1}{\Omega^2}\int h^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})\delta(\mathbf{r_1}+\mathbf{d}^a_\alpha(\boldsymbol{\theta_1})-\hat{\mathbf{r}}_1)\delta(\mathbf{r_2}+\mathbf{d}^b_\beta(\boldsymbol{\theta_2})-\hat{\mathbf{r}}_2)d\mathbf{r_1}d\mathbf{r_2}d\boldsymbol{\theta_1}d\boldsymbol{\theta_2} \quad (3.107)$$

$$Y=\frac{1}{\Omega^2}\int c^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})\delta(\mathbf{r_1}+\mathbf{d}^a_\alpha(\boldsymbol{\theta_1})-\hat{\mathbf{r}}_1)\delta(\mathbf{r_2}+\mathbf{d}^b_\beta(\boldsymbol{\theta_2})-\hat{\mathbf{r}}_2)d\mathbf{r_1}d\mathbf{r_2}d\boldsymbol{\theta_1}d\boldsymbol{\theta_2} \quad (3.108)$$

$$\begin{aligned}Z=&\sum_{c=1}^{M}\frac{\hat{\rho}^c}{\Omega^3}\int c^{ac}(\mathbf{r_1},\mathbf{r_3},\boldsymbol{\theta_1},\boldsymbol{\theta_3})h^{cb}(\mathbf{r_3},\mathbf{r_2},\boldsymbol{\theta_3},\boldsymbol{\theta_2})\times\\&\times\delta(\mathbf{r_1}+\mathbf{d}^a_\alpha(\boldsymbol{\theta_1})-\hat{\mathbf{r}}_1)\delta(\mathbf{r_2}+\mathbf{d}^b_\beta(\boldsymbol{\theta_2})-\hat{\mathbf{r}}_2)d\mathbf{r_3}d\boldsymbol{\theta_3}d\mathbf{r_1}d\mathbf{r_2}d\boldsymbol{\theta_1}d\boldsymbol{\theta_2}\end{aligned} \quad (3.109)$$

The relation (3.107) coincides with the definition (3.103), thus X is defined with the following relation:

$$X\equiv h^{ab}_{\alpha\beta}(\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2) \quad (3.110)$$

Using the RISM assumption (3.102) in the relation (3.108) we have the following:

$$\begin{aligned}Y=&\sum_{\alpha'\beta'}\int\delta(\mathbf{r_1}+\mathbf{d}^a_{\alpha'}(\boldsymbol{\theta_1})-\hat{\mathbf{r}}'_1)\delta(\mathbf{r_2}+\mathbf{d}^b_{\beta'}(\boldsymbol{\theta_2})-\hat{\mathbf{r}}'_2)\cdot c^{ab}_{\alpha'\beta'}(\hat{\mathbf{r}}'_1,\hat{\mathbf{r}}'_2)\times\\&\times\delta(\mathbf{r_1}+\mathbf{d}^a_\alpha(\boldsymbol{\theta_1})-\hat{\mathbf{r}}_1)\delta(\mathbf{r_2}+\mathbf{d}^b_\beta(\boldsymbol{\theta_2})-\hat{\mathbf{r}}_2)d\mathbf{r_1}d\mathbf{r_2}d\boldsymbol{\theta_1}d\boldsymbol{\theta_2}d\hat{\mathbf{r}}'_1d\hat{\mathbf{r}}'_2\end{aligned} \quad (3.111)$$

Let us define the *intramolecular correlation function* $\omega_{\alpha\alpha'}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}'_\mathbf{1})$ in the following way:

$$\omega_{\alpha\alpha'}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}'_\mathbf{1})=\frac{1}{\Omega}\int\delta(\mathbf{r}_\mathbf{1}+\mathbf{d}^a_{\alpha'}(\boldsymbol{\theta}_\mathbf{1})-\hat{\mathbf{r}}'_\mathbf{1})\delta(\mathbf{r}_\mathbf{1}+\mathbf{d}^a_{\alpha}(\boldsymbol{\theta}_\mathbf{1})-\hat{\mathbf{r}}_\mathbf{1})d\mathbf{r}_\mathbf{1}d\boldsymbol{\theta}_\mathbf{1} \tag{3.112}$$

Integration of the first δ-function over the $\mathbf{r}_\mathbf{1}$ gives $\mathbf{r}_\mathbf{1}=\hat{\mathbf{r}}'_\mathbf{1}-\mathbf{d}^a_{\alpha'}(\boldsymbol{\theta}_\mathbf{1})$. Putting this to the second δ-function we have the following:

$$\omega_{\alpha\alpha'}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}'_\mathbf{1})=\frac{1}{\Omega}\int\delta(\mathbf{d}^a_{\alpha\alpha'}(\boldsymbol{\theta}_\mathbf{1})-(\hat{\mathbf{r}}'_\mathbf{1}-\hat{\mathbf{r}}_\mathbf{1}))d\boldsymbol{\theta}_\mathbf{1} \tag{3.113}$$

where $\mathbf{d}^a_{\alpha\alpha'}(\boldsymbol{\theta}_\mathbf{1})=\mathbf{d}^a_{\alpha'}(\boldsymbol{\theta}_\mathbf{1})-\mathbf{d}^a_{\alpha}(\boldsymbol{\theta}_\mathbf{1})$. We can notice, that for any $\hat{\mathbf{r}}_\mathbf{1}$, $\hat{\mathbf{r}}'_\mathbf{1}$ such that $|\hat{\mathbf{r}}'_\mathbf{1}$ - $\hat{\mathbf{r}}_\mathbf{1}|=|\mathbf{d}^a_{\alpha\alpha'}(\boldsymbol{\theta}_\mathbf{1}))|\equiv d^a_{\alpha\alpha'}$ there exists such $\boldsymbol{\theta}_\mathbf{1}$, that $\mathbf{d}^a_{\alpha\alpha'}(\boldsymbol{\theta}_\mathbf{1})=\hat{\mathbf{r}}'_\mathbf{1}-\hat{\mathbf{r}}_\mathbf{1}$. Thus, in this case after integration in (3.113) we have unity. And vice versa, if $|\hat{\mathbf{r}}'_\mathbf{1}$ - $\hat{\mathbf{r}}_\mathbf{1}|\neq d^a_{\alpha\alpha'}$ then integration (3.113) will give zero. Summarizing, we conclude that ω is proportional to $\delta(|\hat{\mathbf{r}}_\mathbf{1}-\hat{\mathbf{r}}'_\mathbf{1}|-d^a_{\alpha'\alpha})$: $\omega_{\alpha\alpha'}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}'_\mathbf{1})=A\delta(|\hat{\mathbf{r}}_\mathbf{1}-\hat{\mathbf{r}}'_\mathbf{1}|-d^a_{\alpha'\alpha})$. This means that the function ω is spherically symmetric. We will use the same definition $\omega^a_{\alpha\alpha'}$ for the radial part of omega, namely: $\omega^a_{\alpha\alpha'}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}'_\mathbf{1})\equiv\omega^a_{\alpha\alpha'}(|\hat{\mathbf{r}}'_\mathbf{1}-\hat{\mathbf{r}}_\mathbf{1}|)$. To determine the constant A we can use a normalization condition for ω. Integration of (3.113) over $\hat{\mathbf{r}}_\mathbf{1}$ will give unity (this follows from the definition of δ-function and constant Ω) . From the normalization rule we have the following:

$$\int\omega^a_{\alpha\alpha}(|\hat{\mathbf{r}}'_\mathbf{1}-\hat{\mathbf{r}}_\mathbf{1}|)d\hat{\mathbf{r}}_\mathbf{1}=A\int\delta(|\hat{\mathbf{r}}_\mathbf{1}-\hat{\mathbf{r}}'_\mathbf{1}|-d^a_{\alpha\alpha'})d\hat{\mathbf{r}}_\mathbf{1}=1 \tag{3.114}$$

By introducing the spherical coordinates with the origin in $\hat{\mathbf{r}}'_\mathbf{1}$ we have $A4\pi(d^a_{\alpha\alpha'})^2=1$. Thus we have $\omega^a_{\alpha\alpha'}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}'_\mathbf{1})=\delta(|\hat{\mathbf{r}}_\mathbf{1}-\hat{\mathbf{r}}'_\mathbf{1}|-d^a_{\alpha\alpha'})/(4\pi(d^a_{\alpha\alpha'})^2)$. We should note, that this formula is correct only for the case $d^a_{\alpha\alpha'}>0$ which is typically equivalent to $\alpha\neq\alpha'$. In case $\alpha=\alpha'$ we have $\mathbf{d}^a_{\alpha\alpha}(\boldsymbol{\theta}_\mathbf{1})=0$, and from (3.113) we immediately have $\omega^a_{\alpha\alpha'}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}'_\mathbf{1})=\delta(\hat{\mathbf{r}}_\mathbf{1}-\hat{\mathbf{r}}'_\mathbf{1})$. Then the full definition of ω is written in the following way:

$$\omega^a_{\alpha\alpha'}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}'_\mathbf{1})=\begin{cases}\dfrac{\delta(|\hat{\mathbf{r}}_\mathbf{1}-\hat{\mathbf{r}}'_\mathbf{1}|-d^a_{\alpha\alpha'})}{4\pi d^a_{\alpha\alpha'}} & ,\alpha\neq\alpha' \\ \delta(|\hat{\mathbf{r}}_\mathbf{1}-\hat{\mathbf{r}}'_\mathbf{1}|) & ,\alpha=\alpha'\end{cases} \tag{3.115}$$

Using the definition (3.112) we can rewrite (3.111) in a following way:

$$Y=\sum_{\alpha'\beta'}\int\omega^a_{\alpha\alpha'}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}'_\mathbf{1})c^{ab}_{\alpha'\beta'}(\hat{\mathbf{r}}'_\mathbf{1},\hat{\mathbf{r}}'_\mathbf{2})\omega^b_{\beta'\beta}(\hat{\mathbf{r}}'_\mathbf{2},\hat{\mathbf{r}}_\mathbf{2})d\hat{\mathbf{r}}'_\mathbf{1}d\hat{\mathbf{r}}'_\mathbf{2} \tag{3.116}$$

Let us consider the relation (3.109). By using (3.102) we rewrite it in the following way:

$$\begin{aligned}Z=\sum_c\frac{\hat{\rho}^c}{\Omega^3}\sum_{\alpha'\gamma'}\int & c^{ac}_{\alpha'\gamma'}(\hat{\mathbf{r}}'_\mathbf{1},\hat{\mathbf{r}}'_\mathbf{3})\utilde{\delta(\mathbf{r}_\mathbf{1}+\mathbf{d}^a_{\alpha'}(\boldsymbol{\theta}_\mathbf{1})-\hat{\mathbf{r}}'_\mathbf{1})}\underline{\delta(\mathbf{r}_\mathbf{3}+\mathbf{d}^c_{\gamma'}(\boldsymbol{\theta}_\mathbf{3})-\hat{\mathbf{r}}'_\mathbf{3})}\times \\ \times & \underline{h^{cb}(\mathbf{r}_\mathbf{3},\mathbf{r}_\mathbf{2},\boldsymbol{\theta}_\mathbf{3},\boldsymbol{\theta}_\mathbf{2})}\utilde{\delta(\mathbf{r}_\mathbf{1}+\mathbf{d}^a_{\alpha}(\boldsymbol{\theta}_\mathbf{1})-\hat{\mathbf{r}}_\mathbf{1})}\underline{\delta(\mathbf{r}_\mathbf{2}+\mathbf{d}^b_{\beta}(\boldsymbol{\theta}_\mathbf{2})-\hat{\mathbf{r}}_\mathbf{2})}\utilde{d\mathbf{r}_\mathbf{1}}\underline{d\mathbf{r}_\mathbf{2}d\mathbf{r}_\mathbf{3}}\utilde{d\boldsymbol{\theta}_\mathbf{1}}\underline{d\boldsymbol{\theta}_\mathbf{2}d\boldsymbol{\theta}_\mathbf{3}}d\hat{\mathbf{r}}'_\mathbf{1}d\hat{\mathbf{r}}'_\mathbf{3}\end{aligned} \tag{3.117}$$

We can notice, that the wavy underlined members correspond to the ω function (3.112) while solid underlined members correspond to the definition of the site-site function $h^{cb}_{\gamma'\beta}$ (3.103). Considering these observations the expression (3.117) can be rewritten in a more compact way:

$$Z = \sum_{c} \hat{\rho}^c \sum_{\alpha'\gamma'} \int \omega^a_{\alpha\alpha'}(\hat{\mathbf{r}}_\mathbf{1}, \hat{\mathbf{r}}'_\mathbf{1}) c^{ac}_{\alpha'\gamma'}(\hat{\mathbf{r}}'_\mathbf{1}, \hat{\mathbf{r}}'_\mathbf{3}) h^{cb}_{\gamma'\beta}(\hat{\mathbf{r}}'_\mathbf{3}, \hat{\mathbf{r}}_\mathbf{2}) d\hat{\mathbf{r}}'_\mathbf{1} d\hat{\mathbf{r}}'_\mathbf{3} \tag{3.118}$$

Putting (3.110), (3.116), (3.118) to (3.106) and assuming the spherical symmetry of the site-site correlation function we obtain the RISM equations, namely:

$$\begin{aligned} h^{ab}_{\alpha\beta}(|\hat{\mathbf{r}}_\mathbf{2} - \hat{\mathbf{r}}_\mathbf{1}|) &= \sum_{\alpha'\beta'} \int \omega^a_{\alpha\alpha'}(|\hat{\mathbf{r}}'_\mathbf{1} - \hat{\mathbf{r}}_\mathbf{1}|) c^{ab}_{\alpha'\beta'}(|\hat{\mathbf{r}}'_\mathbf{2} - \hat{\mathbf{r}}'_\mathbf{1}|) \omega^b_{\beta'\beta}(|\hat{\mathbf{r}}_\mathbf{2} - \hat{\mathbf{r}}'_\mathbf{2}|) d\hat{\mathbf{r}}'_\mathbf{1} d\hat{\mathbf{r}}'_\mathbf{2} \\ &+ \sum_{c} \hat{\rho}^c \sum_{\alpha'\gamma'} \int \omega^a_{\alpha\alpha'}(|\hat{\mathbf{r}}'_\mathbf{1} - \hat{\mathbf{r}}_\mathbf{1}|) c^{ac}_{\alpha'\gamma'}(|\hat{\mathbf{r}}'_\mathbf{3} - \hat{\mathbf{r}}'_\mathbf{1}|) h^{cb}_{\gamma'\beta}(|\hat{\mathbf{r}}_\mathbf{2} - \hat{\mathbf{r}}'_\mathbf{3}|) d\hat{\mathbf{r}}'_\mathbf{1} d\hat{\mathbf{r}}'_\mathbf{3} \end{aligned} \tag{3.119}$$

3.12 Closure relation for the RISM equations

Similarly to the OZ equations (3.87), RISM equations need to be comprised with the closure relation. In the RISM theory it is assumed that sites interact to each other independently. Thus, the closure relation can be written for each pair of sites independently. By analogy to (3.98) we can write the set of site-site closure relations:

$$h^{ab}_{s\alpha}(\mathbf{r}) + 1 = e^{-\beta u^{ab}_{s\alpha}(\mathbf{r}) + h^{ab}_{s\alpha}(\mathbf{r}) - c^{ab}_{s\alpha}(\mathbf{r}) + B^{ab}_{s\alpha}(\mathbf{r})} \tag{3.120}$$

where $u^{ab}_{s\alpha}(\mathbf{r})$ is the site-site interaction potential, $B^{ab}_{s\alpha}(\mathbf{r})$ is the site-site bridge function. By analogy to the six-dimensional HNC approximation (3.97) we can write the site-site HNC closure in the following way [52]:

$$h^{ab}_{s\alpha}(\mathbf{r}) + 1 = e^{-\beta u^{ab}_{s\alpha}(\mathbf{r}) + h^{ab}_{s\alpha}(\mathbf{r}) - c^{ab}_{s\alpha}(\mathbf{r})} \tag{3.121}$$

We note, that this is quite rude approximation. There are two sources of errors: 1) As in the six-dimensional case neglecting the bridge functional leads to some errors 2) The assumption that sites interact independently introduces even more RISM-specific errors. Typically, by using HNC approximation one cannot get good quantitative agreements of the calculated physical quantities to the experimentally measured quantities or quantities calculated using molecular simulations. Moreover: overestimation of the first peaks of the total correlation functions often lead to divergence of the numerical algorithms which solve RISM equations [24]. There are several different bridge approximations. To mention few: Percus-Yevick bridge [106], Martynov-Sarkisov approximation [107], Modified Verlet bridge [108] and others [109]. The search for new bridge functionals is currently actively performed [110, 111](P4)[5]. For some

[5] References given in the parentheses refer to the papers from the list of author's publications which can be found in Appendix B.

specific conditions, for example for Ornstein-Zernike equations for mixtures of Lennard-Jones spheres, the mentioned above bridges can give satisfactorily results. However, the functional which is good for a wide range of different systems is still not known. In the current work it is not our goal to find a universal bridge. However, it is necessary to use the closure which at least warrant the convergence of the algorithm (as was mentioned above, HNC closure is not good in this respect). To improve convergence of the algorithm one can linearize the exponent in the closure relation (3.120) when the argument of the exponent is positive. This method was proposed by Kovalenko and Hirata [24], so this approximation is also usually referenced as KH (Kovalenko-Hirata) approximation. The closure in KH approximation can be written in the following way:

$$c^{ab}_{s\alpha}(r) = \begin{cases} e^{\Xi^{ab}_{s\alpha}(r)} - \gamma_{s\alpha}(r) - 1, & \Xi^{ab}_{s\alpha}(r) < 0 \\ -\beta u^{ab}_{s\alpha}(r), & \Xi^{ab}_{s\alpha}(r) > 0 \end{cases} \tag{3.122}$$

where $\gamma^{ab}_{s\alpha}(r) = h^{ab}_{s\alpha}(r) - c^{ab}_{s\alpha}(r)$, $\Xi^{ab}_{s\alpha}(r) = -\beta u^{ab}_{s\alpha}(r) + \gamma^{ab}_{s\alpha}(r)$. We use the KH approximation in most calculations in our work.

3.13 RISM equations in the Fourier space

Using the properties of the Kronecker δ symbol we can formally write the following relation:

$$c^{ab}_{\alpha'\beta'}(|\hat{\mathbf{r}}'_{\mathbf{2}} - \hat{\mathbf{r}}'_{\mathbf{1}}|)\omega^{b}_{\beta'\beta}(|\hat{\mathbf{r}}_{\mathbf{2}} - \hat{\mathbf{r}}'_{\mathbf{2}}|) = \sum_{c=1}^{M} \delta_{bc} c^{ac}_{\alpha'\beta'}(|\hat{\mathbf{r}}'_{\mathbf{2}} - \hat{\mathbf{r}}'_{\mathbf{1}}|)\omega^{b}_{\beta'\beta}(|\hat{\mathbf{r}}_{\mathbf{2}} - \hat{\mathbf{r}}'_{\mathbf{2}}|) \tag{3.123}$$

In that case we can rewrite the RISM equations in a more compact way, namely:

$$h^{ab}_{\alpha\beta}(|\hat{\mathbf{r}}_{\mathbf{2}} - \hat{\mathbf{r}}_{\mathbf{2}}|) = \sum_{\alpha'\beta'} \int \omega^{a}_{\alpha\alpha'}(|\hat{\mathbf{r}}'_{\mathbf{1}} - \hat{\mathbf{r}}_{\mathbf{1}}|) c^{ab}_{\alpha'\beta'}(|\hat{\mathbf{r}}'_{\mathbf{2}} - \hat{\mathbf{r}}'_{\mathbf{1}}|)\chi^{cb}_{\beta'\beta}(|\hat{\mathbf{r}}_{\mathbf{2}} - \hat{\mathbf{r}}'_{\mathbf{2}}|) d\hat{\mathbf{r}}'_{\mathbf{1}} d\hat{\mathbf{r}}'_{\mathbf{2}} \tag{3.124}$$

where $\chi^{cb}_{\beta'\beta}(|\hat{\mathbf{r}}_{\mathbf{2}} - \hat{\mathbf{r}}'_{\mathbf{2}}|) = \delta_{cb}\omega^{b}_{\beta'\beta}(|\hat{\mathbf{r}}_{\mathbf{2}} - \hat{\mathbf{r}}'_{\mathbf{2}}|) + \hat{\rho}^{c} h^{cb}_{\gamma'\beta}(|\hat{\mathbf{r}}_{\mathbf{2}} - \hat{\mathbf{r}}'_{\mathbf{2}}|)$

Fourier transformation of the function $f(\mathbf{r})$ can be written in the following way:

$$\hat{f}(\mathbf{k}) = \mathcal{T}[\mathbf{f}] = \int_{\mathbb{R}^3} f(\mathbf{r}) e^{i<\mathbf{k},\mathbf{r}>} d\mathbf{r} \tag{3.125}$$

where $< \mathbf{k}, \mathbf{r} >= k_x x + k_y y + k_z z$ is a scalar product, $i = \sqrt{-1}$. The Fourier representations of the functions allow us to use the convolution theorem. Let us multiply both parts of (3.124) by $e^{i<\mathbf{k},\mathbf{r_2}-\mathbf{r_1}>}$ and integrate over $\mathbf{r_2}-\mathbf{r_1}$. Using that $e^{i<\mathbf{k},\mathbf{r_2}-\mathbf{r_1}>} = e^{i<\mathbf{k},\mathbf{r_2}-\mathbf{r_2}'>} e^{i<\mathbf{k},\mathbf{r_2}'-\mathbf{r_1}'>} e^{i<\mathbf{k},\mathbf{r_1}'-\mathbf{r_1}>}$ we have the following relation:

$$\hat{h}^{ab}_{\alpha\beta}(\mathbf{k}) = \\ \sum_{\alpha'\beta'}\int \omega^{a}_{\alpha\alpha'}(|\hat{\mathbf{r}}'_{\mathbf{1}} - \hat{\mathbf{r}}_{\mathbf{1}}|) e^{i\langle\mathbf{k},\mathbf{r_1}'-\mathbf{r_1}\rangle} c^{ab}_{\alpha'\beta'}(|\hat{\mathbf{r}}'_{\mathbf{2}} - \hat{\mathbf{r}}'_{\mathbf{1}}|) e^{i\langle\mathbf{k},\mathbf{r_1}'-\mathbf{r_1}\rangle} \chi^{cb}_{\beta'\beta}(|\hat{\mathbf{r}}_{\mathbf{2}} - \hat{\mathbf{r}}'_{\mathbf{2}}|) e^{i\langle\mathbf{k},\mathbf{r_1}'-\mathbf{r_1}\rangle} d\hat{\mathbf{r}}'_{\mathbf{1}} d\hat{\mathbf{r}}'_{\mathbf{2}} d(\mathbf{r_2} - \mathbf{r_1}) \tag{3.126}$$

Let us fix the coordinates of the point $\mathbf{r_1}$ and re-write the integral in a new coordinates, which are $(\mathbf{r_2}-\mathbf{r_2}')$, $(\mathbf{r_1}'-\mathbf{r_1})$, $(\mathbf{r_2}'-\mathbf{r_1}')$. It can be shown that Jacobian of this transformation is unity. Then in the right hand side of (3.126) we have the product of the Fourier transformations of the functions ω,c, χ. Thus we come to the representation of the RISM equations in the Fourier space:

$$\hat{h}^{ab}_{\alpha\beta}(\mathbf{k}) = \sum_{\alpha'\beta'} \hat{\omega}^{a}_{\alpha\alpha'}(\mathbf{k})\hat{c}^{ab}_{\alpha'\beta'}(\mathbf{k})\hat{\chi}^{cb}_{\beta'\beta}(\mathbf{k}) \tag{3.127}$$

3.14 Reducing the RISM equations to the system of one-dimensional equations

Formally, Fourier transformations in (3.127) are three-dimensional. However, there is no need to numerically calculate the three dimensional Fourier transformation of the spherically symmetric functions. The Fourier transform of a spherically symmetric function $f(|\mathbf{r}|)$ is spherically symmetric itself. It is easy to prove. Let $|\mathbf{k_1}| = |\mathbf{k_2}|$. Then there is a rotation which transforms the vector $\mathbf{k_1}$ into the vector $\mathbf{k_2}$. Let $\mathbf{A}$ be the matrix of this rotation, i.e. $\mathbf{k_2} = \mathbf{A}\mathbf{k_1}$. The Fourier transform of the radially symmetric function $f(|\mathbf{r}|)$ is written in the following way:

$$\hat{f}(\mathbf{k_1}) = \int e^{i\mathbf{k_1}^T\mathbf{r}} f(|\mathbf{r}|)d\mathbf{r} \tag{3.128}$$

Let us rewrite this relation in new coordinates $\mathbf{x} = \mathbf{A}\mathbf{r}$. Due to the properties of rotation matrices the Jacobian of such transformation is $\det(\mathbf{A}) = 1$, thus $d\mathbf{x} = d\mathbf{r}$. Also, it is known that for matrices which represent rotation it holds $\mathbf{A}^{-1} = \mathbf{A}^T$, thus $\mathbf{r} = \mathbf{A}^T\mathbf{x}$. Also, the rotation does not change sizes of vectors, so $|\mathbf{r}| = |\mathbf{A}\mathbf{r}| = |\mathbf{x}|$. Using all this knowledge we rewrite (3.128) in the following way:

$$\hat{f}(\mathbf{k_1}) = \int e^{i\mathbf{k_1}^T\mathbf{A}^T\mathbf{x}} f(|\mathbf{x}|)d\mathbf{x} \tag{3.129}$$

Using that $\mathbf{k_1}^T\mathbf{A}^T = (\mathbf{A}\mathbf{k_1})^T = \mathbf{k_2}^T$ we have the definition of the Fourier transform in the point $\mathbf{k_2}$ in the right hand side of the equation. Thus $\hat{f}(\mathbf{k_1}) = \hat{f}(\mathbf{k_2})$ and $\hat{f}(\mathbf{k})$ is spherically symmetric. This allows us to rewrite (3.127) for radial part $k = |\mathbf{k}|$:

$$\hat{h}^{ab}_{\alpha\beta}(k) = \sum_{\alpha'\beta'} \hat{\omega}^{a}_{\alpha\alpha'}(k)\hat{c}^{ab}_{\alpha'\beta'}(k)\hat{\chi}^{cb}_{\beta'\beta}(k) \tag{3.130}$$

3.15 Bessel-Fourier transformation

The fact that the 3D Fourier transform of spherically symmetric function is spherically symmetric does not give by itself the algorithm how to calculate this Fourier transform in a simple way. Let us obtain this formula. Because the Fourier transformed function is spherically symmetric we only need to know its values along one of the axes. Let us calculate the values $\hat{f}(0, 0, k_z)$. The Fourier transform of the function $f(x, y, z)$ is written in the following way:

$$\hat{f}(\mathbf{k}) = \hat{f}(k_x, k_y, k_z) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f(r)e^{ik_x x+ik_y y+ik_z z}dxdydz \tag{3.131}$$

where $i = \sqrt{-1}$. The values on the k_z axis are:

$$\hat{f}(0, 0, k_z) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f(r)e^{ik_z z}dxdydz \tag{3.132}$$

For the sake of simplicity we denote $k \equiv k_z$, $\hat{f}(0, 0, k_z) \equiv \hat{f}(k)$. Transferring to the spherical coordinates we have the following:

$$\hat{f}(k) = \int\limits_{0}^{\infty}\int\limits_{0}^{\pi}\int\limits_{-\pi}^{\pi} f(r)e^{ikr\cos\theta}r^2\sin\theta drd\theta d\phi \tag{3.133}$$

where $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$, $z = r\cos\theta$, $dxdydz = r^2\sin\theta drd\theta d\phi$. Integration over ϕ gives 2π. Introducing the variable $\xi = cos\theta$, $d\xi = -\sin\theta d\theta$, $\xi \in [\cos 0; \cos\pi] = [1; -1]$, we have the following:

$$\hat{f}(k) = -2\pi\int\limits_{0}^{\infty}\int\limits_{1}^{-1} f(r)e^{ikr\xi}r^2 d\xi dr \tag{3.134}$$

After the integration over ξ we have the following:

$$\hat{f}(k) = 2\pi\int\limits_{0}^{\infty} f(r)\left(\frac{e^{ikr} - e^{-ikr}}{ikr}\right)r^2 dr \tag{3.135}$$

Using the Euler's formula $\sin\alpha = \frac{e^{i\alpha}-e^{-i\alpha}}{2i}$ we obtain the following transformation:

$$\hat{f}(k) = \frac{4\pi}{k}\int\limits_{0}^{\infty} f(r)r\sin krdr \tag{3.136}$$

This transformation is called *the Bessel-Fourier transformation.*

3.16 Inverse Bessel-Fourier transformation

Expression (3.136) can be re-written in a form of the sine transform of the function $rf(r)$, namely:

$$k\hat{f}(k) = 4\pi\int\limits_{0}^{\infty} (rf(r))\sin krdr \tag{3.137}$$

Because the sine transform is inverse to itself the inverse Bessel-Fourier transform can be written in the following way:

$$r'f(r') = A\int_0^\infty \left(k\hat{f}(k)\right)\sin kr'dk \tag{3.138}$$

where A is a normalization constant. It is necessary to note that one need to be cautious in calculations. It is known that the integral over $\sin kr$ has only conditional convergence at the interval $[0;\infty)$. This means that the value of the integral depends on the discretization of the function. In practice one typically uses the discretization on the equispaced grid. Let us assume that the function $f(r)$ has the support $[0;R]$ and is defined at the points with the discretization step $\Delta r = \frac{R}{N}$. Assuming that the function have some finite integral we can write it using the Dirac δ-function:

$$f(r) = \sum_{n=1}^{N} f(r_n)\delta(r-r_n)\Delta r = \frac{R}{N}\sum_{n=1}^{N} f(r_n)\delta(r-r_n) \tag{3.139}$$

Putting (3.139) to the transformation (3.137) we obtain the discrete Bessel-Fourier transform, namely:

$$k\hat{f}(k) = 4\pi\int_0^\infty r\sum_{n=1}^{N} f(r_n)\delta(r-r_n)\Delta r\sin krdr \tag{3.140}$$

Using the properties of the δ-function we have the following:

$$k\hat{f}(k) = 4\pi\frac{R}{N}\sum_{n=1}^{N}\frac{nR}{N}f(\frac{nR}{N})\sin k\frac{nR}{N} \tag{3.141}$$

Let $r' = m\Delta r = \frac{mR}{N}$, $m \in \mathbb{N}$. Let us change the infinite integration limit in the expression (3.138) to the finite number $\frac{\pi N}{R}$. As we show below such a value is necessary for orthogonality of the eigenfunctions $\{\sin k\frac{nR}{N}\}$. Equation (3.138) is transformed to the following expression:

$$\frac{mR}{N}f(\frac{mR}{N}) = A\int_0^{\frac{\pi N}{R}} k\hat{f}(k)\sin k\frac{mR}{N}dk \tag{3.142}$$

Putting (3.141) to (3.142) we have the following:

$$\frac{mR}{N}f(\frac{mR}{N}) = 4\pi\frac{R}{N}A\sum_{n=1}^{N}\frac{nR}{N}f(\frac{nR}{N})\int_0^{\frac{\pi N}{R}}\sin k\frac{nR}{N}\sin k\frac{mR}{N}dk \tag{3.143}$$

The set of the functions $\{\sin\frac{knR}{N}, n \in \mathbb{N}, k \in \mathbb{R}\}$ is orthogonal in a sense of the L_2 scalar product on the interval $k \in [0; \frac{\pi N}{R}]$. If $m \neq n$ we have the following:

$$\begin{gathered}\int\limits_0^{\frac{\pi N}{R}} \sin k\frac{nR}{N}\sin k\frac{mR}{N}dk = \\ \frac{1}{2}\int\limits_0^{\frac{\pi N}{R}}\left(\cos k\frac{(m-n)R}{N} - \cos k\frac{(m+n)R}{N}\right)dk = \\ \frac{1}{2}\frac{N}{m-n}\sin\frac{\pi(m-n)NR}{RN} - \frac{1}{2}\frac{N}{m+n}\sin\frac{\pi(m+n)NR}{RN} = 0\end{gathered} \tag{3.144}$$

When $m = n$ we have the following:

$$\int\limits_0^{\frac{\pi N}{R}} \sin^2\frac{nR}{N}dk = \frac{1}{2}\int\limits_0^{\frac{\pi N}{R}}\left(1 - \cos\frac{2knR}{N}\right)dk = \frac{\pi N}{2R} \tag{3.145}$$

Combining (3.144) and (3.145) we can write the following relation:

$$\int\limits_0^{\frac{\pi N}{R}} \sin k\frac{nR}{N}\sin k\frac{mR}{N}dk = \frac{\pi N}{2R}\delta_{mn} \tag{3.146}$$

where δ_{mn} is the Kronecker delta.

Putting (3.146) into (3.143) and using that $\sum_n \frac{nR}{N}f(\frac{nR}{N})\delta_{mn} = \frac{mR}{N}f(\frac{mR}{N})$ we have the following:

$$\frac{mR}{N}f(\frac{mR}{N}) = 4\pi A\frac{R}{N}\frac{mR}{N}f(\frac{mR}{N})\frac{\pi N}{2R} \tag{3.147}$$

Because this equality is true for any function $f(r)$ we conclude that the following relation holds:

$$1 = 4\pi A\frac{R}{N}\frac{\pi N}{2R} \tag{3.148}$$

From this relation we can express the coefficient A:

$$A = \frac{1}{2\pi^2} \tag{3.149}$$

Taking a limit $N \to \infty$ and dividing both parts of the expression (3.142) by $r' = \frac{mR}{N}$ we obtain the following expression for the inverse Bessel-Fourier transform:

$$f(r') = \frac{1}{2\pi r'}\int\limits_0^{\infty} k\hat{f}(k)\sin kr'dk \tag{3.150}$$

We need to note, that one should be cautious using expressions (3.136),(3.150) because on the infinite gird the results may depend on a grid. To avoid problems it is better to use discrete formulae (3.141),(3.142).

Often for the calculation of the integral over k in the equation (3.142) the zero-order integration approximation is used. In that case to cover the interval $(0; \frac{\pi N}{R})$ one need to choose the grid step in the Fourier space Δk considering the relation $N\Delta k = \frac{\pi N}{R}$, where N is the number of discretization points in both: real and Fourier spaces. This gives the following relation for the Δk

$$\Delta k = \frac{\pi}{R} \tag{3.151}$$

where R is a top boundary of the support of the function $f(r)$. The step of the equispaced N-point grid in the real space is $\Delta r = \frac{R}{N}$. Thus from the expression (3.151) we find the relation between the steps in the real and Fourier spaces:

$$\Delta r \Delta k = \frac{\pi}{N} \tag{3.152}$$

We define $k_m = m\Delta k$, $r_n = n\Delta r$. Using (3.152) we obtain the following:

$$r_n k_m = \frac{\pi m n}{N} \tag{3.153}$$

In these definitions the forward and inverse Bessel-Fourier transforms are defined with the formulae (3.154), (3.155) correspondingly.

$$\hat{f}(k_m) = \frac{4\pi}{k_m} \sum_{n=1}^{N} f(r_n) r_n \sin(\frac{\pi m n}{N}) \Delta r \tag{3.154}$$

$$f(r_m) = \frac{1}{\pi^2 r_n} \sum_{m=1}^{N} \hat{f}(k_m) k_m \sin(\frac{\pi m n}{N}) \Delta k \tag{3.155}$$

There are effective FFT-based algorithms which are able to calculate the expressions (3.154), (3.155) in time proportional to $NlogN$.

3.17 RISM equations in a matrix representation

Let the molecule of type a has K_a sites. We can define the following matrices of the correlation functions:

$$\mathbf{H}_{ab}(k) = [\hat{h}^{ab}_{s\alpha}(k)]_{K_a \times K_b} = \begin{pmatrix} \hat{h}^{ab}_{11}(k) & \dots & \hat{h}^{ab}_{1K_b}(k) \\ \vdots & \ddots & \vdots \\ \hat{h}^{ab}_{K_a 1}(k) & \dots & \hat{h}^{ab}_{K_a K_b}(k) \end{pmatrix} \tag{3.156}$$

$$\mathbf{C}_{ab} = [\hat{c}^{ab}_{s\alpha}(k)]_{K_a \times K_b} = \begin{pmatrix} \hat{c}^{ab}_{11}(k) & \dots & \hat{c}^{ab}_{1K_b}(k) \\ \vdots & \ddots & \vdots \\ \hat{c}^{ab}_{K_a 1}(k) & \dots & \hat{c}^{ab}_{K_a K_b}(k) \end{pmatrix} \tag{3.157}$$

$$\mathbf{W}_a = [\hat{\omega}^a_{s\alpha}(k)]_{K_a \times K_a} = \begin{pmatrix} \hat{\omega}^a_{11}(k) & \dots & \hat{\omega}^a_{1K_a}(k) \\ \vdots & \ddots & \vdots \\ \hat{\omega}^a_{K_a 1}(k) & \dots & \hat{\omega}^a_{K_a K_a}(k) \end{pmatrix} \tag{3.158}$$

Then using the definition of the matrix multiplication we can rewrite the RISM equations (3.130) in the following way:

$$\mathbf{H}_{ab}(k) = \sum_c \mathbf{W}_a(k)\mathbf{C}_{ac}(k)\mathbf{X}_{cb}(k) \tag{3.159}$$

where $\mathbf{X}_{cb}(k) = \delta_{cb}\mathbf{W}_b(k) + \hat{\rho}^c\mathbf{H}_{cb}(k)$

We define the matrices **H**, **C**, **W** in the following way: [6]

$$\mathbf{H}(k) = \begin{pmatrix} \mathbf{H}_{11}(k) & \dots & \mathbf{H}_{1M}(k) \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{M1}(k) & \dots & \mathbf{H}_{MM}(k) \end{pmatrix} \tag{3.160}$$

$$\mathbf{C}(k) = \begin{pmatrix} \mathbf{C}_{11}(k) & \dots & \mathbf{C}_{1M}(k) \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{M1}(k) & \dots & \mathbf{C}_{MM}(k) \end{pmatrix} \tag{3.161}$$

$$\mathbf{W}(k) = \begin{pmatrix} \mathbf{W}_1(k) & 0 & \dots & 0 \\ 0 & \mathbf{W}_2(k) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{W}_M(k) \end{pmatrix} \tag{3.162}$$

Then expression (3.159) can be written in the following matrix-multiplication form:

$$\mathbf{H}(k) = \mathbf{W}(k)\mathbf{C}(k)\mathbf{X}(k) \tag{3.163}$$

where $\mathbf{X}(k) = \mathbf{W}(k) + \mathbf{R}\mathbf{H}(k)$, and the matrix **R** is defined in the following way:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & 0 & \dots & 0 \\ 0 & \mathbf{R}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{R}_M \end{pmatrix} \tag{3.164}$$

[6] I need to note, that there are at least two possible interpretations of the definitions of the matrices **H**, **C**, **W**, which appear due to different understanding of the definitions. In the first interpretation, the matrices **H**, **C**, **W** are the "supermatrices" of size $M \times M$, and the elements of these supermatrices are the matrices $\mathbf{H}_{ij}$, $\mathbf{C}_{ij}$, W_{ij}. In the second interpretation the matrices staying in the elements of these matrices (but not the matrices themselves) are put one near another. So, in that interpretation the elements of the "supermatrix" are still functions (not matrices), and the size of the "supermatrix" is the sum of the sizes of the matrices which form the "supermatrix". I need to note, that during the writing of my thesis I had in mind the second interpretation. However, all the derivations below are correct for both of the interpretations. So the reader can choose the interpretation which seems more natural.

where $\mathbf{R_a}$ is a K_a by K_a diagonal matrix with $\hat{\rho}^a$ on the diagonal. Putting the definition of $\mathbf{X}(k)$ to the expression (3.163) we obtain the following relation:

$$\mathbf{H}(k) = \mathbf{W}(k)\mathbf{C}(k)\left(\mathbf{W}(k) + \mathbf{RH}(k)\right) \tag{3.165}$$

We note that this relation can be interpreted as a recurrent expression where the matrix function $\mathbf{H}(k)$ in the left hand side is expressed by itself. Putting the right hand side of the expression instead of the function $\mathbf{H}(k)$ in the right hand side we come to the following equation:

$$\mathbf{H} = \mathbf{WC}((\mathbf{W} + \mathbf{RWC}\left(\mathbf{W} + \mathbf{RH}\right)) \tag{3.166}$$

After opening the brackets we obtain another recurrent expression for $\mathbf{H}$:

$$\mathbf{H} = \mathbf{WCW} + \mathbf{WCRWCW} + \mathbf{WCRWCRH} \tag{3.167}$$

Repeating the procedure of putting the right hand side of (3.165) instead of the $\mathbf{H}$ in the right hand side of a current recurrent relation we come to the following expression for $\mathbf{H}(k)$

$$\mathbf{H}(k) = \mathbf{W}(k)\mathbf{C}(k)\left(\mathbf{I} + \mathbf{RW}(k)\mathbf{C}(k) + \left(\mathbf{RW}(k)\mathbf{C}(k)\right)^2 + \dots\right)\mathbf{W}(k) \tag{3.168}$$

where $\mathbf{I}$ is the eye matrix.

Assuming that the series converges we can use the formula for the infinite geometric progression of matrices and come to the following expression for the RISM equations:

$$\mathbf{H}(k) = \mathbf{W}(k)\mathbf{C}(k)\left(\mathbf{I} - \mathbf{RW}(k)\mathbf{C}(k)\right)^{-1}\mathbf{W}(k) \tag{3.169}$$

3.18 3DRISM equations

As it was mentioned above, for the solvation free energy calculations one need to know the correlation functions between the solute and solvent molecules where solute is at infinite dilution. To calculate these functions we consider the model system where the single solute molecule is fixed at the origin and surrounded by moving solvent molecules. We will refer to the solute molecule using the superscript index 0. Considering that the solute is infinitely diluted, we can write the OZ equations (3.87) for the solute-solvent correlation functions in the following way:

$$h^{0a}(\mathbf{r_0}, \mathbf{r_2}, \boldsymbol{\theta_0}, \boldsymbol{\theta_2}) = c^{0a}(\mathbf{r_0}, \mathbf{r_2}, \boldsymbol{\theta_0}, \boldsymbol{\theta_2}) + \sum_{c=1}^{M} \frac{\hat{\rho}^c}{\Omega} \int c^{0c}(\mathbf{r_0}, \mathbf{r_3}, \boldsymbol{\theta_0}, \boldsymbol{\theta_3}) h^{ca}(\mathbf{r_3}, \mathbf{r_2}, \boldsymbol{\theta_3}, \boldsymbol{\theta_2}) d\mathbf{r_3} d\boldsymbol{\theta_3} \tag{3.170}$$

where $(\mathbf{r_0}, \boldsymbol{\theta_0})$ are the coordinates of the solute molecule.

