The Journal of Chemical Physics

Iterative integral equation methods for structural coarse-graining

Marvin P. Bernhardt,[1, a)] Martin Hanke,[2, b)] and Nico F.A. van der Vegt[1, c)]

[1)] *Eduard-Zintl-Institut für Anorganische und Physikalische Chemie, Technische Universität Darmstadt, 64287 Darmstadt, Germany*

[2)] *Institut für Mathematik, Johannes Gutenberg-Universität Mainz, 55099 Mainz, Germany*

(Dated: February 1, 2021)

In this paper, new Newton and Gauss-Newton methods for iterative coarse-graining based on integral equation theory are evaluated and extended. In these methods, the potential update is calculated from the current and target radial distribution function, similar to iterative Boltzmann inversion, but gives a potential update of quality comparable with inverse Monte-Carlo. This works well for the coarse-graining of molecules to single beads which we demonstrate for water. We also extend the methods to systems that include coarse-grained bonded interactions and examine their convergence behavior. Finally, using the Gauss-Newton method with constraints, we derive a model for single bead methanol in implicit water which matches the osmotic pressure of the atomistic reference. An implementation of all new methods is provided for the open-source VOTCA package.

[a)] Electronic mail: bernhardt@cpc.tu-darmstadt.de

[b)] Electronic mail: hanke@mathematik.uni-mainz.de

[c)] Electronic mail: vandervegt@cpc.tu-darmstadt.de

I. INTRODUCTION

Structure-based coarse-graining aims at representing structural information of a fine-grained system through a coarse-grained (CG) model with fewer degrees of freedom. The goal of structure-based coarse-graining is the approximation of a CG N-body potential of mean force with pair potentials, that reproduces a set of distribution functions.[1] This can be thought of as an optimization problem: Optimize a potential until it reproduces the radial distribution function (RDF). For isotropic single-particle systems, the Henderson uniqueness theorem states that there is a bijective map between pair potential and RDF, which suggests that an optimal potential exists.[2,3] The two most common structure-based coarse-graining methods are Iterative Boltzmann inversion (IBI)[4] and Inverse Monte Carlo (IMC)[5]. IBI and IMC are applied widely in systems such as ionic liquids, polymers, and biological systems, where they help modelling time and length scales that are too expensive with atomistic MD.[6–10] Both methods iteratively improve pair potentials based on the distance from the current to the target distribution.[11] Delbary et al. have proposed two new Newton-type schemes, HNCN and IHNC, and the Gauss-Newton scheme HNCGN which are based on integral equation theory and conceptual compromises between IBI and IMC.[12]

They showed that their method can retrieve a Lennard-Jones (LJ) potential and generated a potential from experimental Argon data. In this work, we are comparing those new schemes with the IBI and IMC methods for coarse-graining molecular liquids.

The potential update of the Newton scheme is given by

$$u_{k+1} = u_k - \mathbf{J}^{-1}\left(g_k - g_{\mathrm{tgt}}\right). \tag{1}$$

Here, u_k and g_k are the potential and RDF at iteration k, respectively, and $\mathbf{J}$ is the Jacobian with elements $J_{\alpha\gamma} = \frac{\partial g_\alpha}{\partial u_\gamma}$. The analytical form of the map $u(g)$ is unknown, and the usual connection is to calculate g from u by molecular dynamics (MD) or Monte Carlo (MC) simulations. The IBI potential update results from the connection of u and g in the low density approximation, i.e. the direct Boltzmann inversion $g \approx e^{-\beta u}$ with $\beta = \frac{1}{k_{\mathrm{B}}T}$. This makes the IBI update a rough approximation to the Newton scheme if applied to liquids, that sometimes needs hundreds of iterations for convergence.[11,13] In IMC, the Jacobian is calculated from cross-correlations in the distance distribution in the system. With enough sampling the IMC Jacobian approximates the exact Jacobian, which is why the authors call their Ansatz "Newton inversion".[1] For simple systems IMC converges in under 10 iterations.

Lack of convergence, which is typical for Newton methods far from the optimum, can be addressed using regularization.[14] Nonetheless, the convergence can be slow, in particular for systems with multiple components.[9,15] This can make IMC coarse-graining computationally costly since the sampling of the IMC Jacobian needs long trajectories.

While direct Boltzmann inversion (BI) provides a reasonable estimate of the effective pair potential for low densities, this is no longer the case at liquid densities where direct BI leads to multibody contributions included in the effective pair potential. Here, the Ornstein-Zernike (OZ) equation combined with a hypernetted-chain (HNC) or Percus-Yevick (PY) closure provides a better approximation of the effective pair potential.[16] The HNCN method (hypernetted-chain Newton) uses the HNC closure to derive an approximation to the Jacobian, the input to the potential update scheme being the same as for IBI: the RDFs from the current potential.[12] This makes the computational cost per iteration to be comparable to IBI and potentially cheaper than IMC. At the same time, the number of iterations for convergence in single-component systems is comparable to IMC. IBI converges much slower for similar systems due to the crude approximation to the Jacobian.[17]

Using integral equations is not new in the field of molecular coarse-graining. This comes at no surprise as it provides an analytical connection between structure and potentials in liquid systems. Guenza et al. have established non-iterative methods for obtaining potentials for CG polymer melts.[18,19] By approximating the intramolecular distribution function with an analytical function, it is possible to solve the PRISM equation.[20] Due to the analytical nature of the equations, transferability across different resolutions or densities can be induced by fitting general trends in the direct correlation function.[21] Mashayak et al. have demonstrated that one can use integral equations for a good potential guess which can subsequently be improved by IBI.[22]. The method by Levesque et al. uses integral equations in a secant method way.[23,24] It comes closest to the methods discussed in this paper but differs in that the Jacobian is never calculated.

Potentials derived from structural coarse-graining are state point dependent.[25] To improve transferability, multistate-IBI can be used to fit the structure of several state points with one common potential.[26] CG potentials also don't generally represent the thermodynamic properties of their reference system. IBI can be pressure corrected by adding a ramp potential after each iteration.[4,27] Interestingly, we could not find any studies which implemented pressure ramps or constraints with IMC, even though a ramp correction is implemented in both

common IMC codes, VOTCA, and Magic.[28,29] Instead, an extension to constrain the surface tension during IMC was demonstrated.[14] A different approach is to extend the Hamiltonian with a term that only influences the pressure of the CG system, but not the structure, which can be extrapolated to different temperatures.[30,31] For the description of multi-phase systems, this approach was developed to utilize local density potentials which are optimized through the relative entropy method.[32] For the relative entropy method itself it was demonstrated that through the use of Lagrange multipliers the pressure can be constrained during the potential optimization.[33] Another property that can be incorporated as a target is the Kirkwood-Buff integral, as demonstrated for IBI.[34,35] IMC has been previously adapted to incorporate a constraint for the area compressibility of a phospholipid bilayer.[14] To apply constraints into the integral equation methods, Delbary et al. have reformulated their method to a Gauss-Newton method HNCGN.[12] It allows multiple constraints to be incorporated into the updating scheme.

This paper is organized as follows: In the theory section, we shortly recapitulate the basis of the HNCN and the HNCGN methods. We propose a new scheme derived from the Reference Interaction Site Model (RISM) to expand the methods to systems where the CG representation includes bonds. Some variants of the HNCN method and a scheme for RDF extrapolation are defined and examined. We apply IBI, IMC, HNCN, and HNCGN on water, hexane, and naphthalene, where the two latter systems are the test case for CG molecules with bonds. We examine in detail differences in the Jacobians and their physical meaning. Those differences explain the variance in the convergence behavior of the different methods. Finally, we use the HNCGN method with constrained pressure for the coarse-graining of methanol in water to an implicit solvent system with correct osmotic pressure.

II. THEORY

A. Newton iteration

The derivation of the integral equation methods for coarse-graining atomistic systems is described with greater mathematical rigor in the original paper.[12] Here we will just point out the steps important for understanding the method and results. The radially symmetric

OZ equation for a system with one particle type

$$h = c + \rho h * c \tag{2}$$

and the HNC closure

$$u = \frac{1}{\beta}\left(h - \log(g) - c\right). \tag{3}$$

can be used to express the potential as a function of h.[36] Here, $h = g - 1$ is the total correlation function, c is the direct correlation function, ρ is the particle number density, β is $(k_\mathrm{B}T)^{-1}$, and log the natural logarithm. All variables except ρ and β are functions of the particle-particle distance r. The operator $*$ denotes a three-dimensional convolution. To solve the OZ equation for c we switch to Fourier space where the convolution becomes a multiplication. The Fourier and inverse Fourier transform over $\boldsymbol{r}$ in 3D for a radially symmetric function f are defined as

$$\hat{f}(\omega) = \mathcal{F}(f) \quad = \int_{\mathbb{R}^3} f(|\boldsymbol{r}|) e^{-i\boldsymbol{k}\cdot\boldsymbol{r}}\,\mathrm{d}\boldsymbol{r} = \frac{2}{\omega}\int_0^\infty f(r) r \sin(2\pi r\omega)\,\mathrm{d}r \tag{4}$$

$$f(r) = \mathcal{F}^{-1}(\hat{f}) = \frac{2}{r}\int_0^\infty \hat{f}(\omega)\omega \sin(2\pi r\omega)\,\mathrm{d}\omega\,. \tag{5}$$

The resulting expression for $\hat{c}$ from equation (2) is

$$\hat{c} = \frac{\hat{h}}{1 + \rho\hat{h}}. \tag{6}$$

It is transformed back to real space and when inserted in equation (3) gives an expression for the potential as a functional of the RDF

$$u = \frac{1}{\beta}\left(h - \log(g) - \mathcal{F}^{-1}(\hat{c})\right) = \frac{1}{\beta}\left(h - \log(g) - \mathcal{F}^{-1}\left(\frac{\hat{h}}{1+\rho\hat{h}}\right)\right). \tag{7}$$

For the Newton or Gauss-Newton type potential update we need the corresponding approximation of the inverse of the Jacobian, which is the derivative of u by g. From the HNC closure (equation (3)) we obtain

$$\frac{\mathrm{d}u}{\mathrm{d}g} = \frac{1}{\beta}\left(1 - \frac{1}{g} - \frac{\mathrm{d}c}{\mathrm{d}g}\right). \tag{8}$$

Since c is an operator on g, so is the derivative $\frac{\mathrm{d}c}{\mathrm{d}g}$, which can be calculated in Fourier space

$$\frac{\mathrm{d}c}{\mathrm{d}g} = \mathcal{F}^{-1}\left(\frac{\mathrm{d}\hat{c}}{\mathrm{d}\hat{g}}\right)\mathcal{F} = \mathcal{F}^{-1}\left(\frac{1}{(1+\rho\hat{h})^2}\right)\mathcal{F}. \tag{9}$$

For the discrete case this can be written as a matrix, where the Fourier operator becomes a Fourier matrix. The matrix in equation (9) produces the non-diagonal elements of the Jacobian's inverse in equation (1)

$$u_{k+1} = u_k - \frac{1}{\beta}\left[\left(1 - \frac{1}{g_k}\right) - \mathcal{F}^{-1}\left(\frac{1}{(1+\rho\hat{h}_k)^2}\right)\mathcal{F}\right](g_k - g_{\mathrm{tgt}})\,. \tag{10}$$

This is called the hypernetted-chain Newton (HNCN) iteration. The term in the square brackets is the inverse of the Jacobian. It does not need to be explicitly calculated if one takes $g_k - g_{\mathrm{tgt}}$ into Fourier space. In the original paper, it is described that for the HNCN method the potential should be calculated on a longer range to obtain reasonable low-frequency values for $\hat{h}$. The tail of the potential update is then cut off. An alternative is to calculate the Jacobian explicitly from long RDFs and only use the square cutout that describes $\frac{\mathrm{d}u_{\mathrm{short}}}{\mathrm{d}g_{\mathrm{short}}}$. It is then inverted again and equation (10) is applied on the short-range; this we call the HNCN Jacobian cutout (jc) method. The latter method applied to RDFs, where the tails have been extrapolated as will be explained in section II D we call HNCN extrapolated (ex). The straightforward application of equation (10) with the RDF and the potential on the same, short-range we call HNCN short distribution (sd).

Equation (10) can be modified slightly by approximating $-(g_k - g_{\mathrm{tgt}})/g_k$ $\approx \log(1 + (g_{\mathrm{tgt}} - g_k)/g_k) = \log(g_{\mathrm{tgt}}/g_k)$ such that the first term resembles the IBI update, which is then called the inverse hypernetted-chain (IHNC) iteration.[12]

$$u_{k+1} = u_k - \underbrace{\frac{1}{\beta}\log\left(\frac{g_{\mathrm{tgt}}}{g_k}\right)}_{\Delta u_{\mathrm{IBI}}} - \frac{1}{\beta}\left[(g_k - g_{\mathrm{tgt}}) - \mathcal{F}^{-1}\left(\frac{(\hat{g}_k - \hat{g}_{\mathrm{tgt}})}{(1+\rho\hat{h}_k)^2}\right)\right]. \tag{11}$$

B. Gauss-Newton iteration

Delbary et al. introduced a Gauss-Newton type scheme, which has two advantages over a pure Newton scheme: (i) the calculated RDF can be longer ranged than the potential, which can naturally be reflected in a non-square Jacobian, and (ii) the scheme allows for the inclusion of one or multiple constraints.[12] Again, we only show the important parts of the scheme and refer to the original paper for details and mathematical rigor. We switch to a discretized notation with distributions and potentials becoming vectors and operators

becoming matrices. The problem of finding the potential is reformulated as a minimization

$$\arg\min_{\boldsymbol{w_k}} \|\boldsymbol{g}_{\mathrm{tgt}} - \boldsymbol{g_k} - \mathbf{U}^{-1}\mathbf{A_0}\boldsymbol{w_k}\|_2. \tag{12}$$

$\mathbf{U}^{-1}\mathbf{A_0}$ represents the Jacobian matrix and $\boldsymbol{w_k} = \boldsymbol{f_{k+1}} - \boldsymbol{f_k}$ is the force update in iteration k. Writing the iteration in terms of the force allows for a pressure constraint later. The matrix $\mathbf{A_0}$ is Δr times an upper unitriangular matrix stacked on a block of zeros. The unitrianglar part acts as an antiderivative operator that transforms the force update to the potential update

$$\boldsymbol{u_{k+1}} - \boldsymbol{u_k} = \mathbf{A_0}\boldsymbol{w_k}. \tag{13}$$

The block of zeros in $\mathbf{A_0}$ makes it rectangular and causes the potential update to be zero beyond a desired cut-off, but based on the whole range of the input RDF. Matrix $\mathbf{U}$ represents the Jacobian with respect to the potential which we already encountered in equation (10)

$$\mathbf{U}^{-1} = \frac{1}{\beta}\left(\mathrm{diag}\left(1 - \frac{1}{\boldsymbol{g}_{\mathrm{tgt}}}\right) - \mathbf{F}^{-1}\mathrm{diag}\left(\frac{1}{(1+\rho\hat{\boldsymbol{h}}_{\mathrm{tgt}})^2}\right)\mathbf{F}\right). \tag{14}$$

$\mathbf{F}$ is the Fourier matrix.

The Gauss-Newton scheme allows us to add constraints to the potential update. A classic constraint in structure-based coarse-graining is the pressure. The Henderson theorem would predict that only one potential can reproduce a given RDF. Previous studies showed, that certain changes to the potential have little effect on the distribution function.[27] This motivates the scheme, where the pressure is enforced and the RDF is matched as good as possible. The constraint, derived from the virial, has the form

$$\boldsymbol{l^T}\boldsymbol{w_k} = p_{\mathrm{tgt}} - p_k, \tag{15}$$

where p_{tgt} and p_k are the target and current pressure, respectively. The elements of $\boldsymbol{l}$ are given by

$$l_\alpha = \frac{2}{3}\pi\rho^2 \frac{g_{\mathrm{tgt},\alpha} + g_{\mathrm{tgt},\alpha+1}}{2}\frac{r_{\alpha+1}^4 - r_\alpha^4}{4} \tag{16}$$

The constraint is exact if $g_{\mathrm{tgt}} = g_{k+1}$ so the pressure will not be precisely met in early iterations.

C. Extension to symmetric molecules with internal bonds

Here we extend the methodology from section II A and II B to one component systems where the CG molecules consist of n identical beads. The beads have to be bonded in a way

that makes them indistinguishable. Starting point is the RISM-OZ relation[36]

$$\hat{\mathbf{h}} = \hat{\mathbf{\Omega}}\hat{\mathbf{c}}\hat{\mathbf{\Omega}} + \hat{\mathbf{\Omega}}\hat{\mathbf{c}}\boldsymbol{\rho}\hat{\mathbf{h}} = \hat{\mathbf{\Omega}}\hat{\mathbf{c}}\left(\hat{\mathbf{\Omega}} + \boldsymbol{\rho}\hat{\mathbf{h}}\right). \tag{17}$$

where $\hat{\mathbf{h}}$ and $\hat{\mathbf{c}}$ are matrices with elements $\hat{h}_{ij}$ and $\hat{c}_{ij}$ of sites i and j, respectively. Each site in a molecule has a separate index, ignoring indistinguishability for now. Thereby each site has the same number density $\rho_i = \rho$. Ω_{ij} represents the intramolecular distribution function between sites i and j

$$\Omega_{ij} = \rho\left(G_{ij} - g_{ij}\right) + \delta_{ij}\delta(\boldsymbol{r}) \quad \Longleftrightarrow \quad \hat{\Omega}_{ij} = \rho\left(\hat{G}_{ij} - \hat{g}_{ij}\right) + \delta_{ij}. \tag{18}$$

Its Fourier transform $\hat{\Omega}_{ij}$ is the intramolecular structure factor. Note that the delta distribution of vector $\boldsymbol{r}$ (not the same as $\delta(r)$) becomes one by the 3D Fourier transform. G_{ij} is defined similar to a normal RDF, but in addition counts sites that are on the same molecule, except for the reference site. In the appendix we show how it can be expressed in terms of the average intramolecular distribution function under the condition that all combinations of sites in a molecule have the same distance

$$\rho(G_{ij}(r) - g_{ij}(r)) = (1 - \delta_{ij})\frac{n}{n-1}\rho(G(r) - g(r)). \tag{19}$$