It was discussed that due to a high computational complexity the six-dimensional OZ equations are not suitable for practical SFE calculations. On the other hand it is also known that the RISM approximation introduces many additional errors and is unable to treat correctly the molecules' geometry. The compromise between these methods is so-called *3DRISM*-approximation, where the solvent molecules are treated in the RISM approximation, while the solute is a three-dimensional object [69]. To obtain the 3DRISM equations we introduce the *solute-site* total and direct correlation functions, which are averaged over the solvent rotational degrees of freedom molecular correlation functions. By the analogy to the site-site correlation functions (3.103) the *total solute-site correlation function* $h^a_\alpha(\hat{\mathbf{r}})$ of site the α of a molecule of type a is defined in the following way:

$$h^a_\alpha(\mathbf{r}_\alpha) = \frac{1}{\Omega}\int h^{0a}(\mathbf{r_0}, \mathbf{r_2}, \boldsymbol{\theta_0}, \boldsymbol{\theta_2})\delta(\mathbf{r_2} + \mathbf{d}^a_\alpha(\boldsymbol{\theta_2}) - \hat{\mathbf{r}})d\mathbf{r_2}d\boldsymbol{\theta_2} \tag{3.171}$$

where $\mathbf{d}^a_\alpha(\boldsymbol{\theta})$ is a displacement of the site α of the molecule of type a with respect to the center of the molecule if the orientation of the molecule is defined by the Euler angles $\boldsymbol{\theta}$. The main assumption of the 3DRISM theory is that the molecular direct correlation functions c^{0a} can be represented as a sum of solute-site direct correlation functions $c^a_{\alpha'}$, namely:

$$c(\mathbf{r_0}, \mathbf{r_2}, \boldsymbol{\theta_0}, \boldsymbol{\theta_2}) = \sum_{\alpha'}\int c^a_{\alpha'}(\hat{\mathbf{r}}')\delta(\mathbf{r_2} + \mathbf{d}^a_\alpha(\boldsymbol{\theta_2}) - \hat{\mathbf{r}}')d\hat{\mathbf{r}}' \tag{3.172}$$

Multiplying (3.170) by $\Omega^{-1}\delta(\mathbf{r_2} + \mathbf{d}^a_\alpha(\boldsymbol{\theta_2}) - \hat{\mathbf{r}})$, integrating over the $\mathbf{r_2}$, $\boldsymbol{\theta_2}$ and using (3.171), (3.172) we come to the following relation:

$$h^a_\alpha(\hat{\mathbf{r}}) = X + Y \tag{3.173}$$

where X, Y are defined with the following relations:

$$X = \sum_{\alpha'}\frac{1}{\Omega}\int c^a_{\alpha'}(\hat{\mathbf{r}})\delta(\mathbf{r_2} + \mathbf{d}^a_{\alpha'}(\boldsymbol{\theta_2}) - \hat{\mathbf{r}})\delta(\mathbf{r_2} + \mathbf{d}^a_\alpha(\boldsymbol{\theta_2}) - \hat{\mathbf{r}})d\hat{\mathbf{r}}'d\mathbf{r_2}d\boldsymbol{\theta_2} \tag{3.174}$$

$$Y = \sum_c \frac{\hat{\rho}^c}{\Omega^2}\sum_{\gamma'} c^c(\hat{\mathbf{r}}')\delta(\mathbf{r_3} + \mathbf{d}^c_{\gamma'}(\boldsymbol{\theta_3}) - \hat{\mathbf{r}}')h^{ca}(\mathbf{r_3}, \mathbf{r_2}, \boldsymbol{\theta_3}, \boldsymbol{\theta_2})\delta(\mathbf{r_2} + \mathbf{d}^a_\alpha(\boldsymbol{\theta_2}) - \hat{\mathbf{r}})d\hat{\mathbf{r}}'d\mathbf{r_2}d\boldsymbol{\theta_2}d\mathbf{r_3}d\boldsymbol{\theta_3} \tag{3.175}$$

Using the definition of the intramolecular correlation function (3.112) we can express X in the following way:

$$X = \sum_{\alpha'}\omega^a_{\alpha\alpha'}(|\hat{\mathbf{r}}' - \hat{\mathbf{r}}|)c^a_{\alpha'}(\hat{\mathbf{r}}')d\hat{\mathbf{r}}' \tag{3.176}$$

Using the definition of the site-site total correlation function (3.103) we express Y in the following way:

$$Y = \sum_c \hat{\rho}^c \sum_{\gamma'}\int c^c(\hat{\mathbf{r}}')h^{ca}_{\gamma'\alpha}(|\hat{\mathbf{r}}' - \hat{\mathbf{r}}|)d\hat{\mathbf{r}}' \tag{3.177}$$

Putting expressions for X and Y into (3.173) we obtain the 3DRISM equations:

$$h_{\alpha}^{a}(\hat{\mathbf{r}}) = \sum_{c}\sum_{\gamma'} \int c^{c}(\hat{\mathbf{r}}')\chi_{\gamma'\alpha}^{ca}(|\hat{\mathbf{r}}' - \hat{\mathbf{r}}|)d\hat{\mathbf{r}}' \qquad (3.178)$$

where $\chi_{\gamma'\alpha}^{ca}(r) = \delta_{ac}\omega_{\gamma'\alpha}^{a}(r) + \hat{\rho}^{c}h_{\gamma'\alpha}^{ca}(r)$.

Chapter 4

Solvation Free Energy Calculation in RISM and 3DRISM

In this chapter the ways to calculate the solvation free energy in the RISM and 3DRISM approximations are discussed. The main reference for this chapter is Ref. [24]. The description of the semi-empirical solvation free energy expressions can be found in Refs. [66](P6), [64], [112], [65], [74]

4.1 Thermodynamic integration method for calculation of the free energy change

Although we have the formal definition of the Helmholtz free energy (2.79), it is difficult to use it in practice. The partition function (3.7) is formally defined as the $12N$-fold integral, where N is the number of particles in the system, which is a very big number. However, we rarely need to find the total free energy of the system. For most of applications one need to know only relative free energy change in one or another process. To find such changes one can use the *thermodynamic integration method* [113]. In this method it is assumed that the potential energy of the system depends on some parameter λ. When $\lambda = 0$ the potential is the same as the potential in the initial state, when $\lambda = 1$ the potential coincides with the potential of the final state of the system. We would like to find a convenient expression for the free energy change. To do it we find a derivative of the free energy (2.79) over λ. Using the definitions of the configuration integral (3.13) and partition function (3.14) we obtain the following relation:

$$\frac{\partial \mathcal{F}}{\partial \lambda} = \frac{-k_B T}{Q_{N_1 \ldots N_M}} \frac{\partial Q_{N_1 \ldots N_M}}{\partial \lambda} = -k_B T \int (-\beta) \frac{\partial U(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}; \lambda)}{\partial \lambda} \frac{e^{-\beta U(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}; \lambda)}}{Q_{N_1 \ldots N_M}} d\mathbf{r}^{[N]} d\boldsymbol{\theta}^{[N]} \qquad (4.1)$$

Using the definition of the ensemble average we can write the following:

$$\frac{\partial \mathcal{F}}{\partial \lambda} = \left\langle \frac{\partial U}{\partial \lambda} \right\rangle \qquad (4.2)$$

Then the change of the free energy can be found using the following formula:

$$\Delta\mathcal{F} = \mathcal{F}(\lambda = 1) - \mathcal{F}(\lambda = 0) = \int_0^1 \left\langle \frac{\partial U}{\partial \lambda} \right\rangle d\lambda \tag{4.3}$$

This method to calculate the free energy change is called the *thermodynamical integration method*. We note, that strictly speaking we proved the thermodynamical integration method here only for the NVT ensemble. However, thermodynamical integration can be successfully applied for other ensembles as well, which can be proven in a similar manner.

We should also note, that (4.3) can be simplified for the systems with pairwise additive potential. Let the particles in the system interact via the pair-additive potential $U(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]})$, which is defined in a following way:

$$U(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) = \frac{1}{2}\sum_{b=1}^{M}\sum_{c=1}^{M}\sum_{i=1}^{N_b}\sum_{j=1}^{N_c}(1 - \delta_{bc}\delta_{ij})u^{bc}(\mathbf{r}_j^c - \mathbf{r}_i^b, \boldsymbol{\theta}_j^c - \boldsymbol{\theta}_i^c) \tag{4.4}$$

where $u^{bc}(\mathbf{r}, \boldsymbol{\theta})$ is the interaction potential between the particles of type b and c.

Let the potential energy linearly depends on λ. So, the functions $u^{bc}(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}; \lambda)$ are defined in a following way:

$$u^{bc}(\mathbf{r}, \boldsymbol{\theta}; \lambda) = u_0^{bc}(\mathbf{r}, \boldsymbol{\theta}) + \lambda \Delta u(\mathbf{r}, \boldsymbol{\theta}) \tag{4.5}$$

where u_0^{bc}, is the particle interaction potential in the initial system, $\Delta u^{bc} = u_1^{bc} - u_0^{bc}$, $u_1^{bc}(\mathbf{r}, \boldsymbol{\theta})$ is particle interaction potential in the final system. In such considerations we have $\partial U/\partial\lambda = \Delta U$, where ΔU is defined in a following way:

$$\Delta U(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]}) = \frac{1}{2}\sum_{b=1}^{M}\sum_{c=1}^{M}\sum_{i=1}^{N_b}\sum_{j=1}^{N_c}(1 - \delta_{bc}\delta_{ij})\Delta u^{bc}(\mathbf{r}_j^c - \mathbf{r}_i^b, \boldsymbol{\theta}_j^c - \boldsymbol{\theta}_i^c) \tag{4.6}$$

Putting this representation to the thermodynamical integration formula (4.3) we have the following expression:

$$\begin{aligned}&\langle \partial U/\partial\lambda \rangle = \\ &\frac{1}{2}\sum_{b=1}^{M}\sum_{c=1}^{M}\sum_{i=1}^{N_b}\sum_{j=1}^{N_c}(1 - \delta_{bc}\delta_{ij})\int d\mathbf{r}_i^b d\boldsymbol{\theta}_i^b d\mathbf{r}_j^c d\boldsymbol{\theta}_j^c \Delta u^{bc}(\mathbf{r}_j^c - \mathbf{r}_i^b, \boldsymbol{\theta}_j^c - \boldsymbol{\theta}_i^c)\left(\int e^{-\beta U(\mathbf{r}^{[N]}, \boldsymbol{\theta}^{[N]})} d\mathbf{r}^{[N-2]} d\boldsymbol{\theta}^{[N-2]}\right)\end{aligned} \tag{4.7}$$

where $d\mathbf{r}^{[N-2]} \equiv d\mathbf{r}^{[N]}/(d\mathbf{r}_i^b d\mathbf{r}_j^c)$, $d\boldsymbol{\theta}^{[N-2]} \equiv d\boldsymbol{\theta}^{[N]}/(d\boldsymbol{\theta}_i^b d\boldsymbol{\theta}_j^c)$. Using the definition of the pair correlation function (3.22) and changing the sum over identical particles to the product we come to the following expression for the thermodynamic integration process:

$$\Delta\mathcal{F} = \frac{1}{2}\sum_{b=1}^{M}\sum_{c=1}^{M}\int \rho^{bc}(\mathbf{r_2} - \mathbf{r_1}, \boldsymbol{\theta_2} - \boldsymbol{\theta_1})\Delta u^{bc}(\mathbf{r_2} - \mathbf{r_1}, \boldsymbol{\theta_2} - \boldsymbol{\theta_1}) d\mathbf{r_1} d\mathbf{r_2} d\boldsymbol{\theta_1} d\boldsymbol{\theta_2} \tag{4.8}$$

4.2 Thermodynamic integration in the RISM approximation

Our particular task in the current work is calculation of the solvation free energy of a molecule. It was discussed above that solvation free energy corresponds to a process of transfer the solute molecule from a gaseous phase to solution. In the RISM approximation it is assumed that the solvation free energy of a molecule is a sum of solvation free energies of its sites. We use the general formula for the thermodynamic integration (4.3). We assume that the interaction potential depends on a parameter λ, the bulk state corresponds to $\lambda = 0$ and the solvated state corresponds to $\lambda = 1$. For the sake of simplicity we discuss the case of one-component solvent. The final results can be straightforwardly generalized to a case of multi-component solution. We define separate interaction potentials $U_{s\alpha}(\mathbf{r}_1, \ldots, \mathbf{r}_N, \lambda)$ for each site s of a solute molecule and all sites of type α of solvent molecules. For the sake of simplicity let the solute site s has coordinates $\mathbf{r_1}$ and sites α of the solvent molecules have coordinates $\mathbf{r_2}, \ldots, \mathbf{r_N}$ correspondingly, where N is the number of molecules in the system. The solvent molecules are identical, thus the potential $U_{s\alpha}$ is a pairwise-additive function:

$$U_{s\alpha}(\mathbf{r}_1, \ldots, \mathbf{r}_N, \lambda) = \sum_{j=2}^{N} u_{s\alpha}(|\mathbf{r}_j - \mathbf{r}_1|, \lambda) \tag{4.9}$$

where $u_{s\alpha}(r, \lambda)$ is a spherically symmetric interaction potential between the solute site s and the solvent site α. We consider the case of linear dependency of the potential on λ, namely:

$$u_{s\alpha}(r, \lambda) \equiv \lambda u_{s\alpha}(r) \tag{4.10}$$

Putting (4.9) to the relation for the free energy calculation (4.3), considering (4.10) and definition of the ensemble average we have the following:

$$\Delta\mathcal{F}_{RISM} = \sum_{j=2}^{N} \sum_{s\alpha} \int_0^1 d\lambda \int u_{s\alpha}(|\mathbf{r}_j - \mathbf{r}_1|) \int d\mathbf{r}_1 \ldots d\mathbf{r}_N \frac{e^{-\beta U_{s\alpha}(\lambda)}}{Q_N} d\mathbf{r_1} d\mathbf{r}_j \tag{4.11}$$

Here the free energy of the molecule is written as a sum of site-site free energies. Because the solvent sites of the same kind are identical we obtain the sum of $(N-1)$ identical values. The site-site density correlation function $g_{s\alpha}$, by analogy to the six-dimensional density correlation function (3.23), is defined with the following relation:

$$\rho^2 g_{s\alpha}(\mathbf{r}_1, \mathbf{r}_2, \lambda) = N(N-1) \int \frac{e^{-\beta U_{s\alpha}(\lambda)}}{Q_N} d\mathbf{r}_3 \ldots d\mathbf{r}_N \tag{4.12}$$

Using this relation we write the expression for the solvation free energy calculation in a RISM approximation:

$$\Delta\mathcal{F}_{RISM} = \frac{\rho^2}{N} \sum_{s\alpha} \int_0^1 d\lambda \int u_{s\alpha}(|\mathbf{r}_2 - \mathbf{r}_1|) g_{s\alpha}(\mathbf{r}_1, \mathbf{r}_2, \lambda) d\mathbf{r_1} d\mathbf{r}_2 \tag{4.13}$$

So we see that for calculation of the solvation free energy we need to know the site-site correlation functions between the solute and solvent sites. We denote the functions related to the solute molecule with the index u and the functions related to the solvent molecules with the index v. Considering that the concentration of the solute is zero we can write the RISM equations for the solvent-solvent correlation functions in a following form:

$$h^{vv}_{\alpha\beta}(k) = \sum_{\alpha'\beta'} \omega^{v}_{\alpha\alpha'}(k) c^{vv}_{\alpha'\beta'}(k) \omega^{v}_{\beta'\beta}(k) + \hat{\rho}^{v_k} \sum_{\alpha'\gamma'} \omega^{v}_{\alpha\alpha'}(k) c^{vv}_{\alpha'\gamma'}(k) h^{vv}_{\gamma'\beta}(k) \tag{4.14}$$

We see that these equations are the same as the RISM equations (3.127) for the bulk solvent. Typically the number of solvents of interest is not very large. The most interesting are aqueous solutions. Thus one can solve the equations for the bulk solvents of interest separately, and than use the results in calculations for different solute molecules. So in our calculations we consider that the site-site functions are known. To calculate the solute-solvent site-site functions we write the RISM equations (3.127) with the zero solute density $\hat{\rho}^{u}$. We obtain the following equations:

$$h^{uv}_{s\alpha}(k) = \sum_{s'\gamma'} \omega^{u}_{ss'}(k) c^{uv}_{s'\gamma'}(k) \left(\omega_{\gamma'\alpha} + \rho h^{vv}_{\gamma'\alpha}(k) \right) \tag{4.15}$$

where ρ is a solvent density.

4.3 RISM-HNC Solvation Free Energy expression

In a general case solvation free energy calculations with the formula (4.13) require solution of the RISM equations for the series of systems with different values of parameter λ. However, in the case of HNC closure approximation (3.121) the integral over λ can be calculated analytically [52]. To prove this we can show that the integrand $u_{s\alpha}(|\mathbf{r}_2 - \mathbf{r}_1|) g_{s\alpha}(\mathbf{r}_1, \mathbf{r}_2, \lambda)$ in equation (4.13) can be represented as a full derivative over λ. By taking the derivative over λ of the HNC closure (3.120) for $u_{s\alpha}(r;\lambda) = \lambda u_{s\alpha}(r)$ we obtain the following relation:

$$\frac{\partial h_{s\alpha}(r,\lambda)}{\partial \lambda} = e^{-\beta u_{s\alpha}(r;\lambda) + h_{s\alpha}(r,\lambda) - c_{s\alpha}(r,\lambda)} \cdot \left(-\beta u_{s\alpha}(r) + \frac{\partial h_{s\alpha}(r,\lambda)}{\partial \lambda} - \frac{\partial c_{s\alpha}(r,\lambda)}{\partial \lambda} \right) \tag{4.16}$$

Putting the left hand side of the HNC closure to the right hand side of (4.16) and using that $g_{s\alpha}(r,\lambda) = h_{s\alpha}(r,\lambda) + 1$ we come to the following expression:

$$\frac{\partial h_{s\alpha}(r,\lambda)}{\partial \lambda} = -\beta g_{s\alpha}(r,\lambda) u_{s\alpha}(r) + (h_{s\alpha}(r,\lambda) + 1) \frac{\partial}{\partial \lambda} (h_{s\alpha}(r,\lambda) - c_{s\alpha}(r,\lambda)) \tag{4.17}$$

After opening the brackets in the right hand side of this expression we obtain the following:

$$\frac{\partial h_{s\alpha}(r,\lambda)}{\partial \lambda} = -\beta g_{s\alpha}(r,\lambda) u_{s\alpha}(r) + h_{s\alpha}(r,\lambda) \frac{\partial h_{s\alpha}(r,\lambda)}{\partial \lambda} + \frac{\partial h_{s\alpha}(r,\lambda)}{\partial \lambda} - h \frac{\partial c_{s\alpha}(r,\lambda)}{\partial \lambda} - \frac{\partial c_{s\alpha}(r,\lambda)}{\partial \lambda} \tag{4.18}$$

Canceling $\frac{\partial h}{\partial \lambda}$ and using the relation $h\frac{\partial h}{\partial \lambda} = \frac{\partial}{\partial \lambda}(\frac{h^2}{2})$ we have the following:

$$g_{s\alpha}(r,\lambda)u_{s\alpha}(r) = \frac{1}{\beta}\left(\frac{\partial}{\partial\lambda}\left(\frac{h^2_{s\alpha}(r,\lambda)}{2} - c_{s\alpha}(r,\lambda)\right) - h_{s\alpha}(r,\lambda)\frac{\partial c_{s\alpha}(r,\lambda)}{\partial\lambda}\right) \tag{4.19}$$

Putting it to the expression (4.13) and using that $\beta = \frac{1}{k_B T}$ we have the following:

$$\Delta\mathcal{F}_{HNC} = \frac{\rho^2}{N}\sum_{s\alpha}\int\left(\frac{h^2_{s\alpha}(r)}{2} - c_{s\alpha}(r)\right)d\mathbf{r_1}d\mathbf{r}_2 - \sum_{s\alpha}\int d\mathbf{r_1}d\mathbf{r}_2\int_0^1 d\lambda h_{s\alpha}(r,\lambda)\frac{\partial c_{s\alpha}(r,\lambda)}{\partial\lambda} \tag{4.20}$$

where $h_{s\alpha}(r,\lambda=1) \equiv h_{s\alpha}(r)$, $c_{s\alpha}(r,\lambda=1) \equiv c_{s\alpha}(r)$ and $h_{s\alpha}(r,\lambda=0) = c_{s\alpha}(r,\lambda=0) = 0$.

The first summand in (4.20) is already expressed without the integration over λ. To avoid the integration over λ in the second summand we use the integration by parts method and obtain the following relation:

$$\begin{gathered}\sum_{s\alpha}\int d\mathbf{r_1}d\mathbf{r}_2\int_0^1 d\lambda h_{s\alpha}(r,\lambda)\frac{\partial c_{s\alpha}(r,\lambda)}{\partial\lambda} = \\ \sum_{s\alpha}\left(\int d\mathbf{r_1}d\mathbf{r}_2 h_{s\alpha}(r)c_{s\alpha}(r) - \int d\mathbf{r_1}d\mathbf{r}_2\int_0^1 d\lambda c_{s\alpha}(r,\lambda)\frac{\partial h_{s\alpha}(r,\lambda)}{\partial\lambda}\right)\end{gathered} \tag{4.21}$$

Let us show that the following equality holds:

$$\sum_{s\alpha}\int d\mathbf{r_1}d\mathbf{r}_2 h_{s\alpha}(r,\lambda)\frac{\partial c_{s\alpha}(r,\lambda)}{\partial\lambda} = \sum_{s\alpha}\int d\mathbf{r_1}d\mathbf{r}_2 c_{s\alpha}(r,\lambda)\frac{\partial h_{s\alpha}(r,\lambda)}{\partial\lambda} \tag{4.22}$$

To do this we use the RISM equations (4.15) in a real-space representation, namely:

$$h_{s\alpha}(|\mathbf{r_2} - \mathbf{r_1}|,\lambda) = \sum_{s'\alpha'}\int \omega_{ss'}(|\mathbf{r}_1 - \mathbf{r}'|)c_{s'\alpha'}(|\mathbf{r}' - \mathbf{r}''|,\lambda)\chi_{\alpha'\alpha}(|\mathbf{r}'' - \mathbf{r}_2|)d\mathbf{r}'d\mathbf{r}'' \tag{4.23}$$

Using the equation (4.23) we find the left and right hand sides of (4.22) . The left hand side can be expressed in a following way:

$$\begin{gathered}\sum_{s\alpha}\int d\mathbf{r_1}d\mathbf{r}_2 h_{s\alpha}(r,\lambda)\frac{\partial c_{s\alpha}(r,\lambda)}{\partial\lambda} = \\ \sum_{s\alpha}\sum_{s'\alpha'}\int d\mathbf{r}_1 d\mathbf{r}_2 d\mathbf{r}'d\mathbf{r}''\omega_{ss'}(|\mathbf{r}_1 - \mathbf{r}'|)c_{s'\alpha'}(|\mathbf{r}' - \mathbf{r}''|,\lambda)\chi_{\alpha'\alpha}(|\mathbf{r}'' - \mathbf{r}_2|)\frac{\partial c_{s\alpha}(|\mathbf{r}_1 - \mathbf{r}_2|,\lambda)}{\partial\lambda}\end{gathered} \tag{4.24}$$

The right hand side has a following representation:

$$\begin{gathered}\sum_{s\alpha}\int d\mathbf{r_1}d\mathbf{r}_2 c_{s\alpha}(r,\lambda)\frac{\partial h_{s\alpha}(r,\lambda)}{\partial\lambda} = \\ \sum_{s\alpha}\sum_{s'\alpha'}\int d\mathbf{r}_1 d\mathbf{r}_2 d\mathbf{r}'d\mathbf{r}'' c_{s\alpha}(|\mathbf{r}_1 - \mathbf{r}_2|,\lambda)\omega_{ss'}(|\mathbf{r}_1 - \mathbf{r}'|)\frac{\partial c_{s'\alpha'}(|\mathbf{r}' - \mathbf{r}''|,\lambda)}{\partial\lambda}\chi_{\alpha'\alpha}(|\mathbf{r}'' - \mathbf{r}_2|)\end{gathered} \tag{4.25}$$

We can see that after renaming variables ($\mathbf{r}_1 \leftrightarrow \mathbf{r}'$,$\mathbf{r}_2 \leftrightarrow \mathbf{r}''$) the right hand side of expression (4.24) coincides with the right hand side of (4.25) In such a way the relation (4.22) is proved. Putting it to expression (4.21) we have the following:

$$\sum_{s\alpha}\int d\mathbf{r_1}d\mathbf{r}_2\int_0^1 d\lambda h_{s\alpha}(r,\lambda)\frac{\partial c_{s\alpha}(r,\lambda)}{\partial\lambda}=\frac{1}{2}\sum_{s\alpha}\int d\mathbf{r_1}d\mathbf{r_2}h_{s\alpha}(r)c_{s\alpha}(r) \tag{4.26}$$

Putting (4.26) to (4.20) we obtain the following Solvation Free Energy expression for the HNC approximation:

$$\Delta\mathcal{F}_{HNC}=\frac{\rho^2}{N}k_BT\sum_{s\alpha}\int d\mathbf{r_1}d\mathbf{r}_2\left(\frac{h^2_{s\alpha}(r)}{2}-c_{s\alpha}(r)-\frac{1}{2}h_{s\alpha}(r)c_{s\alpha}(r)\right) \tag{4.27}$$

The functions $h_{s\alpha}(r)$, $c_{s\alpha}(r)$ depend only on the relative displacement of molecule sites. Introducing the variable $\mathbf{r}=\mathbf{r_2}-\mathbf{r_1}$ we can avoid the double integration. The integral over $\mathbf{r_1}$ gives the volume V. So, we obtain the following expression:

$$\Delta\mathcal{F}_{HNC}=\rho k_BT\sum_{s\alpha}\int\left(\frac{h^2_{s\alpha}(r)}{2}-c_{s\alpha}(r)-\frac{1}{2}h_{s\alpha}(r)c_{s\alpha}(r)\right)d\mathbf{r} \tag{4.28}$$

Using the spherical symmetry of the functions $h_{s\alpha}(r)$, $c_{s\alpha}(r)$ the integral can be calculated as an integral of the radial part (in spherical coordinates). So the final formula for calculation the SFE in the HNC approximation is a following:

$$\Delta\mathcal{F}_{HNC}=4\pi\rho k_BT\sum_{s\alpha}\int\limits_0^\infty\left(\frac{h^2_{s\alpha}(r)}{2}-c_{s\alpha}(r)-\frac{1}{2}h_{s\alpha}(r)c_{s\alpha}(r)\right)r^2dr \tag{4.29}$$

Doing the similar transformations for the KH closure (3.122) one can obtain the following solvation free energy expression:

$$\Delta\mathcal{F}_{KH}=4\pi\rho k_BT\sum_{s\alpha}\int\limits_0^\infty\left(\frac{h^2_{s\alpha}(r)}{2}\theta(-h_{s\alpha}(r))-c_{s\alpha}(r)-\frac{1}{2}h_{s\alpha}(r)c_{s\alpha}(r)\right)r^2dr \tag{4.30}$$

where $\theta(r)$ is a Heaviside step function.

4.4 Other Solvation Free Energy Expressions

Despite the fact that formulae (4.29) and (4.30) were obtained from the rigorous mathematical transformations, they are quite inaccurate in practical applications [62]. As it was mentioned above the main assumptions of the RISM theory are:

1. The assumption, that the molecular direct correlation function $c(\mathbf{r},\boldsymbol{\theta})$ can be represented as a sum of site-site functions (3.100).

2. The assumption, that the molecular closure relation (3.98) can be substituted by the set of site-site closure relations (3.120)
3. The assumption, that the solvation free energy of a molecule is a sum of solvation free energies of its sites.

The first assumption simply states, that the six-dimensional correlation functions can be reconstructed from the site-site projections. It is true at large distances (where c is proportional to the potential). Of course, substitution of a six-dimensional function with the sum of spherically symmetric projections introduces some errors. However, in my opinion, such a substitution is not a too rough approximation. Indeed, the linear combination of all possible spherically symmetric functions centered at the centers of atoms of the molecule describes a rather wide class of functions. In particular, the more atoms in the molecule there are, the more six-dimensional functions are included in such linear combination. However, it should also be noted that to my knowledge, detailed studies of the accuracy of representation of six-dimensional correlation functions as a sum of spherically symmetric projections were not performed before. Therefore it is difficult to quantify the errors connected with such representation. However, we can describe the effects of the second and third assumptions. These assumptions actually are equivalent to the assumption that the sites of the solvent molecules do no interact to each other, which is not true. Although RISM equations (3.119) are obtained by averaging of the six-dimensional OZ equation, the RISM closure relation (3.120) cannot be obtained by averaging of the six-dimensional closure relation, and thus contradicts it. The same situation occurs with the RISM SFE expressions. In this section we consider the RISM-HNC solvation free energy expression in more details.

In the previous section the HNC expression for RISM approximation was derived. Similarly the HNC solvation free energy expression for the six-dimensional OZ equation can be derived. The six-dimensional analog of the expression (4.27) is written in a following way:

$$\Delta\mathcal{F}_{HNC} = \frac{\rho^2}{N\Omega^2}\int\left(\frac{1}{2}(h(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2}))^2 - c(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2}) - \frac{1}{2}h(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})c(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})\right)d\mathbf{r_1}d\mathbf{r_2}d\boldsymbol{\theta_1}d\boldsymbol{\theta_2} \tag{4.31}$$

where $\Omega = \int d\boldsymbol{\theta}$. For the sake of simplicity this expression can be written in a more compact form:

$$\Delta\mathcal{F}_{HNC} = \frac{\rho^2}{N}\left(\frac{1}{2}X - Y - \frac{1}{2}Z\right) \tag{4.32}$$

where $X = \Omega^{-2}\int (h(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2}))^2 d\mathbf{r_1}d\mathbf{r_2}d\boldsymbol{\theta_1}d\boldsymbol{\theta_2}$, $Y = \Omega^{-2}\int c(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})d\mathbf{r_1}d\mathbf{r_2}d\boldsymbol{\theta_1}d\boldsymbol{\theta_2}$, $Z = \Omega^{-2}\int h(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})c(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})d\mathbf{r_1}d\mathbf{r_2}d\boldsymbol{\theta_1}d\boldsymbol{\theta_2}$. We will use only the first RISM approximation (3.102) to obtain the proper SFE expression. Putting (3.102) to the second summand in (4.32) we get the following relation:

$$Y=\sum_{s\alpha}\int\left(\frac{1}{\Omega}\int\delta(\mathbf{r_1}+\mathbf{d}^u_s(\boldsymbol{\theta_1})-\hat{\mathbf{r}}_\mathbf{1})d\mathbf{r_1}d\boldsymbol{\theta_1}\right)\left(\frac{1}{\Omega}\int\delta(\mathbf{r_2}+\mathbf{d}^v_\alpha(\boldsymbol{\theta_2})-\hat{\mathbf{r}}_\mathbf{2})d\mathbf{r_2}d\boldsymbol{\theta_2}\right)c_{s\alpha}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}_\mathbf{2})d\hat{\mathbf{r}}_\mathbf{1}d\hat{\mathbf{r}}_\mathbf{2} \quad (4.33)$$

Considering that $\int\delta(\mathbf{r}+\mathbf{d}^v_\alpha(\boldsymbol{\theta})-\hat{\mathbf{r}})d\mathbf{r}=1$, $\int d\boldsymbol{\theta}=\Omega$, we get the following expression for Y:

$$Y=\sum_{s\alpha}\int c_{s\alpha}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}_\mathbf{2})d\hat{\mathbf{r}}_\mathbf{1}d\hat{\mathbf{r}}_\mathbf{2} \quad (4.34)$$

Putting (3.102) to the third summand in (4.32) we get the following expression:

$$Z=\\ \sum_{s\alpha}\int\left(\frac{1}{\Omega^2}\int\delta(\mathbf{r_1}+\mathbf{d}^u_s(\boldsymbol{\theta_1})-\hat{\mathbf{r}}_\mathbf{1})\delta(\mathbf{r_2}+\mathbf{d}^v_\alpha(\boldsymbol{\theta_2})-\hat{\mathbf{r}}_\mathbf{2})h(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})d\mathbf{r_1}d\mathbf{r_2}d\boldsymbol{\theta_1}d\boldsymbol{\theta_2}\right)c_{s\alpha}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}_\mathbf{2})d\hat{\mathbf{r}}_\mathbf{1}d\hat{\mathbf{r}}_\mathbf{2} \quad (4.35)$$

Using the definition of the site-site total correlation function (3.103) we get the following expression:

$$Z=\sum_{s\alpha}\int h_{s\alpha}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}_\mathbf{2})c_{s\alpha}(\hat{\mathbf{r}}_\mathbf{1},\hat{\mathbf{r}}_\mathbf{2})d\hat{\mathbf{r}}_\mathbf{1}d\hat{\mathbf{r}}_\mathbf{2} \quad (4.36)$$

We note, that the expressions for the summands (4.34), (4.36) are the same as the expressions for the second and third summand in the RISM-HNC solvation free energy expression (4.27). Thus, this part of the RISM-HNC expression is consistent with the six-dimensional expression. However, it is not so for the first summand. Indeed, the first summand does not contain c-function, thus it could not be straightforwardly reduced to the site-site form. And it is not equal the sum of the site-site summands:

$$X=\int\left(h(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})\right)^2 d\mathbf{r_1}d\mathbf{r_2}d\boldsymbol{\theta_1}d\boldsymbol{\theta_2}\neq\sum_{s\alpha}\int h^2_{s\alpha}(r)d\hat{\mathbf{r}}_\mathbf{1}d\hat{\mathbf{r}}_\mathbf{2} \quad (4.37)$$

This is one of the main sources of errors of the RISM-HNC and RISM-KH expression (4.29), (4.30). For example, it was shown, that RISM-HNC solvation free energy grows linearly with the number of sites in the molecule, even if the sites are artificially introduced and have the same coordinates [114]. Such a behavior is unphysical.

There were several other SFE formulae proposed which try to correct the errors of the RISM-HNC solvation free energy approximations. In Ref. [115] the sources of the errors were analyzed and it was pointed out that the HNC model typically overestimates the hydrogen bond contribution to the solvation free energy. In that work it was proposed to introduce additional correction to the SFE expression which contains the additional repulsing potential. This model is typically called *HNC with the repulsive bridge correction* (HNCB). The solvation free energy in the HNCB model is calculated using the following formula:

$$\Delta\mathcal{F}_{HNCB}=\Delta\mathcal{F}_{HNC}+4\pi\rho k_BT\sum_{s\alpha}(h_{s\alpha}(r)+1)(e^{-b^R_{s\alpha}(r)}-1)r^2dr \quad (4.38)$$

where $b^R_{s\alpha}(r)$ is defined with the following relation:

$$e^{-b^R_{s\alpha}(|\mathbf{r}|)} = \prod_{\beta\neq\alpha} \int_V \omega_{\alpha\beta}(|\mathbf{r}' - \mathbf{r}|) \exp\left(-\beta\epsilon_{s\beta}\left(\frac{\sigma_{s\beta}}{|\mathbf{r}|}\right)^{12}\right) d\mathbf{r}' \tag{4.39}$$

where $\sigma_{s\beta}$, $\epsilon_{s\beta}$ are pair Lennard-Jones parameters of the solute-solvent site-site potential.

Another approximation is Gaussian Fluctuations, (GF) formula which was initially proposed in Ref. [116].

$$\Delta\mathcal{F}_{GF} = 4\pi\rho k_B T \sum_{s\alpha} \int\limits_0^\infty \left(-c_{s\alpha}(r) - \frac{1}{2}h_{s\alpha}(r)c_{s\alpha}(r)\right) r^2 dr \tag{4.40}$$

We see, that this formula simply neglects the first summand ($h^2/2$) of the RISM-HNC expression, and in such a way avoids the linear dependency of the solvation free energy on the number of sites in the molecule. However, neglecting of the "problematic" summand can also introduce additional errors.

More elegant way was proposed in Ref. [62] by Ten-no et al. In this work the first summand in the six-dimensional HNC expression is approximated using the *Partial Wave* (PW) method. In the partial wave method the expression for the solvation free energy calculation has the following form:

$$\Delta\mathcal{F}_{PW} = \Delta\mathcal{F}_{GF} + 2\pi\rho k_B T \sum_{s\alpha} \int_0^\infty \tilde{h}_{s\alpha}(r) h_{s\alpha}(r) r^2 dr \tag{4.41}$$

where the Fourier transform of the functions $\tilde{h}_{s\alpha}(r)$ is defined in a following way:

$$\hat{\tilde{h}}_{s\alpha}(k) = \sum_{s'\alpha'} \hat{\tilde{\omega}}^u_{ss'}(k) h_{s\alpha}(k) \hat{\tilde{\omega}}^v_{\alpha'\alpha}(k) \tag{4.42}$$

where $\hat{\tilde{\omega}}^u_{ss'}(k)$, $\hat{\tilde{\omega}}^v_{\alpha'\alpha}(k)$ are the elements of matrices which are inverse to the matrices of intramolecular functions $\hat{\mathbf{W}}^u$ and $\hat{\mathbf{W}}^v$ (3.158). It was shown that the PW expression is more suitable for the SFE calculations than KH and GF expressions [64].

4.5 Semi-Empirical methods for RISM Solvation Free Energy calculation

As it was discussed above, advanced RISM solvation free energy expressions can correct some systematic errors of the RISM calculations. Nevertheless, accuracy of these methods still often appears to be too low for the practical applications. That's why semi-empirical approaches are used for the accurate SFE calculations with RISM and 3DRISM. One of the ways to develop the semi-empirical SFE calculation method is to parameterize the difference between the experimentally measured SFE ΔG_{exp} and the value ΔG_{RISM} calculated with one of the

RISM SFE expressions [1]. To parameterize this difference one can use some known parameters of the molecule (*descriptors*). The most straightforward is the linear parameterization model where the error is assumed to be linearly proportional to the descriptors. Let $D_1, \ldots, D_M$ be the descriptors. Then the linear parameterization model can be written as follows:

$$\varepsilon = \Delta G_{exp} - \Delta G_{RISM} = \sum_{i=1}^{M} a_i D_i \tag{4.43}$$

where $a_1, \ldots, a_M$ are the free coefficients. The free coefficients $a_1, \ldots, a_M$ can be calculated using the least squares method. To calculate them the *training* set of compounds should be chosen. In principle, the choice of the training set can greatly affect the effectiveness of the fitted formula. Thus, if the training set contains too few molecules or molecules included in this set are not representative enough, the resulting coefficients can be fitted incorrect and the final expression will not have sufficient predictive power. On the other hand, the more molecules are included in the training set, the fewer compounds are left for the test set. Moreover, in this case, the test set will most likely contain compounds with a structure very similar to the structure of some of the compounds in the training set. As a result, it is very difficult to judge the real accuracy of the final formula. Thus, the correct choice of the training and test sets is not an easy task. In many cases, it is necessary to perform a cross-validation of the obtained formula. One can, for example, do the parameterization not once, but many times with different randomly selected training and test sets and then compare the results, determining the mean value and spread of the regression coefficients [66](P6). In this work we will not go into details of the cross-validation methods, as it is beyond the scope of this study and is discussed in the papers on the parametrization of RISM SFE expressions [64, 74, 117].

It is assumed that for the training set of compounds experimentally measured SFEs and the values of all descriptors $D_1, \ldots, D_M$ are known. Let ΔG^k_{RISM}, ΔG^k_{exp} be the calculated and the experimentally measured SFEs of the k^{th} molecule in the training set. Let $D^k_1, \ldots, D^k_M$ be the descriptors of the k^{th} molecule in the training set. Using the linear model (4.43) we can write the following approximate relations for all the molecules in the training set:

$$\varepsilon_k = \Delta G^k_{exp} - \Delta G^k_{RISM} \approx \sum_{i=1}^{M} a_i D^k_i \tag{4.44}$$

Following the least squares method the coefficients $a_1, \ldots, a_M$ should be chosen in the way that

[1] In this section and in the next chapters we use the Gibbs free energy ΔG instead of Helmholtz free energy $\Delta\mathcal{F}$. In the most practical applications for aqueous solutions the change of the volume of the system ΔV is negligible, so the formulae for Helmholtz free energy calculation can be used as well for the calculation of the Gibbs free energy change. The reason why we use the Gibbs free energy is that the experimental data is available for the Gibbs free energy change ΔG.

minimizes the following expression:

$$\sum_{k=1}^{N}\left(\varepsilon_k-\sum_{i=1}^{M}a_iD_i^k\right)^2\to\min \tag{4.45}$$

where N is the number of the molecules in the training set. Using the matrix representation we can rewrite this expression in the following way:

$$\left(\boldsymbol{\varepsilon}-\mathbf{Da}\right)^T\left(\boldsymbol{\varepsilon}-\mathbf{Da}\right)\to\min \tag{4.46}$$

where $\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_N)^T$, $\mathbf{D}=[D_i^k]_{N\times M}$, $\mathbf{a}=(a_1,\ldots,a_M)^T$. The necessary condition of the minimum is equality of the partial derivatives over the all parameters $a_1,\ldots,a_M$ to zero. We obtain the following relations:

$$\begin{cases}\sum\limits_{k=1}^{N}(\varepsilon_k-\sum\limits_{i=1}^{M}a_iD_i^k)\cdot D_j^k=0\\ j=1\ldots M\end{cases} \tag{4.47}$$

This relation in a matrix form can be written in a following way:

$$\left(\boldsymbol{\varepsilon}-\mathbf{Da}\right)^T\mathbf{D}=0 \tag{4.48}$$

From this relation the free coefficients $\mathbf{a}$ can be expressed in a following form:

$$\mathbf{a}^T=\boldsymbol{\varepsilon}^T\mathbf{D}\left(\mathbf{D}^T\mathbf{D}\right)^{-1} \tag{4.49}$$

After calculating of the free coefficients the formula for the semi-empirical SFE calculations is determined. It reads as follows:

$$\Delta G_{corr}=\Delta G_{RISM}+\sum_{i=1}^{M}a_iD_i \tag{4.50}$$

To check accuracy of this expression one needs to perform SFE calculations on the *test* set of compounds and compare results to the experimental ones. The test and the training sets of compounds should not overlap. To check how well is the semi-empirical expression the following quantities can be calculated on the test set of compounds:

- Correlation coefficient between the experimental and calculated values.
- Mean deviation $MD(\Delta G_{corr},\Delta G_{exp})$:

$$MD(\Delta G_{corr},\Delta G_{exp})=\sum_{k=1}^{N_1}\left(\Delta G_{corr}^k-\Delta G_{exp}^k\right)$$

- Root Mean Squared Deviation (RMSD):

$$RMSD(\Delta G_{corr}, \Delta G_{exp}) = \sqrt{\frac{1}{N_1}\sum_{k=1}^{N_1}\left(\Delta G^k_{corr} - \Delta G^k_{exp}\right)^2}$$

- Standard Deviation (SD):

$$SD(\Delta G_{corr}, \Delta G_{exp}) = \sqrt{RMSD(\Delta G_{corr}, \Delta G_{exp})^2 - MD(\Delta G_{corr}, \Delta G_{exp})^2}$$

where N_1 is the number of compounds in the training set.

4.6 Semi-empirical models based on the partial molar volume correction

Predictability of a semi-empirical model strongly depends on the choice of descriptors. The descriptors used in the model should meet the following requirements:

- The values of the descriptors should be known or simply computable for any molecule to which the model is applicable.

- The values of the descriptors should correlate with the errors of the RISM SFE expression used in calculations (otherwise these descriptors are not useful).

One of the perspective descriptors for the RISM SFE calculations is the Partial Molar Volume (PMV) of the molecule. It was shown that the error of the RISM expressions correlate with PMV [41]. Also, PMV can be simply calculated from the site-site correlation functions in both: RISM and 3DRISM theories [24]. In Ref. [41] it was proposed to parameterize RISM SFE expressions with PMV and the number of OH-groups in the molecule. It was shown that using this parameterization method it is possible to predict SFE of the limited number of small organic molecules with the accuracy of about 1 kcal/mol. This approach was developed afterwards in the Structural Descriptor Correction (SDC) method, which includes PMV descriptor and the structural descriptors, e.g. number of double bonds in the molecule, number of branches, number of specific groups etc. [64]. This method was tested on a large set of more than 100 organic compounds. It was shown that the method is able to predict SFE with the accuracy of 1-1.2 kcal/mol. Later on the SDC method was used for the calculation of the SFE of the pollutants [65]. It was shown that for this set of molecules the error of the SDC model is of about 0.9 kcal/mol. In such a way the transferability of the method was proven.