Matrix $\mathbf{\Omega}$ is therefore written as

$$\hat{\mathbf{\Omega}} = \frac{n}{n-1}\rho(\hat{G} - \hat{g})(\mathbf{J_n} - \mathbf{I_n}) + \mathbf{I_n}, \tag{20}$$

where $\mathbf{I_n}$ is the identity matrix and $\mathbf{J_n}$ is the matrix of all ones. The other variables in the RISM-OZ equation are determined straightforwardly. All particle number densities ρ_i and distribution functions g_{ij}, h_{ij} and c_{ij} are equal for all i and j, which we use to express the matrices in terms of $\mathbf{I_n}$ and $\mathbf{J_n}$

$$\boldsymbol{\rho} = \rho\mathbf{I_n} \qquad \hat{\mathbf{g}} = \hat{g}\mathbf{J_n} \qquad \hat{\mathbf{h}} = \hat{h}\mathbf{J_n} \qquad \hat{\mathbf{c}} = \hat{c}\mathbf{J_n}. \tag{21}$$

Inserting equations (20) and (21) in (17) we find that the matrix equation reduces to a single equation because of the identity $\mathbf{J_n}\mathbf{J_n} = n\mathbf{J_n}$

$$\hat{h}\mathbf{J_n} = \left(\frac{n}{n-1}\rho(\hat{G} - \hat{g})(n-1) + 1\right)\hat{c}\left(\left(\frac{n}{n-1}\rho(\hat{G} - \hat{g})(n-1) + 1\right) + n\rho\hat{h}\right)\mathbf{J_n} \tag{22}$$

from which follows

$$\hat{c} = \frac{\hat{h}}{(1 + n\rho(\hat{G} - \hat{g}))^2 + (1 + n\rho(\hat{G} - \hat{g}))n\rho\hat{h}}. \tag{23}$$

Note that this result is similar to the PRISM-OZ relation, which is often employed for polymers.[37] We can use this relation with the HNC closure in equation (3) for an initial potential guess. For a Newton update in the HNCN/IHNC scheme, we take the derivative with respect to the RDF

$$\frac{\mathrm{d}\hat{c}}{\mathrm{d}\hat{g}} = \frac{\hat{1}}{(1 + n\rho(\hat{G} - \hat{g}) + n\rho\hat{h})^2}. \tag{24}$$

Here we ignore that the intramolecular correlation $(\hat{G} - \hat{g})$ is to some degree also a function of $\hat{g}$. We assume that intra- and intermolecular degrees of freedom are uncoupled. With this equation we have all we need for a Newton potential update. For $G - g = 0$, i.e. no intramolecular correlation, equation (24) reduces to equation (9).

D. RDF extrapolation

The sampling of the RDF from a trajectory is, together with the MD simulation, the computationally most demanding task in iterative coarse-graining. It grows significantly with a larger range since the number of particles considered grows roughly with the volume. Here we present a simple scheme for extrapolating an RDF with integral equation theory.

The direct correlation function c decays much faster to zero than the total correlation function h.[36] With the OZ equation we can first calculate c from short-range data for h using equation (6), extrapolate the result with zeros, and then use the OZ equation again to obtain $h^\dagger_{\mathrm{ext},0}$

$$h_{\mathrm{ext},0} = B^{\mathrm{g}} h^\dagger_{\mathrm{ext},0} = B^{\mathrm{g}} \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\mathrm{B}\, c)}{1 - \rho\mathcal{F}(\mathrm{B}\, c)}\right). \tag{25}$$

Here, B is an operator that appends zeros which numerically equals a unit matrix stacked on top of a block of zeros. The ratio of rows to columns equals the factor by which the range is expanded. B^g is the generalized inverse of B which cuts off the tail of a function. The superscript † denotes that a function is defined on the new longer range. Upon a simple application of equation (25) we find that $h^\dagger_{\mathrm{ext},0}$ does generally look like an extrapolation of h but has two issues. It deviates from h on the short range and has a discontinuity at the transition point as depicted in figure 1. The appended tail of c influences all parts of the total correlation function because of the convolution in the OZ equation.

We now aim to find an improved direct correlation function c_{ext} that when plugged in equation (25) will result in $h^\dagger_{\mathrm{ext}}$ which matches h in the first region. We find that equation

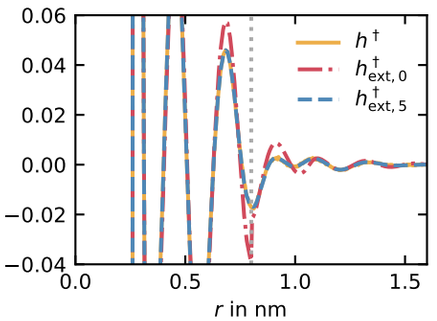


Figure 1: The true long-range total correlation function $h^\dagger$ of the center of mass of SPC/E water and the naive result $h^\dagger_{\text{ext},0}$ from extending the direct correlation function with zeros. Also shown is the extrapolation result $h^\dagger_{\text{ext},5}$ from running five Newton iterations to fit h. The dotted, grey line marks the end of he range of h, from which $h^\dagger_{\text{ext},0}$ and $h^\dagger_{\text{ext},5}$ are calculated.

(25) is not invertible, which makes this an inverse problem. In order to solve it we apply Newton's method where for iteration l we alternately apply

$$\begin{aligned} c_{\text{ext},l+1} &= c_{\text{ext},l} - \left(\frac{\mathrm{d}h_{\text{ext}}}{\mathrm{d}c_{\text{ext}}}\right)^{-1} (h_{\text{ext},l} - h) \\ \frac{\mathrm{d}h_{\text{ext}}}{\mathrm{d}c_{\text{ext}}} &= \mathrm{B}^{\mathrm{g}}\, \mathcal{F}^{-1} \left(\frac{1}{(1-\rho\mathcal{F}(\mathrm{B}\, c_{\text{ext}}))^2}\right) \mathcal{F}\, \mathrm{B} \end{aligned} \tag{26}$$

and equation (25) for obtaining $h_{\text{ext},l+1}$ from $c_{\text{ext},l+1}$. We find this to converge within few iterations and giving the expected results, which are exemplarily shown in figure 1. In order to use this method for the molecular case with n beads as described before, the following two equations are derived from equation (23) for the calculation of h_{ext} and its derivative in the Newton scheme

$$h_{\text{ext}} = \mathrm{B}^{\mathrm{g}}\, \mathcal{F}^{-1} \left(\frac{(1+n\rho\mathcal{F}(\mathrm{B}(G-g)))^2 \mathcal{F}(\mathrm{B}\, c)}{1-n\rho(1+n\rho\mathcal{F}(\mathrm{B}(G-g)))\mathcal{F}(\mathrm{B}\, c)}\right) \tag{27}$$

$$\frac{\mathrm{d}h_{\text{ext}}}{\mathrm{d}c_{\text{ext}}} = \mathrm{B}^{\mathrm{g}}\, \mathcal{F}^{-1} \left(\frac{(1+n\rho\mathcal{F}(\mathrm{B}(G-g)))^2}{(1-n\rho(1+n\rho\mathcal{F}(\mathrm{B}(G-g)))\mathcal{F}(\mathrm{B}\, c_{\text{ext}}))^2}\right) \mathcal{F}\, \mathrm{B}\,. \tag{28}$$

III. METHOD

A. HNCN and HNCGN in VOTCA

The two new coarse-graining methods HNCN and HNCGN described above have been implemented in the Versatile Object-oriented Toolkit for Coarse-graining Applications

(VOTCA)[28,38] software. For HNCN, all variants discussed in section II A are available and can be specified in the VOTCA input file. The usage is similar to IBI and IMC, Python and NumPy are used internally for the vector operations. The code allows for the generation of a potential guess based on integral equations (equations (3) and (7)). A potential guess for molecular systems of identical beads can be generated using equation (23). Potential updates are calculated with the HNCN, IHNC, and HNCGN method (equations (10), (11), and (12)). We also modify VOTCA's `csg_stat` to compute the correlation function $G(r)$, which is needed for the molecular case. An alternative to the HNC closure, the Percus-Yevick closure

$$g = \frac{c}{1 - e^{\beta u}} \tag{29}$$

can be used to derive related update schemes, which are also implemented but not tested thoroughly. Some preliminary tests showed very similar results to the HNC closure for the systems in this paper. The pressure constraint for HNCGN is implemented as an elimination with the algorithm as described by Gander et al..[39] In the current form, the code can be straightforwardly extended to other constraints if they can be expressed in terms of the force $F(r)$ and the RDF $g(r)$.

There is also a new *power* extrapolation scheme for the potential in the core region where the RDF is zero. It fits a power function $U_{\text{fit}} = ar^b$ to the Boltzmann inverse of g_{tgt} from the first r where g_{tgt} is larger than a threshold (default: 1×10^{-10}) to its first maximum. The potential is then extrapolated in the region $r = 0$ to the first r where the potential is convex or g_{tgt} is above 1×10^{-2}.

The code resides currently in a fork of VOTCA at `https://gitlab.com/cpc_group/csg` but we aim to get it into the main repository.

B. Water, hexane, naphthalene

We have tested our methods on simple LJ systems and found similar results as Delbary et al.[12], so we turn to more relevant molecular systems in this work. To test the new coarse-graining schemes we create reference structures by performing all-atom (AA) simulations of water, hexane, and naphthalene. We run NVT simulations of SPC/E water and OPLS/AA models of hexane and naphthalene.[40–42] Long-range electrostatics are accounted for with the particle-mesh-Ewald (PME) method. GROMACS 2019 was used for all simulations.[43]

For all systems, a box with the correct volume is filled randomly with molecules. It is then energy-minimized with the steepest descent algorithm for 10^4 steps. After an equilibration of 10^5 steps a production run of 2×10^6 steps is performed where every 100th step is written to a trajectory file. A Langevin thermostat with a friction constant of $2\,\mathrm{ps}^{-1}$ keeps the temperature constant. Further simulation details and the average pressure from the production run are given in table I.

Table I: Parameters used in the AA NVT simulations and the resulting average pressure.

	water	hexane	naphthalene
Nr. of molecules	5000	2000	2000
Temperature in K	298.15	273.15	373.15
Density in $\mathrm{g\,cm^{-3}}$	0.998	0.67	0.99
Timestep in fs	2	1	1
Cut-off in nm	0.8	1.2	1.2
NVT pressure in bar	259	77	311

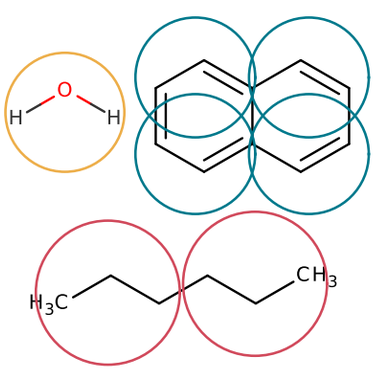


Figure 2: Mapping schemes for water, hexane, and naphthalene. Bead positions are determined by the center of mass of atoms within it. The central carbon atoms in naphthalene belong to two beads at the same time.

The VOTCA package is used for mapping the AA trajectory and calculating the target distribution functions for the coarse-graining. The mapping schemes used are shown in figure 2. In naphthalene, each bead corresponds to the center of mass of two carbon atoms and one of the central carbon atoms with half weight. A snapshot from the equilibrated AA

simulation is taken and mapped to be the starting configuration. The starting non-bonded potentials are generated by equation (7), with the target distribution over a range of twice the cut-off. This ensures that all methods start from the same potential and allows us to compare the potential updates in the first iteration in detail. For the bonded potentials (one bond in hexane, two bond-types, one angle-type, and two dihedral-types in naphthalene) the target distribution is Boltzmann inverted to obtain a starting potential. For each iteration the system is run with a timestep of 4 fs and a cut-off of 0.8 nm for water and 1.2 nm for hexane and naphthalene. MD simulations are run for 7×10^4 (3.1×10^5 for IMC) steps where every 10 steps the positions are saved in a trajectory, of which the first 1×10^4 are discarded. This means, that five times the amount of frames goes into sampling the IMC matrix, compared to the distributions for the other methods. For HNCN, HNCN (jc), and HNCGN the RDF for the potential update at each iteration is determined up to twice the cut-off. In the core region where the potential update is undefined due to the RDF being zero the potential is extrapolated with an exponential function. The simulation settings are otherwise equal to the AA simulations.

While we vary the methods for the non-bonded potential update, we consistently use IBI for the internal degrees of freedom. This ensures that bond and angle distributions are precisely reproduced which simple Boltzmann inversion cannot guarantee. We do however have to scale the potential updates before adding them to the potential for the previous iteration. This is done for naphthalene, where we use 0.5 for angles and 0.25 for dihedrals. The scaling is justified by the interdependence of multiple bonded potentials due to the small ring. In ring-free molecules, bonded potentials are normally relatively independent. However, in a four-ring a small change in all four angle potentials will make the whole ring much stiffer. Therefore, to make the angle distribution narrower the potential needs to be changed only a fraction of the IBI update. If applied without the scaling, we observe divergence in the bonded potentials and the simulation crashes after some iterations.

C. Methanol in water

Methanol-water simulations are prepared with seven different mole fractions of methanol: 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9. The AA simulations employ the OPLS/AA model for methanol and the TIP4P model for water.[41,44] A total of 8000 molecules with eight different

mole fractions are equilibrated and simulated in an NPT ensemble. Molecules are inserted randomly into a box with density $1\,\mathrm{g}^3\,\mathrm{cm}^{-1}$. They are successively energy minimized for 5×10^3 steps, run with a velocity rescaling thermostat with a stochastic term for 1×10^4 steps, and run with an additional Berendsen barostat for 1×10^5 steps. The timestep is 1 fs and the temperature 298.15 K. The cut-off for the LJ interactions is 1.2 nm and PME is used for the long-range electrostatics. The production run is perfomed with a Nosé-Hoover thermostat ($\tau_T = 2\,\mathrm{ps}$) and Parinello-Rahman barostat ($\tau_p = 10\,\mathrm{ps}$). Since we are interested in the methanol-methanol distribution function, we simulate longer for dilute systems. The production run has therefore $\frac{2 \times 10^5}{x(\mathrm{methanol})}$ steps. Positions of the atoms in methanol are written every ten steps after the first 0.5 ns. An RDF is calculated from the center of mass of each methanol molecule using VOTCA.

For each composition, we also determine the osmotic pressure from atomistic simulation. We therefore closely follow the OPAS method.[45] It is based on a simulation of a box elongated in z-direction where the solute is kept in a defined region by two semipermeable membranes normal to the z-axis. From a pre-run with a total pressure of 1000 bar the osmotic pressure is first estimated, such that in the production run the pressure outside the mixed slab can be set to 1 bar. We depart slightly from the original method in that we employ a Parinello-Rahman barostat that scales the box in x- and y-direction during the pre- and the production-run. Thereby the box is scaled perpendicular to any forces exerted by the walls and interference is avoided. Equilibration is performed in the same procedure as described above for the bulk simulations. We test two force constants k_w for the half-harmonic potentials that form the semipermeable membranes, $500\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ and $4000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$. Methanol molecules are attached to a virtual site at their center of mass. The membrane force is determined by the position of the virtual site and the force is distributed on the atoms of the molecule. Inside the mixture slab, some of the methanol molecules adsorb to the semipermeable walls and also push a bit outside of the confined region. The resulting mole fraction and concentration for each osmotic pressure is determined from the unperturbed region in the middle of the slab, by counting molecules in that region.

We develop an implicit solvent model of methanol in water for each mole fraction. Methanol is mapped to a single bead. We use the HNCGN method with and without a pressure constraint to obtain the CG interaction potentials. The configuration is generated from a mapped configuration of a snapshot of an $x(\mathrm{methanol}) = 0.9$ AA system with 7200

methanol molecules, which is scaled to match the methanol density of the NPT run. This results in large boxes for the low mole fractions. Therefore we need again more sampling for the "dilute" systems to obtain meaningful RDFs. We run each coarse-graining iteration for 1×10^4 equilibration steps and $\frac{4 \times 10^4}{x(\text{methanol})}$ production steps of which every tenth is saved.

IV. RESULTS AND DISCUSSION

A. Potential guess

We start by examining the potential obtained from equation (7) and (23) which we use for all performed iterative coarse-graining runs to start from. In figure 3 we compare it to the Boltzmann inverted (BI) target distribution, i.e. the potential of mean force, which is the common choice in coarse-graining for the starting potential. For the HNC potential of hexane

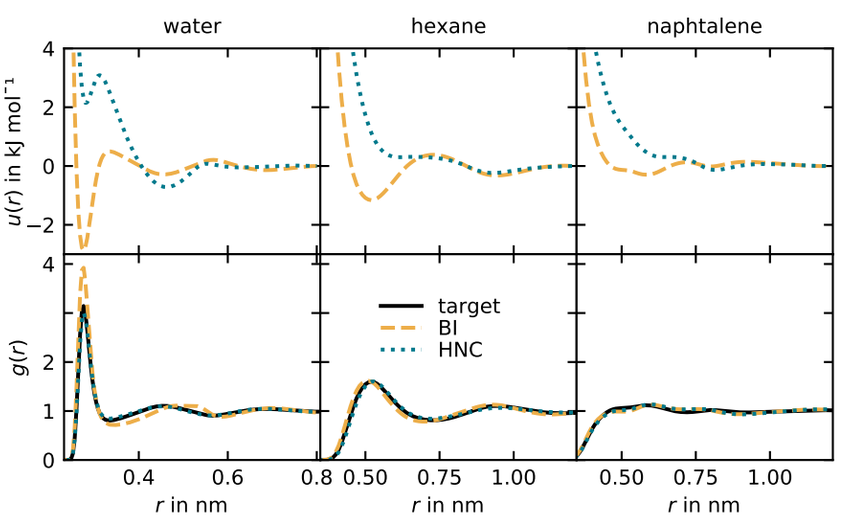


Figure 3: Potentials generated from Boltzmann inversion (BI) and HNC inversion of a given target RDF and the resulting distributions from MD simulations.

and naphthalene, the intramolecular distribution obtained from the mapped AA simulation is used. We find that the HNC potential is always more repulsive for all three molecules compared to the BI potential. It does result in a very good RDF when compared with the BI potential, especially for water, where the BI potential generates too much structuring. For naphthalene both potentials, even though very different in shape, produce almost the same

RDF. This confirms the common observation that while there might be only one optimal potential, the RDF is very insensitive to certain changes in the potential.