Despite of the amazing results, there are some compounds for which the SDC method is hardly applicable. To use the SDC method one needs to calculate the values of all descriptors.

However, it can be non-trivial task for some complicated molecules, because often there are more than one way to divide these molecules into the functional groups. For such molecules the simplified atomic-type correction (ATC) can be applied. In the ATC model the descriptors are PMV and the numbers of atoms of each type in the molecule [66](P6). Typically, atomic type correction model gives worse results in comparison to the SDC model. However, it can be applied to any kind of molecule of arbitrary complexity, while SDC is limited to the molecules which could be simply divided into the molecular groups.

The PMV-based parameterization model was also proposed for 3DRISM. It was shown, that using only two descriptors it is possible to predict SFE with the accuracy of 1 kcal/mol [112]. This model is called Universal Correction (UC) model. UC model was tested for different sets of organic and drug-like compounds and demonstrated quite good accuracy of predictions [74].

In our work we use the ATC model for the RISM and the UC model for the 3DRISM SFE calculations.

Chapter 5

RISM Multi-Grid algorithm for Solvation Free Energy calculations

In this chapter the multi-grid algorithm for solving RISM equations is described. The applicability of the algorithm to the solvation free energy calculation is checked by benchmarking on the set of drug-like compounds. This chapter is based on my recent papers, Refs. [118](P2) and [67](P3).

5.1 RISM equations representation suitable for numerical solution

5.1.1 Indirect correlation functions

The RISM equations in form (3.127) are not suitable for numerical solution. The main causes are: 1) the site-site correlation functions decay slowly, and thus cannot be effectively discretized, 2) the closure relation in the form (3.120) cannot be used to express c-functions due to huge numerical errors. We discuss below these problems in details and also give a more suitable representation of the RISM equations which can be used for the iterative solution. For the sake of simplicity we discuss below the case of one component solvent. The equations can be straightforwardly generalized to a case of for multi-component solvents as well. The following functions are involved in the RISM equation for the infinitely diluted solution: (i) total and direct site-site correlation functions $\{h_{s\alpha}(r)\}$ and $\{c_{s\alpha}(r)\}$ describing correlations between the site s of the solute molecule and sites α of the solvent molecules, (ii) intramolecular correlation functions $\{w_{ss'}(r)\}$ describing the structure of the solute molecule, and (iii) bulk solvent susceptibility functions $\{\chi_{\alpha\alpha'}(r)\}$ describing the structure of the pure solvent. We assume that the solute molecule has M sites and the solvent molecule has K sites. The RISM equations can be written in a matrix form as follows:

$$\hat{\mathbf{H}} = \hat{\mathbf{W}} \cdot \hat{\mathbf{C}} \cdot \hat{\mathbf{X}}, \tag{5.1}$$

where the matrices $\hat{\mathbf{H}}$, $\hat{\mathbf{C}}$, $\hat{\mathbf{W}}$, $\hat{\mathbf{X}}$ are defined as follows: $\hat{\mathbf{H}} = [\hat{h}_{s\alpha}(k)]_{M\times K}$, $\hat{\mathbf{C}} = [\hat{c}_{s\alpha}(k)]_{M\times K}$, $\hat{\mathbf{W}} = [\hat{w}_{ss'}(k)]_{M\times M}$, $\hat{\mathbf{X}} = [\hat{\chi}_{\alpha\alpha'}(k)]_{K\times K}$. Here the hat symbol ($\hat{}$) denotes the Fourier transformed function. The transformation of a spherically symmetric function $f(r)$ is defined by the Bessel-Fourier transform (3.136). Intramolecular correlation functions in the Fourier space $\hat{w}_{ss'}(k)$ are found via the relation

$$\hat{w}_{ss'}(k) = \delta_{ss'} + (1 - \delta_{ss'})\frac{\sin kr_{ss'}}{kr_{ss'}}, \tag{5.2}$$

where $\delta_{ss'}$ is the Kronecker delta and $r_{ss'}$ is the distance between the sites s and s' of the solute molecule. Susceptibility functions of bulk solvent functions are defined as follows:

$$\hat{\chi}_{\alpha\alpha'}(k) = \hat{w}^{\text{solv}}_{\alpha\alpha'}(k) + \rho\hat{h}^{\text{solv}}_{\alpha\alpha'}(k), \tag{5.3}$$

where ρ is the density of the solvent and $\{w^{\text{solv}}_{\alpha\alpha'}(k)\}$ and $\{h^{\text{solv}}_{\alpha\alpha'}(k)\}$ are intramolecular and total correlation functions of the bulk solvent. In the current work we use previously calculated water susceptibility functions [83], therefore we do not discuss these calculations here, and just assume them to be known functions in (5.1). Equation (5.1) is completed by the following closure relation:

$$h_{s\alpha}(r) + 1 = \exp\left(-\beta u_{s\alpha}(r) + h_{s\alpha}(r) - c_{s\alpha}(r) + B_{s\alpha}(r)\right), \tag{5.4}$$

where $\beta = 1/k_B T$, k_B is a Boltzmann constant, T is a temperature, $u_{s\alpha}(r)$ is the site-site potential and $B_{s\alpha}(r)$ is a *bridge* function.

It was discussed above that generally the exact expression for the Bridge function is not known. We use the Kovalenko-Hirata closure relation in our calculations [119]:

$$h_{s\alpha}(r) + 1 = \left\{ \begin{array}{ll} e^{\Xi_{s\alpha}(r)}, & \Xi_{s\alpha}(r) < 0, \\ \Xi_{s\alpha}(r) & \Xi_{s\alpha}(r) > 0, \end{array} \right. \tag{5.5}$$

where $\Xi_{s\alpha}(r) = -\beta u_{s\alpha}(r) + h_{s\alpha}(r) - c_{s\alpha}(r)$ appears in the argument of the exponential function in (5.4).

A typical iteration scheme of solving RISM equations includes two substeps on each iteration step:

1. From the RISM equations (5.1) obtain $h_{s\alpha}$
2. From the closure relation (5.4) obtain $c_{s\alpha}$ functions.

If we express $c_{s\alpha}$ functions from the closure relation (5.4) we come to the following relation:

$$c_{s\alpha}(r) = \ln(h_{s\alpha}(r) + 1) + \beta u_{s\alpha}(r) + h_{s\alpha}(r) + B_{s\alpha}(r) \tag{5.6}$$

The functions $h_{s\alpha}(r)$ are connected to the site-site radial distribution functions $g_{s\alpha}(r)$ in a following way:

$$h_{s\alpha}(r) = g_{s\alpha}(r) - 1 \tag{5.7}$$

Obviously two molecules cannot be simultaneously in the same place. The site-site correlation functions are proportional to the probability to find the sites at the separation r there is some radius r_0 where with a good accuracy the following relations holds:

$$g^{uv}_{s\alpha}(r) = h^{uv}_{s\alpha}(r) + 1 = 0, \qquad r < r_0 \tag{5.8}$$

Substituting (5.8) into (5.6) for the distances $r < r_0$, we have to calculate the logarithm of zero. Of cause, this is not exact zero, only almost a zero, because the probability to find one particle inside the core of other one is non-zero. Nevertheless calculating logarithms of such small numbers is numerically problematic and can cause overflow or at least huge numerical errors. To avoid these problems we define the *indirect correlation functions* $\gamma_{s\alpha}(r) = h_{s\alpha}(r) - c_{s\alpha}(r)$. Putting the expression for the indirect functions to the RISM equations we obtain the following result:

$$\begin{cases} \hat{\gamma}_{s\alpha}(k) = \sum\limits_{s'\alpha'} \hat{\omega}_{ss'}(k)\hat{c}_{s'\alpha'}(k)\hat{\chi}^{\text{solv}}_{\alpha'\alpha}(k) - \hat{c}_{s\alpha}(k) \\ c_{s\alpha}(r) = e^{-\beta u_{s\alpha}(r) + \gamma_{s\alpha}(r) + B_{s\alpha}(r)} - \gamma_{s\alpha}(r) - 1 \\ s = 1 \dots M, \qquad \alpha = 1 \dots K \end{cases} \tag{5.9}$$

Using equations in this form we avoid calculations of the logarithm of the small values and their associated numerical problems.

5.1.2 Long-range approximation of direct correlation functions

Let us consider a smooth transition process from the ideal gas state to the state with interacting particles. The summands which correspond to the ideal gas state are the same in the initial and in the final state, the change of free energy is only due to the exchange part: $\Delta\mathcal{F} = \mathcal{F}^{\text{ex}}$. Assuming that the potential is pairwise-additive we can use expression (4.8). We obtain the following expression for the $\mathcal{F}^{\text{ex}}$:

$$\mathcal{F}^{\text{ex}} = \frac{1}{2}\sum_{c=1}^{M}\sum_{d=1}^{M}\int \rho^{cd}(\mathbf{r}'' - \mathbf{r}', \boldsymbol{\theta''} - \boldsymbol{\theta'}) u^{cd}(\mathbf{r}'' - \mathbf{r}', \boldsymbol{\theta''} - \boldsymbol{\theta'}) d\mathbf{r}' d\mathbf{r}'' d\boldsymbol{\theta'} d\boldsymbol{\theta''} \tag{5.10}$$

In a very rude approximation we may consider that the distributions of different particles are independent. In that case we have $\rho^{cd}(\mathbf{r_2} - \mathbf{r_1}, \boldsymbol{\theta_2} - \boldsymbol{\theta_1}) = \rho^c(\mathbf{r_1}, \boldsymbol{\theta_1})\rho^d(\mathbf{r_2}, \boldsymbol{\theta_2})$. Putting this to the expression for $\mathcal{F}^{\mathrm{ex}}$ we have the following:

$$\mathcal{F}^{\mathrm{ex}} = \frac{1}{2}\sum_{c=1}^{M}\sum_{d=1}^{M}\int \rho^c(\mathbf{r}', \boldsymbol{\theta'})\rho^d(\mathbf{r}'', \boldsymbol{\theta''})u^{cd}(\mathbf{r}'' - \mathbf{r}', \boldsymbol{\theta''} - \boldsymbol{\theta'})d\mathbf{r}'d\mathbf{r}''d\boldsymbol{\theta'}d\boldsymbol{\theta''} \tag{5.11}$$

Taking the functional derivative of this expression we have the following:

$$\frac{\delta \mathcal{F}^{\mathrm{ex}}}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} = \frac{1}{2}\left(\sum_{c=1}^{M}\int \rho^c(\mathbf{r}', \boldsymbol{\theta'})u^{ca}(\mathbf{r_1} - \mathbf{r}', \boldsymbol{\theta_1} - \boldsymbol{\theta'})d\mathbf{r}'d\boldsymbol{\theta'} + \sum_{d=1}^{M}\int \rho^d(\mathbf{r}'', \boldsymbol{\theta''})u^{ad}(\mathbf{r}'') - \mathbf{r_1}, \boldsymbol{\theta''} - \boldsymbol{\theta_1})d\mathbf{r}''d\boldsymbol{\theta''}\right) \tag{5.12}$$

Because $u^{ab}(\mathbf{r}, \boldsymbol{\theta}) = u^{ba}(-\mathbf{r}, -\boldsymbol{\theta})$ both summands are identical, so we have the following:

$$\frac{\delta \mathcal{F}^{\mathrm{ex}}}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})} = \sum_{c=1}^{M}\int \rho^c(\mathbf{r}', \boldsymbol{\theta'})u^{ca}(\mathbf{r_1} - \mathbf{r}', \boldsymbol{\theta_1} - \boldsymbol{\theta'})d\mathbf{r}'d\boldsymbol{\theta'} \tag{5.13}$$

Taking the second derivative of this expression we have the following:

$$\frac{\delta^2 \mathcal{F}^{\mathrm{ex}}}{\delta \rho^a(\mathbf{r_1}, \boldsymbol{\theta_1})\delta \rho^b(\mathbf{r_2}, \boldsymbol{\theta_2})} = u^{ab}(\mathbf{r_2} - \mathbf{r_1}, \boldsymbol{\theta_2} - \boldsymbol{\theta_1}) \tag{5.14}$$

Using the definition of the pair direct correlation function (3.102) we have the following approximation:

$$c^{ab}(\mathbf{r_1}, \mathbf{r_2}, \boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \approx -\beta u^{ab}(\mathbf{r_2} - \mathbf{r_1}, \boldsymbol{\theta_2} - \boldsymbol{\theta_1}) \tag{5.15}$$

We note that this approximation is only valid for the case then the density distributions of particles are independent from each other. This in turn is only valid at large distances. So, expression (5.15) can be used as a long-range approximation of direct correlation functions.

5.1.3 RISM equations for short-range functions

Typically the direct correlation functions decay slowly with a distance which can cause problems with discretization of these functions. However, as it was discussed in the section 5.1.2 we know asymptotic behavior of direct correlation functions, namely:

$$c_{s\alpha}(r) = -\beta u_{s\alpha}(r) \tag{5.16}$$

For charged particles the biggest contribution to the potential at the large distances is the Coulomb interaction potential:[1]

$$u_{s\alpha}(r) \approx \frac{q_s q_\alpha}{r} \qquad r > r_0 \tag{5.17}$$

[1] We use atomic units in our work to avoid scaling coefficients in the Coulomb interaction.

We introduce the short-range and long-range site-site potentials $u^S_{s\alpha}(r)$, $u^L_{s\alpha}(r)$ in the following way:

$$u^L_{s\alpha}(r) = \frac{q_s q_\alpha}{r}\mathrm{erf}(\tau r) \qquad u^S_{s\alpha}(r) = u_{s\alpha}(r) - u^L_{s\alpha}(r) \tag{5.18}$$

where $\frac{q_s q_\alpha}{r}$ is a Coulomb potential, $\mathrm{erf}(r) = (2/\sqrt{\pi})\int_0^r e^{-x^2}dx$ is a Gauss error function, the parameter τ determines the smoothness of the transition between the short-range and long-range functions. The long-range direct correlation functions $c^L_{s\alpha}(r)$ are defined in a following way:

$$c^L_{s\alpha}(r) = -\beta u^L_{s\alpha}(r) \tag{5.19}$$

Short-range direct and indirect correlation functions $c^S_{s\alpha}(r)$ and $\gamma^S_{s\alpha}(r)$ are defined as a difference between the full and long range functions, namely:

$$\begin{array}{c} c^S_{s\alpha}(r) = c_{s\alpha}(r) - c^L_{s\alpha}(r) = c_{s\alpha}(r) + \beta u^L_{s\alpha}(r) \\ \gamma^S_{s\alpha}(r) = h_{s\alpha}(r) - c^S_{s\alpha}(r) = \gamma_{s\alpha}(r) - \beta u^L_{s\alpha}(r) \end{array} \tag{5.20}$$

Short-range functions are convenient for the numerical treatment of the task. They decay rapidly with a distance and can be effectively approximated by functions with a small support. Putting the short-range functions to the closure relation we obtain the following closure for the short-range functions:

$$c^S_{s\alpha}(r) = e^{-\beta u^S_{s\alpha}(r) - \gamma^S_{s\alpha}(r) + B_{s\alpha}(r)} - \gamma^S_{s\alpha}(r) - 1 \tag{5.21}$$

Following Ref. [120] we write the analytical representation of the Bessel-Fourier transform for long-rang potential $u_{s\alpha}$:

$$\hat{u}^L_{s\alpha}(k) = \frac{4\pi q_s q_\alpha}{k^2} e^{\frac{-k^2}{4t^2}} \tag{5.22}$$

We note that this function in the Fourier space is proportional to $\frac{1}{k^2}$ and thus decays rapidly. This enables us to use small-support grids in the Fourier space. Putting (5.20) to equation (5.9) we obtain the following relations for the short-range functions:

$$\hat{\gamma}^S_{s\alpha}(k) = \sum_{s'\nu} \hat{\omega}_{ss'}(k) \cdot \left(\hat{c}^S_{s'\nu}(k) + \hat{u}^L_{s\alpha}(k)\right) \cdot \left(\hat{\omega}^{\mathrm{solv}}_{\alpha\nu}(k) + \rho \hat{h}^{\mathrm{solv}}_{\alpha\nu}(k)\right) - \hat{c}^S_{s\alpha}(k) \tag{5.23}$$

5.1.4 RISM equations in a recurrent form

Equations (5.23), (5.21) can be represented in a matrix form. Let the numbers of the sites in the solute and solvent molecules be M and K respectively. We define the matrix of the short-range indirect correlation functions $\mathbf{\Gamma} = [\gamma^S_{s\alpha}(r)]_{M\times K}$ and a matrix of the Fourier-transformed long range potentials $\hat{\mathbf{U}}^{\mathbf{L}} = [\hat{u}^L_{s\alpha}(k)]_{M\times K}$ in a following way:

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma^S_{11}(r) & \dots & \gamma^S_{1K}(r) \\ \vdots & \ddots & \vdots \\ \gamma^S_{M1}(r) & \dots & \gamma^S_{MK}(r) \end{pmatrix} \tag{5.24}$$

$$\hat{\mathbf{U}}^{\mathbf{L}} = \begin{pmatrix} u^L_{11}(k) & \dots & u^L_{1K}(k) \\ \vdots & \ddots & \vdots \\ u^L_{M1}(k) & \dots & u^L_{MK}(k) \end{pmatrix} \tag{5.25}$$

We define the *closure operator* $C[\mathbf{\Gamma}]$ which is the matrix analog of the closure (5.21) in a following way:

$$C[\mathbf{\Gamma}] = \left[e^{-\beta u_{s\alpha}(r) + \gamma^S_{s\alpha}(r) + B_{s\alpha}(r)} - \gamma^S_{s\alpha}(r) - 1 \right]_{M \times K} \tag{5.26}$$

We introduce the matrices of the solute and solvent intermolecular correlation functions $\hat{\mathbf{W}}^u = [\hat{\omega}^u_{ss'}(k)]_{M\times M}$, $\hat{\mathbf{W}}^v = [\hat{\omega}^v_{\alpha\alpha'}(k)]_{K\times K}$ correspondingly, and also the matrix of the solvent total correlation functions $\hat{\mathbf{H}}^{\text{solv}} = [\hat{h}^{vv}_{ss'}(k)]_{K\times K}$. We introduce the matrix of the solvent susceptibility functions in a following way:

$$\hat{\mathbf{X}} = \hat{\mathbf{W}}^v + \rho \hat{\mathbf{H}}^{\text{solv}} \tag{5.27}$$

where $\hat{\mathbf{X}} = [\hat{\chi}_{\alpha\beta}]_{K\times K}$, $\chi_{\alpha\beta}(k) = \hat{\omega}^v_{\alpha\beta}(k) + \rho \hat{h}^{vv}_{\alpha\beta}(k)$

We use the operators $\mathcal{T}$ and $\mathcal{T}^{-1}$, which perform element-by-element direct and inverse Bessel-Fourier transform correspondingly:

$$\begin{array}{l} \mathcal{T}\left[(\gamma_{s\alpha}(r)]_{M\times K}\right] = [\mathcal{T}\gamma_{s\alpha}(r)]_{M\times K} \\ \mathcal{T}^{-1}\left[(\hat{\gamma}_{s\alpha}(k)]_{M\times K}\right] = [\mathcal{T}^{-1}\hat{\gamma}_{s\alpha}(k)]_{M\times K} \end{array} \tag{5.28}$$

where symbols $\mathcal{T}, \mathcal{T}^{-1}$ in the right hand side of these equations stay for direct and inverse Bessel-Fourier transforms correspondingly. Using these definitions we rewrite the RISM equations (5.23) in a recurrent form:

$$\Gamma = F[\mathbf{\Gamma}] \tag{5.29}$$

where the operator $F[\mathbf{\Gamma}]$ is defined in a following way:

$$F[\Gamma] = \mathcal{T}^{-1}\left(\hat{\mathbf{W}} \cdot (\mathcal{T}\left(C[\Gamma]\right) - \beta \hat{\mathbf{U}}^{\mathbf{L}}) \cdot \hat{\mathbf{X}} \right) - \mathcal{F}\left(C[\Gamma]\right) \tag{5.30}$$

This equation is comprised by the closure relation in the **real** space:

$$C^S = \mathcal{C}[\Gamma] = \left[e^{-\beta u^S_{s\alpha}(r) + \gamma_{s\alpha}(r) + B_{s\alpha}(r)} - \gamma_{s\alpha}(r) - 1 \right]_{M\times K}. \tag{5.31}$$

5.2 Discretization of the problem

In both, real and Fourier space, we discretized the problem on a uniform grid. The grid sizes in the real and Fourier spaces are connected by $\Delta k = \frac{\pi}{\Delta R}$. We denote the real space grid with N points and step size ΔR as $\{N, \Delta R\}$. Each grid is also characterized by the *cutoff distance*, which is the upper limit of the support. For the grid $\{N, \Delta R\}$, the cutoff distance is $R_{\text{cutoff}} = N\Delta R$. The corresponding grid in the Fourier space is denoted by $\{N, \Delta k\}$.

Functions are represented by vectors which contain the values of the functions on the grid points. We denote these *grid functions* by bold letters and indicate their grid in the **superscript**:

$$\boldsymbol{\gamma}_{s\alpha}^{\{N,\Delta R\}} = \left(\gamma_{s\alpha}^{S}(\Delta R), \ldots, \gamma_{s\alpha}^{S}(N\Delta R)\right), \tag{5.32}$$

$$\mathbf{c}_{s\alpha}^{S\{N,\Delta R\}} = \left(c_{s\alpha}^{S}(\Delta R), \ldots, c_{s\alpha}^{S}(N\Delta R)\right), \tag{5.33}$$

$$\mathbf{u}_{s\alpha}^{S\{N,\Delta R\}} = \left(u_{s\alpha}^{S}(\Delta R), \ldots, u_{s\alpha}^{S}(N\Delta R)\right). \tag{5.34}$$

The grid functions in the Fourier space are indicated by the hat symbol ($\hat{}$):

$$\hat{\boldsymbol{\gamma}}_{s\alpha}^{\{N,\Delta k\}} = \left(\hat{\gamma}_{s\alpha}^{S}(\Delta k), \ldots, \hat{\gamma}_{s\alpha}^{S}(N\Delta k)\right), \tag{5.35}$$

$$\hat{\mathbf{c}}_{s\alpha}^{S\{N,\Delta k\}} = \left(\hat{c}_{s\alpha}^{S}(\Delta k), \ldots, \hat{c}_{s\alpha}^{S}(N\Delta k)\right), \tag{5.36}$$

$$\hat{\mathbf{w}}_{ss'}^{\{N,\Delta k\}} = \left(\hat{w}_{ss'}(\Delta k), \ldots, \hat{w}_{ss'}(N\Delta k)\right), \tag{5.37}$$

$$\hat{\boldsymbol{\chi}}_{\alpha\beta}^{\{N,\Delta k\}} = \left(\hat{\chi}_{\alpha\beta}(\Delta k), \ldots, \hat{\chi}_{\alpha\beta}(N\Delta k)\right), \tag{5.38}$$

$$\hat{\mathbf{u}}_{s\alpha}^{L\{N,\Delta k\}} = \left(\hat{u}_{s\alpha}^{L}(\Delta k), \ldots, \hat{u}_{s\alpha}^{L}(N\Delta k)\right). \tag{5.39}$$

Similarly, matrix-valued grid functions carry a **subscript** denoting the grid. Matrices without a hat sign symbolize real-space functions: $\mathbf{\Gamma}_{\{N,\Delta R\}} = [\boldsymbol{\gamma}_{s\alpha}^{\{N,\Delta R\}}(k)]_{M\times K}$. We use the hat symbol ($\hat{}$) to denote the matrices of functions in the Fourier space: $\hat{\mathbf{\Gamma}}_{\{N,\Delta k\}} = [\hat{\boldsymbol{\gamma}}_{s\alpha}^{\{N,\Delta k\}}(k)]_{M\times K}$, $\hat{\mathbf{C}}_{\{N,\Delta k\}}^{S} = [\hat{\mathbf{c}}_{s\alpha}^{S\{N,\Delta k\}}]_{M\times K}$, $\hat{\mathbf{W}}_{\{N,\Delta k\}} = \left[\hat{\mathbf{w}}_{ss'}^{\{N,\Delta k\}}\right]_{M\times M}$, $\hat{\mathbf{U}}_{\{N,\Delta k\}}^{L} = [\hat{\mathbf{u}}_{s\alpha}^{L\{N,\Delta k\}}]_{M\times K}$, and $\hat{\mathbf{X}}_{\{N,\Delta k\}} = [\hat{\boldsymbol{\chi}}_{\alpha\beta}^{\{N,\Delta k\}}]_{K\times K}$.

To map the grid functions from the real to the Fourier space and back, we use the discrete forward and inverse Bessel-Fourier transformations $\mathcal{T}_{\{N,\Delta R\}}[\cdot]$, $\mathcal{T}_{\{N,\Delta k\}}^{-1}[\cdot]$ respectively:

$$\hat{\mathbf{f}}^{\{N,\Delta k\}} = \mathcal{T}_{\{N,\Delta R\}}[\mathbf{f}^{\{N,\Delta R\}}], \tag{5.40}$$

$$\mathbf{f}^{\{N,\Delta R\}} = \mathcal{T}_{\{N,\Delta k\}}^{-1}[\hat{\mathbf{f}}^{\{N,\Delta k\}}]. \tag{5.41}$$

Vectors $\mathbf{f}^{\{N,\Delta R\}}$, $\hat{\mathbf{f}}^{\{N,\Delta k\}}$ are defined as

$$\mathbf{f}^{\{N,\Delta R\}} = \left(f(r_1), \ldots, f(r_N)\right), \qquad r_n = n\Delta R, \tag{5.42}$$

$$\hat{\mathbf{f}}^{\{N,\Delta k\}} = \left(\hat{f}(k_1), \ldots, \hat{f}(k_N)\right), \qquad k_m = m\Delta k, \tag{5.43}$$

and components of these vectors are connected via the relations

$$\hat{f}(k_m) = \frac{4\pi}{k_m} \sum_{n=1}^{N} f(r_n) r_n \sin(\frac{\pi m n}{N}) \Delta R, \tag{5.44}$$

$$f(r_n) = \frac{1}{2\pi^2 r_n} \sum_{m=1}^{N} \hat{f}(k_m) k_m \sin(\frac{\pi m n}{N}) \Delta k. \tag{5.45}$$

Discrete analogues of equation (5.30) and the closure relation (5.31) are formulated as

$$\begin{aligned}&\hat{\boldsymbol{\Gamma}}_{\{N,\Delta R\}} = \\ &\hat{\mathbf{W}}_{\{N,\Delta R\}} \cdot \left(\hat{\mathbf{C}}^{\mathbf{S}}_{\{N,\Delta R\}} - \beta \hat{\mathbf{U}}^{\mathbf{L}}_{\{N,\Delta R\}}\right) \cdot \hat{\mathbf{X}}_{\{N,\Delta R\}} - \hat{\mathbf{C}}^{\mathbf{S}}_{\{N,\Delta R\}},\end{aligned} \tag{5.46}$$

$$\begin{aligned}&\mathbf{C}^{\mathbf{S}}{}_{\{N,\Delta R\}} = \mathcal{C}[\boldsymbol{\Gamma}_{\{N,\Delta R\}}] = \\ &\left[e^{-\beta \mathbf{u}^{\mathbf{S}}{}_{s\alpha} + \boldsymbol{\gamma}_{s\alpha} + \mathbf{B}_{s\alpha}} - \boldsymbol{\gamma}_{s\alpha} - 1\right]_{M \times K},\end{aligned} \tag{5.47}$$

where the mathematical operations between vectors are understood to be entry-wise.

5.3 Picard iteration

Combining (5.46) and(5.47) , we can define the iterative operator $\mathcal{K}_{\{N,\Delta R\}}[\cdot]$:

$$\begin{aligned}&\mathcal{K}_{\{N,\Delta R\}}[\boldsymbol{\Gamma}_{\{N,\Delta R\}}] \equiv \mathcal{T}^{-1}_{\{N,\Delta k\}}\Big[\\ &\hat{\mathbf{W}}_{\{N,\Delta k\}}\Big(\mathcal{T}_{\{N,\Delta R\}}\left[\mathcal{C}[\boldsymbol{\Gamma}_{\{N,\Delta R\}}]\right] - \beta \hat{\mathbf{U}}^{\mathbf{L}}_{\{N,\Delta k\}}\Big)\hat{\mathbf{X}}_{\{N,\Delta k\}} \\ &-\mathcal{T}_{\{N,\Delta R\}}\left[\mathcal{C}[\boldsymbol{\Gamma}_{\{N,\Delta R\}}]\right]\Big],\end{aligned} \tag{5.48}$$

We consider the generalized task

$$\boldsymbol{\Gamma}_{\{N,\Delta R\}} = \mathcal{K}_{\{N,\Delta R\}}[\boldsymbol{\Gamma}_{\{N,\Delta R\}}] + \mathbf{f}_{\{N,\Delta R\}} \tag{5.49}$$

for a given right-hand side vector $\mathbf{f}_{\{N,\Delta R\}}$. Problem (5.46) - (5.47) corresponds to the case $\mathbf{f}_{\{N,\Delta R\}} = \mathbf{0}$. The necessity of introducing the generalized problem will be described bellow during the description of the multi-grid method. The n-th iterate of an iterative scheme is denoted by $\boldsymbol{\Gamma}^{(n)}_{\{N,\Delta R\}}$. The damped Picard iteration with the damping parameter λ is defined as

$$\boldsymbol{\Gamma}^{(n+1)}_{\{N,\Delta R\}} = (1-\lambda)\boldsymbol{\Gamma}^{(n)}_{\{N,\Delta R\}} + \lambda \boldsymbol{\Gamma'}_{\{N,\Delta R\}}, \tag{5.50}$$

where $\boldsymbol{\Gamma'}_{\{N,\Delta R\}}$ abbreviates

$$\boldsymbol{\Gamma'}_{\{N,\Delta R\}} = \mathcal{K}_{\{N,\Delta R\}}[\boldsymbol{\Gamma}^{(n)}_{\{N,\Delta R\}}] + \mathbf{f}_{\{N,\Delta R\}}. \tag{5.51}$$

We use a short notation for this operator:

$$\Upsilon_{\{N,\Delta R\}}[\boldsymbol{\Gamma}^{(n)}_{\{N,\Delta R\}}, \mathbf{f}_{\{N,\Delta R\}}] \equiv (1-\lambda)\boldsymbol{\Gamma}^{(n)}_{\{N,\Delta R\}} + \lambda \boldsymbol{\Gamma'}_{\{N,\Delta R\}}. \tag{5.52}$$

One iteration step of the algorithm consists of the partial steps

$$\begin{aligned}&\boldsymbol{\Gamma}^{(n)}_{\{N,\Delta R\}} \xrightarrow[(5.47)]{closure} \mathbf{C}^{\mathbf{S}}{}_{\{N,\Delta R\}} \xrightarrow[(5.40)]{BFT} \hat{\mathbf{C}}^{\mathbf{S}}_{\{N,\Delta k\}} \xrightarrow[(5.46)]{RISM} \\ &\hat{\boldsymbol{\Gamma}}_{\{N,\Delta k\}} \xrightarrow[(5.41)]{IBFT} \boldsymbol{\Gamma'}_{\{N,\Delta R\}} \xrightarrow[(5.50)]{damping} \boldsymbol{\Gamma}^{(n+1)}_{\{N,\Delta R\}}.\end{aligned} \tag{5.53}$$

As a measure of accuracy we use the L_2 norm between two successive iterates averaged over all site-site functions:

$$\begin{gathered}\left\|\mathbf{\Gamma}^{(n+1)}_{\{N,\Delta R\}} - \mathbf{\Gamma}^{(n)}_{\{N,\Delta R\}}\right\| = \\ \frac{1}{MK}\sum_{s\alpha}\sqrt{\sum_{m=1}^{N}\left(\gamma^{(n+1)}_{s\alpha}(m\Delta R) - \gamma^{(n)}_{s\alpha}(m\Delta R)\right)^2 \Delta R.}\end{gathered} \tag{5.54}$$

We stop the iteration when the iterates differ by less than a given threshold ε:

$$\left\|\mathbf{\Gamma}^{(n+1)}_{\{N,\Delta R\}} - \mathbf{\Gamma}^{(n)}_{\{N,\Delta R\}}\right\| \leq \varepsilon. \tag{5.55}$$

By $n(N, \Delta R, \varepsilon)$ we denote the minimal number n such that (5.55) holds. So, using the operator power notation, the iterative process to obtain the solution $\mathbf{\Gamma}^{\varepsilon}_{\{N,\Delta R\}}$ with accuracy ε can be written as

$$\mathbf{\Gamma}^{\varepsilon}_{\{N,\Delta R\}} = \left(\Upsilon_{\{N,\Delta R\}}\right)^{n(N,\Delta R,\varepsilon)} [\mathbf{\Gamma}^{(0)}_{\{N,\Delta R\}}, \mathbf{f}_{\{N,\Delta R\}}]. \tag{5.56}$$

5.4 Moving between the grids

In the multi-scale methods which we discuss below, several grids are used. Below we define operators, which map grid functions from one grid to another: the *restriction operator* $r[\cdot]$, the *interpolation operator* $p[\cdot]$ and the *extension operator* $e[\cdot]$

The *restriction operator* $r[\cdot]$ maps the matrix of grid functions to the coarser grid with doubled grid size and, therefore, half the number of grid points. For example, the restriction maps from the grid $\{2N, \Delta R\}$ to the grid $\{N, 2\Delta R\}$:

$$r[\mathbf{\Gamma}_{\{2N,\Delta R\}}] = \mathbf{\Gamma}_{\{N,2\Delta R\}}. \tag{5.57}$$

In the current work we use the *trivial injection* as a restriction operator. Let $\mathbf{\Gamma}_{\{2N,\Delta R\}} = [\boldsymbol{\gamma}^{\{2N,\Delta R\}}_{s\alpha}]_{M\times K}$, where

$$\boldsymbol{\gamma}^{\{2N,\Delta R\}}_{s\alpha} = (\gamma_{s\alpha}(\Delta R), \gamma_{s\alpha}(2\Delta R)\ldots, \gamma_{s\alpha}(2N\Delta R)) \tag{5.58}$$

and $\mathbf{\Gamma}_{\{N,2\Delta R\}} = [\boldsymbol{\gamma}^{\{N,2\Delta R\}}_{s\alpha}]_{M\times K} = r[\mathbf{\Gamma}_{\{2N,\Delta R\}}]$. Then the vectors $\boldsymbol{\gamma}^{\{N,2\Delta R\}}_{s\alpha}$ are defined by

$$\boldsymbol{\gamma}^{\{N,2\Delta R\}}_{s\alpha} = (\gamma^{S}_{s\alpha}(2\Delta R), \gamma^{S}_{s\alpha}(4\Delta R), \ldots, \gamma^{S}_{s\alpha}(2N\Delta R)). \tag{5.59}$$

We should mention that the restriction operator $r[\cdot]$ is linear:

$$\begin{gathered}r[a\mathbf{\Gamma}'_{\{2N,\Delta R\}} + b\mathbf{\Gamma}''_{\{2N,\Delta R\}}] = \\ a\cdot r[\mathbf{\Gamma}'_{\{2N,\Delta R\}}] + b\cdot r[\mathbf{\Gamma}''_{\{2N,\Delta R\}}].\end{gathered} \tag{5.60}$$

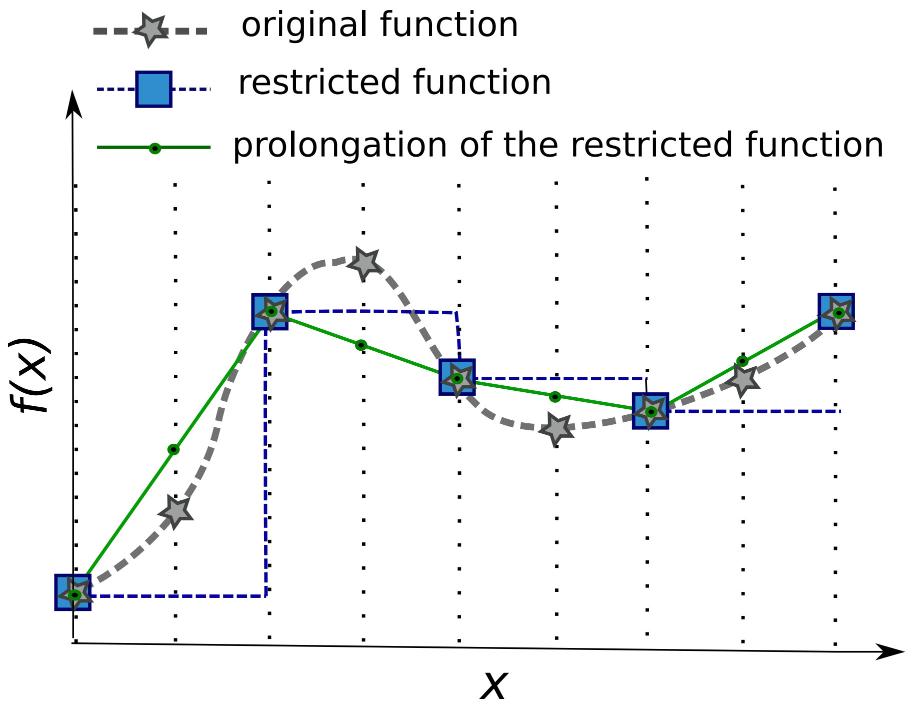


Figure 5.1: Illustration of the restriction and the prolongation operators. The values of the functions on different grids are shown with markers. The lines which connect these markers are used for better visibility only. Three functions are demonstrated: initial function $f(x)$ (star markers, thick dashed line), restricted function $r[f(x)]$ (square markers, thin dashed line) and prolongation of the restricted function $p[r[f(x)]]$ (round markers, solid line).

The *interpolation operator* $p[\cdot]$ maps the matrix of grid functions to the grid with half the grid size and, thereby, the double number of grid points. For example, the interpolation maps from the grid $\{N, 2\Delta R\}$ to the grid $\{2N, \Delta R\}$:

$$p[\mathbf{\Gamma}_{\{N,2\Delta R\}}] = \tilde{\mathbf{\Gamma}}_{\{2N,\Delta R\}} = [\tilde{\boldsymbol{\gamma}}_{s\alpha}^{\{2N,\Delta R\}}]. \tag{5.61}$$

In the current work we use *cubic spline interpolation* as interpolation operator. Let $\mathbf{\Gamma}_{\{N,2\Delta R\}} = [\boldsymbol{\gamma}_{s\alpha}^{\{N,2\Delta R\}}]_{M\times K}$, where vectors $\{\boldsymbol{\gamma}_{s\alpha}^{\{N,2\Delta R\}}\}$ are defined by (5.59) . Then $\tilde{\boldsymbol{\gamma}}_{s\alpha}^{\{2N,\Delta R\}}$ in (5.61) are defined by

$$\begin{aligned}\tilde{\boldsymbol{\gamma}}_{s\alpha}^{\{2N,\Delta R\}} = \; & (\tilde{\gamma}_{s\alpha}^{S}(\Delta R), \gamma_{s\alpha}^{S}(2\Delta R), \tilde{\gamma}_{s\alpha}^{S}(3\Delta R), \ldots \\ & \ldots, \tilde{\gamma}_{s\alpha}^{S}((2N-1)\Delta R), \gamma_{s\alpha}^{S}(2N\Delta R)),\end{aligned} \tag{5.62}$$

where the values $\tilde{\gamma}_{s\alpha}^{S}((2k-1)\Delta R)$ are obtained from the values $\gamma_{s\alpha}^{S}(2k\Delta R)$ using the cubic spline interpolation.

We note that $r[\cdot]$ is the left-inverse of $p[\cdot]$, i.e., $r[p[\mathbf{\Gamma}_{\{N,2\Delta R\}}]] = \mathbf{\Gamma}_{\{N,2\Delta R\}}$, but not the right-inverse:

$$\tilde{\mathbf{\Gamma}}_{\{2N,\Delta R\}} = p[r[\mathbf{\Gamma}_{\{2N,\Delta R\}}]] \neq \mathbf{\Gamma}_{\{2N,\Delta R\}}. \tag{5.63}$$

This fact is demonstrated in Figure 5.1. However, sufficiently smooth functions satisfy

$$\tilde{\mathbf{\Gamma}}_{\{2N,\Delta R\}} = p[r[\mathbf{\Gamma}_{\{2N,\Delta R\}}]] = \mathbf{\Gamma}_{\{2N,\Delta R\}} + \boldsymbol{O}(\Delta R). \tag{5.64}$$

The *extension operator* $e[\cdot]$ does not change the grid size, but doubles the number of points. For example, the extension maps from the grid $\{N, \Delta R\}$ to the grid $\{2N, \Delta R\}$:

$$e[\mathbf{\Gamma}_{\{N,\Delta R\}}] = \mathbf{\Gamma}_{\{2N,\Delta R\}} = [\boldsymbol{\gamma}_{s\alpha}^{\{2N,\Delta R\}}]_{M\times K} \tag{5.65}$$

Indirect correlation functions decay fast to zero as the distance increases. Thus, it is natural to extend them by zeros yielding the *zero extension operator* which we use in the current work. Let $\mathbf{\Gamma}_{\{N,\Delta R\}} = [\boldsymbol{\gamma}_{s\alpha}^{\{N,\Delta R\}}]_{M\times K}$ with vectors $\{\boldsymbol{\gamma}_{s\alpha}^{\{N,\Delta R\}}\}$ defined by (5.32). Then the functions $\{\boldsymbol{\gamma}_{s\alpha}^{\{2N,\Delta R\}}\}$ in (5.65) are defined by

$$\boldsymbol{\gamma}_{s\alpha}^{\{2N,\Delta R\}} = (\gamma_{s\alpha}(\Delta R), \ldots, \gamma_{s\alpha}(N\Delta R), \underbrace{0, \ldots, 0}_{N}). \tag{5.66}$$

5.5 Nested Picard iteration

Having at hand different grids, the idea of the nested Picard iteration [87] is straightforward: use as an initial guess the (approximate) solution from the coarse grid with a smaller number of grid points. Here we exploit that computations in the coarse grid are cheaper. Below we describe the scheme of the *nested Picard iteration.* Consider two grids: the "coarse" grid $\{N, 2\Delta R\}$ and the "fine" grid $\{2N, \Delta R\}$. We start from the coarse-grid solution $\mathbf{\Gamma}^{(0)}_{\{N,2\Delta R\}}$. We perform an iteration process of type (5.56) to obtain a solution with accuracy ε on the coarse grid, interpolate it to the fine grid and use it as the initial approximation for the fine-grid iteration.

The scheme for performing the two-grid nested Picard iteration is written as follows:

$$\begin{aligned} &\mathbf{\Gamma}^{(0)}_{\{N,2\Delta R\}} \xrightarrow{\Upsilon_{\{N,2\Delta R\}}} \mathbf{\Gamma}^{\varepsilon}_{\{N,2\Delta R\}} \xrightarrow{p} \\ &\rightarrow \mathbf{\Gamma}^{(0)}_{\{2N,\Delta R\}} \xrightarrow{\Upsilon_{\{2N,\Delta R\}}} \mathbf{\Gamma}^{\varepsilon}_{\{2N,\Delta R\}}. \end{aligned} \tag{5.67}$$

The nested Picard iteration scheme for more than two grids $\{N, 2^L\Delta R\}$, $\{2N, 2^{L-1}\Delta R\}$, ..., $\{2^L N, \Delta R\}$ with the same cutoff distance $R_{\text{cutoff}} = 2^L N\Delta R$ is

$$\begin{aligned} &\mathbf{\Gamma}^{(0)}_{\{N,2^L\Delta R\}} \xrightarrow{\Upsilon_{\{N,2^L\Delta R\}}} \mathbf{\Gamma}^{\varepsilon}_{\{2N,2^{L-1}\Delta R\}} \\ &\xrightarrow{p} \mathbf{\Gamma}^{(0)}_{\{2N,2^{L-1}\Delta R\}} \xrightarrow{\Upsilon_{\{2N,2^{L-1}\Delta R\}}} \ldots \\ &\ldots \xrightarrow{p} \mathbf{\Gamma}^{(0)}_{\{2^L N,\Delta R\}} \xrightarrow{\Upsilon_{\{2^L N,\Delta R\}}} \mathbf{\Gamma}^{\varepsilon}_{\{2^L N,\Delta R\}}. \end{aligned} \tag{5.68}$$

To obtain the solution on a grid with larger cutoff distance, we may continue the process in a similar way using the extension operator $e[\cdot]$:

$$\begin{aligned} &\mathbf{\Gamma}^{(0)}_{\{2^{L+1}N,\Delta R\}} = e[\mathbf{\Gamma}^{\varepsilon}_{\{2^{L}N,\Delta R\}}], \\ &\mathbf{\Gamma}^{\varepsilon}_{\{2^{L+1}N,\Delta R\}} = \left(\Upsilon_{\{2^{L+1}N,\Delta R\}}\right)^{n(2^{L+1}N,\Delta R,\varepsilon)}[\mathbf{\Gamma}^{(0)}_{\{2^{L+1}N,\Delta R\}}]. \end{aligned} \tag{5.69}$$

The same process can be defined for multiple grids $\{2^{L}N,\Delta R\}$, $\{2^{L+1}N,\Delta R\}$, ..., $\{2^{L+P}N,\Delta R\}$:

$$\begin{aligned} &\mathbf{\Gamma}^{\varepsilon}_{\{2^{L}N,\Delta R\}} \xrightarrow{e} \mathbf{\Gamma}^{(0)}_{\{2^{L+1}N,\Delta R\}} \xrightarrow{\Upsilon_{\{2^{L+1}N,\Delta R\}}} \mathbf{\Gamma}^{\varepsilon}_{\{2^{L+1}N,\Delta R\}} \xrightarrow{e} \\ &\ldots \xrightarrow{e} \mathbf{\Gamma}^{(0)}_{\{2^{L+P}N,\Delta R\}} \xrightarrow{\Upsilon_{\{2^{L+P}N,\Delta R\}}} \mathbf{\Gamma}^{\varepsilon}_{\{2^{L+P}N,\Delta R\}}. \end{aligned} \tag{5.70}$$

5.6 Two- and multi-grid iterations

Although the nested Picard iteration scheme is able to essentially enhance the performance of numerical iteration, there is a drawback which limits its efficiency. Consider two grids $\{N, 2\Delta R\}$ and $\{2N, \Delta R\}$, to which we refer below as *coarse* and *fine* grid, respectively. Denote the exact solutions $\mathbf{\Gamma}^{*}_{\text{coarse}}$, $\mathbf{\Gamma}^{*}_{\text{fine}}$ on the respective coarse and fine grids by

$$\begin{aligned} \mathbf{\Gamma}^{*}_{\text{coarse}} &= \mathcal{K}_{\text{coarse}}[\mathbf{\Gamma}^{*}_{\text{coarse}}] + r[\mathbf{f}_{\text{fine}}], \\ \mathbf{\Gamma}^{*}_{\text{fine}} &= \mathcal{K}_{\text{fine}}[\mathbf{\Gamma}^{*}_{\text{fine}}] + \mathbf{f}_{\text{fine}}. \end{aligned} \tag{5.71}$$

We note, that the restricted fine grid solution **is not** the coarse grid solution:

$$\mathbf{\Gamma}^{*}_{\text{coarse}} \neq r[\mathbf{\Gamma}^{*}_{\text{fine}}] \tag{5.72}$$

Due to this fact, the coarse-grid iteration is not able to give a very good approximation of the fine-grid solution, which limits the performance of the nested iterative schemes. The multi-grid scheme is able to overcome this limitation. For the complete description of the different multi-grid schemes we refer to the book [87]. Below we briefly describe the multi-grid-based algorithm for solving the RISM equation which is used in the current work.

The following *grid difference operator* $G[\cdot]$ applies to fine-grid functions and indicates the difference of $\mathcal{K}_{\text{fine}}$ and $\mathcal{K}_{\text{coarse}}$:

$$G[\mathbf{\Gamma}_{\text{fine}}] = r[\mathcal{K}_{\text{fine}}[\mathbf{\Gamma}_{\text{fine}}]] - \mathcal{K}_{\text{coarse}}[r[\mathbf{\Gamma}_{\text{fine}}]]. \tag{5.73}$$

Let us consider the task

$$\mathbf{\Gamma}_{\text{coarse}} = \mathcal{K}_{\text{coarse}}[\mathbf{\Gamma}_{\text{coarse}}] + r[\mathbf{f}_{\text{fine}}] + G[\mathbf{\Gamma}^{*}_{\text{fine}}]. \tag{5.74}$$

Substituting here the restricted exact fine grid solution $\mathbf{\Gamma}_{\text{coarse}} = r[\mathbf{\Gamma}^{*}_{\text{fine}}]$ and using the definition (5.73) , we have:

$$r[\mathbf{\Gamma}^{*}_{\text{fine}}] = r[\mathcal{K}_{\text{fine}}[\mathbf{\Gamma}^{*}_{\text{fine}}]] + r[\mathbf{f}_{\text{fine}}]. \tag{5.75}$$

This equality holds due to the linearity of the restriction operator (5.60) and second equality in (5.71). One can see, that the task (5.74) is of the form of (5.49) where $\mathbf{f}_{\text{coarse}} = r[\mathbf{f}_{\text{fine}}] + G[\mathbf{\Gamma}^*_{\text{fine}}]$. This shows the role of the vector $\mathbf{f}_{\text{coarse}}$: it accumulates the grid differences between the finer and coarser grids during the multi-grid iteration.

The task (5.74) can be used to find the solution $r[\mathbf{\Gamma}^*_{\text{fine}}]$. We do not know the exact difference $G[\mathbf{\Gamma}^*_{\text{fine}}]$. However, even if the approximate solution is far from the exact solution, the value $G[\mathbf{\Gamma}^{(0)}_{\text{fine}}]$ can be accurate enough:

$$\left\| G[\mathbf{\Gamma}^{(0)}_{\text{fine}}] - G[\mathbf{\Gamma}^*_{\text{fine}}] \right\| \ll \left\| \mathbf{\Gamma}^{(0)}_{\text{fine}} - \mathbf{\Gamma}^*_{\text{fine}} \right\| . \tag{5.76}$$

In the two-grid scheme one performs a small number ν_1 of fine-grid iteration steps before solving the coarse grid task and uses the coarse-grid solution to eliminate the low-frequency errors of fine-grid iterate. However, for some operators the interpolation of the coarse-grid solution may be not smooth enough. That is why one may need to perform some additional number ν_2 of so-called *smoothing* fine-grid iteration steps. Let $\mathbf{\Gamma}^{(n)}_{\text{fine}}$ be the fine-grid approximation on the n-th step of two-grid iteration. The two-grid iteration process can be written in a following way:

$$\mathbf{\Gamma}^{(n+1)}_{\text{fine}} = \mathcal{T}[\mathbf{\Gamma}^{(n)}_{\text{fine}}, \mathbf{f}_{\text{fine}}] \tag{5.77}$$

where the two-grid operator $\mathcal{T}[\cdot, \cdot]$ is defined by the following algorithm:

Algorithm 5.1 *RISM Two-grid Operator*

Input: $\mathbf{\Gamma}^{(n)}_{\text{fine}}$, $\mathbf{f}_{\text{fine}}$
Output: $\mathbf{\Gamma}^{(n+1)}_{\text{fine}}$

1. Perform ν_1 fine-grid iteration steps:

$$\mathbf{\Gamma}'_{\text{fine}} = (\Upsilon_{\text{fine}})^{\nu_1}\,[\mathbf{\Gamma}^{(n)}_{\text{fine}}, \mathbf{f}_{\text{fine}}].$$

2. Define a coarse-grid analogue of $\mathbf{\Gamma}'_{\text{fine}}$:

$$\mathbf{\Gamma}^{(0)}_{\text{coarse}} = r[\mathbf{\Gamma}'_{\text{fine}}].$$

3. Calculate the grid correction $G[\mathbf{\Gamma}'_{\text{fine}}]$:

$$G[\mathbf{\Gamma}'_{\text{fine}}] = r[\mathcal{K}_{\text{fine}}[\mathbf{\Gamma}'_{\text{fine}}]] - \mathcal{K}_{\text{coarse}}[\mathbf{\Gamma}^{(0)}_{\text{coarse}}]$$

4. Determine the solution $\mathbf{\Gamma}^{*}_{\text{coarse}}$ the coarse grid problem

$$\mathbf{\Gamma}_{\text{coarse}} = \mathcal{K}_{\text{coarse}}[\mathbf{\Gamma}_{\text{coarse}}] + G[\mathbf{\Gamma}'_{\text{fine}}] + r[\mathbf{f}_{\text{fine}}]. \tag{5.78}$$

5. Add the coarse-grid correction:

$$\mathbf{\Gamma}''_{\text{fine}} = \mathbf{\Gamma}'_{\text{fine}} + p[\mathbf{\Gamma}^{*}_{\text{coarse}} - \mathbf{\Gamma}^{(0)}_{\text{coarse}}].$$

6. Perform ν_2 smoothing fine-grid iteration steps:

$$\mathbf{\Gamma}^{(n+1)}_{\text{fine}} = (\Upsilon_{\text{fine}})^{\nu_2}\,[\mathbf{\Gamma}''_{\text{fine}}, \mathbf{f}_{\text{fine}}].$$

In the two-grid algorithm it is not specified how the coarse-grid equation of step 4 is solved. If the algorithm described above is used recursively for solving the coarse-grid problem, we obtain the *multi-grid iterative scheme* [87]. Assume that we have the grids $\{N, 2^L\Delta R\}$, $\{2N, 2^{L-1}\Delta R\}$, ..., $\{2^L N, \Delta R\}$. We will use the subscript *grid* to refer to any of these grids. On each grid we define the multi-grid iteration with iterative operator $\mathcal{M}^{level}_{grid}[\mathbf{\Gamma}^{(n)}_{grid}, \mathbf{f}_{grid}]$:

$$\mathbf{\Gamma}^{(n+1)}_{grid} = \mathcal{M}^{level}_{grid}[\mathbf{\Gamma}^{(n)}_{grid}, \mathbf{f}_{grid}], \tag{5.79}$$

where $\mathbf{\Gamma}^{(n)}_{grid}$ is the n-th multi-grid iterate, $\mathbf{f}_{grid}$ is given, and the superscript *level* indicates the number of recursions which are done while calculating the operator. The multi-grid iteration converges to the solution

$$\mathbf{\Gamma}_{grid} = \mathcal{K}_{grid}[\mathbf{\Gamma}_{grid}] + \mathbf{f}_{grid}. \tag{5.80}$$

The multi-grid operator at level zero is a single-grid solver of (5.80) on the coarsest grid.

In our work we use n steps of the damped Picard iteration:

$$\begin{gathered}\mathcal{M}^{0}_{\{N,2^L\Delta R\}}[\mathbf{\Gamma}^{(0)}_{\{N,2^L\Delta R\}},\mathbf{f}_{\{N,2^L\Delta R\}}]=\\ \left(\Upsilon_{\{N,2^L\Delta R\}}\right)^n[\mathbf{\Gamma}^{(0)}_{\{N,2^L\Delta R\}},\mathbf{f}_{\{N,2^L\Delta R\}}]\end{gathered}\tag{5.81}$$

The proper choice of the number n of iteration steps on the coarsest grid is discussed in Section 5.8 . For the sake of brevity, below we use the subscript "fine" to refer to the grid $\{2^{\ell}N,2^{L-\ell}\Delta R\}$ and the subscript "coarse" to refer to the grid $\{2^{\ell-1}N,2^{L-\ell+1}\Delta R\}$. The multi-grid operator $\mathcal{M}^{\ell}_{\text{fine}}[\mathbf{\Gamma}^{(n)}_{\text{fine}},\mathbf{f}_{\text{fine}}]$ of level $\ell>0$ is defined by the following algorithm:

Algorithm 5.2 *RISM Multi-Grid Operator*

```
Input: $\mathbf{\Gamma}^{(n)}_{\text{fine}}$, $\mathbf{f}_{\text{fine}}$, $\ell$
Output: $\mathbf{\Gamma}^{(n+1)}_{\text{fine}}$

1. Perform $\nu_1$ steps of the fine-grid Picard iteration:
       $\mathbf{\Gamma}'_{\text{fine}}=(\Upsilon_{\text{fine}})^{\nu_1}[\mathbf{\Gamma}^{(n)}_{\text{fine}},\mathbf{f}_{\text{fine}}].$
2. Define a coarse-grid analogue of $\mathbf{\Gamma}'_{\text{fine}}$ and use it as the initial guess of the iteration
   in Step 4:
       $\mathbf{\Gamma}^{(0)}_{\text{coarse}}=r[\mathbf{\Gamma}'_{\text{fine}}].$
3. Calculate grid correction $G[\mathbf{\Gamma}'_{\text{fine}}]$:
       $G[\mathbf{\Gamma}'_{\text{fine}}]=r[\mathcal{K}_{\text{fine}}[\mathbf{\Gamma}'_{\text{fine}}]]-\mathcal{K}_{\text{coarse}}[\mathbf{\Gamma}^{(0)}_{\text{coarse}}].$
4. Perform, recursively, $\mu$ steps of the coarse-grid multi-grid iteration of level $(\ell-1)$:
       $\mathbf{\Gamma}^{(\mu)}_{\text{coarse}}=\left(\mathcal{M}^{\ell-1}_{\text{coarse}}\right)^{\mu}[\mathbf{\Gamma}^{(0)}_{\text{coarse}},r[\mathbf{f}_{\text{fine}}]+G[\mathbf{\Gamma}'_{\text{fine}}]].$
5. Add the coarse-grid correction:
       $\mathbf{\Gamma}''_{\text{fine}}=\mathbf{\Gamma}'_{\text{fine}}+p[\mathbf{\Gamma}^{(\mu)}_{\text{coarse}}-\mathbf{\Gamma}^{(0)}_{\text{coarse}}].$
6. Perform $\nu_2$ steps of the fine-grid Picard iteration:
       $\mathbf{\Gamma}^{(n+1)}_{\text{fine}}=(\Upsilon_{\text{fine}})^{\nu_2}[\mathbf{\Gamma}''_{\text{fine}},\mathbf{f}_{\text{fine}}].$
```

If in Step 4 the number of the multi-grid iteration steps is $\mu=1$, the multi-grid iteration is called a *V-cycle*. If $\mu=2$, the iteration is called *W-cycle* [87]. In the current work we use $\mu=1$. In our case the iterative operator $\mathcal{K}[\cdot]$ is smooth enough, thus for the multi-grid iteration we use $\nu_1=1$ and $\nu_2=0$ on the steps 1, 6 of the algorithm, which is standard for the multi-grid of the second kind [87].

Now we assume that grids with different cutoff distances are given: $\{2^L N, \Delta R\}$, $\{2^{L+1} N, \Delta R\}$, …, $\{2^{L+P} N, \Delta R\}$. We can use a scheme similar to (5.70) for the multi-grid iteration: having solved the task on the grid $\{2^L N, \Delta R\}$ by the multi-grid iteration up to accuracy ε, extend the solution to the grid $\{2^{L+1} N, \Delta R\}$ and use it as initial guess for a next multi-grid iteration and so on. We denote by $\mathbf{\Gamma}^{(0)}_{grid}$ the initial approximation on the grid *grid* and by $\mathbf{\Gamma}^{\varepsilon}_{grid}$ the solution with the L_2-norm accuracy ε, obtained via the iterative process (5.79) . The multi-grid iteration with the grid extension can be written schematically as follows:

$$\begin{aligned}
&\mathbf{\Gamma}^{(0)}_{\{2^L N,\Delta R\}} \xrightarrow{\mathcal{M}^L_{\{2^L N,\Delta R\}}} \mathbf{\Gamma}^{\varepsilon}_{\{2^L N,\Delta R\}} \overset{e}{\to} \\
&\mathbf{\Gamma}^{(0)}_{\{2^{L+1} N,\Delta R\}} \xrightarrow{\mathcal{M}^L_{\{2^{L+1} N,\Delta R\}}} \dots \\
&\dots \overset{e}{\to} \mathbf{\Gamma}^{(0)}_{\{2^{L+P} N,\Delta R\}} \xrightarrow{\mathcal{M}^L_{\{2^{L+P} N,\Delta R\}}} \mathbf{\Gamma}^{\varepsilon}_{\{2^{L+P} N,\Delta R\}}.
\end{aligned} \tag{5.82}$$

5.7 Nested multi-grid

The nested iteration technique suggests one to use an approximate coarse-grid solution as initial guess for the fine-grid iteration. In the multi-grid iteration the fine-grid initial guess can be found via the multi-grid iteration on the coarser grid. Typically, it is enough to perform a single multi-grid iteration on the coarser grid to obtain a good initial guess for the fine-grid iteration. Assume grids $\{N, 2^L \Delta R\}$, $\{2N, 2^{L-1} \Delta R\}$, …, $\{2^L N, \Delta R\}$ and an initial guess $\mathbf{\Gamma}^{(0)}_{\{N,2^L \Delta R\}}$ on the coarsest grid are given. We find an initial guess $\mathbf{\Gamma}^{(0)}_{\{2^L N,\Delta R\}}$ on the fine grid $\{2^L N, \Delta R\}$ by the following algorithm:

Algorithm 5.3 *Nested multi-grid*

```
Input: $\mathbf{\Gamma}^{(0)}_{\{N,2^L \Delta R\}}$
Output: $\mathbf{\Gamma}^{(0)}_{\{2^L N,\Delta R\}}$

for $\ell$=0...L-1:

1. Perform one multi-grid iteration on the coarse grid:

   $$\mathbf{\Gamma}^{(1)}_{\{2^\ell N,2^{L-\ell}\Delta R\}} = \mathcal{M}^{\ell}_{\{2^\ell N,2^{L-\ell}\Delta R\}}(\mathbf{\Gamma}^{(0)}_{\{2^\ell N,2^{L-\ell}\Delta R\}})$$

2. Interpolate the result to the finer grid:

   $$\mathbf{\Gamma}^{(0)}_{\{2^{\ell+1} N,2^{L-\ell-1}\Delta R\}} = p[\mathbf{\Gamma}^{(1)}_{\{2^\ell N,2^{L-\ell}\Delta R\}}]$$
```

The same can be written schematically:

$$
\begin{aligned}
&\mathbf{\Gamma}^{(0)}_{\{N,2^L\Delta R\}} \xrightarrow{\mathcal{M}^0_{\{N,2^L\Delta R\}}} \mathbf{\Gamma}^{(1)}_{\{N,2^L\Delta R\}} \xrightarrow{p} \\
&\mathbf{\Gamma}^{(0)}_{\{2N,2^{L-1}\Delta R\}} \xrightarrow{\mathcal{M}^1_{\{2N,2^{L-1}\Delta R\}}} \dots \\
&\dots \xrightarrow{\mathcal{M}^{L-1}_{\{N,2^L\Delta R\}}} \mathbf{\Gamma}^{(1)}_{\{2^{L-1}N,2\Delta R\}} \xrightarrow{p} \\
&\mathbf{\Gamma}^{(0)}_{\{2^L N,\Delta R\}}
\end{aligned}
\tag{5.83}
$$

Having the initial guess $\mathbf{\Gamma}^{(0)}_{\{2^L N,\Delta R\}}$ one can use the multi-grid iteration process (5.79) to obtain the approximate solution on the grid $\{2^L N, \Delta R\}$ with a given accuracy ε. After that, using scheme (5.82) , one may extend the solution to the grid $\{2^{L+P}N, \Delta R\}$. In the current paper we call the process (5.83) - (5.79)- (5.82) the *nested multi-grid iteration*, and compare its performance to the multi-grid iteration (5.79) - (5.82), to the nested Picard iteration (5.68)-(5.70) , and to the one-level Picard iteration (5.53).

5.8 Determining the optimal number of coarse-grid iteration steps

In the multi-grid algorithm on the coarsest grid we solve the task of type (5.78) with correction $G[\mathbf{\Gamma}'_{\text{fine}}]$. Because $G[\mathbf{\Gamma}'_{\text{fine}}]$ is only an approximation of the $G[\mathbf{\Gamma}^*_{\text{fine}}]$ there is no need to solve this task with accuracy better than the accuracy $\varepsilon_{G[\mathbf{\Gamma}'_{\text{fine}}]}$ of calculation of $G[\mathbf{\Gamma}'_{\text{fine}}]$, which is defined as follows:

$$\varepsilon_{G[\mathbf{\Gamma}'_{\text{fine}}]} = \|G[\mathbf{\Gamma}'_{\text{fine}}] - G[\mathbf{\Gamma}^*_{\text{fine}}]\| \,. \tag{5.84}$$

The value of $\varepsilon_{G[\mathbf{\Gamma}'_{\text{fine}}]}$ can be estimated using the expression

$$\varepsilon_{G[\mathbf{\Gamma}'_{\text{fine}}]} \approx \left\| G[\mathbf{\Gamma}'_{\text{fine}}] - G[\mathbf{\Gamma}^{(n+1)}_{\text{fine}}] \right\| \tag{5.85}$$

Let us assume that the error $\varepsilon(n)$ of the solution decays exponentially with the number n of the coarse-grid iteration steps:

$$\varepsilon(n) = \left\| \Gamma^{(n)}_{\text{coarse}} - \Gamma^*_{\text{coarse}} \right\| = \varepsilon(0) \cdot \delta^n \tag{5.86}$$

If we assume that the error decay rate δ is constant, then the value of δ can be found from the previous expression using two first iteration steps:

$$\delta = \frac{\varepsilon(1)}{\varepsilon(0)} \tag{5.87}$$

We may estimate $\varepsilon(1)$ and $\varepsilon(0)$ by the expressions

$$
\begin{aligned}
\varepsilon(1) &\approx \left\| \Gamma^{(1)}_{\text{coarse}} - \Gamma^{(n)}_{\text{coarse}} \right\| , \\
\varepsilon(0) &\approx \left\| \Gamma^{(0)}_{\text{coarse}} - \Gamma^{(n)}_{\text{coarse}} \right\| .
\end{aligned}
\tag{5.88}
$$

For the optimal number n_{opt} of iteration steps we obtain

$$\varepsilon(n_{\rm opt}) = \varepsilon(0) \cdot \delta^{n_{\rm opt}} = \varepsilon_{G[\mathbf{\Gamma}'_{\rm fine}]} \tag{5.89}$$

Let the actual number of the iteration steps be n.
Dividing (5.86) by (5.89) we get

$$\frac{\varepsilon(n)}{\varepsilon_{G[\mathbf{\Gamma}']}} = \frac{\varepsilon(0)\delta^n}{\varepsilon_{G[\mathbf{\Gamma}']}} = \delta^{n-n_{\rm opt}} \tag{5.90}$$

and find the optimal number of iteration steps as

$$n_{opt} = \log_\delta \frac{\varepsilon_{G[\mathbf{\Gamma}']}}{\varepsilon(0)} \tag{5.91}$$

where δ, $\varepsilon_{G[\mathbf{\Gamma}']}$, $\varepsilon(0)$ are estimated through (5.87), (5.85), (5.88), respectively. So after each multi-grid iteration step we may estimate the optimal number of iteration steps on the coarsest grid and use this number in the next multi-grid iteration step. To avoid fast change of n from one multi-grid iteration step to another, we start from some number $n^{(0)}$ of coarse-grid iteration steps and use a damped iteration process for the number $n^{(k)}$ of the coarsest-grid iteration steps on the k-th multi-grid iteration step:

$$n^{(k+1)} = \alpha n^{(k)} + (1-\alpha) n_{\rm opt} \tag{5.92}$$

where $0 < \alpha < 1$ is the damping parameter.

5.9 Choice of optimal grid parameters for Hydration Free Energy calculations

In the sections above we explained how to solve the RISM equations (5.46) - (5.47) numerically. During the numerical solution we perform iteration steps on several grids with different grid sizes and cutoff distances. In principle, we are free to choose the parameters of the iterations. However, we plan to apply the RISM for calculations of the Hydration Free Energy (HFE). Thus, we would like to choose parameters which yield HFE values with given numerical accuracy at minimal computational cost. HFE can be calculated using the Kirkwood's thermodynamic integration formula [113]. In the RISM approximation, the HFE of a molecule is found as the sum of the partial HFEs of the sites. In the scope of the HNC approximation, the thermodynamic integration can be performed analytically and HFE (ΔG) may be found explicitly from the solutions of the RISM equation [52]:

$$\Delta G_{HNC} = 2\pi\rho kT \sum_{s\alpha} \int_0^\infty \left[-2c_{s\alpha}(r) + \gamma_{s\alpha}(r)\left(c_{s\alpha}(r) + \gamma_{s\alpha}(r)\right)\right] r^2 \mathrm{d}r \tag{5.93}$$

where ρ is the bulk number density, k is the Boltzmann constant and T is the temperature.

Usually, HFE is measured in kilo-calories per mole (kcal/mol). The accuracy of experimental HFE measurements for bioactive compounds is $\gtrsim 0.1$ kcal/mol [93]. To make some theoretical and statistical investigations, we typically need to obtain a computational accuracy of about 100 times higher than the experimental one. That means that we should choose the grid parameters which allow us to calculate the expression (5.93) with accuracy of at least 0.001 kcal/mol. We use a uniform grid which can be described by the grid size ΔR and cutoff distance R_{cutoff}. First, we try to determine an appropriate grid size ΔR. We perform series of RISM calculations with same cutoff distance and different grid sizes, and for each grid size we calculate the free energy of solvation using (5.93) . We assume that the solution on the finest grid yields an almost exact value of ΔG_{HNC}. Let us denote by $\Delta G_{HNC}^{\Delta R}$ the HFE-value calculated on the grid with step ΔR, and by ΔG_{HNC}^{best} the value of the HFE calculated on the finest grid. We can estimate the error of the HFE calculations as difference of $\Delta G_{HNC}^{\Delta R}$ and the best value ΔG_{HNC}^{best}:

$$Error(\Delta G_{HNC}^{\Delta R}) = \left|\Delta G_{HNC}^{\Delta R} - \Delta G_{HNC}^{best}\right| \tag{5.94}$$

In our calculations, we measure distances in atomic units (Bohr, 1 Bohr ≈ 0.52918 Å) as they are the most natural for the atomic scale. As the base grid size we choose $\Delta R_0 =$ 0.1 Bohr, and then make the solvation free energy calculations for the different grid sizes $\Delta R = \frac{\Delta R_0}{2^k}, k = -3\ldots 6$ (grid sizes from 0.8 Bohr to $\frac{1}{640}$ Bohr). The value of $\frac{1}{640}$ Bohr is taken to be an approximation ΔG_{HNC}^{best} of the exact HFE-value. The cutoff distance is 204.8 Bohr. Calculations on all grids are performed up to L_2-norm accuracy $\varepsilon = 10^{-10}$, which is near the limit caused by the numerical errors of the calculations. The parameters of the best grid are chosen to represent the solution of the original non-discretized RISM equation with a high accuracy. The cutoff distance of 204.8 Bohr ($\approx 106\mathring{A}$) is more than 30 times larger than the size of the solvent (water) molecule. Typically, the most of fluctuations of the correlation functions are in the first-second solvation shells (3-7Åfor water), and the fluctuations after the 10th solvation shell (30Å) are negligible. Using the cutoff distance 3 times larger (>100 Å) we can be sure that this cutoff distance does not affect much the results of the calculations. The grid size of $\frac{1}{640}$ Bohr is taken to be just very small grid step, but still feasible to compute.

To find the optimal cutoff distance R_{cutoff}, we perform calculations with the fixed ΔR but different R_{cutoff}. We estimate the error of HFE calculations by taking the integral in (5.93) over the interval $(R_{\mathrm{cutoff}}, \infty)$, because this is the part of the axis which we omit while using the function with finite support:

$$Error(\Delta G_{HNC}) = 2\pi\rho kT \sum_{s\alpha} \int_{R_{\mathrm{cutoff}}}^{\infty} [-2c_{s\alpha}(r) + \gamma_{s\alpha}(r)\,(c_{s\alpha}(r) + \gamma_{s\alpha}(r))]\, r^2 dr \tag{5.95}$$

To evaluate the infinite integral (5.95) , we can calculate functions $\gamma_{s\alpha}(r)$ and $c_{s\alpha}(r)$ on the grid with a very large cutoff distance $R^{\infty}_{\text{cutoff}}$ and calculate the integral over the interval (R_{cutoff},$R^{\infty}_{\text{cutoff}}$). As large cutoff distance we use $R^{\infty}_{\text{cutoff}} = 409.6$ Bohr. The cutoff value of 409.6 Bohr is chosen to be twice larger than the largest cutoff of the calculations (204.8 Bohr). This allows us to estimate the errors of different cutoff distances (including the large distances, like 204.8 Bohr) with a high accuracy.

The most of computational time in the multi-grid iteration is spent on the coarsest grid. The coarsest-grid solution should give a good approximation to the low-frequent part of the exact solution. That means that using the coarsest-grid solution, we as well should be able to roughly approximate chemical properties of the system, in particular the free energy of hydration. HFE-values for a wide class of compounds lie in the range from -5 kcal/mol to +5 kcal/mol. That means that to obtain some qualitative information about the value of HFE we need to have an accuracy in the calculations of at least 1-2 kcal/mol. In the current work grid size and cutoff distance of the coarsest grid are chosen to give the numerical error in HFE calculations $\leq$1 kcal/mol.

It is known that the solution of the RISM equations behave differently for neutral and charged systems. That is why for determining the optimal grid parameters we have chosen five different systems: single uncharged atom (Argon), simple charged ion pair (Sodium chloride), uncharged molecule (methane), polar molecule with high partial charges of atoms (methanol) and water.

We can find how much faster are nested Picard iteration, multi-grid and nested multi-grid than the one-grid Picard iteration. To do the comparison we use the speed-up factors:

$$S_{\text{NMG}}(\varepsilon) = \frac{t_{\text{one-grid}}(\varepsilon)}{t_{\text{NMG}}(\varepsilon)} \tag{5.96}$$

$$S_{\text{MG}}(\varepsilon) = \frac{t_{\text{one-grid}}(\varepsilon)}{t_{\text{MG}}(\varepsilon)} \tag{5.97}$$

$$S_{\text{Nest}}(\varepsilon) = \frac{t_{\text{one-grid}}(\varepsilon)}{t_{\text{NP}}(\varepsilon)} \tag{5.98}$$

where $t_{\text{NMG}}(\varepsilon)$, $t_{\text{MG}}(\varepsilon)$, $t_{\text{NP}}(\varepsilon)$, $t_{\text{one-grid}}(\varepsilon)$ are the computer times to solve the RISM equations by the nested multi-grid method, multi-grid method, nested Picard iteration method and one-level Picard iteration method up to the L_2-norm accuracy ε respectively. The values $t_{\text{NP}}(\varepsilon)/t_{\text{MG}}(\varepsilon)$ and $t_{\text{NP}}(\varepsilon)/t_{\text{NMG}}(\varepsilon)$ show how many times faster the multi-grid and nested multi-grid methods are than the nested Picard iteration respectively.

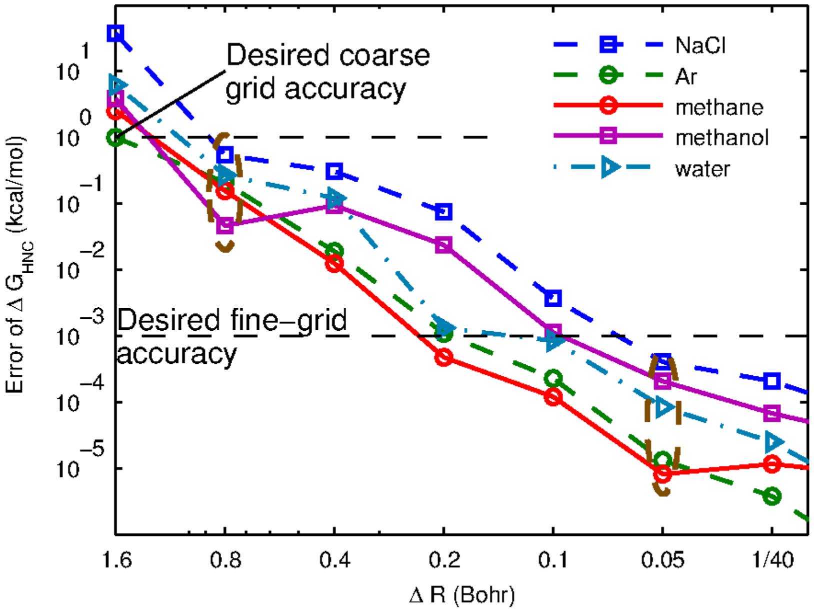


Figure 5.2: Dependencies of the error in hydration free energy calculations on the grid step ΔR (logarithmic scale). Calculations are done for the cutoff distance R_{cutoff}=204.8 Bohr. Groups of points which satisfy the minimal requirements for the desired coarse-grid and fine-grid accuracies are marked with the dashed ovals.

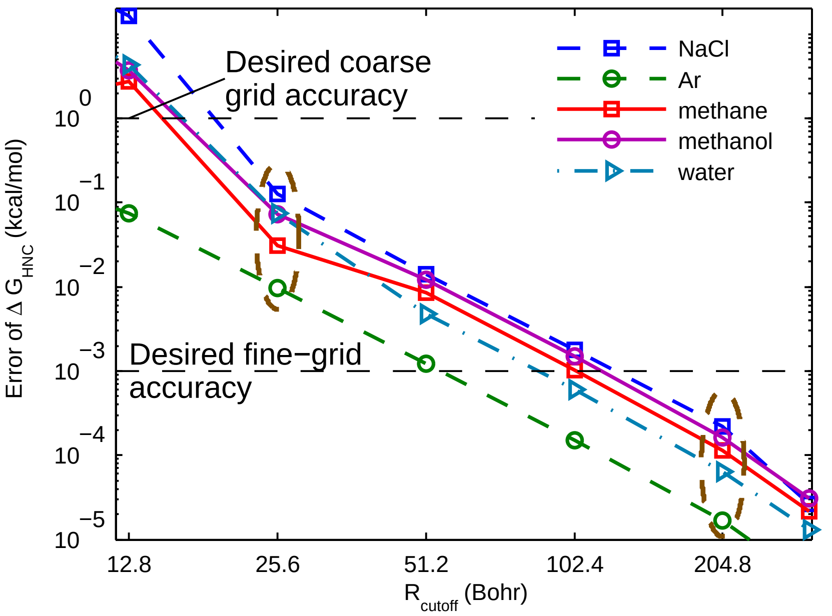


Figure 5.3: Dependencies of the error in hydration free energy calculations on the cutoff distance R_{cutoff} (logarithmic scale). Groups of points which satisfy the minimal requirements for the desired core-grid and fine-grid accuracies are marked with the dashed ovals.

5.10 Optimal grid parameters

To determine the appropriate grid parameters, the RISM Hydration Free Energy calculations were performed for grids with different fine-grid sizes and different cutoff distances. In Figure 5.2 one can see how the error of HFE calculations depends on the grid size. For all investigated systems starting from the grid size $\Delta R = 0.05$ Bohr, the error is smaller than the desired threshold 0.001 kcal/mol. We take the grid with $\Delta R = 0.05$ Bohr as the fine-grid for the numerical solution of the RISM equations. To determine the optimal cutoff distance, we performed the RISM calculations for the systems with very large cutoff distance $R^{\infty}_{\text{cutoff}} = 409.6$ Bohr and calculated the numerical error of HFE calculations with (5.95) for argon, sodium chloride, methane, methanol and water. We use grids with 2^p $(p = 8 \dots 13)$ points which gives cutoff distances from 12.8 Bohr to 409.6 Bohr. The results of the calculations are presented in Figure 5.3. We can see that to achieve the desired accuracy of HFE calculations one should use a cutoff distance 204.8 Bohr, which corresponds to 4096 grid points with the grid size 0.05 Bohr.

Also, from Figures 5.2 and 5.3 one can see that a grid with ΔR_{coarse}=0.8 Bohr, $R^{\text{coarse}}_{\text{cutoff}}$=25.6 Bohr is enough to give a numerical error in the HFE calculations less than 1 kcal/mol. Thus, this grid size and cutoff distance are used in the current work as parameters of the coarsest-grid in the multi-grid algorithm.

5.11 The RISM-MOL solver

In the current work the calculations of the RISM solute-solvent correlation functions were performed with the RISM-MOL program package for fast solution of the RISM integral equations developed by Maxim V. Fedorov and Volodymyr P. Sergiievskyi in the Computational Physical Chemistry and Biophysics group of the Max-Planck-Institute for Mathematics in the Sciences.

The multi-grid method has been implemented in the RISM-MOL program for 1D RISM calculations [121]. Using this program, the HFE calculations for the largest molecule in the set (42 atoms) took about 30 seconds on one PC. The average time of the Hydration Free Energy calculations was 17 sec.

As the input data the RISM-MOL program takes the coordinates, parameters of the Lennard-Jones potential and partial charges q_s of the atoms of the solute molecule. The parameters of the solvent molecules, as well as pre-calculated bulk-solvent correlation functions $h_{s\alpha}^{\mathrm{bulk}}(r)$ are embedded to the program. Using the atomic parameters, the site-site interaction potentials between the solute sites s and solvent sites α are calculated:

$$u_{s\alpha}(r) = u_{s\alpha}^{LJ}(r) + u_{s\alpha}^{C}(r) \tag{5.99}$$

where $u_{s\alpha}^{C}(r)$ is the Coulomb potential

$$u_{s\alpha}^{C}(r) = \frac{q_s q_\alpha}{r} \tag{5.100}$$

and $u_{s\alpha}^{LJ}(r)$ is a Lennard-Jones potential

$$u_{s\alpha}^{LJ}(r) = 4\epsilon_{s\alpha}\left(\left(\frac{\sigma_{s\alpha}}{r}\right)^{12} - \left(\frac{\sigma_{s\alpha}}{r}\right)^{6}\right) \tag{5.101}$$

The pair Lennard-Jones parameters $\sigma_{s\alpha}$, $\epsilon_{s\alpha}$ are calculated via the combining rules. By default the Lorentz-Berthelot rules are used:

$$\sigma_{s\alpha} = \frac{\sigma_s + \sigma_\alpha}{2} \qquad \epsilon_{s\alpha} = \sqrt{\epsilon_s \epsilon_\alpha} \tag{5.102}$$

Other combining rules can be defined by the user.

In the RISM-MOL program, it is possible to vary the number of grids, the number of grid points, the number of iterations and, hence, the accuracy of the calculation. In the current study the six-grid iterations were used. The final solution was obtained on the grid with 4096 grid points and 0.05 Bohr step size with L_2 -norm accuracy $\varepsilon = 10^{-4}$.

The fast implementation of the algorithm for the numerical solution of the RISM equations together with the presented possibilities for accurate hydration free energy calculations makes the RISM-MOL solver a robust tool for investigating the thermodynamics of solution. The program can be obtained for academic users free of charge from Maxim V. Fedorov by request.

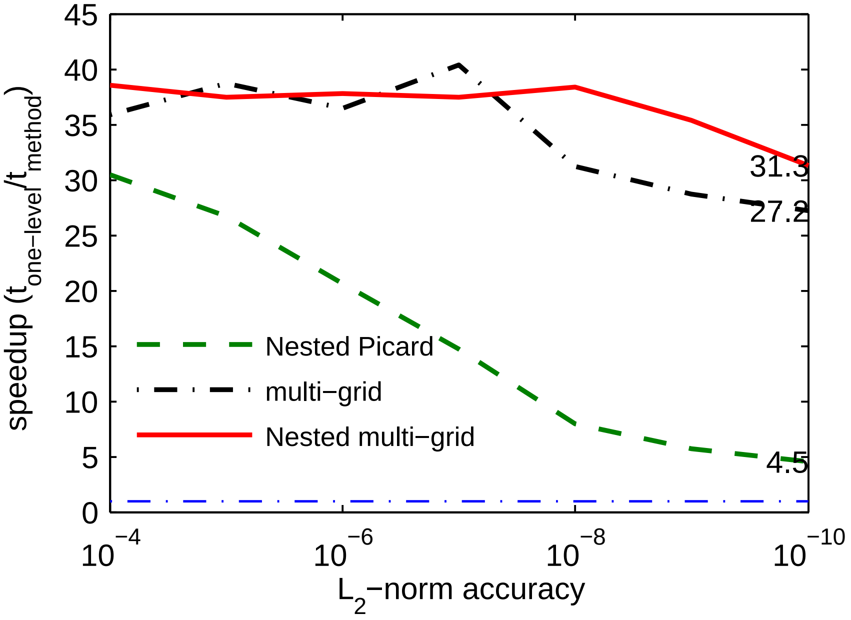


Figure 5.4: Dependencies of the nested multi-grid speedup $S_{\text{NMG}}(\varepsilon) = t_{\text{one-grid}}(\varepsilon)/t_{\text{NMG}}(\varepsilon)$, the multi-grid speedup $S_{\text{MG}}(\varepsilon) = t_{\text{one-grid}}(\varepsilon)/t_{\text{MG}}(\varepsilon)$ and the nested Picard iteration speedup $S_{\text{NP}}(\varepsilon) = t_{\text{one-grid}}(\varepsilon)/t_{\text{NP}}(\varepsilon)$ on the L_2-norm accuracy of calculations.

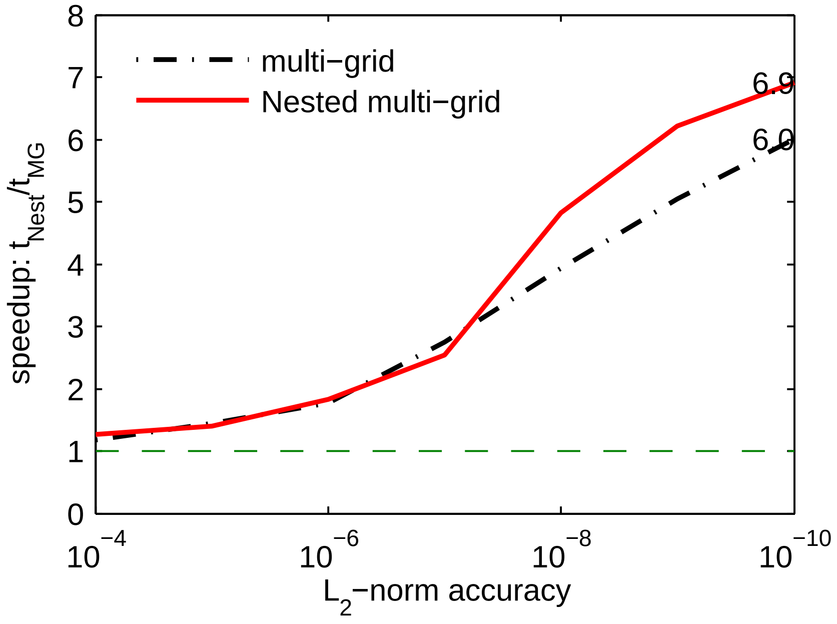


Figure 5.5: Dependencies of the nested multi-grid and multi-grid speedup with regards to the nested Picard iterations ($t_{\text{NP}}(\varepsilon)/t_{\text{NMG}}(\varepsilon)$, $t_{\text{NP}}(\varepsilon)/t_{\text{MG}}(\varepsilon)$) on the L_2-norm accuracy of calculations.

Table 5.1: Comparison between the multi-grid and the nested Picard iteration. Number of iteration steps and percent of total computational time, spent on each grid.