B. Jacobian comparison

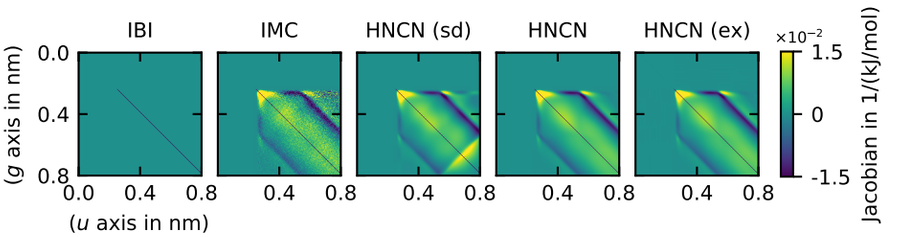


Figure 4: Jacobian $\frac{\mathrm{d}g}{\mathrm{d}u}$ for the first iteration of the SPCE-water coarse-graining from four different methods. For HNCN and HNCN (ex) the Jacobian extends till 1.6 nm in both directions and only the top left quarter is presented.

To analyze the quality of the update schemes, we compare the Jacobians of each Newton method. In figure 4 the Jacobians from the first iteration of water with the respective methods are shown. Each element represents the derivative $\frac{\partial g_\alpha}{\partial u_\gamma}$ for the change of the RDF at α by changes in the potential at γ. In the first iteration, the Jacobian contains all information about the update, since for all methods the distance $g_k - g_{\mathrm{tgt}}$ is the same. For IBI it is a diagonal matrix with values

$$\mathrm{diag}(\mathbf{J}_{\mathrm{IBI}}) = \beta((g_k - g_{\mathrm{tgt}})/\log(g_{\mathrm{tgt}}/g_k)) \tag{30}$$

on the diagonal and zero on the off-diagonal elements. The diagonality reflects that IBI can only update the potential based on local information about structure mismatch. For IMC usually the IMC matrix $\mathbf{A}_{\mathrm{IMC}}$ is written down in terms of a function $S(r) = 4\pi r^2 N^2/(2V)g(r)$ where N is the number of particles and V is the volume. The elements of the IMC matrix are calculated by

$$\mathbf{A}_{\mathrm{IMC}\alpha\gamma} = \frac{\partial S_\alpha}{\partial u_\gamma} = \beta\left(\langle S_\alpha\rangle\langle S_\gamma\rangle - \langle S_\alpha S_\gamma\rangle\right). \tag{31}$$

For comparison purpose, the Jacobian with respect to g can be retrieved by

$$\mathbf{J}_{\mathrm{IMC}\alpha\gamma} = \frac{\partial g_\alpha}{\partial u_\gamma} = (4\pi r_\alpha^2 N^2/(2V))^{-2}\mathbf{A}_{\mathrm{IMC}\alpha\gamma} \tag{32}$$

The Jacobian for the HNCN method is the term in square brackets in equation (10), inverted.

The diagonal elements of all Jacobians are negative since a positive perturbation of the potential at α will result in a negative impact in the RDF at the same distance. Except for IBI, other features of the three Jacobians are very similar too, for example the positive region close to the diagonal at about the first peak of the RDF. Physically, this can be thought of as an effect of the potential well. A positive potential perturbation in that region leads to a higher sampling of neighboring points because the remaining potential well is favored. Another motif that is present in both methods is a pair of diagonal stripes of negative values, which are located $\pm 0.27\,\mathrm{nm}$ from the diagonal. That distance is equivalent to the position of the first peak r_{fp} in the RDF. It represents that a perturbation of the potential at any distance will make it less likely to find a second particle at that distance which will indirectly make the presence of a third particle at that distance $\pm r_{\mathrm{fp}}$ less likely. The IMC Jacobian is not perfectly smooth, where the noise depends on the number of frames used for sampling the IMC matrix. It is symmetric by construction, see equation (31) and fading out to the bottom due to the normalization with $\frac{1}{r^2}$ in equation (32). The Jacobians $\mathbf{J}_{\mathrm{HNCN\ (sd)}}$, $\mathbf{J}_{\mathrm{HNCN}}, \mathbf{J}_{\mathrm{HNCN\ (ex)}}$ are much smoother but seem to retain the same structure and symmetry.

The similarity in structure indicates that through integral equations it is well possible to construct a Jacobian just from RDFs. $\mathbf{J}_{\mathrm{HNCN\ (sd)}}$ shows an interesting artifact in the bottom right corner: a stripe of positive values orthogonal to the matrix diagonal. The same motif is present in $\mathbf{J}_{\mathrm{HNCN}}$ and $\mathbf{J}_{\mathrm{HNCN\ (ex)}}$, but outside the shown sector of the matrix. On closer inspection, there are more, weaker but similar motifs that go perpendicular to the diagonal. We account those artifacts to the numerical deconvolution of c and h. In the OZ equation (2) the total correlation function h is input and output of the convolution. One aspect of the convolution of two distributions is that the output will have a larger range than the inputs. The situation with the OZ equation is best illustrated by its recursive expansion

$$h = c + \rho h * c = c + \rho c * c + \rho^2 c * c * c + \dots \tag{33}$$

So even if the direct correlation function decays very fast, which it usually does, it will generate non-negligible values for the total correlation function on a longer range than c has. Therefore, by calculating c from h with a short-ranged h one will produce an erroneous estimate for c. The same argument can be made for the calculation of $\frac{\mathrm{dc}}{\mathrm{d}g}$ from h. We think this effect is also the reason why Heinen calculates the structure factor in a way that uses effectively the whole simulation box.[24] The original HNCN method avoids this by calculating

a longer RDF and cutting away the possibly erroneous part of the potential. The HNCN (jc) method as introduced in section II A also avoids the artifact by cutting the Jacobian, before multiplication with the RDF distance. The Jacobian of the HNCN and the HNCN (jc) method are equal, so we only show the former in figure 4. Finally, the variant where the RDF has been extrapolated, HNCN (ex), shows a Jacobian indistinguishable from the HNCN (jc) method.

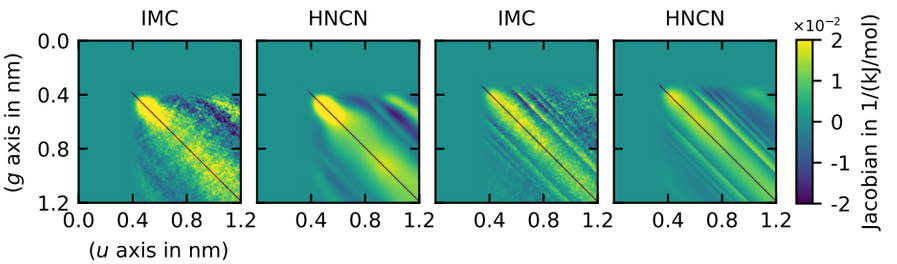


Figure 5: Jacobians for the first iteration of the hexane (left two) and naphthalene (right two) coarse-graining with IMC or the HNCN method. For HNCN the Jacobians extend till 2.4 nm in both directions and only the top left quarters are presented.

To confirm the newly derived update scheme for symmetric molecules we also compare the Jacobians of hexane and naphthalene with IMC in figure 5. Due to the absence of noise, finer details are recognizable in the HNCN Jacobian, such as sharp lines, that originate from the peak like intramolecular distribution function. The HNCN Jacobian is indeed very similar to the IMC result. This confirms, that a Newton update based on the RISM equation does sufficiently approximate the exact Newton update. It also implies that the approximation made in equation (19) for naphthalene, namely the equivalence of off-diagonal elements in G_{ij}, is reasonable.

We cannot present a Jacobian calculated from HNCN (ex) at this point. The reason is that the Newton scheme for RDF extrapolation, as described by equations (27) and (28), fails to converge for the distribution generated by the potential guess for both hexane and naphthalene. Interestingly, it converges for the target distribution in both cases. We believe this is related to the inadequateness of the assumption of a fast decay of the direct correlation function for the molecular case. The RISM-OZ equation (17), in contrast to the simple OZ equation (2), is known to not be fully consistent with a short-ranged direct correlation function.[36,46]

C. Convergence

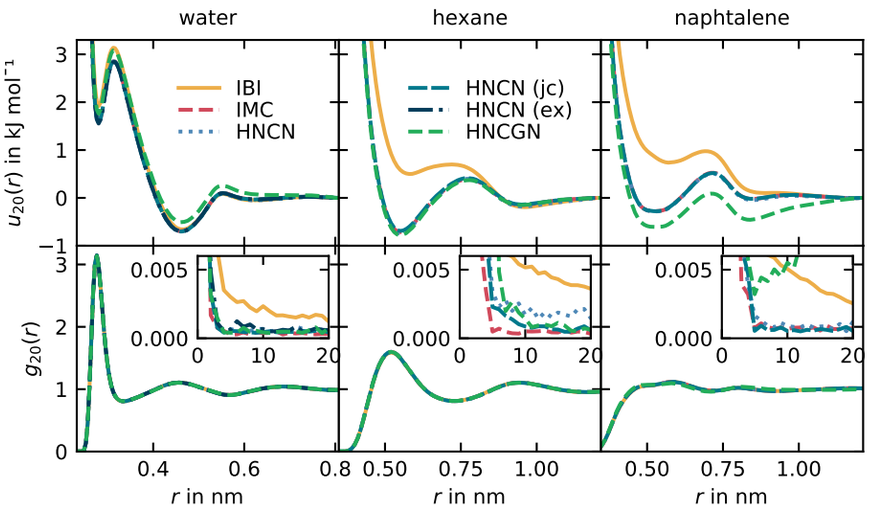


Figure 6: Potentials and distributions up to the cut-off after 20 iterations of four different coarse-graining methods. The insets in the bottom row show the convergence of the data fit δ versus the iteration number. The target RDF is also plotted, but is not visible because it is hidden by the graphs of the other distributions.