Grid			**Number of iteration steps**		**% of time spent on level**	
Grid Points	ΔR (Bohr)	R_{cutoff} (Bohr)	multi-grid	Nested Picard	multi-grid	Nested Picard
4096	0.05	204.8	2	8	0.9%	0.5%
2048	0.05	102.4	4	394	0.9%	8.0%
1024	0.05	51.2	18	3624	2.0%	**33.0%**
512	0.05	40.6	52	3868	2.9%	**18.8%**
406	0.1	40.6	52	4406	1.5%	11.1%
128	0.2	40.6	57	6399	1.1%	11.7%
64	0.4	40.6	4843	7895	**27.2%**	8.3%
32	0.8	40.6	11336	10422	**63.6%**	8.7%

5.12 Comparison of performance of the one-grid Picard iteration, nested Picard iteration, multi-grid and nested multi-grid

We compare the numerical performance of the one-grid iteration (5.53) , the nested Picard iteration (5.68) -(5.70) , the multi-grid iteration (5.79) - (5.82) and the nested multi-grid iteration (5.83) - (5.79) (5.82) .

In the experiments, the coarsest grid has 32 grid points, grid size $\Delta R = 0.8$ Bohr and cutoff distance $R_{\text{cutoff}} = 25.6$ Bohr. The finest grid has 4096 grid points, grid size $\Delta R = 0.05$ Bohr and cutoff distance $R_{\text{cutoff}} = 204.8$ Bohr. The one-level iteration was performed on the finest grid only.

To compare the efficiency of the methods, the RISM equations were solved numerically for the same molecule (methane), with the same accuracy, using the one-level Picard iteration, nested Picard iteration and the proposed multi-grid-based algorithms.

The dependencies of the speed-up factors (5.97), (5.98) on the accuracy of the calculations are presented in Figure 5.4.

As one can see, the speed-up decreases when accuracy increases. This can be explained by the fact that for higher accuracies high frequencies of the solution are essential, so we need to perform more time-consuming Picard iteration steps on the fine grids. Nevertheless, even an accuracy of $\varepsilon = 10^{-10}$ multi-grid methods are about 30 times faster than one-level iteration. One can see, that for low accuracies multi-grid and nested multi-grid have almost the same performance, while for the high accuracies nested multi-grid is slightly faster. The speedup of the nested Picard iteration is lower and decreases faster. For $\varepsilon = 10^{-10}$, the nested Picard iteration is only 4.5 times faster than the one-level iteration. Figure 5.5 presents the dependency of the multi-grid speedup $t_{\mathrm{NP}}(\varepsilon)/t_{\mathrm{MG}}(\varepsilon)$ and nested multi-grid speedup $t_{\mathrm{NP}}(\varepsilon)/t_{\mathrm{NMG}}(\varepsilon)$ with regard to the nested Picard iteration.

We see that for the low accuracy regime, the nested Picard iteration is less than 1.5 times slower than the multi-grid iteration, but when the accuracy increases, the efficiency of the nested Picard iteration becomes worse than the efficiency of the multi-grid methods. Indeed, for an accuracy of $\varepsilon = 10^{-10}$ the nested multi-grid iteration is almost 7 times faster than the nested iteration method. This happens because in the nested Picard iteration there is no correction for the difference between accurate solutions on different grids. While the accuracy of the solution is less than the difference between the accurate solutions on different grids, the multi-grid and nested Picard iteration have similar efficiency. But for the high accuracy regime, using the nested Picard iteration process it is not possible to produce a correct result on the coarse grid, thus the number of expensive fine-grid iteration steps increases and efficiency goes down. If we look at the Table 5.1 , we see that for an accuracy of $\varepsilon = 10^{-10}$ multi-grid performs most of elementary iteration steps and spends most of the time on the coarse grids, while nested Picard iteration method performs a large number of iteration steps on the fine grids.

5.13 Calculation of Hydration Free Energy of drug-like compounds

One of the main applications of the RISM multi-grid method described above is calculation of the solvation free energy of drug-like compounds. Below we demonstrate an example of solvation free energy calculations for drug-like molecules based on the RISM Multi-grid algorithm. We note that the proposed parameterization scheme described below is given mostly to demonstrate basic concepts of RISM calculations and parameterization of the RISM results. More information about the parameterization of solvation free energies calculated with RISM and 3DRISM one can find in Refs. [64–66].

For our investigation we choose the SAMPL1 molecule set published in Ref. [93]. The set consists of 63 drug-like bioactive compounds. Table 5.2 contains the following information: 1) the list of the compounds, 2) experimentally-measured solvation free energies 3) code names of the compounds (Cup08001-Cup08063).

Table 5.2: Experimentally measured solvation free energies for the 63 compounds from the SAMPL1 molecule set

Code name	**Chemical name**	**Chemical formula**	$\Delta\mathcal{F}_{hydr}$(kcal/mol)
cup08001	nitroglycol	$C_2H_4N_2O_6$	-5.73±0.10
cup08002	1,2-dinitroxypropane	$C_3H_6N_2O_6$	-4.95±0.10
cup08003	butyl nitrate	$C_4H_9NO_3$	-2.09±0.10
cup08004	2-butyl nitrate	$C_4H_9NO_3$	-1.82±0.10
cup08005	isobutyl nitrate	$C_4H_9NO_3$	-1.88±0.10
cup08006	ethyleneglycol mononitrate	$C_2H_5NO_4$	-8.18±0.10
cup08007	alachlor	$C_{14}H_{20}NO_2Cl$	-8.21±0.29
cup08008	aldicarb	$C_7H_{14}N_2O_2S$	-9.84±0.10
cup08009	ametryn	$C_9H_{17}N_5S$	-7.65±0.45
cup08010	azinphosmethyl	$C_{10}H_{12}N_3O_3PS_2$	-10.03±1.37
cup08011	enefin	$C_{13}H_{16}N_3O_4F_3$	-3.51±1.93
cup08012	ensulfuron	$C_{16}H_{18}N_4O_7S$	-17.17±1.93
cup08013	bromacil	$C_9H_{13}N_2O_2Br$	-9.73±1.93
cup08014	captan	$C_9H_8NO_2SCl_3$	-9.01±1.93
cup08015	carbaryl	$C_{12}H_{11}NO_2$	-9.45±0.10
cup08016	carbofuran	$C_{12}H_{15}NO_3$	-9.61±0.30
cup08017	carbophenothion	$C_{11}H_{16}O_2PS_3Cl$	-6.50±0.83
cup08018	hlordane	$C_{10}H_6Cl_8$	-3.44±0.10
cup08019	chlorfenvinphos	$C_{12}H_{14}O_4PCl_3$	-7.07±1.37
cup08020	chlorimuronethyl	$C_{15}H_{15}N_4O_6SCl$	-14.01±1.93
cup08021	chloropicrin	CNO_2Cl_3	-1.45±0.10
cup08022	chlorpyrifos	$C_9H_{11}NO_3PSCl_3$	-5.04±0.21
cup08023	dialifor	$C_{14}H_{17}NO_4PS_2Cl$	-5.74±1.93
cup08024	diazinon	$C_{12}H_{21}N_2O_3PS$	-6.48±0.13
cup08025	dicamba	$C_8H_6O_3Cl_2$	-9.86±1.93
cup08026	dichlobenil	$C_7H_3NCl_2$	-4.71±1.93
cup08027	dinitramine	$C_{11}H_{13}N_4O_4F_3$	-5.66±1.93
cup08028	dinoseb	$C_{10}H_{12}N_2O_5$	-6.23±1.93
cup08029	endosulfan alpha	$C_9H_6O_3SCl_6$	-4.23±0.26
cup08030	endrin	$C_{12}H_8OCl_6$	-4.82±0.10
cup08031	ethion	$C_9H_{22}O_4P_2S_4$	-6.10±1.37
cup08032	fenuron	$C_9H_{12}N_2O$	-9.13±1.93
cup08033	heptachlor	$C_{10}H_5Cl_7$	-2.55±0.10
cup08034	isophorone	$C_9H_{14}O$	-5.18±1.37
cup08035	lindane	$C_6H_6Cl_6$	-5.44±0.10
cup08036	malathion	$C_{10}H_{19}O_6PS_2$	-8.15±0.21
cup08037	methomyl	$C_5H_{10}N_2O_2S$	-10.65±1.93
cup08038	methyparathion	$C_8H_{10}NO_5PS$	-7.19±0.10
cup08039	metsulfuronmethyl	$C_{14}H_{15}N_5O_6S$	-15.54±1.93

Continued on the the next page

Code name	Chemical name	Chemical formula	$\Delta\mathcal{F}_{hydr}$(kcal/mol)
cup08040	nitralin	$C_{13}H_{19}N_3O_6S$	-7.98±1.93
cup08041	nitroxyacetone	$C_3H_5NO_4$	-5.99±0.10
cup08042	oxamyl	$C_7H_{13}N_3O_3S$	-10.18±1.93
cup08043	parathion	$C_{10}H_{14}NO_5PS$	-6.74±0.10
cup08044	pebulate	$C_{10}H_{21}NOS$	-3.63±1.93
cup08045	phorate	$C_7H_{17}O_2PS_3$	-4.37±0.10
cup08046	profluralin	$C_{14}H_{16}N_3O_4F_3$	-2.45±1.37
cup08047	prometryn	$C_{10}H_{19}N_5S$	-8.43±0.10
cup08048	propanil	$C_9H_9NOCl_2$	-7.78±1.93
cup08049	pyrazon	$C_{10}H_8N_3OCl$	-16.43±1.93
cup08050	simazine	$C_7H_{12}N_5Cl$	-10.22±0.10
cup08051	sulfometuron-methyl	$C_{15}H_{16}N_4O_5S$	-20.25±1.93
cup08052	terbacil	$C_9H_{13}N_2O_2Cl$	-11.14±1.93
cup08053	terbutryn	$C_{10}H_{19}N_5S$	-6.68±0.42
cup08054	thifensulfuron	$C_{12}H_{13}N_5O_6S_2$	-16.23±1.93
cup08055	trichlorfon	$C_4H_8O_4PCl_3$	-12.74±1.93
cup08056	trifluralin	$C_{13}H_{16}N_3O_4F_3$	-3.25±0.10
cup08057	vernolate	$C_{10}H_{21}NOS$	-4.13±1.36
cup08058	4-amino-4’-nitroazobenzene	$C_{12}H_{10}N_4O_2$	-11.24±0.44
cup08059	1-amino-4-anilino-anthraquinone	$C_{20}H_{14}N_2O_2$	-7.44±1.93
cup08060	1,4,5,8-tetramino-anthraquinone	$C_{14}H_{12}N_4O_2$	-8.94±1.37
cup08061	1-amino-anthraquinone	$C_{14}H_9NO_2$	-7.97±1.37
cup08062	4-dimethylamino-azobenzene	$C_{14}H_{15}N_3$	-6.66±0.22
cup08063	pirimor (pirimicarb)	$C_{11}H_{18}N_4O_2$	-9.41±1.93

We assigned the OPLS2005 force-field parameters [122] to the molecules listed in Table 5.2. For all of the 63 molecules the solvation free energies was calculated using the RISM algorithm with the KH closure (3.122). The calculated solvation free energies were compared to the experimental results. The standard deviation, root mean square deviation and correlation coefficient were calculated. The parameterization of the calculated results was performed. The parameterization formula included the partial molar volume(PMV) corrections and corrections for different types of atoms in the molecules. The set of compounds was divided into the training and test sets. Using the training set by the least squares method the parameterization coefficients were found. Using the parameterization formula the solvation free energies for the test set of compounds were predicted. We compared the predicted values to experiment and estimated the quality of the model.

The OPLS2005 force-field parameters were assigned with the *ffld_srv* utility of the MCPRO+ of the Schrödinger Maestro LLC program package [123].

It is known that using the standard charges of the OPLS2005 force-field it is impossible to accurately estimate the solvation free energy [124]. That's why in our calculations we also used the charges obtained form the quantum-mechanical calculations and compared the results for both types of charges. We performed the B3LYP calculations with the 6-31G(d,p) basis using the Gaussian 03 program [125]. The detailed discussion of the quantum mechanical methods used for these calculations is beyond of scope of the current work. More detailed description of B3LYP method can be found in Ref. [126]. The partial charges of the molecules were calculated using the CHELPG (CHarges form Electrostatic Potential, Grid method), which assumes that the partial charges were assigned in such a way that the electrical potential in some chosen points is equal in classical and quantum mechanical approximations [127]. In the standard procedure for calculating CHELPG charges there are no parameters for the Bromine atom. In our quantum calculations we changed the Bromine atom in the molecule Cup08013 to the Chlorine atom. For the numerical solution of the RISM equations (5.1) one need to know solute intermolecular functions $\omega_{ss'}(k)$, solvent susceptibility functions $\chi_{\alpha\alpha'}(k)$ and pairwise site-site potentials $u_{s\alpha}(r)$. The intermolecular functions can be simply calculated from the 3D structure of the solute molecule. In the Fourier space the expressions for these functions can be represented in a following way:

$$\omega_{ss'}(k) = \frac{\sin kr_{ss'}}{kr_{ss'}} \tag{5.103}$$

where $r_{ss'}$ is the distance between the atoms s and s' of the solute molecule. We perform RISM and SFE calculations for the temperature $T = 300K$. We use the susceptibility functions from Ref. [83] where these functions were calculated by using the wavelet-based algorithm for solving RISM equations [56, 84, 85]. The site-site potential used in this work is a superposition of the Coulomb potential $u^{C}_{s\alpha}(r)$ and Lennard-Jones potential $u^{LJ}_{s\alpha}(r)$, namely:

$$u_{s\alpha}(r) = u^{C}_{s\alpha}(r) + u^{LJ}_{s\alpha}(r) \tag{5.104}$$

The Coulomb potential is defined in a following way:

$$u^{C}_{s\alpha}(r) = \frac{q_s q_\alpha}{r} \tag{5.105}$$

where q_s, q_α are partial charges of the atoms s and α. We use atomic units to avoid using of the scaling coefficient in Coulomb potential. The unite charge is the positron charge (e) and distance between the atoms r is given in Bohr units. A pair Lennard-Jones potential is defined with the following relation:

$$u^{LJ}_{s\alpha}(r) = 4\epsilon_{s\alpha}\left(\left(\frac{\sigma_{s\alpha}}{r}\right)^{12} - \left(\frac{\sigma_{s\alpha}}{r}\right)^{6}\right) \tag{5.106}$$

Pairwise parameters $\sigma_{s\alpha}$, $\epsilon_{s\alpha}$ are calculated from the atomic OPLS2005 parameters σ_s, σ_α, ϵ_s, ϵ_α using the Lorentz-Berthelot mixing rules:

$$\sigma_{s\alpha} = \frac{\sigma_s + \sigma_\alpha}{2} \qquad \epsilon_{s\alpha} = \sqrt{\epsilon_s \epsilon_\alpha} \tag{5.107}$$

We note, that the standard OPLS2005 mixing rules differ from the Lorentz-Berthelot rules. In OPLS2005 force-field it is assumed that $\sigma_{s\alpha} = \sqrt{\sigma_s \sigma_\alpha}$. The reason why we are using the Lorentz-Berthelot rules in our work is that for standard OPLS2005 rules the RISM equation solver diverges for some molecules. In our work we use the modified SPC/E water model (MSPC/E) [128]. The MSPC/E water model in contrast to the standard SPCE model [129] has non-zero LJ potential for the Hydrogen atom. In our work we use the following Hydrogen LJ parameters: $\sigma_H = 0.8$Å, $\quad \epsilon_H = 0.046$ kcal/mol. The RISM and SFE calculations were performed for the aqueous solutions at the temperature $T = 300K$ and water density $\rho = 33.7$ particles/nm^3. The calculations were performed with the RISM-MOL multi-grid solver [121]. In our work the eight-level multi-grid method was used. The correlation functions were obtained on the equispaced grid with the grid size 0.265Å and 4096 discretization points. The KH-closure was used for calculations. The calculations were performed for all the molecules listed in Table 5.2. Two calculations for each molecule were performed: for OPLS2005 and CHELPG partial charges correspondingly. Average computation time was 15 sec/molecule. After solving the RISM equations the solvation free energy was calculated using four different expression: KH, HNCB, GF, and PW.

The results of the RISM SFE calculations were compared to the experimentally measured values. We calculated the mean error, standard deviation (SD) and root mean squared deviation(RMSD). The mean value and the standard deviation were calculated with the following expressions:

$$M(\Delta G - \Delta G_{\text{exp}}) = \frac{1}{N} \sum_{i \in S} \left(\Delta G^{(i)} - \Delta G^{(i)}_{exp} \right) \tag{5.108}$$

$$\begin{aligned} &SD(\Delta G - \Delta G_{\text{exp}}) = \\ &\sqrt{\frac{1}{N} \sum_{i \in S} \left(\Delta G^{(i)} - \Delta G^{(i)}_{exp} - M(\Delta G - \Delta G_{\text{exp}}) \right)^2} \end{aligned} \tag{5.109}$$

The RMSD can be calculated with the following formula:

$$\begin{aligned} RMSD(\Delta G, \quad &\Delta G_{\text{exp}})^2 = \\ &M(\Delta G - \Delta G_{\text{exp}})^2 + SD(\Delta G - \Delta G_{\text{exp}})^2 \end{aligned} \tag{5.110}$$

In Ref. [41] it is shown that the partial molar volume (PMV) correction can essentially increase the accuracy of the RISM SFE calculations for simple non-polar compounds. This suggests that the parameterization of the SFE for the compounds from the SAMPL1 set can also be useful.

In the RISM approximation the partial molar volume can be calculated as a limiting case for the infinite dilution using the general formula for the partial molar volume [130]. This gives the following expression for the PMV of the solute molecule:

$$V_{ex} = \frac{1}{\rho} + \frac{4\pi}{N_{solute}} \sum_s \int\limits_0^\infty \left(h_{oo}^{\mathrm{solv}}(r) - h_{so}(r) \right) r^2 dr. \tag{5.111}$$

where h_{oo}^{solv} is a total oxygen-oxygen correlation function of pure water taken from Ref. [42] where it was calculated using the dielectrically consistent RISM, $h_{so}(r)$ is a total correlation function between the solute site s and water oxygen. For the parameterization it is convenient to use dimensionless value. So we use the dimensionless value ρV_{ex} where ρ is a water density. In Ref. [41] the systematic overestimation of the hydrogen-bond contribution to the SFE in the RISM calculations for the molecules with highly charged groups is discussed. The authors introduce the hydrogen-bond correction which accounts the number of hydroxyl groups in the molecule. In our work we use the similar method. However, because the compounds in the SAMPL1 set are much more complicated than ones in Ref. [41] we do not see the simple way to divide these compounds into functional groups. Thus we simply introduce corrections for each type of atoms in the molecule. The molecules in Table 5.2 contain Hydrogen, Carbon, Oxygen, Nitrogen, Sulfur, Phosphorus, Chlorine, Fluorine and Bromine. To reduce the number of parameterization coefficients we do not distinguish Fluorine, Chlorine and Bromine and introduce one Halogen correction (F, Cl, and Br). We use the following parameterization formula:

$$\Delta G_{corr}(\mathbf{b}) = \Delta G_{RISM} + b_V \rho V_{ex} + \sum_j b_j n_j \tag{5.112}$$

where summation is done over all atom types $j \in \{H,\, C,\, N,\, O,\, Hal,\, P,\, S\}$ (Hal means halogen), n_j is a number of atoms of type j in the molecule, $\mathbf{b} = \{\, b_V,\, b_H,\, b_C,\, b_N,\, b_O,\, b_{Hal},\, b_P,\, b_S\}$ are the parameterization coefficients. To calculate the parameterization coefficients the set of compounds from Table 5.2 was divided into the training and test sets. The training set of compounds is composed from all the compounds those codes end up with an odd digit (Cup08001, Cup08003, ..., Cup08063), the training set of compounds includes all the compounds those numbers end up with a even digit (Cup08002, Cup08004, ..., Cup08062). Coefficients $b = \{\, b_V$, b_H, b_C, b_N, b_O, b_{Hal}, b_P, $b_S\}$ were calculated with the least squares method on the training set of compounds. In such a way the exact expression for prediction the solvation free energies was determined. Using this formula the SFE of the compounds from the test set were calculated. The results of the calculations were compared to the experimental values.

The results of the free energy calculations for different SFE expressions and OPLS2005 and CHELPG charges are presented in Table 5.3. Sorting the results in ascending RMSD order

Table 5.3: Results of the RISM solvation free energy calculations without the parameterization. In the table the data for the differences between the experimental and calculated data are presented: RMSD, mean deviation (M), standard deviation (SD) and correlation coefficient. Data is given in kcal/mol.

Expression	**Charges**	**RMSD**	**M**	**SD**	**Correlation Coefficient**
PW	OPLS2005	11.366	9.978	5.442	0.637
GF	OPLS2005	11.932	-6.778	9.820	0.603
HNCB	OPLS2005	14.084	-5.753	12.855	0.803
KH	OPLS2005	75.665	72.660	21.112	0.119
PW	CHELPG	12.758	-12.153	3.883	0.690
GF	CHELPG	8.421	4.139	7.333	0.651
KH	CHELPG	78.844	-75.362	23.173	0.065
HNCB	CHELPG	11.985	2.870	11.636	0.794

we can see that the accuracy of the KH expression calculations is the lowest, HNCB is the next one, the error of PW and GF are approximately equal. We note that in case of PW expression the most contribution to RMSD gives the systematic mean error shift while in GF formula the most contribution is due to dispersion. All the expressions do not show a very big correlation with experiment. However for all expression except KH the correlation coefficient is greater than 0.5. The largest correlation with experiment is for the HNCB expression. This can be explained by the fact that the additional repulsive potential introduced in (4.39) can be regarded as some kind of partial molar volume correction. And this in turn means that despite the fact that HNCB results correlate with the experimental measurements they could not be essentially improved by the partial molar volume parameterization (this is shown below). Comparing results for CHELPG and OPLS2005 charges we see that CHELPG charges provide better results. Although the mean value of the error (M) and RMSD weakly depend on the partial charges. But in case of the CHELPG charges standard deviation of the error is essentially smaller, which suggests that the data calculated with CHELPG charges can be effectively parameterized.

The results of the calculations were parameterized using the expression (5.112) for KH, HNCB, GF, PW solvation free energy expressions. In Table 5.4 the values of RMSD, mean error and standard deviation of error on the test set of compounds are presented. To prove additionally the necessity of RISM calculations we also performed "pure chemoinformatic" parameterization without calculating solvation free energies. To do this in formula (5.112) we set ΔG to zero and perform the parameterization. We see that the best of OPLS2005 results is calculated with PW formula. The RMSD for this method is 3.1 kcal/mol, while the pure chemoinformatic parameterization for the same compounds gives an error 2.8 kcal/mol. So

Table 5.4: Results of parameterization for KH, HNCB, GF, PW expressions. In the table the following data for differences between the calculated and experimental values are given: RMSD, standard deviation (SD), mean error (M), correlation coefficient. Units in all cases are kcal/mol. The line with the expression "none" correspond to the pure chemoinformatic parameterization (neglecting ΔG_{RISM}).

Expression	**Charges**	**RMSD**	**M**	**SD**	**Correlation coefficient**
PW	OPLS2005	3.075	-0.173	3.070	0.803
GF	OPLS2005	3.690	-0.378	3.671	0.791
KH	OPLS2005	5.541	-1.132	5.424	0.775
HNCB	OPLS2005	6.902	-1.706	6.688	0.767
none	OPLS2005	2.837	0.350	2.815	0.651
PW	CHELPG	1.913	0.382	1.875	0.927
GF	CHELPG	2.542	0.375	2.514	0.912
KH	CHELPG	5.024	-0.318	5.013	0.825
HNCB	CHELPG	6.197	-0.680	6.159	0.819
none	CHELPG	2.845	0.342	2.824	0.638

OPLS2005 charges should not be used for SFE prediction.

In contrast to unsatisfactory results for OPLS2005 charges, the parameterization for CHELPG charges allows to predict the SFE with RMSD=1.9 kcal/mol for the PW expression, which is almost 1.5 times better than the best "pure chemoinformatic" result. Result for GF formula (RMSD=2.5 kcal/mol) still can slightly improve the pure chemoinformatic result. KH and HNCB expression show unsatisfactorily results (RMSD > 5 kcal/mol). We note that although HNCB method without parameterization shows better results than KH, after the parameterization KH becomes more effective. Coefficients $b = \{\ b_V,\ b_H,\ b_C,\ b_N,\ b_O,\ b_{Hal},\ b_P,\ b_S\}$ for the CHELPG/PW method are given in Table 5.5. Analyzing the coefficients we see that the largest contribution to the correction is from the Partial Molar Volume term while other corrections are in average 4-5 times smaller.

5.14 Conclusions

We described a new multi-grid-based algorithm for solving RISM equations. We adopted a general non-linear multi-grid scheme for solving the RISM equations. We also proposed an extension of the algorithm for the grids with different cutoff distances. Additionally we investigated the efficiency of the coarse-grid solver and proposed an adaptive algorithm for calculating the optimal number of iteration steps on the coarsest grid. We performed numerical investigations to optimize the algorithm parameters to give the required numerical accuracy of the

Table 5.5: Values of the parameterization coefficients for CHELPG/PW method. Units - kcal/mol

Coefficient	**Value**
b_V	-7.773
b_H	0.797
b_C	1.578
b_N	1.466
b_O	1.863
b_{Hal}	2.851
b_P	3.152
b_S	4.982

Hydration Free Energy calculations by RISM. The investigated parameters were: fine-grid cutoff distance, fine-grid discretization step, coarsest grid cutoff distance and coarsest grid size. By numerical tests on polar and non-polar simple and multi-atom molecules it was shown that RISM calculations with a fine grid $\{4096, 0.05\ \text{ Bohr}\}$ are able to a numerical accuracy of the Hydration Free Energy value of 0.001 kcal/mol, which is satisfactory for most of the chemical applications. It was shown that the multi-grid calculations with coarsest grid $\{32, 0.8\ \text{Bohr}\}$ are optimal.

The proposed multi-grid methods were compared to the one-grid Picard iteration scheme, which is the reference algorithm, and to the nested Picard iteration algorithm, which is the most straightforward implementation of the multi-scale scheme. It was shown that for high accuracies the proposed methods are about 30 times faster than the single-grid Picard iteration, and almost 7 times faster than the nested Picard iteration. The solvation free energy calculations for the SAMPL1 set of drug-like compounds from the paper [93] were performed with the RISM multi-grid algorithm. The force-field for the calculation included LJ parameters from the OPLS2005 force-field. Two types of the partial charges (OPLS2005 and CHELPG) were used. Solvation free energy calculations were performed using the KH, HNCB, GF, PW expressions. The calculated results for all expressions except KH have a recognizable correlaton with the experimental data. However, the results without corrections cannot be considered as satisfactorily (RMSD $>$ 5 kcal/mol). The parameterization of the calculation results was performed. The parameterization formula included corrections for partial molar volume and number of atoms of each type in the mole The parameterization results showed that calculations with OPLS2005 charges do not give satisfactorily results (RMSD for the best method is 3.1 kcal/mol). In contrast, parameterization for the CHELPG charges showed that the RMSD for the best method (CHELPG/PW) is 1.9 kcal/mol which is almost 1.5 fold better than the verification “pure chemoinformatic” parameterization. Thus we conclude that although RISM solvation free energy calculations do not immediately give the perfect results, after proper parameterization one can get the expression which is able to predict solvation free energies with reasonable accuracy (1.9 kcal/mol). We also note that results depend on the correct choice of partial charges calculation and also on RISM SFE expression.

Chapter 6

3DRISM Multi-grid Algorithm for Fast Solvation Free Energy Calculations

In this chapter the multi-grid algorithm for solving 3DRISM equations is described and tested on a set of organic compounds. The chapter is based on my paper Ref. [131](P1).

6.1 Iterative solution of the 3DRISM equations

In our work we use the Kovalenko-Hirata formulation of the 3D RISM theory [119, 132] in order to describe infinitely diluted solutions of small organic solute molecules. Solvent (water) molecules are described by the RISM approximation, while a solute molecule is a three-dimensional object. Structure of the solvent is described by the total and direct correlation functions $h_\alpha(\mathbf{r})$, $c_\alpha(\mathbf{r})$ where α indicates a solvent site. The 3DRISM equations are written in the following way:

$$h_\alpha(\mathbf{r}) = \sum_{\xi=1}^{N_{solvent}} \int_{\mathbb{R}^3} c_\xi(\mathbf{r}')\chi_{\xi\alpha}(\mathbf{r}-\mathbf{r}')d\mathbf{r}' \tag{6.1}$$

where N_{solvent} is the number of solvent sites, $\chi_{\xi\alpha}(\mathbf{r})$ is the solvent susceptibility function for sites ξ and α. Solvent susceptibility functions $\chi_{\xi\alpha}(\mathbf{r})$ are defined as following:

$$\chi_{\xi\alpha}(\mathbf{r}) = \omega_{\xi\alpha}(r) + \rho h^{\text{solv}}_{\xi\alpha}(r), \tag{6.2}$$

where $r = |\mathbf{r}|$, $\omega_{\xi\alpha}(r) = \delta_{\xi\alpha} + (1-\delta_{\xi\alpha})\delta(r - r_{\xi\alpha})/(4\pi r^2_{\xi\alpha})$, $r_{\xi\alpha}$ is the distance between the sites ξ and α of a solvent molecule, $h^{\text{solv}}_{\xi\alpha}(r)$ is the total site-site correlation function of the solvent sites ξ and α, $\delta_{\xi\alpha}$ is the Kronecker delta and $\delta(r)$ is the Dirac delta function. We used the functions $h^{\text{solv}}_{\xi\alpha}(r)$ calculated in [42].

(6.1) is completed by closure relations:

$$h_\alpha(\mathbf{r}) = e^{-\beta U_\alpha((r))+h_\alpha(\mathbf{r})-c_\alpha(\mathbf{r})+B_\alpha(\mathbf{r})} - 1, \tag{6.3}$$

where $\beta = 1/k_B T$, k_B is the Boltzmann constant, T is the temperature, $U_\alpha(\mathbf{r})$ is the interaction potential corresponding to a solute site α, $B_\alpha(\mathbf{r})$ is the bridge functional.

To use iterative solvers we rewrite (6.1) in the following form [133]:

$$\gamma_\alpha(\mathbf{r}) = \sum_{\xi=1}^{N_{\text{solvent}}} \int_{\mathbb{R}^3} \mathcal{C}[\gamma_\xi(\mathbf{r}' - \mathbf{r})] \cdot \chi_{\xi\alpha}(\mathbf{r}') d\mathbf{r}' + \theta_\alpha(\mathbf{r}) - \mathcal{C}[\gamma_\alpha(\mathbf{r})] \tag{6.4}$$

where $\gamma_\alpha(\mathbf{r}) = h_\alpha(\mathbf{r}) - c_\alpha^S(\mathbf{r})$, $c_\alpha^S(\mathbf{r}) = c_\alpha(\mathbf{r}) + \beta U_\alpha^L(\mathbf{r})$, $U_\alpha(\mathbf{r}) = U_\alpha^S(\mathbf{r}) + U_\alpha^L(\mathbf{r})$, $U_\alpha^S(\mathbf{r})$ is a short range potential, $U_\alpha^L(\mathbf{r})$ is a long range potential, $\theta_\alpha(\mathbf{r}) = -\beta \sum_\xi \int_{\mathbb{R}^3} U_\xi^L(\mathbf{r} - \mathbf{r}')\chi_{\xi\alpha}(\mathbf{r}')d\mathbf{r}'$, $\mathcal{C}[\cdot]$ is a closure (bridge) functional.

We use interaction potentials which are superpositions of the site-site interaction potentials:

$$U_\alpha^S(\mathbf{r}) = \sum_{s=1}^{N_{\text{solute}}} u_{s\alpha}^S(|\mathbf{r} - \mathbf{r_s}|); \tag{6.5}$$

$$U_\alpha^L(\mathbf{r}) = \sum_{s=1}^{N_{\text{solvent}}} u_{s\alpha}^L(|\mathbf{r} - \mathbf{r_s}|); \tag{6.6}$$

where $\mathbf{r}_s$ is the position of a solute site s with respect to the center of a molecule, N_{solute} is the number of solute sites. In our work the site-site potentials contain Lennard-Jones and Coulomb part. Pair Lennard-Jones parameters are obtained from the atomic LJ parameters by using the Lorentz-Berthelot mixing rules:

$$\sigma_{s\alpha} = \frac{1}{2}(\sigma_s + \sigma_\alpha) \qquad \epsilon_{s\alpha} = \sqrt{\epsilon_s \epsilon_\alpha} \tag{6.7}$$

To avoid divergence of the algorithm due to the long range behavior of the interaction potentials we separate the short range and the long range of the potentials that we then treat separately by using the Ng procedure [120]. We use the atomic units for distance and energy Bohr and Hartree. This allows us to avoid scaling coefficients in the representation of the Coulomb potential. Thus expressions for the short-range and long-range potentials are written as following:

$$u_{s\alpha}^S(r) = u_{s\alpha}^{LJ(short)}(r) + u_{s\alpha}^C(r)(1 - \text{erf}(\tau r)) \tag{6.8}$$

$$u_{s\alpha}^L(r) = u_{s\alpha}^{LJ(long)}(r) + u_{s\alpha}^C(r)\text{erf}(\tau r) \tag{6.9}$$

where $u_{s\alpha}^C(r)$ is the Coulomb component of the site-site potential, $\text{erf}(r) = \int_{-\infty}^r e^{-t^2} dt$, τ=0.5 Bohr^{-1}, $u_{s\alpha}^{LJ(short)}(r)$, $u_{s\alpha}^{LJ(long)}(r)$ are short-range and long-range components of the Lennard-Jones potential respectively. The latter are defined by the following relations:

$$u_{s\alpha}^{LJ(short)}(r) = \begin{cases} u_{s\alpha}^{LJ}(r) - u_{s\alpha}^{LJ}(R_{cut}) & \text{when } \; r < R_{cut} \\ 0 & \text{otherwise} \end{cases} \tag{6.10}$$

$$u_{s\alpha}^{LJ(long)}(r) = u_{s\alpha}^{LJ}(r) - u_{s\alpha}^{LJ(short)}(r) \tag{6.11}$$

where $u_{s\alpha}^{LJ}(r)$ is a Lennard-Jones component of a site-site potential, R_{cut}=8Å .

In the article we use the Kovalenko-Hirata (KH) closure, which is defined as following [134]:

$$\mathcal{C}[\gamma_\alpha(\mathbf{r})] = \begin{cases} e^{-\beta U_\alpha^S(\mathbf{r})+\gamma_\alpha(\mathbf{r})} - \gamma_\alpha(\mathbf{r}) - 1 & \text{when} \quad -\beta U_\alpha^S(\mathbf{r}) + \gamma_\alpha(\mathbf{r}) > 0 \\ -\beta U_\alpha^S(\mathbf{r}) & \text{otherwise} \end{cases} \tag{6.12}$$

In the numerical representation of (6.4) the functions $\gamma_\alpha(\mathbf{r})$, $\chi_{\xi\alpha}(\mathbf{r})$, $\theta_\alpha(\mathbf{r})$ are defined by their values in the grid points of an uniform Cartesian grid. A grid is defined by two parameters: *spacing* and *buffer*. *Spacing* is the smallest distance between the grid points and *buffer* is the minimal distance from the solute atoms to the boundaries of the grid (see Figure 6.1 for explanations). At first glance, such parameterization may seem to be inconvenient from a theoretical point of view because the same buffer and spacing parameters may give different grids for different solutes. However, our work is mostly oriented towards future practical applications of the method and in practical applications we are interested in the accuracy of calculations for different cutoff distances of the correlation functions; and these cutoff distances for a Cartesian grid are defined by the buffer parameter. Using the same buffer parameter we can adjust the size and the shape of the grid preserving a constant cutoff of the solvent correlation functions for different solutes. That provides us a straightforward way to control the accuracy of calculations.

We denote the forward and the inverse Fourier transforms on the grid $\mathcal{G}$ as $\mathcal{T}_\mathcal{G}[\cdot]$, $\mathcal{T}_\mathcal{G}^{-1}[\cdot]$ correspondingly. Then a discrete analogue of (6.4) reads as:

$$\mathbf{\Gamma}^\mathcal{G} = \mathcal{T}_\mathcal{G}^{-1}\left[\hat{\mathbf{X}} \cdot \mathcal{T}_\mathcal{G}\left[\mathcal{C}\left[\mathbf{\Gamma}^\mathcal{G}\right]\right]\right] + \mathbf{\Theta}^\mathcal{G} - \mathcal{C}\left[\mathbf{\Gamma}^\mathcal{G}\right] \tag{6.13}$$

where $\mathbf{\Gamma}^\mathcal{G} = \left(\boldsymbol{\gamma}_1^\mathcal{G}, \ldots, \boldsymbol{\gamma}_{N_\text{solvent}}^\mathcal{G}\right)^T$, $\mathbf{\Theta}^\mathcal{G} = \left(\boldsymbol{\theta}_1^\mathcal{G}, \ldots, \boldsymbol{\theta}_{N_\text{solvent}}^\mathcal{G}\right)^T$, $\hat{\mathbf{X}}^\mathcal{G} = [\hat{\boldsymbol{\chi}}_{\boldsymbol{\xi\alpha}}^\mathcal{G}]_{N_\text{solvent} \times N_\text{solvent}}$, $\hat{\boldsymbol{\chi}}_{\boldsymbol{\xi\alpha}}^\mathcal{G} = \mathcal{F}_\mathcal{G}[\boldsymbol{\chi}_{\xi\alpha}]$, upper index $\mathcal{G}$ means that functions are given by their values in the grid points of the grid $\mathcal{G}$.

Equation (6.13) can be written in a more compact way:

$$\mathbf{\Gamma}^\mathcal{G} = F[\mathbf{\Gamma}^\mathcal{G}] \tag{6.14}$$

where $F[\mathbf{\Gamma}^\mathcal{G}] = \mathcal{T}_\mathcal{G}^{-1}\left[\hat{\mathbf{X}} \cdot \mathcal{T}_\mathcal{G}\left[\mathcal{C}\left[\mathbf{\Gamma}^\mathcal{G}\right]\right]\right] + \mathbf{\Theta}^\mathcal{G} - \mathcal{C}\left[\mathbf{\Gamma}^\mathcal{G}\right]$.

The Picard iteration method is defined by the following recurrent formula:

$$\mathbf{\Gamma}_{n+1}^\mathcal{G} = (1-\lambda)\mathbf{\Gamma}_n^\mathcal{G} + \lambda F[\mathbf{\Gamma}_n^\mathcal{G}] \tag{6.15}$$

where $\mathbf{\Gamma}_n^\mathcal{G}$ is the n-th step approximation, λ is the coupling parameter.

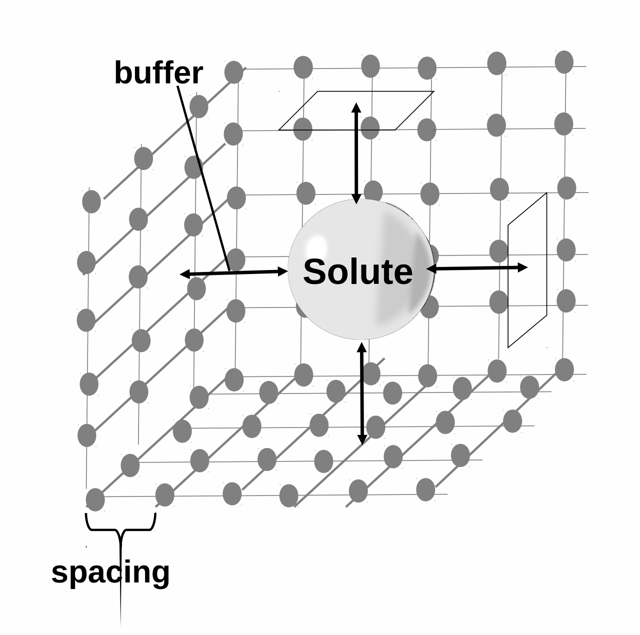


Figure 6.1: *Spacing* is the minimal distance between the grid points, *buffer* is the minimal distance from the solute atoms to the boundaries of the grid

6.2 DIIS and MDIIS iteration

Direct inverse in the iterative subspace (DIIS) method is an iteration method initially introduced to improve convergence of Schrödinger equation solvers [135]. Later modified DIIS (MDIIS) method was applied to the 3DRISM equations [77]. In the DIIS method on the n-th iteration step one finds an approximate solution $\mathbf{\Gamma}_*^{\mathcal{G}}$ which is a linear combination of the approximations on the k previous iteration steps:

$$\mathbf{\Gamma}_*^{\mathcal{G}} = \sum_{i=1}^{k} C_i \mathbf{\Gamma}_{n-k+i}^{\mathcal{G}} \tag{6.16}$$

Below we describe the DIIS and MDIIS algorithms which solve the 3DRISM equations in the form (6.14). We also plan to use the MDIIS algorithm in our multi-grid scheme. This requires one to consider a generalized task in the following form:

$$\mathbf{\Gamma}^{\mathcal{G}} = F[\mathbf{\Gamma}^{\mathcal{G}}] + \mathbf{D}^{\mathcal{G}} \tag{6.17}$$

where $\mathbf{D}^{\mathcal{G}} = \left(\mathbf{d}_1^{\mathcal{G}}, \ldots, \mathbf{d}_{N_{\text{solvent}}}^{\mathcal{G}}\right)^T$ is an arbitrary vector of corrections. The vector of corrections will be calculated during the multi-grid algorithm when we move from one grid to another one. This procedure is described in the next section. In the current section we describe one-grid solvers where vector $\mathbf{D}^{\mathcal{G}}$ is given. Below we describe the DIIS and MDIIS algorithms for a general case of an arbitrary vector $\mathbf{D}$ having in mind that the 3DRISM equations (6.14) correspond to the case $\mathbf{D}^{\mathcal{G}} \equiv 0$.

In the DIIS method the coefficients C_i in (6.16) are chosen to minimize the norm of the residue $\Delta_*^{\mathcal{G}} = \mathbf{\Gamma}_*^{\mathcal{G}} - F[\mathbf{\Gamma}_*^{\mathcal{G}}] - \mathbf{D}^{\mathcal{G}}$. If one assumes linearity of the operator F (which for smooth operators is locally true) then the task reduces to the following system of linear equations [135]:

$$\begin{pmatrix} a_{11} & \dots & a_{1k} & -1 \\ \vdots & \ddots & \vdots & -1 \\ a_{k1} & \dots & a_{kk} & -1 \\ 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} C_1 \\ \vdots \\ C_k \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \tag{6.18}$$

where $a_{ij} = \int_{\mathbb{R}^3} \Delta_i(\mathbf{r})\Delta_j(\mathbf{r})d\mathbf{r}$, $\Delta_i(\mathbf{r}) = \mathbf{\Gamma}_{n-k+i}^{\mathcal{G}} - F[\mathbf{\Gamma}_{n-k+i}^{\mathcal{G}}] - \mathbf{D}^{\mathcal{G}}$. In the DIIS method $\mathbf{\Gamma}_*^{\mathcal{G}}$ is used as a solution approximation on the (n+1)-st iteration step. However, such a procedure can lead to a linearly dependent system of equations. The MDIIS iteration method avoids this problem by adding a weighted residue to the (n+1)-st step approximation [77]:

$$\mathbf{\Gamma'}^{\mathcal{G}} = \mathbf{\Gamma}_*^{\mathcal{G}} + \eta \left(F[\mathbf{\Gamma}_*^{\mathcal{G}}] + \mathbf{D}^{\mathcal{G}} - \mathbf{\Gamma}_*^{\mathcal{G}}\right) \tag{6.19}$$

where η is a weight for the residue. In combination with the standard damping technique the solution approximation on the (n+1)-st step $\mathbf{\Gamma}_{n+1}^{\mathcal{G}}$ in the MDIIS method can be found by using the following formula:

$$\mathbf{\Gamma}_{n+1}^{\mathcal{G}} = (1-\lambda)\Gamma_n^{\mathcal{G}} + \lambda\mathbf{\Gamma}_*^{\mathcal{G}} + \lambda\eta \left(F[\mathbf{\Gamma}_*^{\mathcal{G}}] + \mathbf{D}^{\mathcal{G}} - \mathbf{\Gamma}_*^{\mathcal{G}}\right) \tag{6.20}$$

In our work we use $\lambda = 0.5$, $\eta = 0.3$. These values are sub-optimal and allow one to ensure stability of the algorithm and in the same time retain reasonable performance. Detailed description of the dependence of the computation time on λ and η parameters is given below.