We now compare the output of six different coarse-graining methods: (i) IBI, (ii) IMC, (iii) HNCN, (iv) HNCN (jc), (v) HNCN (ex), and (vi) HNCGN. Bonded potentials have in any case been updated using IBI and have converged within few iterations in any case, so we are not showing them here. In figure 6 we show the potentials and distributions obtained after 20 iterations. The convergence of the RDF, given by the data fit measure

$$\delta = \sqrt{\frac{1}{r_{\mathrm{co}}} \int_0^{r_{\mathrm{co}}} (g_k - g_{\mathrm{tgt}})^2 dr}, \tag{34}$$

is also shown.[28] The upper integration limit is the cut-off r_{co}. HNCN (ex) results are only present for water since, as mentioned in the last section, the RDF extrapolation fails for hexane and naphthalene. Results for the HNCN (sd) method are missing for all three molecules because the iterative method would produce erroneous potentials: those would be strongly oscillating over their whole range or have very deep minima which would at some point crash the MD simulation. Since the other HNCN methods converge, we contribute the convergence failure to the artifacts in the Jacobian as discussed previously. This limits

the applicability of the naive HNCN (sd) method. If short cut-offs are wanted, and this is typically the case in coarse-graining, an Ansatz with longer RDF for the Jacobian has to be used to obtain meaningful results.

All methods produce the same distribution within line thickness after 20 iterations, the potentials however differ. The IBI potentials are much more repulsive in hexane and naphthalene, but we did not test if they would eventually converge to the IMC potential. The IMC, HNCN, and HNCN (jc) results are almost identical, and also converge to a similar δ value. The convergence speed is fastest with IMC only requiring 3-8 iterations till δ flattens out. HNCN, HNCN (jc), and HNCN (ex) converge within 4-11 iterations, a bit slower than IMC but much faster than IBI. This includes hexane and naphthalene which shows that RISM based coarse-graining works for the examined molecular systems.

We find that for hexane HNCN achieves less accuracy than HNCN (jc), i.e. the converged δ value is higher. HNCN (jc) only uses a cut-out of the Jacobian, while HNCN uses the inverse of the full matrix including the artifact. Even though the tail of the potential update is cut off, it seems the artifact still influences the short-ranged potential update.

One can compare the performance by looking at the amount of time used for each iteration. For water we find 1.5 min for IBI, HNCN (sd), and HNCN (ex), 6 min for IMC, and 7 min for HNCN, HNCN (jc), and HNCGN, on a 24 core AMD Opteron 6174. The bottlenecks are the MD simulation, the RDF calculation, and the IMC matrix calculation. HNCN and HNCN (jc) were performed with the RDF calculated on double the range than HNCN (sd), which has a large performance impact. HNCN (ex) has the same input as HNCN (sd) and extrapolates the RDFs to the doubled range on the fly. This results in HNCN (ex) being the fastest method to converge in total computational time, beating IMC by a factor of four.

When looking in more detail at the convergence behavior we find that the RDF oscillates slightly around the target distribution. Oscillations around the solution are expected from a Newton method with a slight error in the derivative. For IMC this error in the Jacobian is statistical. For HNCN (jc) it is statistical, caused by noise in the distributions, and systematic due to the HNC closure not being exact. It follows that if one is interested in improving the RDF to a very precise degree, the way to go is using IMC updates with sufficient statistics in the IMC matrix. For practical purposes, the precision of HNCN (jc) should be more than satisfying.

The HNCGN results are less uniform. We find that it converges similarly fast as HNCN

for water and hexane, but to slightly different potentials, once more repulsive, once more attractive. For naphthalene, the situation is worse and the potential oscillates largely and never converges. To understand if artifacts in the Jacobian can be the source of this, we have to know which parts of it are used. HNCGN by default can use the RDF on a longer range than the potential update is made on through matrix $\mathbf{A_0}$ in equation (12). Due to the block of zeros in the lower part of $\mathbf{A_0}$ it cuts out parts from the right of the Jacobian when multiplied with it. This reflects the idea of the Gauss-Newton Ansatz, where a short-ranged potential is updated based on a longer RDF and non-square shape allows for the addition of constraints. We depict this in figure 7. The artifacts, which are visible as "waves" perpendicular to the

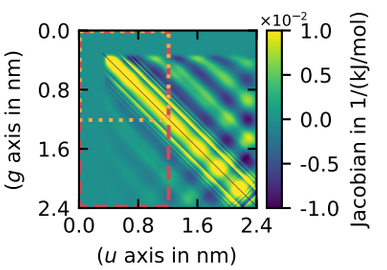
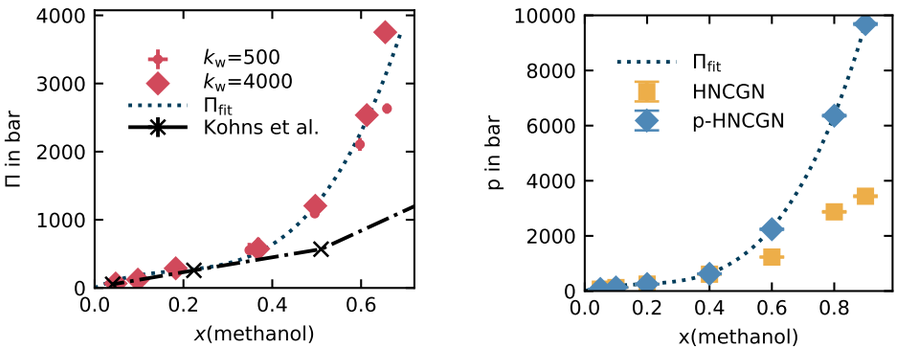


Figure 7: The Jacobian obtained from integral equations for the first coarse-graining iteration of naphthalene. The orange dashed rectangle shows the region that is used for the HNCGN update, whereas the yellow dotted square shows the cut out that HNCN (jc) is using.

diagonal and emerging from the lower right corner, are present in the HNCGN region. We cannot with full certainty link the issues in convergence with HNCGN in naphthalene to these artifacts in the Jacobian, but since they have shown to be the root cause with the HNCN (sd) method it seems likely. The main artifact lies outside of the region used, which probably explains why HNCGN fails to converge only sometimes. A concluding test would be to calculate the Jacobian on an even larger range, such that the part that is used by HNCGN is uninfluenced by the artifacts, but this we have not implemented. The difference in the potentials obtained for water and hexane can be explained by that HNCN and HNCGN are optimizing different residues. HNCGN finds the best potential for the RDF on a range twice as long as HNCN does. So the difference in the potential is to be expected since more and different target information goes into the HNCGN method.

D. Pressure matching



(a) (b)

Figure 8: (a) Osmotic pressures obtained by the OPAS method from all-atom slab simulations with two different wall force constants k_w in kJ mol^{-1} nm^{-2}. The polynomial fit function (35) is included as dotted line. The literature data from Kohns et al. are shown as crosses and connected by dashed-dotted lines to guide the eye. (b) The pressure values resulting from NVT simulations of CG models of methanol in implicit water.

We now present the results for the methanol-water mixtures. Figure 8a displays the osmotic pressures Π determined by the OPAS method. The osmotic pressure is found to increase strongly up to several thousand bars when the fraction of methanol goes up. The two different wall force constants seem to have little systematic influence on the osmotic pressure, but our error bars calculated from block averaging over time are probably too low. Literature values from Kohns et al. are significantly lower for higher mole fractions.[47] However, they use the TIP4P/2005 water model and a united-atom methanol model whereas we use the TIP4P water model and the OPLS-AA methanol model, so differences have to be expected. We take the data for Π from both force constants and fit them with a polynomial of third order without zero order term,

$$\Pi_{\text{fit}}(x) = ax^3 + bx^2 + cx. \tag{35}$$

The resulting fit parameters are $a = 24\,070$ bar, $b = -12\,960$ bar, and $c = 2930$ bar; this fit is also shown in figure 8a.

To illustrate the ability of our scheme of matching these pressures we compare (i) HNCGN without constraints and (ii) HNCGN with pressure constraint (p-HNCGN) on the potential

updates. We use the beforementioned polynomial Π_{fit} to set the target pressure for the p-HNCGN method at the given mole fractions. Osmotic pressures for $x > 0.65$ are extrapolated values because of the lack of corresponding OPAS data for higher mole fractions.

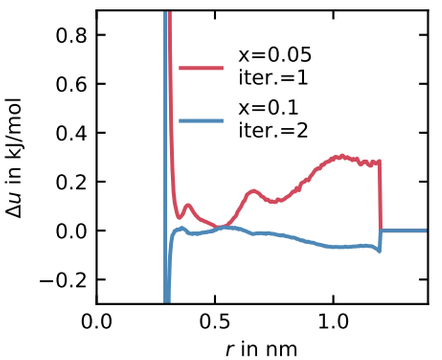


Figure 9: Two exemplary potential updates that show spikes near the core region and jumps before the cut-off arising from using applying the p-HNCGN method without post-processing.