To make notations shorter we introduce the MDIIS operator $\Xi[\cdot,\cdot]$:

$$\Xi[\Gamma_n^{\mathcal{G}}, \mathbf{D}^{\mathcal{G}}] = (1-\lambda)\Gamma_n^{\mathcal{G}} + \lambda\mathbf{\Gamma}_*^{\mathcal{G}} + \lambda\eta \left(F[\mathbf{\Gamma}_*^{\mathcal{G}}] + \mathbf{D}^{\mathcal{G}} - \mathbf{\Gamma}_*^{\mathcal{G}}\right) \tag{6.21}$$

6.3 Multi-grid

We use the multi-grid technique in order to decrease the computation time spent on solving the 3DRISM equations. General description of the multi-grid theory can be found in the book [87]. Here we give only short description of the multi-grid method applied to the 3DRISM equations. More information on the theoretical background of the method can be found in the previous chapter.

In the multi-grid method the numerical task is discretized on several grids with the same buffer but different spacing. Grids with smaller numbers of points and larger spacing are called *coarse* grids, grids with larger number of the points and smaller spacing are called *fine* grids. In our work we consider grids where number of points differ by the factor of 2^n, where $n = 0, 1, 2, ...$

We introduce operators $p[\cdot]$, $r[\cdot]$, which convert a coarse grid to a finer one and vice versa. We introduce an operator $R[\cdot]$ which map a fine-grid function to a coarse grid.

$$R[\mathbf{\Gamma}^{\mathcal{G}}] = \mathbf{\Gamma}^{r[\mathcal{G}]} \tag{6.22}$$

Also we introduce an operator $P[\cdot]$ which interpolates a coarse-grid function to a fine grid:

$$P[\mathbf{\Gamma}^{r[\mathcal{G}]}] = \mathbf{\Gamma}^{\mathcal{G}}_{\mathbf{1}} \tag{6.23}$$

We use the linear interpolation operator.

To make notations simpler we introduce an operator $\Lambda[\cdot;\cdot]$:

$$\Lambda[\mathbf{\Gamma}_{\mathcal{G}}; \mathbf{D}^{\mathcal{G}}] = (1-\lambda)\mathbf{\Gamma}^{\mathcal{G}} + \lambda\left(F_{\mathcal{G}}[\mathbf{\Gamma}^{\mathcal{G}}] + \mathbf{D}^{\mathcal{G}}\right). \tag{6.24}$$

A multi-grid iterative algorithm which solves the task (6.17) can be written in the following form:

$$\mathbf{\Gamma}^{\mathcal{G}}_{n+1} = \mathcal{M}^{l}_{\mathcal{G}}\left[\mathbf{\Gamma}^{\mathcal{G}}_{n}; \mathbf{D}^{\mathcal{G}}\right], \tag{6.25}$$

where $\mathbf{\Gamma}^{\mathcal{G}}_{n}$ is the n-th step approximation, $\mathcal{M}^{l}_{\mathcal{G}}[\cdot;\cdot]$ is a multi-grid operator which performs one multi-grid iteration step of the depth l on the grid $\mathcal{G}$. To calculate the multi-grid operator of the depth $l = 0$ one performs m_0 one-grid iteration steps on the grid $\mathcal{G}$. The multi-grid technique can be applied to both: the Picard and the MDIIS iteration methods. We define a generalized operator $\Phi[\cdot;\cdot]$ in the following way:

$$\Phi[\mathbf{\Gamma}^{\mathcal{G}}_{n}; \mathbf{D}^{\mathcal{G}}] = \left\{ \begin{array}{ll} \Lambda[\mathbf{\Gamma}^{\mathcal{G}}_{n}; \mathbf{D}^{\mathcal{G}}] & \text{for MG-Picard method} \\ \Xi[\mathbf{\Gamma}^{\mathcal{G}}_{n}; \mathbf{D}^{\mathcal{G}}] & \text{for MG-MDIIS method} \end{array} \right. \tag{6.26}$$

Then the multi-grid operator of the depth $l = 0$ is defined as:

$$\mathcal{M}^{0}_{\mathcal{G}}\left[\mathbf{\Gamma}^{\mathcal{G}}; \mathbf{D}^{\mathcal{G}}\right] = \Phi^{m_0}\left[\mathbf{\Gamma}^{\mathcal{G}}; \mathbf{D}^{\mathcal{G}}\right] \tag{6.27}$$

For $l > 0$, given the n-th step approximation $\mathbf{\Gamma}^{\mathcal{G}}_{n}$ and the correction vector $\mathbf{D}^{\mathcal{G}}$, the multi-grid operator $\mathcal{M}^{l}_{\mathcal{G}}[\cdot;\cdot]$ is calculated by the following algorithm:

Algorithm 6.1 *3DRISM Multi-Grid Operator*

Input: $\mathbf{\Gamma}^{\mathcal{G}}_{n}$, $\mathbf{D}^{\mathcal{G}}$, l
Output: $\mathbf{\Gamma}^{\mathcal{G}}_{n+1} = \mathcal{M}^{l}_{\mathcal{G}}[\mathbf{\Gamma}^{\mathcal{G}}_{n}; \mathbf{D}^{\mathcal{G}}]$

1. Perform ν_1 Picard iteration steps on the fine grid (in our work $\nu_1 = 5$):

$$\mathbf{\Gamma}'^{\mathcal{G}} = \Lambda^{\nu_1}\left[\mathbf{\Gamma}^{\mathcal{G}}_{n}; \mathbf{D}^{\mathcal{G}}\right]$$

2. Move to the coarse grid $r[\mathcal{G}]$:

$$\mathbf{\Gamma}^{r[\mathcal{G}]}_{(0)} = R[\mathbf{\Gamma}'^{\mathcal{G}}];$$

3. Calculate the coarse-grid correction:

$$\mathbf{E}^{r[\mathcal{G}]} = R\left[F[\mathbf{\Gamma}'^{\mathcal{G}}]\right] - F[\mathbf{\Gamma}^{r[\mathcal{G}]}_{(0)}]$$

4. Perform recursively μ multi-grid iteration steps of depth $l-1$ on the coarse-grid (in our work μ=1):

$$\mathbf{\Gamma}^{r[\mathcal{G}]}_{(\mu)} = \left(\mathcal{M}^{l-1}_{r[\mathcal{G}]}\right)^{\mu}\left[\mathbf{\Gamma}^{r[\mathcal{G}]}_{(0)}; R[\mathbf{D}^{\mathcal{G}}] + \mathbf{E}^{r[\mathcal{G}]}\right]$$

5. Correct the fine-grid solution using the coarse-grid results:

$$\mathbf{\Gamma}''^{\mathcal{G}} = \mathbf{\Gamma}'^{\mathcal{G}} + P\left[\mathbf{\Gamma}^{r[\mathcal{G}]}_{(\mu)} - \mathbf{\Gamma}^{r[\mathcal{G}]}_{(0)}\right]$$

6. Perform ν_2 Picard iteration steps on the fine grid (in our work $\nu_2 = 0$):

$$\mathbf{\Gamma}^{\mathcal{G}}_{n+1} = \Lambda^{\nu_2}\left[\mathbf{\Gamma}''^{\mathcal{G}}; \mathbf{D}^{\mathcal{G}}\right]$$

The number of the iteration steps m_0 in the multi-grid operator of the depth $l = 0$ depends on the number of the multi-grid iteration step n: $m_0 = m_0(n)$. We define $m_0(n)$ in such a way that after $m_0(n)$ iteration steps, a residue decays by the factor K_n:

$$K_n||\Phi^{m_0(n)}[\mathbf{\Gamma}^{\mathcal{G}}_{n}; \mathbf{D}^{\mathcal{G}}] - \Phi^{m_0(n)+1}[\mathbf{\Gamma}^{\mathcal{G}}_{n}; \mathbf{D}^{\mathcal{G}}]|| < ||\mathbf{\Gamma}^{\mathcal{G}}_{n} - \Phi[\mathbf{\Gamma}^{\mathcal{G}}_{n}; \mathbf{D}^{\mathcal{G}}]|| \tag{6.28}$$

We call the value K_n *the decay factor*.

Constant decay factor may lead to a non-smooth decay of residue from one multi-grid iteration step to another which in turn leads to increasing of the number of the idle coarse-grid iteration steps (see Figure 6.2, solid line). To achieve a smoother decay of the error, in our work we change K_n by the following recursive formula:

$$K_{n+1} = \begin{cases} \max(\frac{1}{\alpha}K_n, K_{\min}) & \text{if} \quad ||\mathbf{\Gamma}^{\mathcal{G}}_{n,m_0} - \Phi[\mathbf{\Gamma}^{\mathcal{G}}_{n,m_0}; \mathbf{D}^{\mathcal{G}}]|| < ||\mathbf{\Gamma}^{\mathcal{G}}_{n+1} - \Phi[\mathbf{\Gamma}^{\mathcal{G}}_{n+1}; \mathbf{D}^{\mathcal{G}}]|| \\ \min(\beta K_n, K_{\max}) & \text{otherwise} \end{cases} \tag{6.29}$$

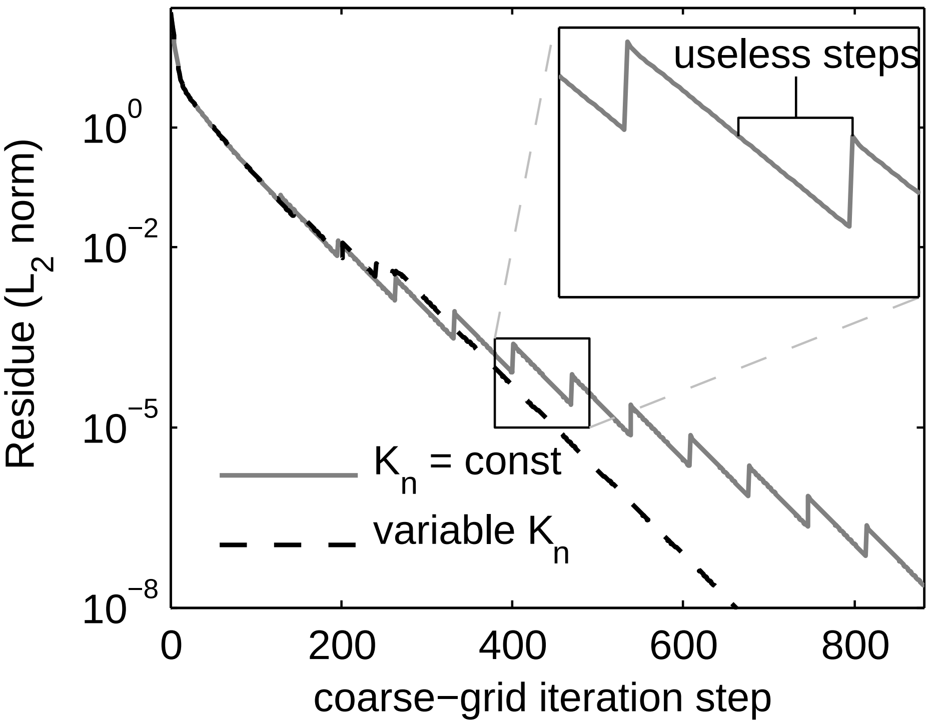


Figure 6.2: Coarse-grid residue decays with the number of the iteration steps in the multi-grid method. Two cases are shown: constant decay factor $K_n = 10$ (solid line), and variable decay factor K_n (dashed line). System: argon aqueous solution, spacing 0.1Å , buffer 6.4Å Peaks on the saw-shaped line (K_n=const) correspond to the boundaries of multi-grid iteration steps. The coarse-grid correction is re-calculated when iteration returns from the coarse grid to the fine grid. Saw-shaped line means that iteration steps on a coarse grid are performed even after the desired accuracy of the coarse-grid correction calculation has been achieved. Thus, a significant number of coarse-grid iteration steps are actually idle because they do not improve the final result. Introducing a variable decay factor allows one to adjust the accuracy of the coarse-grid calculations and to avoid the idle iteration steps.

where $\mathbf{\Gamma}^{\mathcal{G}}_{n,m_0} = (\Phi_{\mathcal{G}})^{m_0(n)}[\mathbf{\Gamma}^{\mathcal{G}}_n; \mathbf{D}^{\mathcal{G}}]$, $\alpha = 2$, $\beta = 1.2$. For the MG-Picard method we use $K_0 = 10$, $K_{\min} = 5$, $K_{\max} = 100$, for the MG-MDIIS method we use $K_0 = 100$, $K_{\min} = 10$, $K_{\max} = 100$. This allows us to smooth the decay of error and to reduce the total number of the iteration steps (see Figure 6.2, dashed line).

Usually iterative algorithms stop when the norm of the residue is less than some threshold. However, this method has its own disadvantages. The first one is that a small residue between two iteration steps does not necessarily imply a small distance from the current approximation to the exact solution. The second one is that a threshold is typically given in dimensionless values which have no physical meaning and thus one has no guidelines to choose an appropriate threshold. In the current work we use another criteria to stop iteration steps. Multi-grid iteration stops on the n-th iteration step if the following condition is satisfied:

$$||\mathbf{\Gamma}_n - \mathbf{\Gamma}_{n+m}|| < \varepsilon_{\text{tres}} \tag{6.30}$$

where m is such that

$$||\mathbf{\Gamma}^{\mathcal{G}}_{n+m} - \mathbf{\Gamma}^{\mathcal{G}}_{n+m+1}|| < 0.01||\mathbf{\Gamma}^{\mathcal{G}}_n - \mathbf{\Gamma}^{\mathcal{G}}_{n+1}|| \tag{6.31}$$

We use such a condition because usually $\mathbf{\Gamma}^{\mathcal{G}}_{n+m}$ is a good approximation of the exact solution. We use a norm based on the Solvation Free Energy calculations:

$$||\mathbf{\Gamma}^{\mathcal{G}}_1 - \mathbf{\Gamma}^{\mathcal{G}}_2|| = |\Delta G_{KH}(\mathbf{\Gamma}_1) - \Delta G_{KH}(\mathbf{\Gamma}_2)| \tag{6.32}$$

The solvation free energy is calculated in the 3DRISM-KH approximation [24]:

$$\Delta G_{KH}(\mathbf{\Gamma}^{\mathcal{G}}) = \rho k_B T \sum_{\alpha}^{N_{\text{solvent}}} \int_{\mathbb{R}^3} \theta(-h_\alpha(\mathbf{r}))h_\alpha(\mathbf{r}) - \frac{1}{2}c_\alpha(\mathbf{r})h_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r})d\mathbf{r} \tag{6.33}$$

where $\theta(\cdot)$ is the Heaviside step function. Because of such definition our threshold has well-defined physical meaning and is measured in energy units. In our work we use $\varepsilon_{\text{tres}}$=0.001 kcal/mol.

To make the calculations faster, in addition to the multi-grid technique we use several grids with the same spacing but different buffers. We introduce a grid-enlargement operator $e[\cdot]$ which enlarges the buffer of a grid. We introduce an operator $E[\cdot]$ which extrapolates a solution $\mathbf{\Gamma}^{\mathcal{G}}$ to a grid $e[\mathcal{G}]$.

$$E[\mathbf{\Gamma}^{\mathcal{G}}] = \mathbf{\Gamma}^{e[\mathcal{G}]} \tag{6.34}$$

Because the functions $\gamma_\alpha(\mathbf{r})$ tend to zero when $|\mathbf{r}| \to \infty$, operator $E[\cdot]$ extrapolates functions by adding zeros at those parts of the grid $e[\mathcal{G}]$ which do not belong to the grid $\mathcal{G}$. The scheme of the iteration can be written in the following way:

$$\mathbf{\Gamma}^{\mathcal{G}}_0 \xrightarrow{\text{solve 3DRISM eqs.}} \mathbf{\Gamma}^{\mathcal{G}}_* \xrightarrow{E[\cdot]} \mathbf{\Gamma}^{e[\mathcal{G}]}_0 \xrightarrow{\text{solve 3DRISM eqs.}} \mathbf{\Gamma}^{e[\mathcal{G}]}_* \xrightarrow{E[\cdot]} \dots \tag{6.35}$$

We start from a zero approximation $\mathbf{\Gamma}_0^{\mathcal{G}}$ on the grid $\mathcal{G}$ with a small buffer and using the scheme (6.35) after several steps we obtain a solution on a grid with a large buffer.

We performed 3DRISM calculations for infinitely diluted aqueous solutions of argon, methane, methanol and dimethyl ether (DME). For the partial charges and Lennard-Jones (LJ) parameters of the solute molecules we used the OPLS-AA force-field parameters [136]. We used the MSPC-E water model [42] to describe the solvent. In the 3DRISM calculations we used total site-site correlation functions of water which were initially calculated by the dielectrically consistent RISM technique [42]. Pairwise σ Lennard-Jones parameters were calculated as an arithmetic mean of atomic parameters, pairwise ϵ Lennard-Jones parameters were calculated as a geometric mean of atomic parameters:

$$\sigma_{12} = \frac{\sigma_1 + \sigma_2}{2}; \qquad \epsilon_{12} = \sqrt{\epsilon_1 \cdot \epsilon_2} \tag{6.36}$$

6.4 Implementation details

The algorithm for solving 3DRISM equations described above was implemented as a computer program. As a programming language was used C++. On the one hand, this language is an object-oriented high-level language that allows one to describe the algorithms in an abstract way. On the other hand, C++ allows one to use low-level methods such as direct memory access, specialized external libraries (like LAPACK), which allow one to create an effective program code. One of important features of the algorithm's implementation is the block structure of the source code and high independence of individual components. For example, multi-grid methods and MDIIS iterations are implemented at a high level of abstraction, so one can use them for a wide class of problems on grids of any kind and of arbitrary dimensionality. For calculation of the direct and inverse fast Fourier transforms the FFTW3 library was used [137]. The source code of the algorithm is open and can be downloaded free of charge [138].

6.5 Finding the optimal grid parameters

We performed 3DRISM calculations for infinitely diluted aqueous solutions of four solutes: argon, methane, methanol and DME. To determine the optimal grid parameters we performed solvation free energy (SFE) calculations on grids with different spacing parameters and different buffers. In Figure 6.3 the dependence of calculation errors on the spacing parameter is shown. For the calculations we used several different grids with the fixed buffer of 8Å and different spacing which vary from 0.1Å to 2Å . Errors were calculated as absolute values of the differences between SFEs calculated on a current grid and SFEs calculated on the very fine grid a the spacing of 0.05Å and the buffer of 8Å . The results show that the grid with a spacing of

0.2Å provides an error that is less than 0.1 kcal/mol for all solutes, which is acceptable for most chemical applications.

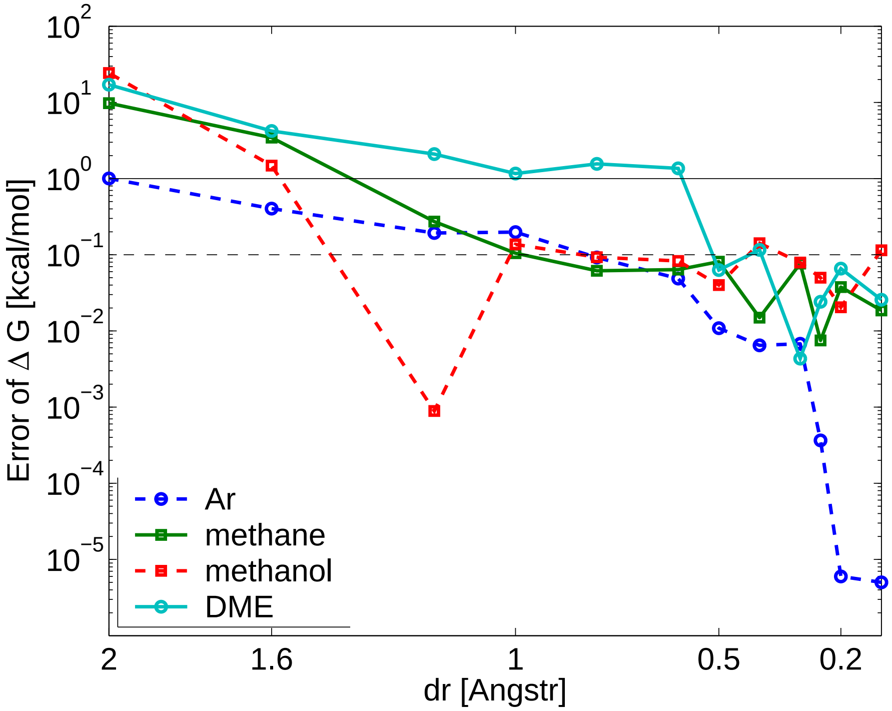


Figure 6.3: Dependency of the calculation errors on the grid spacing at constant buffer (8Å)

In Figure 6.4 we show the dependency of calculation errors on the grid buffer. The calculations were performed on grids with fixed spacing of 0.2Å and different buffers varying from 8Å to 20Å. Errors were calculated as differences between the SFEs calculated on the current grid and the SFEs calculated on the very fine grid with a spacing of 0.2Å and a buffer of 30Å. Figure 6.4 shows that the grid with the buffer of 15Å is sufficient for an SFE accuracy of $\leqslant$ 0.1kcal/mol.

6.6 Dependencies of the computational time on η and λ

In Figures 6.5-6.6 the dependencies of the computational time on λ for MG-Picard and MG-MDIIS methods are shown. We can see that generally for the both methods the computational time decrease with increasing of λ. However, for $\lambda > 0.9$ the MG-MDIIS method diverges for some of the systems. Also, for $\lambda < 0.3$ the method is non-stable (this is because when λ is small the vectors in the DIIS matrix become nearly-linear dependent, which makes the method less stable). In the both methods – MG-Picard and MG-MDIIS - we use the sub-optimal value λ=0.5. This allows us to ensure the convergence and to have a reasonable

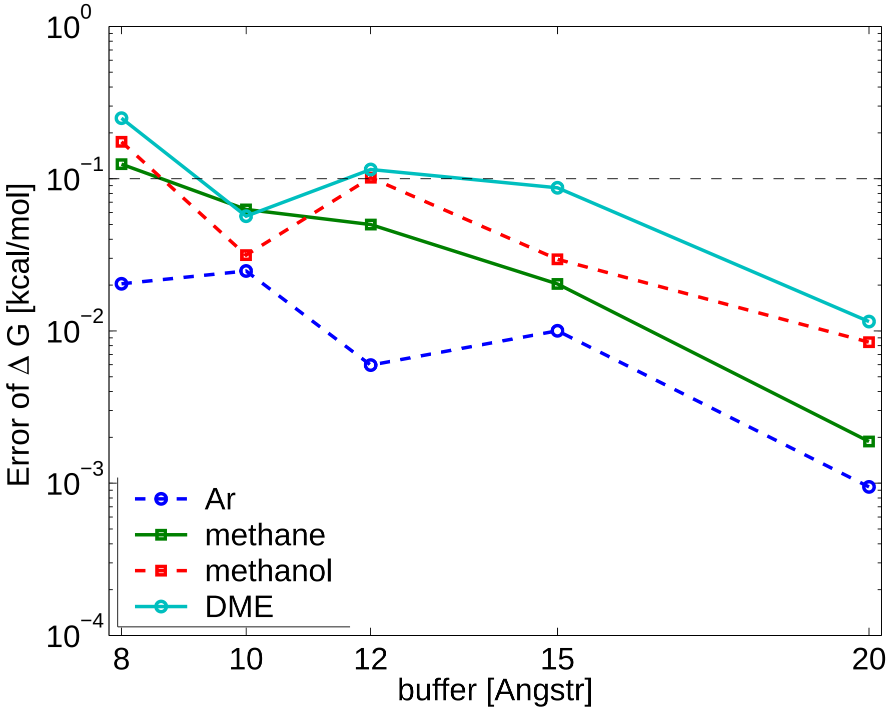


Figure 6.4: Dependency of calculation errors on the grid buffer with constant spacing of the grid (0.2Å).

computational performance. In Figure 6.7 the dependency of the computational time of the MG-MDIIS iteration on η is shown. We see that the computational time weakly depends on this parameter. This can be explained by the specificity of the MDIIS algorithm, where the next solution approximation is the linear combination of several previous approximations. Thus only the linear independence of the solutions matters, but not the value of the scaling coefficient η. In our work we use the value η=0.3, as it was used in the original paper Ref. [77].

6.7 Computational benchmarks of different 3DRISM solvers

To check the numerical performance of the proposed multi-grid algorithm we performed 3DRISM calculations for infinitely diluted aqueous solutions of argon, methane, methanol and DME using the Picard iteration, the MDIIS, the MG-Picard and the MG-MDIIS methods. For the Picard and the MDIIS methods the grid with spacing of 0.2Å and buffer of 15Å was used. For the multi-grid methods (MG-Picard, MG-DIIS) we used the scheme of Eq. (6.35) with two enlargements: we start from the grid with the buffer of 7.65Å , then move to the grid with the buffer of 10.71Å and finish iteration on the grid with the buffer of 15Å . Solutions on the grids with smaller buffers were used as initial guesses for the grids with larger buffers. For each

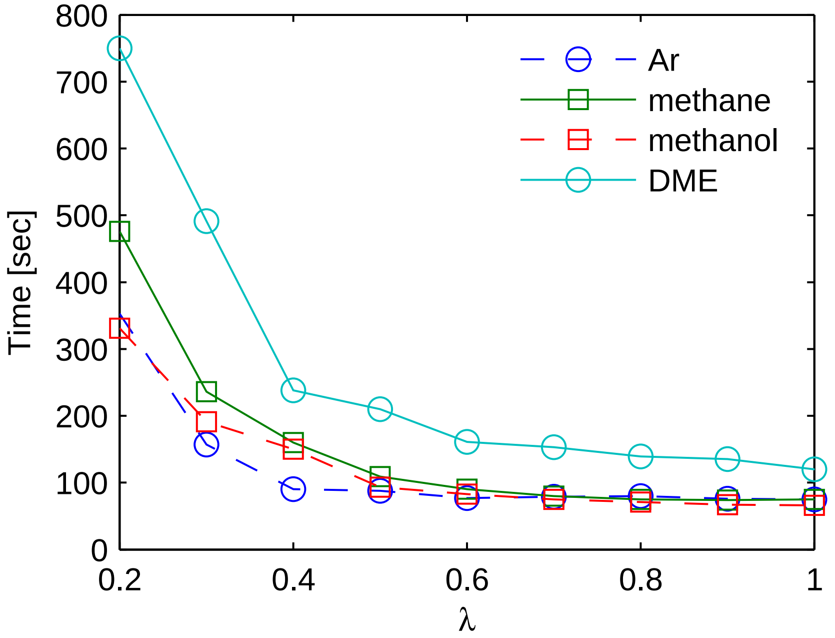


Figure 6.5: Dependency of the computational time on λ for the MG-Picard method

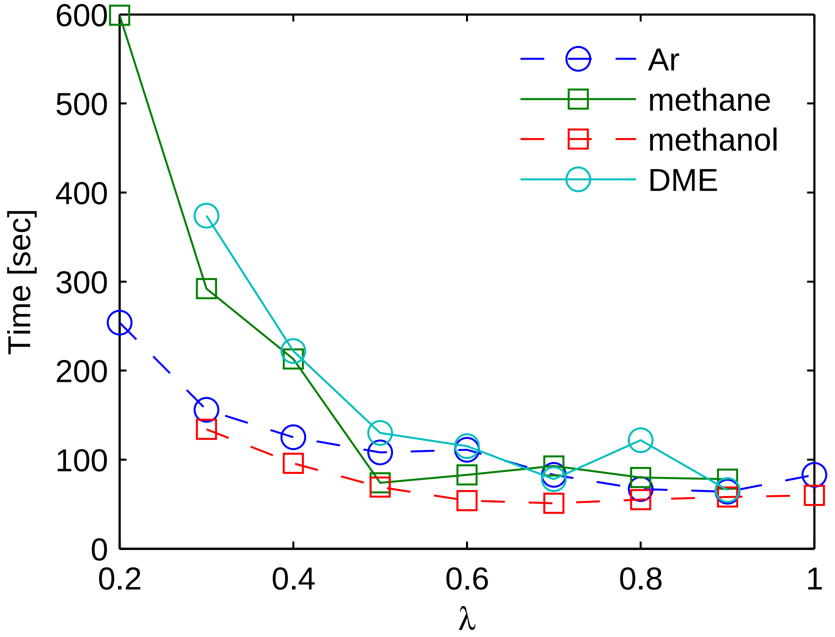


Figure 6.6: Dependency of the computational time on λ for the MG-MDIIS method

buffer we used the multi-grid algorithm with 3 different grids (depth $l = 2$).

Computational expenses in solving 3DRISM equations for each of the investigated four

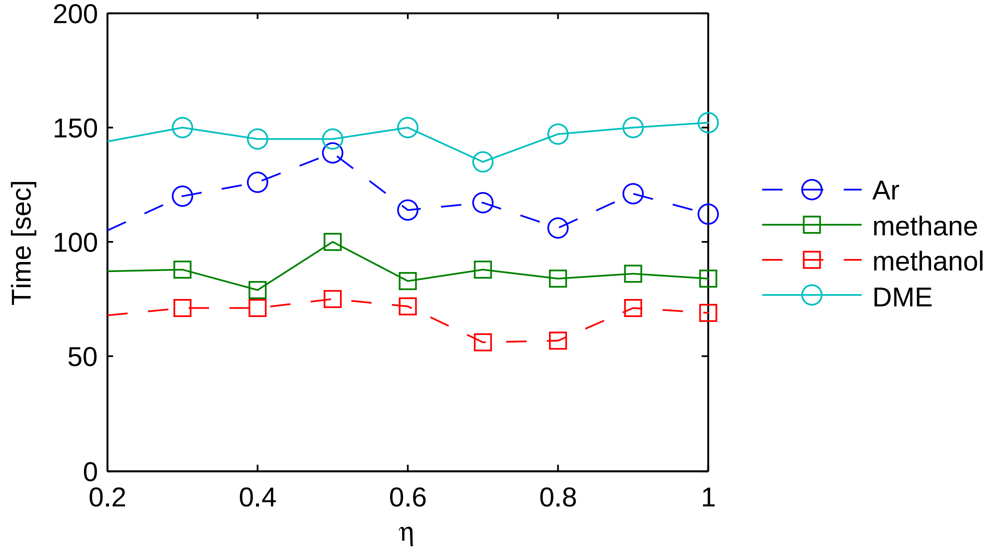


Figure 6.7: Dependency of the computational time on η for the MG-MDIIS method

Compound	**Picard iteration**	**MDIIS**	**MG-Picard**	**MG-DIIS**
argon	1148 sec	167 sec	46 sec	50 sec
methane	1484 sec	154 sec	149 sec	82 sec
methanol	1857 sec	416 sec	165 sec	83 sec
dimethyl ether	4462 sec	509 sec	241 sec	133 sec

Table 6.1: Computation expenses of 3DRISM calculations with the Picard iteration, MDIIS, MG-Picard and MG-MDIIS methods.

methods are presented in Table 6.1. These results show that the Picard iteration is the least efficient method, while the most efficient is the MG-MDIIS method. We note that the multi-grid methods in all investigated cases are more efficient than one-grid methods.

Figure 6.8 compares computational performance of the MDIIS, the MG-Picard and the MG-MDIIS with the Picard iteration method. The figure shows that for all four compounds multi-grid methods give more a factor of 10 speedup while for three of these four compounds the MG-MDIIS method is more than 20 times faster than the Picard iteration method. Average speedup factors with respect to the plain Picard method for the MDIIS, the MG-Picard and the MG-MDIIS methods are correspondingly 7.4, 16.2 and 24.2. The most effective is the MG-MDIIS method that is in average about 3.5 faster than the MDIIS method. Difference between the multi-grid methods is not very large: the MG-MDIIS method is in average only 1.5 times faster than the MG-Picard method. The results show that the multi-grid scheme can be effectively used in combination with different types of coarse-grid iteration methods for

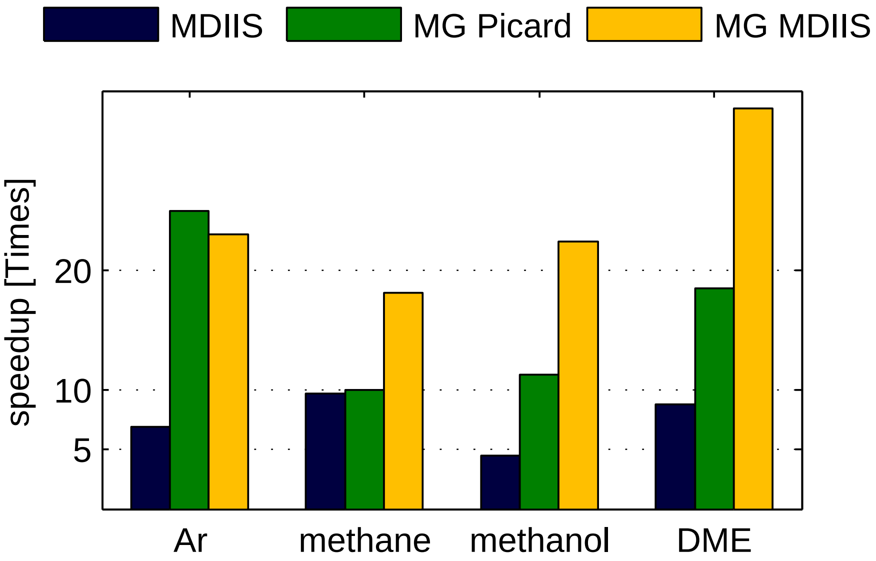


Figure 6.8: Speed up of the calculations by using the MDIIS, the MG-Picard and the MG-MDIIS methods as compared to the Picard iteration method.

solving the 3DRISM equations for aqueous solutions of small non-charged molecules.

6.8 Computational benchmarks on a large set of organic molecules

The main goal of this part of our study was to investigate the overall efficiency of the new method in a view of large-scale practical applications, for example physical-chemical profiling of large sets of organic compounds. We performed an additional benchmark and tested the efficiency of the new algorithm on a set of organic molecules as well as the accuracy of the SFE prediction. We estimate average computational expenses for the 3DRISM calculations and also check whether numerical accuracy of the calculations is sufficient for accurate estimation of SFEs.

We have chosen a set of 99 organic molecules. This set of molecules is a part of the set used in [74]. The list of the molecules in the set is given in the Table 6.2 and Table 6.3.The experimental free energy values were taken from the Ref. [74]. This set includes alkanes, ketones, alkyl-benzenes, alcohols, alkyl-phenols, ethers and other (polyfunctional) molecules. The number of atoms in molecules of the set varies from 5 to 31. An average number of atoms in the molecule is 16. The Antechamber tool [139] from the Amber Tools 1.4 Package [140] was used for molecular structure optimization and assigning Force-Field parameters. Structures of the molecules were optimized by using the AM1 method [141]. Atomic partial charges were calculated by using the bond charge correction (BCC) method [142, 143]. LJ parameters from the General Amber Force Field (GAFF) [144] were assigned to the solutes. The benchmark calculations for simple molecules reported above show that the most effective is a combination of the multi-grid and MDIIS (MG-MDIIS) methods. Therefore, we use the MG-MDIIS algorithm in our benchmarking of the overall efficiency of the method.

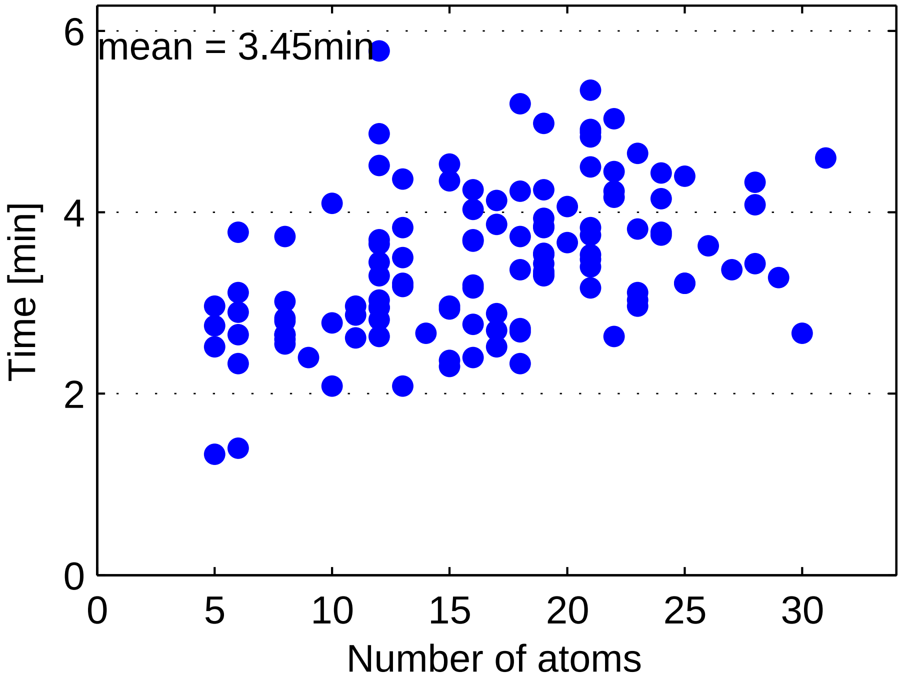


Figure 6.9: Dependency of the computational time spent on MG-MDIIS calculations on the molecule number of atoms for 99 organic molecules from the chosen molecule set.

Figure 6.9 shows dependency of the computational time spent on MG-MDIIS 3D-RISM calculations as a function of the number of atoms in a molecule. The plot shows that the computational time can vary for different molecules even if the molecules have the same number of atoms. However, this somehow counterintuitive result has a straightforward explanation. Indeed, convergence of the algorithm depends not only on the number of atoms but also on the chemical composition of a molecule and its structure, particularly on the distribution of atomic partial charges and the molecule surface accessible area. This is illustrated by the results shown in Figure 6.3 and Figure 6.4 that show different error dependencies for polar and non-polar molecules. Also, even if two different molecules have the same number of atoms, they may still have rather different shapes. This can result in different grid sizes for them. More compact molecules need smaller grids than the less compact molecules, even if they have the same buffer and the same number of atoms. Therefore, combination of these two factors causes this significant spread of computational time for molecules of the same number of atoms. However, the computational time for any molecule in the set is still less than 6 minutes. Average computational time is some 3.5 minutes (3 min 27 sec).

We used the 3D RISM correlation functions calculated by MG-MDIIS method as an input for SFE calculations for all molecules from the above mentioned set of 99 organic compounds. We used 25 molecules as a training set and the rest of the set (74 molecules) as a test set (see Table 6.2 and Table 6.3 for the full list of compounds in the training and test sets).

Table 6.2: Compounds in the training set. The values of the solvation free energies are given in kcal/mol.

Compound	ΔG_{GF}	ρV	ΔG_{calc}	ΔG_{exp}	$\Delta G_{exp} - \Delta G_{calc}$
1_1-dichloroethane	7.439	4.113	-0.452	-0.846	-0.395
1_1_2-trichloroethane	6.075	4.661	-3.037	-1.991	1.046
1_2-dichloropropane	8.267	4.941	-1.472	-1.269	0.203
1_2_3_5-tetrachlorobenzene	10.951	6.796	-2.925	-1.623	1.302
1_3-dichlorobenzene	10.720	5.671	-0.647	-0.982	-0.336
1_4-dichlorobenzene	10.672	5.672	-0.697	-1.009	-0.312
2-ethyltoluene	12.532	7.063	-1.940	-1.037	0.902
2-methylpentan-2-ol	8.645	6.506	-4.583	-3.927	0.656
2-methylstyrene	12.453	6.770	-1.364	-1.240	0.124
2_3-dimethylbuta-1_3-diene	12.618	5.587	1.439	0.394	-1.045
2_4-dimethylphenol	6.742	6.477	-6.421	-6.013	0.407
3-methylhexane	18.250	7.246	3.371	2.713	-0.658
4-chlorophenol	4.416	5.304	-6.132	-7.036	-0.904
4-methylpentan-2-one	9.505	6.228	-3.104	-3.054	0.049
cis-1_2-dichloroethene	6.263	3.742	-0.801	-1.174	-0.372
heptan-2-one	11.839	7.198	-2.934	-3.040	-0.106
hexan-3-ol	8.773	6.591	-4.646	-4.063	0.583
methane	5.426	2.050	2.138	1.991	-0.147
n-nonane	22.854	9.117	3.801	3.136	-0.665
o-xylene	10.611	6.295	-2.147	-0.901	1.246
octanal	14.751	8.161	-2.170	-2.292	-0.122
pentachloroethane	11.806	5.852	0.037	-1.391	-1.428
propan-1-ol	2.044	4.000	-5.595	-4.854	0.741
tert-butylbenzene	15.590	7.844	-0.623	-0.437	0.186
trans-hept-2-ene	16.954	6.996	2.632	1.678	-0.954

Table 6.3: Compounds in the test set. The values of the solvation free energies are given in kcal/mol.