Upon straightforward implementation of the p-HNCGN method, we embrace two problems that can cause unphysical potentials. The first is a jump in the potential update right before the cut-off as seen in the potential update of iteration one for $x = 0.05$ in figure 9. This is related to the pressure constraint since the sudden jump in the potential creates a huge force at that distance, which increases the pressure while having a small effect on the structure. This effect is stronger in the more dilute systems. To prevent this, we implement a simple fix, where the whole potential is shifted, such that it is continuous at the cut-off. However, we do not use it here, since we find the final potentials only have this jump for $x = 0.05$ and it was relatively small. Secondly, there is a negative spike that appears when the pressure is corrected down. It appears at the end of the core region, where the current and target RDF are close to zero. We show an example in figure 9. This can create a very narrow and deep potential well, which crashes the MD simulation. We circumvent this by extrapolating the potential, e.g. by the *power* scheme described in section III A. This scheme, contrary to the standard scheme, substitutes a small part of the repulsive region of the calculated potential. With this tweak, we find the coarse-graining iterations to be stable and to converge after approximately five iterations.

The potentials and distributions obtained from the p-HNCGN method after 20 iterations

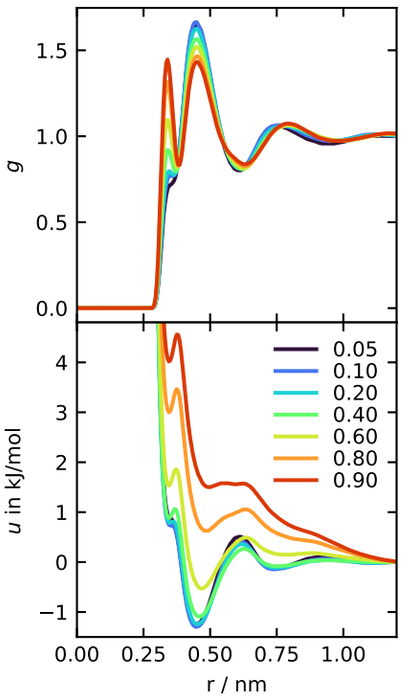


Figure 10: RDFs and potentials for CG methanol obtained with the p-HNCGN method.

Table II: Pressures in bar of CG models of methanol water mixtures from two different methods.

x(methanol)	target	HNCGN	p-HNCGN
0.05	117	67(1)	75(1)
0.1	187	125(3)	129(3)
0.2	260	251(6)	253(6)
0.4	639	611(13)	623(12)
0.6	2291	1232(18)	2240(18)
0.8	6373	2872(22)	6360(23)
0.9	9686	3439(26)	9685(26)

are shown in figure 10. The potentials differ and get more repulsive with a growing mole fraction of methanol. For implicit solvent systems, one cannot expect there to be a common potential for all concentrations, since the influence of the water is incorporated in the potential.

However, we find that for concentrations lower than $x \leq 0.4$ there are only small differences in the potentials. This indicates that below this concentration methanol molecules interact in a similar way, independent from the concentration. Above $x = 0.4$ we find the potentials to become much more repulsive and increasingly showing two distinct minima, one at 0.35 nm and one at 0.45 nm. The pressures obtained from the HNCGN and p-HNCGN potential are shown in 8b and in table II. The error estimates are smaller for the more dilute systems because they have been simulated longer. From the graphic representation, we can see that the target pressure is met very well, where the unconstrained HNCGN method results in too low pressure.

V. CONCLUSION AND OUTLOOK

We have demonstrated the applicability of the Newton methods based on integral equation theory for molecular coarse-graining. The results show that from the OZ equation the Jacobian for a Newton update can be deduced which is equivalent to the IMC Jacobian. The new HNCN methods have properties similar to IMC: rapid convergence in under 10 iterations and the same resulting potential, but it does not require the sampling of the IMC matrix. For short potential cut-offs, which are prevailing in molecular coarse-graining, we find the HNCN Jacobian to have artifacts if the RDF is not sampled on a longer range. They are explained by the numerical deconvolution of the OZ equation. Those artifacts are found to prohibit convergence of the method and distributions have to be evaluated beyond up to around the double cut-off to construct a valid Jacobian. Using HNCN with a longer RDFs makes it as slow as IMC, but when we use a physically motivated scheme to extrapolate the RDFs we obtain four times faster convergence than IMC for water.

An extension to multiple-site representations of molecular liquids based on the RISM-OZ equation has been proposed and demonstrated to work well for CG models of liquid hexane and naphthalene. Again, the resulting potentials are similar to IMC results and convergence speed and accuracy is comparable. This is a first step towards applying iterative integral equation coarse-graining methods in general systems. The same approach could also be used for coarse-graining of polymer melts if the intramolecular distribution function is approximated to be equal between all bead combinations. Existing uses of the OZ equation in polymer coarse-graining might profit from a Newton formulation for computing improved

pair potentials.

The HNCGN method has been used to derive models for a single bead CG methanol model in implicit water at different concentrations. The p-HNCGN method can match the osmotic pressure that was previously determined by OPAS simulations. The implementation of pressure constraints by elimination from the Gauss-Newton problem proves to be a powerful tool to enforce additional conditions on the potentials with only minor numerical pitfalls. The formulation could in this form also be used with the IMC method and would probably be more consistent than the common ramp corrections.

We note that IBI has been extended to inhomogeneous systems which improves the structure of the phase boundary.[48] This extension is non-trivial for the methods studied in this work as homogeneity is a requirement for integral equation theory.

ACKNOWLEDGMENTS

The authors acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 233630050 - TRR 146 on "Multiscale Simulation Methods for Soft Matter Systems". The authors thank Viktor Klippenstein and Swaminath Bharadwaj (both Technische Universität Darmstadt) for discussions on aspects of this work.

DATA AVAILABILITY

The computer code that supports the findings of this study is openly available at https://dx.doi.org/10.5281/zenodo.4290469.

APPENDIX: INTRAMOLECULAR DISTRIBUTION FUNCTION

Here we show how in a molecule of n identical beads the $G_{ij} - g_{ij}$ relates to $G - g$. We first write down the summation over N molecules with indices g and h for the explicit case

$$\rho g_{ij}(r) = \left\langle \frac{1}{N} \sum_{g}^{N} \sum_{h}^{N} (1 - \delta_{gh}) \delta(r - |\boldsymbol{r}_{gi} - \boldsymbol{r}_{hj}|) \right\rangle \tag{36}$$

$$\rho G_{ij}(r) = \left\langle \frac{1}{N} \sum_{g}^{N} \sum_{h}^{N} (1 - \delta_{gh}\delta_{ij}) \delta(r - |\boldsymbol{r}_{gi} - \boldsymbol{r}_{hj}|) \right\rangle. \tag{37}$$

Here, $\boldsymbol{r}_{gi}$ is the position of site i on molecule g. Therefore we obtain for the difference of the two distribution functions

$$\begin{aligned}\rho(G_{ij}(r) - g_{ij}(r)) &= \left\langle \frac{1}{N}\sum_{g}^{N}\sum_{h}^{N} -\delta_{gh}(\delta_{ij}-1)\delta(r-|\boldsymbol{r}_{gi}-\boldsymbol{r}_{hj}|)\right\rangle \\ &= (1-\delta_{ij})\langle\delta(r-|\boldsymbol{r}_i-\boldsymbol{r}_j|)\rangle_{\mathrm{mol}}.\end{aligned} \tag{38}$$

$\langle\ \rangle_{\mathrm{mol}}$ indicates an average over all molecules. This reflects that the intramolecular distribution function is independent of intermolecular correlations. If all combinations i and j have the same distance distribution $e(r)$ we can simplify to

$$\rho(G_{ij}(r) - g_{ij}(r)) = (1-\delta_{ij})e(r) \tag{39}$$

Now we assume molecules with n equal sites. The density of those sites is $n\rho$. To obtain the site independent expression we need to sum over all sites

$$\begin{aligned}n\rho(G(r) - g(r)) &= \left\langle \frac{1}{nN}\sum_{g}^{N}\sum_{h}^{N}\sum_{i}^{n}\sum_{j}^{n} -\delta_{gh}(\delta_{ij}-1)\delta(r-|\boldsymbol{r}_{gi}-\boldsymbol{r}_{hj}|)\right\rangle \\ &= \left\langle \frac{1}{n}\sum_{i}^{n}\sum_{j}^{n}(1-\delta_{ij})\delta(r-|\boldsymbol{r}_i-\boldsymbol{r}_j|)\right\rangle_{\mathrm{mol}}.\end{aligned} \tag{40}$$

Again we assume all distance distributions to be equal $e(r)$ and obtain

$$n\rho(G(r) - g(r)) = (n-1)e(r). \tag{41}$$

By comparing equations (39) and (41) we find

$$\rho(G_{ij}(r) - g_{ij}(r)) = (1-\delta_{ij})\frac{n}{n-1}\rho(G(r) - g(r)) \tag{42}$$

Note that this is an approximation, if not all sites have the same distance distribution. For example in a rectangular molecule G_{12} would be different from G_{13}. Equation (42) is therefore only exact for molecules where all sites have the same internal distances, i.e. a dumbbell, equilateral triangle, and tetrahedron.

REFERENCES

[1]A. Lyubartsev, A. Mirzoev, L. Chen, and A. Laaksonen, Faraday Discuss. **144**, 43 (2010).

[2]R. L. Henderson, Phys. Lett. A **49**, 197 (1974).