Compound	ΔG_{GF}	ρV	ΔG_{calc}	ΔG_{exp}	$\Delta G_{exp} - \Delta G_{calc}$
1_1-dichloroethene	8.287	3.786	1.125	0.246	-0.878
1_1_1-trichloroethane	9.588	4.745	0.287	-0.191	-0.478
1_1_1_2-tetrachloroethane	10.361	5.297	-0.171	-1.281	-1.110
1_1_2_2-tetrachloroethane	9.666	5.216	-0.686	-2.469	-1.783
1_2-dichlorobenzene	10.182	5.630	-1.092	-1.365	-0.273
1_2-dichloroethane	7.107	4.097	-0.749	-1.785	-1.037
1_2_3-trichlorobenzene	10.557	6.224	-2.042	-1.240	0.801
1_2_3-trimethylbenzene	11.479	7.049	-2.961	-1.214	1.747
1_2_3_4-tetrachlorobenzene	10.514	6.825	-3.425	-1.336	2.089
1_2_4-trichlorobenzene	11.147	6.328	-1.685	-1.119	0.567
1_2_4-trimethylbenzene	11.741	7.133	-2.886	-0.858	2.028
1_2_4_5-tetrachlorobenzene	11.094	6.818	-2.830	-1.336	1.494
1_3-dichloropropane	7.767	4.961	-2.014	-1.895	0.119
1_3_5-trichlorobenzene	11.637	6.316	-1.168	-0.777	0.391
1_3_5-trimethylbenzene	12.261	7.227	-2.575	-0.901	1.674
2-butoxyethanol	7.764	6.977	-6.515	-6.260	0.255
2-chlorophenol	2.923	5.256	-7.518	-4.555	2.963
2-ethoxyethanol	3.065	5.211	-7.275	-6.697	0.578
2-methylbut-2-ene	12.450	5.190	2.157	1.310	-0.848
2-methylbuta-1_3-diene	10.555	4.806	1.118	0.681	-0.437
2-methylbutan-2-ol	6.349	5.639	-4.946	-4.431	0.515
2-methylpentan-3-ol	8.800	6.482	-4.375	-3.886	0.488
2-methylpentane	16.142	6.405	3.139	2.510	-0.629
2-methylpropan-1-ol	4.129	4.826	-5.353	-4.500	0.853
2-phenylethanol	5.607	6.496	-7.600	-6.793	0.807
2-propoxyethanol	5.621	6.099	-6.700	-6.410	0.290
2_2-dimethylpentane	17.996	7.114	3.410	2.878	-0.532
2_3-dimethylpentane	17.926	7.113	3.344	2.524	-0.820
2_3-dimethylphenol	6.698	6.427	-6.355	-6.164	0.191
2_3_4-trimethylpentane	19.810	7.855	3.572	2.565	-1.008
2_5-dimethylphenol	6.431	6.478	-6.734	-5.918	0.816
2_6-dimethylphenol	6.860	6.400	-6.132	-5.265	0.866
3-hydroxybenzaldehyde	1.655	5.527	-9.390	-9.505	-0.115
3-methylpentane	15.983	6.357	3.086	2.510	-0.577
3-phenylpropanol	5.872	7.389	-9.327	-6.929	2.398
3_4-dimethylphenol	5.490	6.381	-7.459	-6.506	0.953
4-ethyltoluene	13.157	7.229	-1.683	-0.954	0.729
4-hydroxybenzaldehyde	0.718	5.515	-10.300	-8.836	1.464
4-methoxyacetophenone	9.543	7.321	-5.505	-4.405	1.100
benzyl_alcohol	3.087	5.576	-8.068	-6.628	1.440

Continued on next page

Compound	ΔG_{GF}	ρV	ΔG_{calc}	ΔG_{exp}	$\Delta G_{exp} - \Delta G_{calc}$
buta-1_3-diene	8.528	3.998	0.892	0.614	-0.278
chlorobenzene	9.777	5.139	-0.401	-1.119	-0.717
decan-2-one	18.358	9.834	-2.295	-2.345	-0.050
dichloromethane	5.009	3.223	-0.897	-1.310	-0.413
ethane	7.529	2.960	2.209	1.828	-0.381
heptanal	12.527	7.276	-2.419	-2.672	-0.253
hexa-1_5-diene	13.075	5.761	1.508	1.009	-0.499
hexan-1-ol	8.618	6.631	-4.890	-4.405	0.485
hexanal	10.297	6.391	-2.675	-2.808	-0.134
isobutylbenzene	16.383	8.037	-0.261	0.163	0.423
m-xylene	10.703	6.329	-2.130	-0.832	1.298
methanol	-2.399	2.229	-6.087	-5.100	0.986
n-butane	11.710	4.680	2.555	2.072	-0.483
n-heptane	18.470	7.363	3.330	2.672	-0.658
nonan-1-ol	15.346	9.301	-4.118	-3.886	0.232
nonan-2-one	16.218	8.956	-2.476	-2.495	-0.020
nonanal	16.968	9.044	-1.921	-2.072	-0.151
oct-1-ene	19.109	7.882	2.812	1.924	-0.888
octan-1-ol	13.084	8.418	-4.410	-4.092	0.318
octan-2-one	14.014	8.079	-2.722	-2.878	-0.155
p-xylene	10.661	6.321	-2.154	-0.805	1.349
pent-1-ene	12.477	5.232	2.091	1.678	-0.413
penta-1_4-diene	11.058	4.893	1.428	0.927	-0.500
pentan-1-ol	6.490	5.775	-5.107	-4.570	0.538
pentan-2-ol	6.184	5.705	-5.257	-4.391	0.867
pentan-3-ol	5.965	5.694	-5.453	-4.350	1.103
propan-2-ol	1.713	3.984	-5.891	-4.747	1.145
propene	8.024	3.475	1.556	1.322	-0.235
sec-butylbenzene	16.153	7.993	-0.393	-0.449	-0.056
tetrachloroethene	10.552	4.921	0.858	0.096	-0.762
tetrachloromethane	9.188	4.439	0.570	0.081	-0.489
trans-1_2-dichloroethene	7.498	3.763	0.388	-0.777	-1.165
trichloroethene	8.741	4.347	0.329	-0.437	-0.767
trichloromethane	7.225	3.831	-0.038	-1.078	-1.040

For accurate SFE calculations we used the Universal Correction (UC) method that was introduced in recent papers [73, 74]. We tested two modifications of the UC method. The first one (UC-KH method) is based on the Kovalenko-Hirata (KH) Free Energy functional (6.33). The second one (UC-GF method) is based on the Free Energy calculations using the Gaussian

Fluctuations (GF) formula [116]:

$$\Delta G_{GF} = \rho k_B T \sum_{\alpha=1}^{N_{\text{site}}} \int_{\mathbb{R}^3} (-\frac{1}{2} h_\alpha(\mathbf{r}) c_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r})) d\mathbf{r} \tag{6.37}$$

where k_B is the Boltzmann constant, T is the temperature, ρ is the number density of a bulk solvent. In the UC-GF method (ΔG_{UC}^{GF}) and in the UC-KH method (ΔG_{UC}^{KH}) SFE is calculated by using the following relations:

$$\Delta G_{UC}^{GF} = \Delta G_{GF} + a_{GF} \rho V + b_{GF} \tag{6.38}$$

$$\Delta G_{UC}^{KH} = \Delta G_{KH} + a_{KH} \rho V + b_{KH} \tag{6.39}$$

where V is the partial molar volume of the molecule, a_{GF}, b_{GF}, a_{KH}, b_{KH} are calculated by using the linear regression method to fit experimental data. Partial molar volume of a molecule was calculated by the following formula [37, 145]:

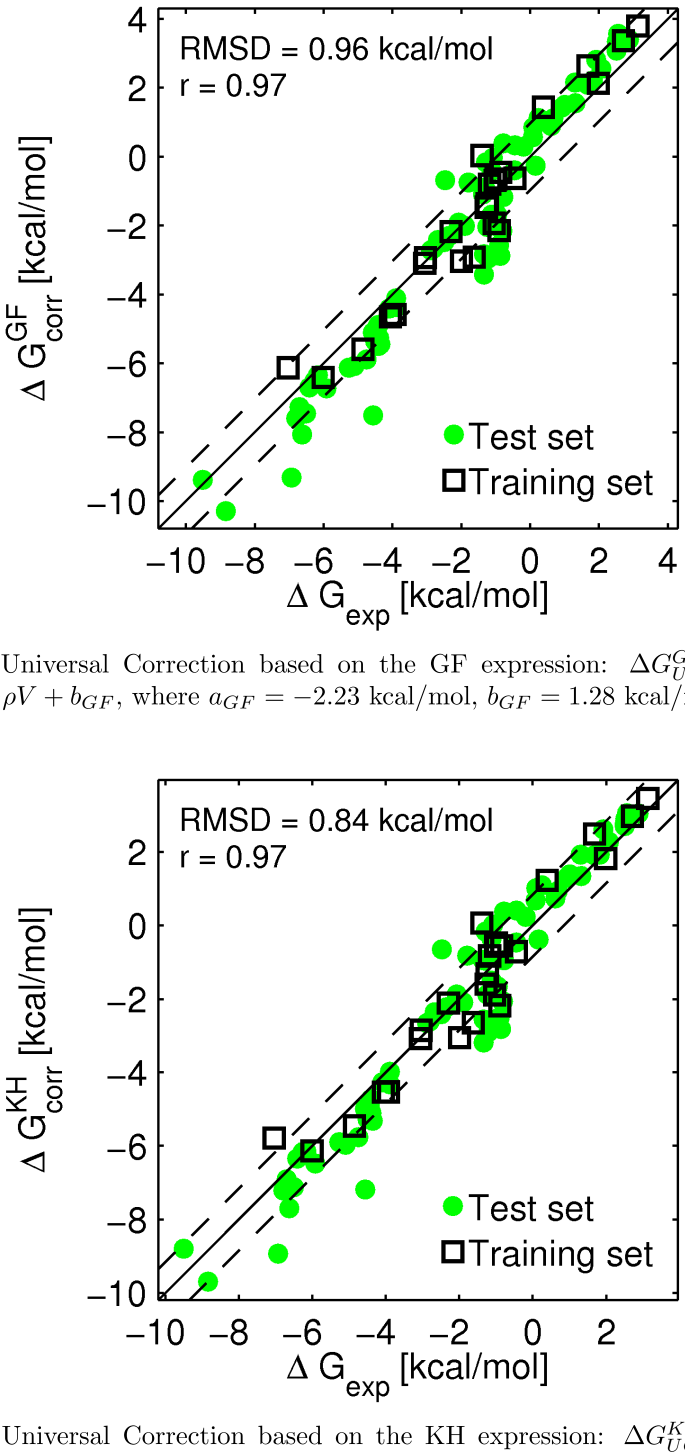
$$V = \left(\frac{1}{\rho} + 4\pi \int\limits_0^\infty (g_{oo}(r) - 1) r^2 dr \right) \left(1 - \rho \sum_{\alpha=1}^{N_{\text{site}}} \int_{\mathbb{R}^3} c_\alpha(\mathbf{r}) d\mathbf{r} \right) \tag{6.40}$$

where $g_{oo}(r)$ is the oxygen-oxygen RDF of bulk water from Ref. [42] where it was calculated using the dielectrically consistent RISM.

Using the training set of compounds, the following values of coefficients were obtained by the linear regression fitting procedure [64, 74] : a_{GF}=-2.23 kcal/mol b_{GF}=1.28 kcal/mol for the UC-GF method and a_{KH} = -3.51 kcal/mol, b_{KH} = 0.81 kcal/mol for the UC-KH. Figure 6.10(a) and Figure 6.10(b) shows the correlation between the experimental values ΔG_{exp} and the calculated values $\Delta G_{\text{UC}}^{GF}$ and ΔG_{UC}^{KH}. The correlation coefficient is 0.97 for the both methods. Root mean square deviation (RMSD) on a test set is 0.96 kcal/mol for the UC-GF method and 0.84 kcal/mol for the UC-KH method. Accuracy of predictions is comparable with accuracies of experimental methods [65,93] and corresponds to accuracies of current state-of-the-art methods for SFE calculations by molecular dynamics [11, 124, 146, 147] and other advanced molecular theories (e.g. the energy representation method by Matubayasi and Nakahara) [9, 12, 63, 148, 149]. Thus we show that the numerical accuracy of the algorithm is enough for SFE calculations and parameterization of calculation results.

6.9 Conclusions

We proposed a new multi-grid based method which solves the 3DRISM equations. To determine the optimal grid parameters we performed 3DRISM calculations for infinitely diluted aqueous solutions of argon, methane, methanol, and dimethyl ether. We showed that on the grid



(a) Universal Correction based on the GF expression: $\Delta G_{UC}^{GF} = \Delta G_{GF} + a_{GF}\rho V + b_{GF}$, where $a_{GF} = -2.23$ kcal/mol, $b_{GF} = 1.28$ kcal/mol

(b) Universal Correction based on the KH expression: $\Delta G_{UC}^{KH} = \Delta G_{KH} + a_{KH}\rho V + b_{KH}$, where $a_{KH} = -3.51$ kcal/mol, $b_{KH} = 0.81$ kcal/mol

Figure 6.10: Correlation of experimentally measured SFEs with the SFEs values calculated by the Universal Correction method for the investigated set of organic molecules.

with a spacing of 0.2Å and buffer of 15Å , the maximal error is less than 0.1 kcal/mol. We tested two modifications of the multi-grid algorithm: MG-Picard and MG-MDIIS methods. We compared the numerical efficiency of the multi-grid algorithms with the numerical efficiency of the standard Picard iteration method and the MDIIS method. We showed that the MG-MDIIS algorithm is more than 24 times faster than the Picard iteration method and more than 3.5 times faster than the MDIIS method.

In turn, efficiencies of the MG-DIIS and MG-Picard methods do not differ very much. The MG-DIIS method is about 1.5 times faster than the MG-Picard method. We suggest that the most effective MG-MDIIS method can be used in the future as a fast tool for calculations of Solvation Free Energy for organic molecules. To support this statement we performed 3DRISM calculations for aqueous solutions of 99 organic compounds. For all compounds in the set the computational time does not exceed 6 minutes per one molecule while the average computational time is only 3.5 minutes per one molecule on a standard personal computer. We calculated solvation free energies by using GF and KH expressions with the universal partial molar volume corrections (UC-GF and UC-KH methods). We showed that calculated and experimental values of solvation free energy are strongly correlated to each other (correlation coefficient is 0.97). RMSD error for the test set of compounds is less than 1 kcal/mol for both UC-GF and UC-KH methods. The performed tests show that the proposed algorithm can be used for fast and accurate predictions of aqueous solvation free energies of neutral molecules.

Chapter 7

Conclusions and further work

7.1 Conclusions

The goal of the current work was the development of fast and accurate methods for solvation free energy calculations that are reliable for practical applications. The RISM and the 3DRISM molecular models in combination with the semi-empirical methods for solvation free energy calculations were used to achieve high accuracy of the calculations. The multi-grid technology was used to accelerate the calculations. The following results were achieved in the current research project:

- Multi-grid based numerical methods for solving the RISM and the 3DRISM problems were developed and implemented in a form of computer programs [121, 138].

- The methods were optimized for fast solvation free energy calculations. To determine the optimal grid parameters the test calculations for simple molecules on various different grids were performed. Optimal parameters which allow one to calculate the SFE with a required accuracy in the minimal computation time were determined.

- Computational performance of the proposed methods was compared to the performance of standard approaches:

  - The RISM multi-grid algorithm was compared to the one-grid Picard iteration scheme, which is the reference algorithm, and to the nested Picard iteration algorithm, which is the most straightforward implementation of the multi-scale scheme. It was shown that at the same accuracy level the proposed method is about 30 times faster than the single-grid Picard iteration, and almost 7 times faster than the nested Picard iteration.

  - For the 3DRISM multi-grid algorithm there were proposed two modifications: multi-grid Picard (MG-Picard) and multi-grid MDIIS (MG-MDIIS) schemes. It was shown that the MG-MDIIS algorithm is more than 24 times faster than the plain-Picard iteration method and more than 3.5 times faster than the plain MDIIS method.

- Correlation functions calculated with the RISM and 3DRISM methods were used for solvation free energy calculations. Semi-empirical Solvation Free Energy calculation methods were used to increase the accuracy of the SFE calculations. The methods were benchmarked on extended sets of organic compounds from different chemical groups, including polyfunctional drug-like molecules.

  - To test the effectiveness of the developed RISM-based method for calculation of SFE of bioactive drug-like molecules a set of 63 compounds from Ref. [93] was used. It was shown that using the atom type correction method with the CHELPG charges and PW expression it is possible to get an accuracy of 1.9 kcal/mol which is almost 1.5 fold better than the verification "pure chemoinformatic" parameterization. The average computation time on this set of compounds was about 15 sec/molecule.

  - For testing of the 3DRISM-based method a set of 99 organic compounds from Ref. [74] was chosen. For these compounds the correlation functions with the 3DRISM MG-MDIIS algorithm were calculated. The Universal Correction (UC) model was used for the SFE calculations. For all compounds in the set the computational time of 3DRISM calculations did not exceed 6 minutes per molecule while the average computational time was only 3.5 minutes per molecule on a standard personal computer. The solvation free energies were calculated by using GF and KH expressions with the universal partial molar volume corrections (UC-GF and UC-KH methods). It was shown that calculated and experimental values of solvation free energy are strongly correlated to each other (correlation coefficient is 0.97). RMSD error for the test set of compounds was less than 1 kcal/mol for both UC-GF and UC-KH methods.

The performed tests show that the proposed methods are suitable for fast and accurate calculations of solvation free energies of organic compounds form different chemical classes in aqueous solutions. Therefore, we conclude that all the goals of the thesis were successfully achieved.

7.2 Future work

7.2.1 Approaches for solving 6D OZ equation using low-rank representations of multidimensional functions

As it was shown, the methods based on RISM and 3DRISM theories can be successfully used for prediction of SFE of bioactive compounds. However, in some cases RISM and 3DRISM are not able to give adequate description of the system. It is known that in RISM approximation the structure of the solvent molecule is not represented correctly. In general RISM and 3DRISM can give a good description for systems where the solvent molecule is relatively small (e.g. water, supercritical CO_2) while for the systems where the solvent molecule is a more complicated compound (e.g. octanol, toluene, ionic liquids) RISM is not able to give a accurate description of the solvent structure which also should have effect on the accuracy of SFE calculations. The most straightforward method to overcome the disadvantages of RISM theory is to use six-dimensional OZ equations instead, which are suitable for description of solvents of arbitrary complexity. It is a challenging task to develop an algorithm which solves six dimensional integral equations. Despite this fact there are some examples of successful solving the six-dimensional OZ equations for a few different systems [54, 55, 150]. Unfortunately till now no universal algorithm which is applicable to any system was proposed and tested. Thus there is still a place for investigations.

The most of the methods for solving OZ equations use explicitly or implicitly some kind of tensor product representation for the correlation functions. This mean that to approximate the function $f(\mathbf{r}, \boldsymbol{\theta})$ the following representation is used:

$$f(\mathbf{r}, \boldsymbol{\theta}) = \sum_{j=1}^{n} f_j(\mathbf{r}) \phi_j(\boldsymbol{\theta}) \tag{7.1}$$

To represent the angular dependencies of the six-dimensional functions different basis functions can be used including wavelets, rotational invariants and others. We propose to use a harmonic basis set for representing the angular dependencies. There are at least two reasons for this. Firstly, at the large distances dipole-dipole interaction is proportional to $r^{-3} \cos\theta$ there θ is the angle between the dipole axes. It is known, that the interaction of two molecules can be represented as a sum of multipole interactions, and dipole-dipole interaction is the leading term in this representation. Thus at large distances the interaction of the molecules can be effectively described by the harmonic series with a one-two members. Secondly, a harmonic basis is convenient from the computational point of view. Many operations in harmonic basis can be reduced to the Fourier transformation which in turn can be calculated with the FFT algorithm. So we propose to use the following representation for all of the six-dimensional functions which occur in OZ equation:

$$f(\mathbf{r},\boldsymbol{\theta}) = \sum_{k_1,k_2,k_3} f_{k_1k_2k_3}(\mathbf{r})e^{ik_1\theta_1+ik_2\theta_2+ik_3\theta_3} \tag{7.2}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$, $k_1 = 0 \dots N_1(\mathbf{r})$, $k_2 = 0 \dots N_2(\mathbf{r})$, $k_3 = 0 \dots N_3(\mathbf{r})$, $N_1(r)$, $N_2(r)$, $N_3(r)$ are the numbers of the discretization points of angular components at the spatial point $\mathbf{r}$. We propose to use a equispaced grid for the spatial components and non-equispaced grid for the angular components. Without the proper numerical experiments it is difficult to predict which values of $N_i(\mathbf{r})$ one need to use to achieve the reasonable accuracy of the calculations. However it is reasonable to assume that at the distances of $5 - 10$Å $\quad N_i(\mathbf{r}) \approx 1 - 2$, while near the solute molecule these values can rise up to $N_i \approx 10 - 30$. In that case the general number of discretization points will not exceed much the number of discretization points for the 3DRISM algorithm. The detailed description of the proposed format is out of the scope of the current section. More detailed description and discussion of the format can be found in the Appendix A. Here we only note that the most of operations in the format (7.2) require not more than $O(N \log N)$ operations, where $N = \sum_{\mathbf{r}}(N_1(\mathbf{r}) + N_2(\mathbf{r}) + N_3(\mathbf{r}))$ is the total number of discretization points. Considering that at the large distances $N_i(\mathbf{r}) \approx 1 - 2$, the number N should not exceed much the number of the discretization points in 3DRISM equations. Thus the computational expenses for solution of the 6D equations in the proposed format should be comparable to the computational expenses for 3DRISM method. So we think that 6D equation should not be regarded any more as some unfeasible task and we hope that the six dimensional methods can become a routine tool in chemical investigations in the next 3-5 years.

7.2.2 Proper closure for the OZ equations

We note that the development of the effective OZ equations solver is not enough to obtain the correct computational results. As mentioned in the previous chapters it is impossible to solve the OZ equation without the closure relation. And the existing closure relations, such as HNC or KH, are not able to give correct results. In this section we propose the way to obtain more accurate closure relation. We recall the procedure for obtaining the HNC closure from the chapter 3 section 3.10. Let's consider a smooth transition from the gas phase to solution. We consider the coordinate system connected to one of the molecules of type a and obtain the following expression for the g^{ab} functions:

$$g^{ab}(\mathbf{r_0}, \mathbf{r_1}, \boldsymbol{\theta_0}, \boldsymbol{\theta_1}) = \exp\left(-\beta u^{ab}(\mathbf{r_1}, \boldsymbol{\theta_1}) + c^b(\mathbf{r_1}, \boldsymbol{\theta_1})\right) \tag{7.3}$$

where $(\mathbf{r_0}, \boldsymbol{\theta_0}) \equiv (\mathbf{0}, \mathbf{0})$. Assuming the linear change of the density from initial to the final state $\rho^b(\mathbf{r_1}, \boldsymbol{\theta_1}; \lambda) = \rho^b(\mathbf{r_1}, \boldsymbol{\theta_1}; \lambda = 0) + \lambda\Delta\rho^b(\mathbf{r_1}, \boldsymbol{\theta_1})$ and using the definition of the pair direct correlation function (3.102) we obtain the following expression:

$$c^b(\mathbf{r_1},\boldsymbol{\theta_1}) = c^b(\mathbf{r_1},\boldsymbol{\theta_1};\lambda=0) + \sum_c \int\limits_0^1 d\lambda \int c^{cb}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2};\lambda)\Delta\rho^c(\mathbf{r_2},\boldsymbol{\theta_2})d\mathbf{r_2}d\boldsymbol{\theta_2} \tag{7.4}$$

In the HNC model we assume that the state $\lambda = 0$ corresponds to the gaseous state and that c^{ab} function does not change with lambda i.e. $c^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2};\lambda) \equiv c^{ab}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})$. As a result we can avoid integration over λ and come to the HNC expression (3.97). In fact this means that we make the transition from the gaseous phase to solution "in one jump", which is indeed rather a crude approximation. However we can perform a smoother transition from the gas to the solution. Let us smoothly (linearly) change the particle interaction potential from zero to the final one:

$$u^{ab}(\mathbf{r_1},\boldsymbol{\theta_1};\xi) = \xi \cdot u(\mathbf{r_1},\boldsymbol{\theta_1}) \tag{7.5}$$

where $\xi = 0$ corresponds to the gaseous phase, $\xi = 1$ corresponds to the solvated phase. We consider the states of the system at the points $(\xi_1,\ldots,\xi_n)$ where $\xi_k = k/n$. We define with the index k the correlation functions which correspond to the value ξ_k, e.g. c_k^{ab}, h_k^{ab}, etc. Using the equation (7.4) where $\lambda = 0$ corresponds to $\xi = \xi_k$, $\lambda = 1$ corresponds to $\xi = \xi_{k+1}$ we obtain the following relation:

$$c_{k+1}^b(\mathbf{r_1},\boldsymbol{\theta_1}) = c_k^b(\mathbf{r_1},\boldsymbol{\theta_1}) + \sum_c \int\limits_0^1 d\lambda \int c^{cb}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2};\lambda)\Delta\rho^c(\mathbf{r_2},\boldsymbol{\theta_2})d\mathbf{r_2}d\boldsymbol{\theta_2} \tag{7.6}$$

where $\Delta\rho = \rho_{k+1} - \rho_k$. We will use the approximation $c^{cb}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2};\lambda) \equiv c_{k+1}^{cb}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})$. We note that this approximation is much better than the HNC approximation because in our case the step from ξ_k to ξ_{k+1} is relatively small. Using that $\Delta\rho^c(\mathbf{r_2},\boldsymbol{\theta_2}) = \rho_{k+1}^c - \rho_k^c = \rho_0^c(h_{k+1}^{ac} - h_k^{ac})$ we obtain the following expression:

$$\begin{aligned}
&c_{k+1}^b(\mathbf{r_1},\boldsymbol{\theta_1}) = c_k^b(\mathbf{r_1},\boldsymbol{\theta_1}) + \\
&\sum_c \rho_0^c \int c_{k+1}^{cb}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})h_{k+1}^{ac}(\mathbf{r_0},\mathbf{r_2},\boldsymbol{\theta_0},\boldsymbol{\theta_2})d\mathbf{r_2}d\boldsymbol{\theta_2} - \\
&\sum_c \rho_0^c \int c_{k+1}^{cb}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})h_k^{ac}(\mathbf{r_0},\mathbf{r_2},\boldsymbol{\theta_0},\boldsymbol{\theta_2})d\mathbf{r_2}d\boldsymbol{\theta_2}
\end{aligned} \tag{7.7}$$

Substituting this expression into (7.3) and using the OZ equations for the function of the $(k+1)^{st}$ step we obtain the following closure for the OZ equations:

$$\begin{aligned}
&h_{k+1}^{ab}(\mathbf{r_0},\mathbf{r_1},\boldsymbol{\theta_0},\boldsymbol{\theta_1}) + 1 = \\
&\exp(-\beta u^{ab}(\mathbf{r_1},\boldsymbol{\theta_1};\xi_{k+1}) + c_k^b(\mathbf{r_1},\boldsymbol{\theta_1}) + h_{k+1}^{ab}(\mathbf{r_0},\mathbf{r_1},\boldsymbol{\theta_0},\boldsymbol{\theta_1}) - c_{k+1}^{ab}(\mathbf{r_0},\mathbf{r_1},\boldsymbol{\theta_0},\boldsymbol{\theta_1}))\times \\
&\times \exp(-\sum_c \rho_0^c \int c_{k+1}^{cb}(\mathbf{r_1},\mathbf{r_2},\boldsymbol{\theta_1},\boldsymbol{\theta_2})h_k^{ac}(\mathbf{r_0},\mathbf{r_2},\boldsymbol{\theta_0}\boldsymbol{\theta_2})d\mathbf{r_2}d\boldsymbol{\theta_2})
\end{aligned} \tag{7.8}$$

Moving step by step from $\xi_0 = 0$ to $\xi_n = 1$ at the $(k+1)^{st}$ step the functions from the k^{th} step are known. So the only unknown functions in the proposed closure relation are h^{ab}_{k+1}, c^{ab}_{k+1}. This means that using this closure relation together with the OZ equations it is possible to calculate the correlation functions on the $(k+1)^{st}$ step, and moving to $\xi_n = 1$ obtain the solution of the OZ equations for the fully-solvated state. We should note that this method requires much more computational expenses that the HNC method, because in our case one needs to solve the OZ equations n times. However, such a calculations should be much more accurate and there is a hope that the result or OZ calculations can be as accurate as the MD simulations are.

7.2.3 Finding of the most probable binding positions of solvent molecules

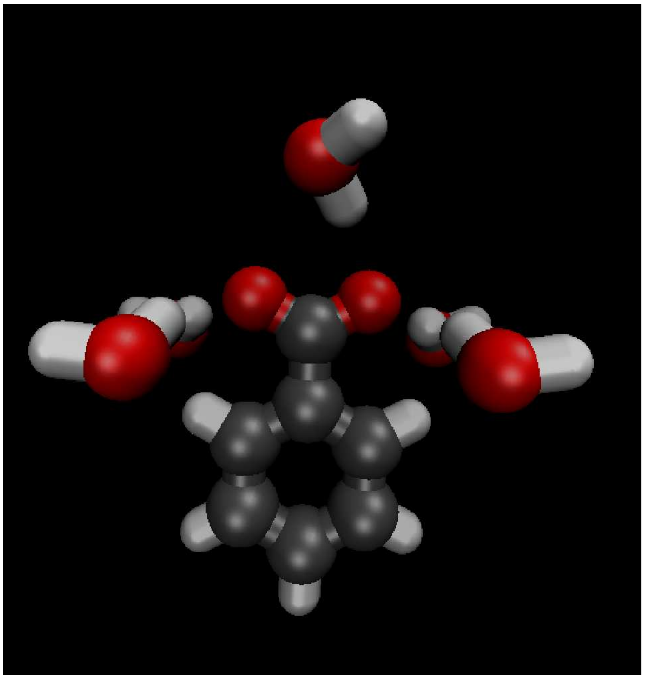


Figure 7.1: Water binding sites for the deprotonated benzoic acid in water

One of the promising applications of the IETL is finding the most probable binding site of solvent molecules. Such algorithms for simple cases was proposed in the literature [47, 151] However development in this area is still required. Probably the most straightforward way to obtain the probable binding positions is to solve OZ equation and analyze 6D correlation functions. However as it was discussed this approach requires a lot of theoretical and computational efforts. Below we describe less computationally demanding approach based on the 3DRISM calculations. We denote as $\mathbf{d}^a_\alpha(\boldsymbol{\theta})$ the displacement of the site α with respect to the center of the molecule of type a provided that it has orientation $\boldsymbol{\theta}$. Having the site distribution functions

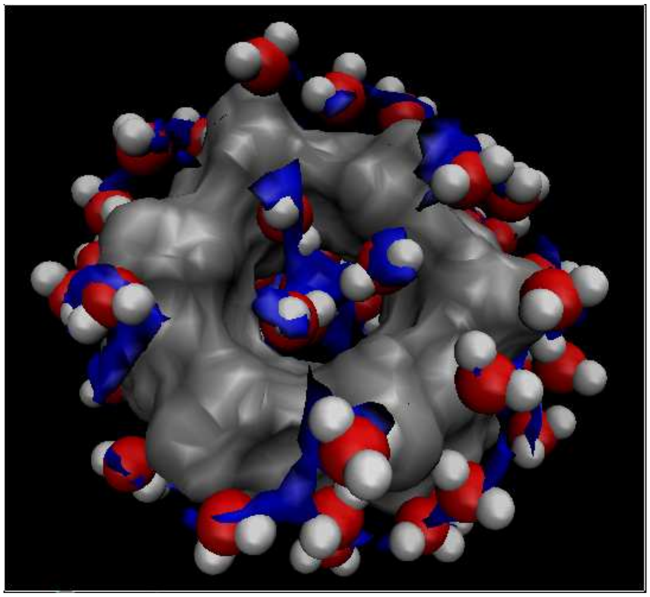


Figure 7.2: Water binding sites for the α-cyclodextrin in water

it is possible to estimate how probable is it to find the molecule in a certain position $(\mathbf{r}, \boldsymbol{\theta})$. We denote the site distribution functions as $g^a_\alpha(\mathbf{r})$. We define the function $G^a(\mathbf{r}, \boldsymbol{\theta})$ in a following way:

$$G^a(\mathbf{r}, \boldsymbol{\theta}) = \left(\prod_{\alpha}^{K_a} g^a_\alpha(\mathbf{r} + \mathbf{d}^a_\alpha(\boldsymbol{\theta})) \right)^{1/K_a} \tag{7.9}$$

where K_a is the number of sites in the molecule a. The functions g^a_α are proportional to the probability to find the site α in a certain position. The function $G^a(\mathbf{r}, \boldsymbol{\theta})$ is proportional to the probability to find the molecule a in a certain position $(\mathbf{r}, \boldsymbol{\theta})$. We define the functions G^a_{max}, $\boldsymbol{\Theta}^a_{\mathrm{max}}$ in a following way:

$$G^a_{\mathrm{max}}(\mathbf{r}) = \max_{\boldsymbol{\theta}} G^a(\mathbf{r}, \boldsymbol{\theta}) \tag{7.10}$$

$$\boldsymbol{\Theta}^a_{\mathrm{max}}(\mathbf{r}) = \mathrm{argmax}_{\boldsymbol{\theta}} G^a(\mathbf{r}, \boldsymbol{\theta}) \tag{7.11}$$

The function $\boldsymbol{\Theta}^a_{\mathrm{max}}(\mathbf{r})$ defines the most probable orientation of the molecule at the point $\mathbf{r}$ and the function $G^a_{\mathrm{max}}(\mathbf{r})$ is proportional to the maximal probability to find the molecule in the point $\mathbf{r}$. If we choose a certain threshold G_{tres} we can visualize the molecules in the positions $\mathbf{r_1}, \ldots, \mathbf{r_n}$ such that $G^a_{\mathrm{max}}(\mathbf{r_i}) > G_{\mathrm{tres}}$. The orientation of the visualized molecule is defined by the function $\boldsymbol{\Theta}^a(\mathbf{r_i})$. To achieve a better visual effect it is reasonable to visualize only one of the overlapping molecules in neighboring position, namely the molecule which have the maximum value of $G_{\mathrm{max}}(\mathbf{r})$ among all of the overlapping molecules. This method for visualizing the binding sites was implemented as a plug-in for the Visual Molecular Dynamics program [152]. This plug-in is an open-source program and is available for the online download [153]. To illustrate the possibilities of the algorithm we performed the water binding site search for the benzoic acid in a deprotonated state ($C_6H_5COO^-$). The results of the calculations are presented in Figure 7.1. We see that the hydrogen atoms of water are oriented towards the negatively charged oxygen atoms of COO^- group which matches both: the common sense and the intuition. In Figure 7.2 is presented the result of the water binding site calculations for α-cyclodextrin.

It is necessary to remember that RISM and 3DRISM models give a reasonably accurate description of small-molecule solvents. Due to this fact it is reasonable to assume that the described above method for finding binding sites also works well only for small molecular solvents. Of course for the solvents with larger molecules it is possible to use the 6D OZ equations which immediately give the $G^a(\mathbf{r},\boldsymbol{\theta})$ functions. However, there is an alternative method. We can divide a big solvent molecule into several fragments, for each of the fragments solve the 3DRISM equations and calculate the $G^a(\mathbf{r},\boldsymbol{\theta})$ functions. Let $\mathbf{D}_a(\boldsymbol{\theta})$ be the displacement of the fragment a with respect to the center of the solvent molecule in orientation $\boldsymbol{\theta}$. We define the function $G(\mathbf{r},\boldsymbol{\theta})$ in a following way:

$$G(\mathbf{r},\boldsymbol{\theta}) = \left(\prod_{a=1}^{M} G^a_\alpha(\mathbf{r}+\mathbf{D}_a(\boldsymbol{\theta}))\right)^{1/M} \tag{7.12}$$

where M is the number of fragments in the solvent molecule. The function $G(\mathbf{r},\boldsymbol{\theta})$ is proportional to the probability to find a solvent molecule in the position $(\mathbf{r},\boldsymbol{\theta})$. Similarly as it was done for $G^a(\mathbf{r},\boldsymbol{\theta})$ function we can find and visualize the most probable positions of big solvent molecules. Of course this method is less accurate than the methods based on the OZ equations. However it is much simpler from the computational point of view and we think it can be useful for practical applications.

7.3 Concluding remarks

Summarizing the current work and the future plans we can say that currently IETL is a promising method which in principle in many cases can substitute computationally expansive MD and MC simulations. In our work and in the works of other researchers it was shown that the integral equation theory can be of use for the accurate solvation free energy calculation [64–66, 73], predicting the self-assembling behavior [43] and many other applications [133, 154, 155]. On the other hand, despite its potential the theory is yet not widely used in practical applications. This can be explained by the lack of developed theoretical and computational methods. We mentioned three ways for future development of the theory and computational methods, which are: 1) development of the universal six-dimensional OZ equation solver; 2) search for new closures and bridge functionals; 3) development of new solvent binding methods. However, there are plenty other topics for investigation, which include for example the theory for confined and non-uniform liquids, non-equilibrium density functional theory, theory for flexible molecules etc. We hope that rapid development of the integral equation theory of liquids will result in a future in powerful methods which could be of use in variety of applications in computational chemistry.

Appendix A

Low rank representation of multidimensional functions

In this appendix the low-rank format for representation of multidimensional functions is described. The low-rank format is quite general and is not necessarily be used for the representation of the six-dimensional correlation functions which appear in the integral equation theory of liquids. However, in this section we discuss the complexity of those operations with the multidimensional functions in the low-rank format which are particularly relevant to the numerical solution of the system of six-dimensional MOZ equations.

Low rank format

The idea of the low-rank format is to use different angle-grid resolutions at different distances to the solute molecule. There are different ways to do that but the common idea may be illustrated on the simple example and then generalized to more complicated systems. Let us consider a simplified case when the correlation functions depend only on two variables: distance between the molecules r and angle between the molecules θ. Such system for example describes the interaction of the ball solute with the diatomic solvent. The function $f(r, \theta)$ can be represented in a low-rank format. At the large distances all the functions in the OZ and closure equations are smooth and tend to zero. This means, that at large distances one needs smaller number of the basis functions to represent the initial function $f(r, \theta)$. This can be rewritten in such a way: for each point r there is a rank $M(r)$, which decays at large distances:

$$f(r, \theta) = \sum_{m=1}^{M(r)} f_m(r)\phi_m(\theta) \tag{A.1}$$

where $\phi_m(\theta)$ are some basis functions, $M(r) \in \mathbb{N}$ and $M(r) \to 0$ when $r \to \infty$.

Operation in the low rank format

Operations which we need to perform with the functions are:

- Converting to the low-rank format
- Reconstructing the function from the low-rank format
- Multiplication of the functions
- Calculation of the convolutions

Converting a function to the low rank format

To represent a function in the format (A.1) one needs to calculate the function only in certain number of points. Let at the distance r the rank is $M(r)$. We calculate the values of the function f at the points (r, θ_n) where $n = 1 \dots M(r)$. Using representation (A.1) we have:

$$\begin{aligned} f(r,\theta_1) &= \sum_{m=1}^{M(r)} f_m(r)\phi_m(\theta_1) \\ \dots \\ f(r,\theta_{M(r)}) &= \sum_{m=1}^{M(r)} f_m(r)\phi_m(\theta_{M(r)}) \end{aligned} \tag{A.2}$$

This is the system of linear equations with respect to the unknown coefficients $f_1^r \dots f_{M(r)}^r$. The matrix of the system $\mathbf{\Phi}$ is the following:

$$\mathbf{\Phi} = \left(\begin{array}{ccc} \phi_1(\theta_1) & \dots & \phi_{M(r)}(\theta_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\theta_{M(r)}) & \dots & \phi_{M(r)}(\theta_{M(r)}) \end{array} \right) \tag{A.3}$$

We use the following definitions:

$$\mathbf{b} = (f(r,\theta_1), \dots, f(r,\theta_{M(r)}))^T, \tag{A.4}$$

$$\mathbf{f} = (f_1(r), \dots, f_{M(r)}(r))^T \tag{A.5}$$

Then the system (A.2) can be rewritten in the matrix notation:

$$\mathbf{\Phi f} = \mathbf{b} \tag{A.6}$$

The coefficients $f_1, \dots, f_{M(r)}$ can be found as

$$\mathbf{f} = \mathbf{\Phi}^{-1}\mathbf{b} \tag{A.7}$$

Because the matrix $\mathbf{\Phi}$ consists of the values of the basis functions in the certain points $\theta_1, \ldots, \theta_{M(r)}$ the inverse matrix $\mathbf{\Phi}^{-1}$ can be pre-computed before the calculations start, and thus the calculation of the coefficients $\mathbf{f}$ at the point r will need $M(r)^2$ operations. If the distance r is discretized on the grid with N grid points, the total cost of the decomposition the function $f(r, \theta)$ to the format (A.1) is $\sum_{n=1}^{N} M(r_n)^2$ operations and only $\sum_{n=1}^{N} M(r)$ function calculations.

Reconstruction of the function from the low-rank format

To reconstruct the function from the low-rank format we should be able having the representation (A.1) to calculate the value of the function at any given point (r, θ). To do this, one can simply use the formula (A.1). The computational cost of the calculation the value of the function at the point (r, θ) is $M(r)$ operations.

Multiplication in the low-rank format

Let we have two functions in the format (A.1):

$$\begin{aligned} f(r,\theta) &= \sum_{m_1=1}^{M_1(r)} f_{m_1}(r)\phi_{m_1}(\theta) \\ g(r,\theta) &= \sum_{m_2=1}^{M_2(r)} g_{m_2}(r)\phi_{m_2}(\theta) \end{aligned} \tag{A.8}$$

We may formally multiply the representations:

$$f(r,\theta)g(r,\theta) = \sum_{m_1=1}^{M_1(r)} \sum_{m_2=1}^{M_2(r)} f_{m_1}(r)g_{m_2}(r)\phi_{m_1}(\theta)\phi_{m_2}(\theta) \tag{A.9}$$

Now we need to convert the product to the format (A.1). To do this we can find the representation of the products $\phi_{m_1}(\theta)\phi_{m_2}(\theta)$ in the following form:

$$\phi_{m_1}(\theta)\phi_{m_2}(\theta) = \sum_{m_3=1}^{M_3} a_{m_1 m_2 m_3}\phi_{m_3}(\theta) \tag{A.10}$$

If such representation is known then the product (A.9) can be written as follows:

$$f(r,\theta)g(r,\theta) = \sum_{m_3=1}^{M_3} \left(\sum_{m_1=1}^{M_1(r)} \sum_{m_2=1}^{M_2(r)} a_{m_1 m_2 m_3} f_{m_1}(r) g_{m_2}(r) \right) \cdot \phi_{m_3}(\theta) \tag{A.11}$$

which gives the representation in the format (A.1):

$$f(r,\theta)g(r,\theta) = \sum_{m_3=1}^{M_3} P_{m_3}(r)\phi_{m_3}(\theta) \tag{A.12}$$

where $P_{m_3}(r) = \sum_{m_1=1}^{M_1(r)} \sum_{m_2=1}^{M_2(r)} a_{m_1 m_2 m_3} f_{m_1}(r) g_{m_2}(r)$

This method needs $M_3 \cdot M_1(r) \cdot M_2(r)$ operations to calculate the coefficients at the certain distance r, and thus $\sum_{n=1}^{N} M_3 \cdot M_1(r_n) \cdot M_2(r_n)$ operation in total to perform the multiplication.

However, there is a simpler method. We assume, that the rank of the product at the point r is fixed, and denote it $M_3(r)$. In that case we can just calculate for each r the values of the product at the points $(r, \theta_1), \ldots, (r, \theta_{M_3(r)})$ and reconstruct the coefficients using the formula (A.7).

The computational cost of calculating of the product coefficients at the distance r includes:

- calculating the values of the functions $f(r, \theta)$ and $g(r, \theta)$ and their product at the points $(r, \theta_1), \ldots, (r, \theta_{M_3(r)})$. This takes $M_3(r)(M_1(r) + M_2(r) + 1)$ operations

- reconstructing the coefficients of the product from the values at the points $(r, \theta_1), \ldots, (r, \theta_{M_3(r)})$. This takes $M_3(r)^2$ operations.

So, in total, if there are N discretization points of the grid in distance direction the computational cost is $\sum_{n=1}^{N} M_3(r)(M_1(r) + M_2(r) + M_3(r) + 1)$ operations.

Calculating a convolution

Let we have two functions in the format (A.1):

$$\begin{array}{c} f(r,\theta) = \sum_{m_1=1}^{M_1(r)} f_{m_1}(r)\phi_{m_1}(\theta) \\ g(r,\theta) = \sum_{m_2=1}^{M_2(r)} g_{m_2}(r)\phi_{m_2}(\theta) \end{array} \tag{A.13}$$

We may formally calculate the convolution of these functions:

$$\int\int f(r'-r, \theta'-\theta) g(r', \theta') dr' d\theta' = \sum_{m_1=1}^{M_1(r)} \sum_{m_2=1}^{M_2(r)} \int f_{m_1}(r'-r) g_{m_2}(r') dr' \int \phi_{m_1}(\theta'-\theta)\phi_{m_2}(\theta') d\theta' \tag{A.14}$$

Now we need to convert the convolution to the format (A.1). To do this we can find the representation of the convolutions $\int \phi_{m_1}(\theta'-\theta)\phi_{m_2}(\theta')d\theta'$ in the following form:

$$\int \phi_{m_1}(\theta'-\theta)\phi_{m_2}(\theta')d\theta' = \sum_{m_3=1}^{M_3} b_{m_1 m_2 m_3} \phi_{m_3}(\theta) \tag{A.15}$$

If such a representation is known, the convolution (A.14) can be written as follows:

$$\int f(r'-r, \theta'-\theta) g(r', \theta') dr' d\theta' = \sum_{m_3=1}^{M_3} \left(\sum_{m_1=1}^{M_1(r)} \sum_{m_2=1}^{M_2(r)} b_{m_1 m_2 m_3} \int f_{m_1}(r'-r) g_{m_2}(r') dr' \right) \cdot \phi_{m_3}(\theta) \tag{A.16}$$

which gives the representation in the format (A.1):

$$f(r,\theta)g(r,\theta)=\sum_{m_3=1}^{M_3}C_{m_3}(r)\phi_{m_3}(\theta) \tag{A.17}$$

where $C_{m_3}(r)=\sum_{m_1=1}^{M_1(r)}\sum_{m_2=1}^{M_2(r)}b_{m_1m_2m_3}\int f_{m_1}(r'-r)g_{m_2}(r')dr'$

This method needs to calculate convolutions of each pair of functions $f_{m_1}(r)$ and $g_{m_2}(r)$. Each convolution takes $const \cdot N\log(N)$ operations. Because the convolution is a non-local operation, the procedure will need $const \cdot \max(M_1(r))\max(M_2(r))N\log(N)$ operations.

Fourier low-rank representation

In the previous sections we did not make any assumptions about the basis set $\phi_1(\theta),\ldots,\phi_M(r)(\theta)$. However, by using the special basis sets one can essentially reduce the computational expanses. One of the convenient basis sets for our calculations is the Fourier basis $\phi_m(\theta)=e^{-im\theta}$. However, for this format it is more natural to let m be negative: $m\in\mathbb{Z}$ So, finally the Fourier analogue of the format (A.1) is the following:

$$f(r,\theta)=\sum_{-M(r)/2}^{M(r)/2}f_m(r)e^{-im\theta} \tag{A.18}$$

The format is convenient by two reasons:

- Firstly, at large distances the angular component of the functions in the Ornstein-Zernike equation can be approximated by the sin/cos functions with the high accuracy.

- Secondly, the Fourier basis essentially simplifies the calculations and allows us in many cases to use the efficient FFT algorithm.

Below we will see how the Fourier basis can simplify the basic operations.

Converting a function to the Fourier low rank format

For each distance r we may write the analogue of the equations (A.2). In our case it is convenient to choose the equidistant values of θ_k: $\theta_k = 2\pi k/M(r)$. So, we have:

$$\begin{aligned} f(r,\theta_1) &= \sum_{m=-M(r)/2}^{M(r)/2} f_m(r) e^{-2\pi i m/M(r)} \\ \ldots \\ f(r,\theta_k) &= \sum_{m=-M(r)/2}^{M(r)/2} f_m(r) e^{-2\pi i m k/M(r)} \\ \ldots \\ f(r,\theta_{M(r)}) &= \sum_{m=-M(r)/2}^{M(r)/2} f_m(r) e^{-2\pi i m} \end{aligned} \tag{A.19}$$

We see that the expressions in the right hand sides of the equations form the Fourier series. The coefficients $f_1(r), \ldots, f_{M(r)}(r)$ can be found using the Discrete Fourier Transform and FFT algorithm. This will take $const \cdot M(r) \log M(r)$ operations. In general, if we have N discretization points in the distance direction, converting to the low rank format will take $const \cdot \sum_{n=1}^{N} M(r_n) \log M(r_n)$

Reconstruction values of the function from the Fourier format

If we need to reconstruct the value of the function in the single point (r,θ) we can use the formula (A.18), and this will need $M(r)$ operations. However, if we need to calculate the values of the function at the certain distance r and at the range of angles $\theta_1, \ldots, \theta_{M(r)}$ where $\theta_k = 2\pi k/M(r)$ we may use the FFT algorithm, which will take $M(r) \log M(r)$ operations. So in general to reconstruct all the values of the function in the grid points we will need $const \cdot \sum_{n=1}^{N} M(r_n) \log M(r_n)$ operations.

Multiplication of the functions in the Fourier format

The analogue of the expression (A.9) in the Fourier format is:

$$f(r,\theta) \cdot g(r,\theta) = \sum_{m_1=-M_1(r)/2}^{M_1(r)/2} \sum_{m_2=-M_2(r)/2}^{M_2(r)/2} f_{m_1}(r) g_{m_1}(r) e^{-i m_1 \theta} e^{-i m_2 \theta} \tag{A.20}$$

Using that $e^{-im_1\theta} e^{-im_2\theta} = e^{-i(m_1+m_2)\theta}$ we may introduce the variable $m_3 = m_1 + m_2$. In such a definition the product is rewritten as:

$$f(r,\theta) \cdot g(r,\theta) = \sum_{m_3=-M_3(r)/2}^{M_3(r)/2} \left(\sum_{m_1=-M_1(r)/2}^{M_1(r)/2} f_{m_1}(r) g_{m_3-m_1}(r) \right) e^{-i m_3 \theta} \tag{A.21}$$

where $M_3(r) = M_1(r) + M_2(r)$. The expression (A.21) gives the representation in the format (A.18)

$$f(r, \theta) \cdot g(r, \theta) = P_{m_3}(r) e^{-im_3\theta} \tag{A.22}$$

where $P_{m_3}(r) = \sum_{m_1=-M_1(r)/2}^{M_1(r)/2} f_{m_1}(r) g_{m_3-m_1}(r)$ is the discrete convolution of the coefficients f_{m_1} and g_{m_2}. This convolution can be calculated using the FFT algorithm, which will take $const \cdot M_3(r) \log M_3(r)$ operations. In general, the calculation of the product will take $const \cdot \sum_{n=1}^{N} M_3(r_n) \log M_3(r_n)$ operations.

Calculation of the convolution in the Fourier format

The analogue of the formula (A.14) in the Fourier format is written as follows:

$$\begin{gathered}\int\limits_{-\infty}^{\infty}\int\limits_{-\pi}^{\pi} f(r'-r, \theta'-\theta) g(r', \theta') dr' d\theta' = \\ \sum_{m_1=-M_1(r)/2}^{M_1(r)/2} \sum_{m_2=-M_2(r)/2}^{M_2(r)/2} \int\limits_{-\infty}^{\infty} f_{m_1}(r'-r) g_{m_2}(r') dr' \int\limits_{-\pi}^{\pi} e^{-im_1(\theta'-\theta)} e^{-im_2\theta'} d\theta'\end{gathered} \tag{A.23}$$

Let us find the convolution of the basis functions:

$$\int\limits_{-\pi}^{\pi} e^{-im_1(\theta'-\theta)} e^{-im_2\theta'} d\theta' = e^{im_1\theta} \int\limits_{-\pi}^{\pi} e^{-i(m_1+m_2)\theta'} d\theta' \tag{A.24}$$

Because the integral is taken over the full period of the function $e^{-i(m_1+m_2)\theta'}$, the integral is non zero only when $m_1 + m_2 = 0$. This can be compactly written using the Kronecker delta function:

$$\int\limits_{-\pi}^{\pi} e^{-im_1(\theta'-\theta)} e^{-im_2\theta'} d\theta' = 2\pi e^{im_1\theta} \delta_{m_1,-m_2} \tag{A.25}$$

Putting this expression to the (A.23) and using the properties of the delta function we have:

$$\int\limits_{-\infty}^{\infty}\int\limits_{-\pi}^{\pi} f(r'-r, \theta'-\theta) g(r', \theta') dr' d\theta' = \sum_{m_2=-M_3(r)/2}^{M_3(r)/2} \left(2\pi \int\limits_{-\infty}^{\infty} f_{-m_2}(r'-r) g_{m_2}(r') dr' \right) e^{-im_2\theta} \tag{A.26}$$

where $M_3(r) = \min(M_1(r), M_2(r))$ This expression gives a representation in the format (A.18):

$$\int\limits_{-\infty}^{\infty}\int\limits_{-\pi}^{\pi} f(r'-r, \theta'-\theta) g(r', \theta') dr' d\theta' = \sum_{m_3=-M_3(r)/2}^{M_3(r)/2} C_{m_3}(r) e^{-im_3\theta} \tag{A.27}$$

where $C_{m_3}(r) = 2\pi \int\limits_{-\infty}^{\infty} f_{-m_2}(r'-r) g_{m_2}(r') dr'$. Because the convolution in distance direction is a non-local operation, we need to pre-compute the coefficients $C_{m_3}(r)$ and then use them at any distance r. However, we need to calculate the convolutions only of the selected pairs of the functions $f_{m_1}(r)$ and $g_{m_2}(r)$, namely of such of them that $m_1 = -m_2$. This will give only $M_3 = \min(\max_r M_1(r), \max_r M_2(r))$ convolution calculations. Moreover, because the ranks $M_1(r)$ and $M_2(r)$ decay at the large distances, that means that the support of the functions $f_{m_1}(r)$ and $g_{m_2}(r)$ will decay with the growth of the coefficients m_1 and m_2. And the functions with the smaller support need smaller number of discretization points. We denote $N(f(r))$ the number of points which is needed to discretize the function $f(r)$. Then the total computational cost to calculate the convolution is $const \cdot \sum_{m3=-M_3/2}^{M_3/2} N_{m3} \log N_{m3}$, where $N_{m3} = N(f_{-m_3}(r)) + N(g_{m_3}(r))$

Appendix B

List of peer-reviewed publications resulted from this PhD project

(**P1**) *Sergiievskyi, V.P. Fedorov, M. V.* , “3DRISM Multi-grid Algorithm for Fast Solvation Free Energy Calculations” , J. Chem. Theor. Comput. 2012, 8, pp 2062-2070

(**P2**) *Sergiievskyi, V.P.; Hackbusch, W.; Fedorov, M.V.*, “Multigrid Solver for the Reference Interaction Site Model of Molecular Liquids Theory”, J. Comput. Chem., 2011, 32, pp 1982-1992

(**P3**) *Sergiievskyi, V.P.*, “Model for Calculating the Free Energy of Hydration of Bioactive Compounds Based on Integral Equations of the Theory of Liquids”, Russian J. Phys. Chem. B, 2011, 5, pp. 326-331.

(**P4**) *Sergievskii, V. P.; Frolov, A. I.*, “A universal bridge functional for infinitely diluted solutions: A case study for Lennard-Jones spheres of different diameters”, Russian J. Phys. Chem. A., 2012, 8, pp. 1254-1260

(**P5**) *Ratkova, E. L.; Chuev, G. N.; Sergiievskyi, V. P.; Fedorov, M. V.*, “An Accurate Prediction of Hydration Free Energies by Combination of Molecular Integral Equations Theory with Structural Descriptors”, J. Phys. Chem. B, 2010, 114, pp 12068-12079

(**P6**) *Palmer, D. S.; Sergiievskyi, V. P.; Jensen, F.; Fedorov, M. V.* “Accurate calculations of the hydration free energies of drug-like molecules using the reference interaction site model”, J. Chem. Phys., 2010, 133, 044104

(**P7**) *Kolombet V.A.; Sergievskii V.P.* “The Special Features of the Thermodynamic Characteristics of Hydration of Univalent Ions According to the Reference Interaction Site Model”, Russian J. Phys. Chem. A, 2010, 84, pp. 1467-1472

List of oral and poster conference presentations

(**C1**) *Sergiievskyi V.P., Fedorov M.V.* “Fast RISM and 3DRISM algorithms for biochemical applications”. European Molecular Liquids Group (EMLG) Meeting, Eger, Hungary, September 2012

(**C2**) *Sergiievskyi V.P.* “Bridge functionals in Ornstein-Zernike and RISM equations”. DUEL Workshop, Leipzig, Germany, 8 October 2011

(**C3**) *Volodymyr P. Sergiievskyi, Ekaterina L. Ratkova, Andrey I. Frolov, David S. Palmer and Maxim V. Fedorov*, “Fast Multi-grid Solvers of RISM and 3D-RISM Equations for Accurate Calculations of Hydration Free Energy”, European Molecular Liquids Group (EMLG) Meeting, Warsaw, Poland, September 2011

(**C4**) *Sergiievskyi V.P.* “Fast multi-grid solver for Hydration Free Energy calculations”, Physical Chemical Aspect of Biomolecular Solvation, Leipzig, Germany, 23-24 May 2011

(**C5**) *Sergiievskyi V.P.; Fedorov M.V.* “Ornstein-Zernike and RISM equations and their numerical solution”, DUEL workshop, Hagen, Germany, 1 Oct 2010

(**C6**) *Volodymyr P. Sergiievskyi, W. Hackbusch, Maxim V. Fedorov* “Fast Milti-Grid Algorithm for Prediction Hydration Free Energies of the Drug-Like Molecules”, European Molecular Liquids Group (EMLG) Meeting, Lviv, Ukraine, 5-9 Sep 2010

(**C7**) *Sergiievskyi V.P.; Fedorov M.V.* “Application of multigrid techniques to the reference interaction site model of molecular liquids”, Physics of Liquid Matter: Modern Problems (PLMMP), Kyiv, Ukraine, 21-24 May 2010

Bibliography

[1] C. A. Reynolds, P. M. King, and W. G. Richards. Free-energy calculations in molecular biophysics. *Molecular Physics*, 76(2):251–275, June 1992.

[2] P. Kollman. Free-energy calculations - applications to chemical and biochemical phenomena. *Chemical Reviews*, 93(7):2395–2417, November 1993.

[3] G.L. Perlovich and A. Bauer-Brandl. Solvation of drugs as a key for understanding partitioning and passive transport exemplified by nsaids. *Current Drug Delivery*, 1(3):213–226, July 2004.

[4] G. L. Perlovich, T. V. Volkova, and A. Bauer-Brandl. Towards an understanding of the molecular mechanism of solvation of drug molecules: A thermodynamic approach by crystal lattice energy, sublimation, and solubility exemplified by paracetamol, acetanilide, and phenacetin. *Journal of Pharmaceutical Sciences*, 95(10):2158–2169, October 2006.

[5] G. L. Perlovich, L. K. Hansen, T. V. Volkova, S. Mirza, A. N. Manin, and A. Bauer-Brandl. Thermodynamic and structural aspects of hydrated and unhydrated phases of 4-hydroxybenzamide. *Crystal Growth & Design*, 7(12):2643–2648, December 2007.

[6] L. D. Hughes, D. S. Palmer, F. Nigsch, and J. B. O. Mitchell. Why are some properties more difficult to predict than others? a study of qspr models of solubility, melting point, and log p. *Journal of Chemical Information and Modeling*, 48(1):220–232, January 2008.

[7] D. S. Palmer, A. Llinas, I. Morao, G. M. Day, J. M. Goodman, R. C. Glen, and J. B. O. Mitchell. Predicting intrinsic aqueous solubility by a thermodynamic cycle. *Molecular Pharmaceutics*, 5(2):266–279, March 2008.

[8] W. L. Jorgensen and J. TiradoRives. Free energies of hydration for organic molecules from monte carlo simulations. *Perspectives in Drug Discovery and Design*, 3:123–138, 1995.

[9] N. Matubayasi and M. Nakahara. Theory of solutions in the energetic representation. i. formulation. *Journal of Chemical Physics*, 113(15):6070–6081, October 2000.

[10] N. Matubayasi and M. Nakahara. An approach to the solvation free energy in terms of the distribution functions of the solute-solvent interaction energy. *Journal of Molecular Liquids*, 119(1-3):23–29, May 2005.

[11] M. R. Shirts and V. S. Pande. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *Journal of Chemical Physics*, 122(13):134508, April 2005.

[12] N. Matubayasi. Free-energy analysis of solvation with the method of energy representation. *Frontiers in Bioscience*, 14:3536–3549, January 2009.

[13] J. L. Knight and C. L. Brooks. lambda-dynamics free energy simulation methods. *Journal of Computational Chemistry*, 30(11):1692–1700, August 2009.

[14] J. Tomasi and M. Persico. Molecular-interactions in solution - an overview of methods based on continuous distributions of the solvent. *Chemical Reviews*, 94(7):2027–2094, November 1994.

[15] B. Roux and T. Simonson. Implicit solvent models. *Biophysical Chemistry*, 78(1-2):1–20, April 1999.

[16] D. Bashford and D. A. Case. Generalized born models of macromolecular solvation effects. *Annual Review of Physical Chemistry*, 51:129–152, 2000.

[17] J. Tomasi, B. Mennucci, and R. Cammi. Quantum mechanical continuum solvation models. *Chemical Reviews*, 105(8):2999–3094, August 2005.

[18] M. B. Ulmschneider, J. P. Ulmschneider, M. S. P. Sansom, and A. Di Nola. A generalized born implicit-membrane representation compared to experimental insertion free energies. *Biophysical Journal*, 92(7):2338–2349, April 2007.

[19] A. V. Marenich, C. J. Cramer, and D. G. Truhlar. Performance of sm6, sm8, and smd on the sampl1 test set for the prediction of small-molecule solvation free energies. *Journal of Physical Chemistry B*, 113(14):4538–4543, March 2009.

[20] A. Klamt, F. Eckert, and M. Diedenhofen. Prediction of the free energy of hydration of a challenging set of pesticide-like compounds. *Journal of Physical Chemistry B*, 113(14):4508–4510, March 2009.

[21] T. Sulea, D. Wanapun, S. Dennis, and E. O. Purisima. Prediction of sampl-1 hydration free energies using a continuum electrostatics-dispersion model. *Journal of Physical Chemistry B*, 113(14):4511–4520, March 2009.

[22] P. A. Monson and G. P. Morriss. Recent progress in the statistical-mechanics of interaction site fluids. *Advances in Chemical Physics*, 77:451–550, 1990.

[23] J.-P. Hansen and I. R. McDonald. *Theory of Simple Liquids, 3rd ed.* Academic Press, London, 1991.

[24] F. Hirata, editor. *Molecular theory of solvation.* Kluwer Academic Publishers, Dordrecht, Netherlands, 2003.

[25] D. Frenkel and B. Smit. *Understanding molecular simulation.* Academic Press, 2002.

[26] M. V. Fedorov, J. M. Goodman, and S. Schumm. The effect of sodium chloride on poly-l-glutamate conformation. *Chemical Communications*, (8):896–898, 2009.

[27] M. V. Fedorov, J. M. Goodman, and S. Schumm. To switch or not to switch: The effects of potassium and sodium ions on alpha-poly-l-glutamate conformations in aqueous solutions. *Journal of the American Chemical Society*, 131(31):10854–10856, August 2009.

[28] Benedetta Mennucci and Roberto Cammi, editors. *Continuum Solvation Models in Chemical Physics.* Wiley, Chippenham, Wiltshire, UK, 2007.

[29] L. Blum and A. J. Torruella. Invariant expansion for two-body correlations: Thermodynamic functions, scattering, and the Ornstein-Zernike equation. *Journal of Chemical Physics*, 56(1):303–310, 1972.

[30] K. Amano and M. Kinoshita. Entropic insertion of a big sphere into a cylindrical vessel. *Chemical Physics Letters*, 488(1-3):1–6, March 2010.

[31] D. Chandler and H. C. Andersen. Optimized cluster expansions for classical fluids. 2. Theory of molecular liquids. *Journal of Chemical Physics*, 57(5):1930–1937, 1972.

[32] F. Hirata, B. M. Pettitt, and P. J. Rossky. Application of an extended RISM equation to dipolar and quadrupolar fluids. *Journal of Chemical Physics*, 77(1):509–520, 1982.

[33] B. M. Pettitt and P. J. Rossky. Integral-equation predictions of liquid-state structure for waterlike intermolecular potentials. *Journal of Chemical Physics*, 77(3):1451–1457, 1982.

[34] M. Kinoshita, Y. Okamoto, and F. Hirata. Calculation of solvation free energy using RISM theory for peptide in salt solution. *Journal of Computational Chemistry*, 19(15):1724–1735, November 1998.

[35] M. Kinoshita, Y. Okamoto, and F. Hirata. First-principle determination of peptide conformations in solvents: Combination of monte carlo simulated annealing and rism theory. *Journal of the American Chemical Society*, 120(8):1855–1863, March 1998.

[36] M. Kinoshita, Y. Okamoto, and F. Hirata. Analysis on conformational stability of c-peptide of ribonuclease a in water using the reference interaction site model theory and monte carlo simulated annealing. *Journal of Chemical Physics*, 110(8):4090–4100, February 1999.

[37] T. Imai, M. Kinoshita, and F. Hirata. Salt effect on stability and solvation structure of peptide: An integral equation study. *Bulletin of the Chemical Society of Japan*, 73(5):1113–1122, May 2000.

[38] T. Imai, R. Hiraoka, A. Kovalenko, and F. Hirata. Water molecules in a protein cavity detected by a statistical-mechanical theory. *Journal of the American Chemical Society*, 127(44):15334–15335, November 2005.

[39] N. Yoshida, S. Phongphanphanee, Y. Maruyama, T. Imai, and F. Hirata. Selective ion-binding by protein probed with the 3d-rism theory. *Journal of the American Chemical Society*, 128(37):12042–12043, September 2006.

[40] N. Yoshida, S. Phongphanphanee, and F. Hirata. Selective Ion Binding by Protein Probed with the Statistical Mechanical Integral Equation Theory. *Journal of Physical Chemistry B*, 111(17):4588–4595, 2007.

[41] G.N. Chuev, M.V. Fedorov, and J. Crain. Improved estimates for hydration free energy obtained by the reference interaction site model. *Chemical Physics Letters*, 448(4-6):198–202, 2007.

[42] M. V. Fedorov and A. A. Kornyshev. Unravelling the solvent response to neutral and charged solutes. *Molecular Physics*, 105(1):1–16, January 2007.

[43] G. N. Chuev and M. V. Fedorov. Reference interaction site model study of self-aggregating cyanine dyes. *The Journal of Chemical Physics*, 131:074503, 2009.

[44] T. Imai, Y. Harano, M. Kinoshita, A. Kovalenko, and F. Hirata. Theoretical analysis on changes in thermodynamic quantities upon protein folding: Essential role of hydration. *Journal of Chemical Physics*, 126(22):225102, June 2007.

[45] T. Imai, S. Ohyama, A. Kovalenko, and F. Hirata. Theoretical study of the partial molar volume change associated with the pressure-induced structural transition of ubiquitin. *Protein Science*, 16(9):1927–1933, September 2007.

[46] D. Yokogawa, H. Sato, T. Imai, and S. Sakaki. A highly parallelizable integral equation theory for three dimensional solvent distribution function: Application to biomolecules. *Journal of Chemical Physics*, 130(6):064111, February 2009.

[47] T. Imai, K. Oda, A. Kovalenko, F. Hirata, and A. Kidera. Ligand mapping on protein surfaces by the 3d-rism theory: Toward computational fragment-based drug design. *Journal of the American Chemical Society*, 131(34):12430–12440, September 2009.

[48] Y. Kiyota, R. Hiraoka, N. Yoshida, Y. Maruyama, I. Imai, and F. Hirata. Theoretical study of co escaping pathway in myoglobin with the 3d-rism theory. *Journal of the American Chemical Society*, 131(11):3852–3853, March 2009.

[49] K. Nishiyama, T. Yamaguchi, and F. Hirata. Solvation dynamics in polar solvents studied by means of rism/mode-coupling theory. *Journal of Physical Chemistry B*, 113(9):2800–2804, March 2009.

[50] J.-P. Hansen and I. R. McDonald. *Theory of Simple Liquids, 4th ed.* Elsevier Academic Press, Amsterdam, The Netherlands, 2000.

[51] E. Meeron. Series expansion of distribution functions in multicomponent fluid systems. *Journal of Chemical Physics*, 27(6):1238–1246, 1957.

[52] S. J. Singer and D. Chandler. Free-energy functions in the extended rism approximation. *Molecular Physics*, 55(3):621–625, 1985.

[53] L. Gendre, R. Ramirez, and D. Borgis. Classical density functional theory of solvation in molecular solvents: Angular grid implementation. *Chemical Physics Letters*, 474(4-6):366–370, June 2009.

[54] R. Ramirez, R. Gebauer, M. Mareschal, and D. Borgis. Density functional theory of solvation in a polar solvent: Extracting the functional from homogeneous solvent simulations. *Physical Review E*, 66(3):031206, September 2002.

[55] T. Urbic, V. Vlachy, Y. V. Kalyuzhnyi, and K. A. Dill. Orientation-dependent integral equation theory for a two-dimensional model of water. *Journal of Chemical Physics*, 118(12):5516–5525, March 2003.

[56] M. V. Fedorov, H. J. Flad, G. N. Chuev, L. Grasedyck, and B. N. Khoromskij. A structured low-rank wavelet solver for the Ornstein-Zernike integral equation. *Computing*, 80(1):47–73, May 2007.

[57] S. Ten-no, F. Hirata, and S. Kato. A hybrid approach for the solvent effect on the electronic-structure of a solute based on the rism and hartree-fock equations. *Chemical Physics Letters*, 214(3-4):391–396, November 1993.

[58] S. Ten-no, F. Hirata, and S. Kato. Reference interaction site model self-consistent-field study for solvation effect on carbonyl-compounds in aqueous-solution. *Journal of Chemical Physics*, 100(10):7443–7453, May 1994.

[59] H. Sato, F. Hirata, and S. Kato. Analytical energy gradient for the reference interaction site model multiconfigurational self-consistent-field method: Application to 1,2-difluoroethylene in aqueous solution. *Journal of Chemical Physics*, 105(4):1546–1551, July 1996.

[60] D. Yokogawa, H. Sato, and S. Sakaki. New generation of the reference interaction site model self-consistent field method: Introduction of spatial electron density distribution to the solvation theory. *Journal of Chemical Physics*, 126(24):244504, June 2007.

[61] D. Yokogawa, H. Sato, and S. Sakaki. The position of water molecules in bacteriorhodopsin: A three-dimensional distribution function study. *Journal of Molecular Liquids*, 147(1-2):112–116, July 2009.

[62] S. Ten-no. Free energy of solvation for the reference interaction site model: Critical comparison of expressions. *Journal of Chemical Physics*, 115(8):3724–3731, August 2001.

[63] Y. Karino, M. V. Fedorov, and N. Matubayasi. End-point calculation of solvation free energy of amino-acid analogs by molecular theories of solution. *Chemical Physics Letters*, 496(4-6):351–355, August 2010.

[64] E. L. Ratkova, G. N. Chuev, V. P. Sergiievskyi, and M. V. Fedorov. An accurate prediction of hydration free energies by combination of molecular integral equations theory with structural descriptors. *Journal of Physical Chemistry B*, 114(37):12068–12079, 2010.

[65] E. L. Ratkova and M. V. Fedorov. Combination of rism and cheminformatics for efficient predictions of hydration free energy of polyfragment molecules: Application to a set of organic pollutants. *Journal of Chemical Theory and Computation*, 7(5):1450–1457, 2011.

[66] D. S. Palmer, V. P. Sergiievskyi, F. Jensen, and M. V. Fedorov. Accurate calculations of the hydration free energies of druglike molecules using the reference interaction site model. *Journal of Chemical Physics*, 133(4):044104, July 2010.

[67] V. P. Sergiievskyi. Model for calculating the free energy of hydration of bioactive compounds based on integral equations of the theory of liquids. *Russian Journal of Physical Chemistry B*, 5(2):326–331, 2011.

[68] D. Chandler, J. D. Mccoy, and S. J. Singer. Density functional theory of nonuniform polyatomic systems. 1. General formulation. *Journal of Chemical Physics*, 85(10):5971–5976, 1986.

[69] D. Beglov and B. Roux. Numerical-solution of the hypernetted-chain equation for a solute of arbitrary geometry in 3 dimensions. *Journal of Chemical Physics*, 103(1):360–364, July 1995.

[70] T. Imai. Molecular theory of partial molar volume and its applications to biomolecular systems. *Condensed Matter Physics*, 10(3):343–361, 2007.

[71] T. Luchko, S. Gusarov, D. R. Roe, C. Simmerling, D. A. Case, J. Tuszynski, and A. Kovalenko. Three-dimensional molecular theory of solvation coupled with molecular dynamics in amber. *Journal of Chemical Theory and Computation*, 6(3):607–624, March 2010.

[72] M. C. Stumpe, N. Blinov, D. Wishart, A. Kovalenko, and V. S. Pande. Calculation of local water densities in biological systems: A comparison of molecular dynamics simulations and the 3d-rism-kh molecular theory of solvation. *Journal of Physical Chemistry B*, 115(2):319–328, January 2011.

[73] D. S. Palmer, G. N. Chuev, E. L. Ratkova, and M. V. Fedorov. In silico screening of bioactive and biomimetic solutes by integral equation theory. *Current Pharmaceutical Design*, 17(17):1695–1708, 2011.

[74] A. I. Frolov, E. L. Ratkova, D. S. Palmer, and M. V. Fedorov. Hydration thermodynamics using the reference interaction site model: Speed or accuracy? *The Journal of Physical Chemistry B*, 115(19):6011–6022, 2011.

[75] G. Zerah. An efficient newtons method for the numerical-solution of fluid integral-equations. *Journal of Computational Physics*, 61(2):280–285, 1985.

[76] Michael J. Booth, A.G. Schlijper, L.E. Scales, and A.D.J. Haymet. Efficient solution of liquid state integral equations using the newton–gmres algorithm. *Computer Physics Communications*, 119:122–134, 1999.

[77] A. Kovalenko, S. Ten-No, and F. Hirata. Solution of three-dimensional reference interaction site model and hypernetted chain equations for simple point charge water by modified method of direct inversion in iterative subspace. *Journal of Computational Chemistry*, 20(9):928–936, July 1999.

[78] M. Kawata, C. M. Cortis, and R. A. Friesner. Efficient recursive implementation of the modified broyden method and the direct inversion in the iterative subspace method: Acceleration of self-consistent calculations. *Journal of Chemical Physics*, 108(11):4426–4438, March 1998.

[79] H.H.H. Homeier, S. Rast, and H. Krienke. Iterative solution of the Ornstein-Zernike equation with various closures using vector extrapolation. *Computer Physics Communications*, 92(2-3):188–202, 1995.

[80] M. J. Gillan. New method of solving the liquid structure integral-equations. *Molecular Physics*, 38(6):1781–1794, 1979.

[81] S. Labik, A. Malijevsky, and P. Vonka. A rapidly convergent method of solving the Ornstein-Zernike equation. *Molecular Physics*, 56(3):709–715, 1985.

[82] S. Woelki, H.H. Kohler, H. Krienke, and G. Schmeer. Improvements of DRISM calculations: symmetry reduction and hybrid algorithms. *Physical Chemistry Chemical Physics*, 10(6):898–910, 2008.

[83] G. N. Chuev and M. V. Fedorov. Wavelet algorithm for solving integral equations of molecular liquids. a test for the reference interaction site model. *Journal of Computational Chemistry*, 25(11):1369–1377, August 2004.

[84] G. N. Chuev and M. V. Fedorov. Wavelet treatment of structure and thermodynamics of simple liquids. *Journal of Chemical Physics*, 120(3):1191–1196, January 2004.

[85] M. V. Fedorov and G. N. Chuev. Wavelet method for solving integral equations of simple liquids. *Journal of Molecular Liquids*, 120(1–3):159–162, June 2005.

[86] C. T. Kelley and B. M. Pettitt. A fast solver for the Ornstein-Zernike equations. *Journal of Computational Physics*, 197(2):491–501, JUL 2004.

[87] W. Hackbusch. *Multi-grid methods and Applications*. Springer-Verlag, Berlin, 1985.

[88] H. Y. Wang, W. Jiang, and Y. N. Wang. Implicit and electrostatic particle-in-cell/monte carlo model in two-dimensional and axisymmetric geometry: I. analysis of numerical techniques. *Plasma Sources Science & Technology*, 19(4):045023, August 2010.

[89] M. Heiskanen, T. Torsti, M.J. Puska, and R.M. Nieminen. Multigrid method for electronic structure calculations. *Physical Review B*, 63(24):245106, JUN 15 2001.

[90] W. Janke and T. Sauer. Multicanonical multigrid Monte-Carlo method. *Physical Review E*, 49(4, Part B):3475–3479, APR 1994.

[91] F. Gygi and G. Galli. Real-space adaptive-coordinate electronic-structure calculations. *Physical Review B*, 52(4):R2229–R2232, 1995.

[92] M. V. Fedorov and W. Hackbusch. A multigrid solver for the integral equations of the theory of liquids. Preprint 88, Max-Planck-Institut fuer Mathematik in den Naturwissenschaften, 2008.

[93] J. P. Guthrie. A blind challenge for computational solvation free energies: Introduction and overview. *Journal of Physical Chemistry B*, 113(14):4501–4507, April 2009.

[94] L. D. Landau and E. M. Lifshitz. *Statistical Physics, Third Edition, Part 1: Volume 5 of Course of Theoretical Physics*, volume 5. Butterworth-Heinemann, 3 edition, January 1980.

[95] V.I. Kalikmanov. *Statistical physics of fluids: basic concepts and applications.* Berlin; New York: Springer, 2001.

[96] L. D. Landau and E. M. Lifshitz. *Mechanics, Third Edition: Volume 1 of Course of Theoretical Physics*, volume 1. Butterworth-Heinemann, 3 edition, 1976.

[97] David E. Shaw, Ron O. Dror, John K. Salmon, J.P. Grossman, Kenneth M. Mackenzie, Joseph A. Bank, Cliff Young, Brannon Batson Martin M. Deneroff, Kevin J. Bowers, Edmond Chow, Michael P. Eastwood, Douglas J. Ierardi, John L. Klepeis, Jeffrey S. Kuskin, Richard H. Larson, Kresten Lindorff-Larsen, Paul Maragakis, Mark A. Moraes, Stefano Piana, Yibing Shan, and Brian Towles. Millisecond-scale molecular dynamics simulations on anton. *Proceedings of the ACM/IEEE Conference on Supercomputing (SC09) (Portland, Oregon: ACM): 1–11. doi:10.1145/1654059.1654099*, 2009.

[98] D.C. Rapaport. Multibillion-atom molecular dynamics simulation: Design considerations for vector-parallel processing. *Computer Physics Communications*, 174:521–529, 2006.

[99] Kai Kadau, Timothy C. Germann, and Peter S. Lomdahl. Molecular dynamics comes of age:. 320 billion atom simulation on bluegene/l. *International Journal of Modern Physics C*, 17(12):1755–1761, 2006.

[100] R.K. Pathria and Paul D. Beale. *Statistical mechanics. 3rd. expanded ed.* Boston, MA: Academic Press. 744 p. , 2011.

[101] A letter from the late reverent mr. thomas bayes f.r.s to john canton, m.a. abd f.r.s, http://www.york.ac.uk/depts/maths/histstat/letter.pdf.

[102] Lagrange multipliers. encyclopedia of mathematics., URL: http://www.encyclopediaofmath.org/index.php?title=Lagrange_multipliers.

[103] Eric W. Weisstein. Gaussian integral. from mathworld–a wolfram web resource. http://mathworld.wolfram.com/gaussianintegral.html.

[104] V. M. Alekseev, V. M Tikhomirov, and S. V. Fomin. *Optimal Control*. New York : Consultants Bureau, 1987.

[105] A. Ben-Naim. *Molecular Theory of Solutions*. Oxford University Press, USA, 2006.

[106] J. K. Percus and G. J. Yevick. Analysis of classical statistical mechanics by means of collective coordinates. *Physical Review*, 110(1):1–13, 1958.

[107] G. A. Martynov and G. N. Sarkisov. Exact equations and the theory of liquids .5. *Molecular Physics*, 49(6):1495–1504, 1983.

[108] S. Labik, A. Malijevsky, and W. R. Smith. An accurate integral-equation for molecular fluids .2. hard heteronuclear diatomics. *Molecular Physics*, 73(3):495–502, June 1991.

[109] H. C. Andersen. The structure of liquids. *Annual Review of Physical Chemistry*, 26:145–166, 1975.

[110] V. N. Bondarev. Ising-like criticality derived from the theory of fluids. *Physical Review E*, 77(5):050103, May 2008.

[111] V. P. Sergievskii and A. I. Frolov. Universal bridge functional for infinitely diluted solutions: a case study for lennard-jones spheres of different diameter. *Russian Journal of Physical Chemistry A*, 86(8):1254–1260, 2012. preprint: arXiv:1111.2257v1.

[112] S. Palmer, A. I. Frolov, E. L. Ratkova, and M. V. Fedorov. Towards a universal method to calculate hydration free energies: a 3d reference interaction site model with partial molar volume correction. *Journal of Physics: Condensed Matter*, 22(49):492101, 2010.

[113] J. G. Kirkwood. Statistical mechanics of fluid mixtures. *Journal of Chemical Physics*, 3:300–313, 1935.

[114] S. Ten-no and S. Iwata. On the connection between the reference interaction site model integral equation theory and the partial wave expansion of the molecular Ornstein-Zernike equation. *Journal of Chemical Physics*, 111(11):4865–4868, 1999.

[115] A. Kovalenko and F. Hirata. Hydration free energy of hydrophobic solutes studied by a reference interaction site model with a repulsive bridge correction and a thermodynamic perturbation method. *Journal of Chemical Physics*, 113:2793–2805, 2000.

[116] D. Chandler, Y. Singh, and D. M. Richardson. Excess electrons in simple fluids .1. general equilibrium-theory for classical hard-sphere solvents. *Journal of Chemical Physics*, 81(4):1975–1982, 1984.

[117] D. S. Palmer, A. I. Frolov, E. L. Ratkova, and M. V. Fedorov. Toward a universal model to calculate the solvation thermodynamics of druglike molecules: The importance of new experimental databases. *Molecular Pharmaceutics*, 8(4):1423–1429, 2011.

[118] V. P. Sergiievskyi, W. Hackbusch, and M. V. Fedorov. Multigrid solver for the reference interaction site model of molecular liquids theory. *Journal of Computational Chemistry*, 32(9):1982–1992, 2011.

[119] A. Kovalenko and F. Hirata. Potential of mean force between two molecular ions in a polar molecular solvent: A study by the three-dimensional reference interaction site model. *Journal of Physical Chemistry B*, 103:7942–7957, 1999.

[120] K. C. Ng. Hypernetted chain solutions for the classical one-component plasma up to $/gamma$ =7000. *Journal of Chemical Physics*, 61(7):2680–2689, 1974.

[121] Rism-mol solver: the program for fast and accurate rism solvation free energy calculations, http://compchemmpi.wikispaces.com/RISM-MOL.

[122] W. L. Jorgensen and J. Tirado-Rives. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proceedings of the National Academy of Sciences of the United States of America*, 102(19):6665–6670, May 2005.

[123] W. L. Jorgensen and J. Tirado-Rives. Molecular modeling of organic and biomolecular systems using boss and mcpro. *Journal of Computational Chemistry*, 26(16):1689–1700, December 2005.

[124] D. Shivakumar, J. Williams, Y. J. Wu, W. Damm, J. Shelley, and W. Sherman. Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the opls force field. *Journal of Chemical Theory and Computation*, 6(5):1509–1519, May 2010.

[125] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, Montgomery, Jr., J. A., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. *Gaussian 03*. Gaussian, Inc., Wallingford, CT, 2004.

[126] http://www.gaussian.com/g_tech/g_ur/k_dft.htm.

[127] C. M. Breneman and K. B. Wiberg. Determining atom-centered monopoles from molecular electrostatic potentials. the need for high sampling density in formamide conformational-analysis. *Journal of Computational Chemistry*, 11(3):361–373, April 1990.

[128] L. Lue and D. Blankschtein. Liquid-state theory of hydrocarbon water-systemsapplication to methane, ethane, and propane. *Journal of Physical Chemistry*, 96(21):8582–8594, October 1992.

[129] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *Journal of Physical Chemistry*, 91(24):6269–6271, November 1987.

[130] J. G. Kirkwood and F. P. Buff. The statistical mechanical theory of solutions .1. *Journal of Chemical Physics*, 19(6):774–777, 1951.

[131] Volodymyr .P Sergiievskyi and Maxim V. Fedorov. 3drism multigrid algorithm for fast solvation free energy calculations. *Journal of Chemical Theory and Computation*, 8(6):2062–2070, 2012.

[132] A. Kovalenko and F. Hirata. Potentials of mean force of simple ions in ambient aqueous solution. II. Solvation structure from the three-dimensional reference interaction site model approach, and comparison with simulations. *Journal of Chemical Physics*, 112(23):10403–10417, June 2000.

[133] J. S. Perkyns, G. C. Lynch, J. J. Howard, and B. M. Pettitt. Protein solvation from theory and simulation: Exact treatment of coulomb interactions in three-dimensional theories. *Journal of Chemical Physics*, 132(6):064106, February 2010.

[134] A. Kovalenko and F. Hirata. Self-consistent description of a metal-water interface by the kohn-sham density functional theory and the three-dimensional reference interaction site model. *Journal of Chemical Physics*, 110:10095–10112, 1999.

[135] P. Pulay. Convergence acceleration of iterative sequences - the case of scf iteration. *Chemical Physics Letters*, 73(2):393–398, 1980.

[136] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, November 1996.

[137] M. Frigo. A fast fourier transform compiler. *Acm Sigplan Notices*, 34(5):169–180, May 1999.

[138] Rism-mol3d: Multi-grid 3drism equations solver, http://compchemmpi.wikispaces.com/RISM-MOL-3D.

[139] J. M. Wang, W. Wang, P. A. Kollman, and D. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics & Modelling*, 25(2):247–260, October 2006.

[140] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, December 2005.

[141] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart. Am1 - a new general-purpose quantum-mechanical molecular-model. *Journal of the American Chemical Society*, 107(13):3902–3909, 1985.

[142] A. Jakalian, B. L. Bush, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method. *Journal of Computational Chemistry*, 21(2):132–146, January 2000.

[143] A. Jakalian, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. am1-bcc model: Ii. parameterization and validation. *Journal of Computational Chemistry*, 23(16):1623–1641, December 2002.

[144] J. M. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, July 2004.

[145] T. Imai, Y. Harano, A. Kovalenko, and F. Hirata. Theoretical study for volume changes associated with the helix-coil transition of peptides. *Biopolymers*, 59(7):512–519, December 2001.

[146] M. Udier-Blagovic, P. M. De Tirado, S. A. Pearlman, and W. L. Jorgensen. Accuracy of free energies of hydration using cm1 and cm3 atomic charges. *Journal of Computational Chemistry*, 25(11):1322–1332, August 2004.

[147] A. S. Paluch, J. K. Shah, and E. J. Maginn. Efficient solvation free energy calculations of amino acid analogs by expanded ensemble molecular simulation. *Journal of Chemical Theory and Computation*, 7(5):1394–1403, 2011.

[148] N. Matubayasi and M. Nakahara. Theory of solutions in the energy representation. ii. functional for the chemical potential. *Journal of Chemical Physics*, 117(8):3605–3616, August 2002.

[149] N. Matubayasi and M. Nakahara. Theory of solutions in the energy representation. iii. treatment of the molecular flexibility. *Journal of Chemical Physics*, 119(18):9686–9702, November 2003.

[150] S Zhao, Z. Jin, and J. Wu. Dft for rapid prediction of solvation free energy of ions in water. *Journal of the American Chemical Society*, 2011.

[151] Daniel J. Sindhikara, Norio Yoshida, and Fumio Hirata. Placevent: An algorithm for prediction of explicit solvent atom distribution—application to hiv-1 protease and f-atp synthase. *Journal of Computational Chemistry*, 33(18):1536–1543, 2012.

[152] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, February 1996.

[153] Rism-vmol: Vmd pluging for visualizing binding site of solvent molecules, http://compchemmpi.wikispaces.com/RISM-VMol.

[154] T. Imai, N. Miyashita, Y. Sugita, A. Kovalenko, F. Hirata, and A. Kideras. Functionality mapping on internal surfaces of multidrug transporter acrb based on molecular theory of solvation: Implications for drug efflux pathway. *Journal of Physical Chemistry B*, 115(25):8288–8295, June 2011.

[155] T. Yamazaki and A. Kovalenko. Spatial decomposition of solvation free energy based on the 3d integral equation theory of molecular liquid: Application to miniproteins. *The Journal of Physical Chemistry B*, 115(2):310–318, 2011.