[3]F. Frommer, M. Hanke, and S. Jansen, J. Math. Phys. **60**, 093303 (2019).

[4]D. Reith, M. Pütz, and F. Müller-Plathe, J. Comput. Chem. **24**, 1624 (2003).

[5]A. P. Lyubartsev and A. Laaksonen, Phys. Rev. E **52**, 3730 (1995).

[6]Y.-L. Wang, A. Lyubartsev, Z.-Y. Lu, and A. Laaksonen, Phys. Chem. Chem. Phys. **15**, 7701 (2013).

[7]B. L. Peters, K. M. Salerno, A. Agrawal, D. Perahia, and G. S. Grest, J. Chem. Theory Comput. **13**, 2890 (2017).

[8]A. Gooneie, S. Schuschnigg, and C. Holzer, Polymers **9**, 16 (2017).

[9]A. P. Lyubartsev, A. Naômé, D. P. Vercauteren, and A. Laaksonen, J. Chem. Phys. **143**, 243120 (2015).

[10]H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole, and S. J. Marrink, Wiley Interdiscip. Rev. Comput. Mol. Sci. **4**, 225 (2014).

[11]D. Rosenberger, M. Hanke, and N. F. A. van der Vegt, Eur. Phys. J. Spec. Top. **225**, 1323 (2016).

[12]F. Delbary, M. Hanke, and D. Ivanizki, Inverse Probl. Sci. Eng. **28**, 1166 (2020).

[13]C. D. Williams and M. Lísal, 2D Mater. **7**, 025025 (2020).

[14]T. Murtola, E. Falck, M. Karttunen, and I. Vattulainen, J. Chem. Phys. **126**, 075101 (2007).

[15]S. Mortezazadeh, Y. Jamali, H. Naderi-Manesh, and A. P. Lyubartsev, PLOS One **14**, e0214673 (2019).

[16]Q. Wang, D. J. Keffer, D. M. Nicholson, and J. B. Thomas, Phys. Rev. E **81**, 061204 (2010).

[17]S. Jain, S. Garde, and S. K. Kumar, Ind. Eng. Chem. Res. **45**, 5614 (2006).

[18]E. J. Sambriski, G. Yatsenko, M. A. Nemirovskaya, and M. G. Guenza, J. Chem. Phys. **125**, 234902 (2006).

[19]A. J. Clark, J. McCarty, and M. G. Guenza, J. Chem. Phys. **139**, 124906 (2013).

[20]M. G. Guenza, J. Phys.: Conf. Ser. **640**, 012009 (2015).

[21]M. G. Guenza, M. Dinpajooh, J. McCarty, and I. Y. Lyubimov, J. Phys. Chem. B **122**, 10257 (2018).

[22]S. Y. Mashayak, L. Miao, and N. R. Aluru, J. Chem. Phys. **148**, 214105 (2018).

[23]D. Levesque, J. J. Weis, and L. Reatto, Phys. Rev. Lett. **54**, 451 (1985).

[24]M. Heinen, J. Comput. Chem. **39**, 1531 (2018).

[25]J. Ghosh and R. Faller, Mol. Simul. **33**, 759 (2007).

[26]T. C. Moore, C. R. Iacovella, and C. McCabe, J. Chem. Phys. **140**, 224104 (2014).

[27]H. Wang, C. Junghans, and K. Kremer, Eur. Phys. J. E **28**, 221 (2009).

[28]V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko, J. Chem. Theory Comput. **5**, 3211 (2009).

[29]A. Mirzoev and A. P. Lyubartsev, J. Chem. Theory Comput. **9**, 1512 (2013).

[30]A. Das and H. C. Andersen, J. Chem. Phys. **132**, 164106 (2010).

[31]D. Rosenberger and N. F. A. van der Vegt, Phys. Chem. Chem. Phys. **20**, 6617 (2018).

[32]M. R. DeLyser and W. G. Noid, J. Chem. Phys. **147**, 134111 (2017).

[33]A. Moradzadeh, M. H. Motevaselian, S. Y. Mashayak, and N. R. Aluru, J. Chem. Theory Comput. **14**, 3252 (2018).

[34]P. Ganguly, D. Mukherji, C. Junghans, and N. F. A. van der Vegt, J. Chem. Theory Comput. **8**, 1802 (2012).

[35]T. E. de Oliveira, P. A. Netz, K. Kremer, C. Junghans, and D. Mukherji, J. Chem. Phys. **144**, 174106 (2016).

[36]J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids* (Elsevier, 1990).

[37]K. S. Schweizer and J. G. Curro, in *Atomistic Modeling of Physical Properties*, Advances in Polymer Science, edited by L. Monnerie and U. W. Suter (Springer, Berlin, Heidelberg, 1994) pp. 319–377.

[38]S. Y. Mashayak, M. N. Jochum, K. Koschke, N. R. Aluru, V. Rühle, and C. Junghans, PLOS One **10**, e0131754 (2015).

[39]W. Gander, M. J. Gander, and F. Kwok, *Scientific Computing - An Introduction Using Maple and MATLAB*, Texts in Computational Science and Engineering (Springer International Publishing, 2014).

[40]H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, J. Phys. Chem. **91**, 6269 (1987).

[41]W. L. Jorgensen and J. Tirado-Rives, J. Am. Chem. Soc. **110**, 1657 (1988).

[42]L. S. Dodda, I. Cabeza de Vaca, J. Tirado-Rives, and W. L. Jorgensen, Nucleic Acids Res. **45**, W331 (2017).

[43]M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, SoftwareX **1-2**, 19 (2015).

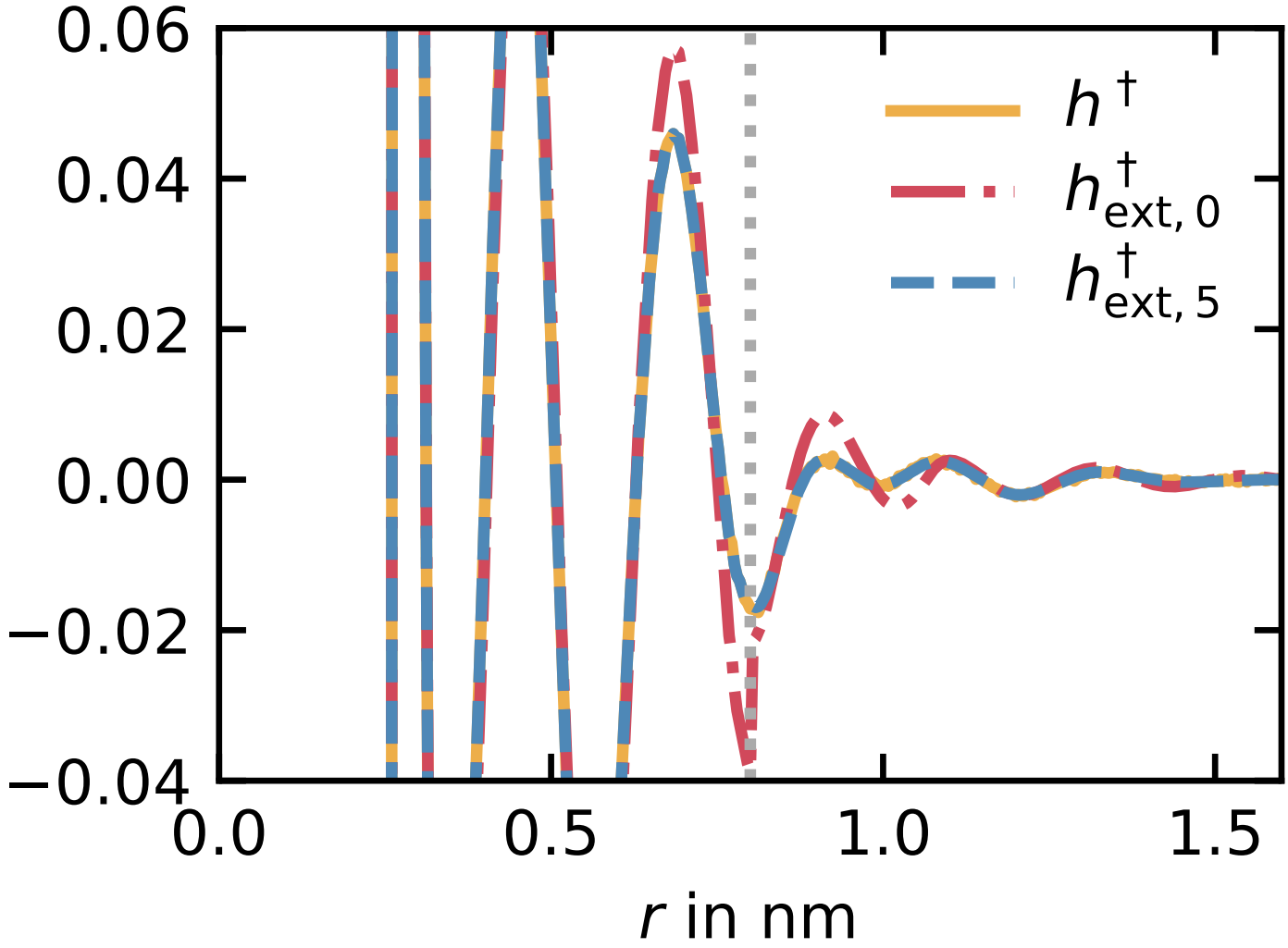
[44]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).

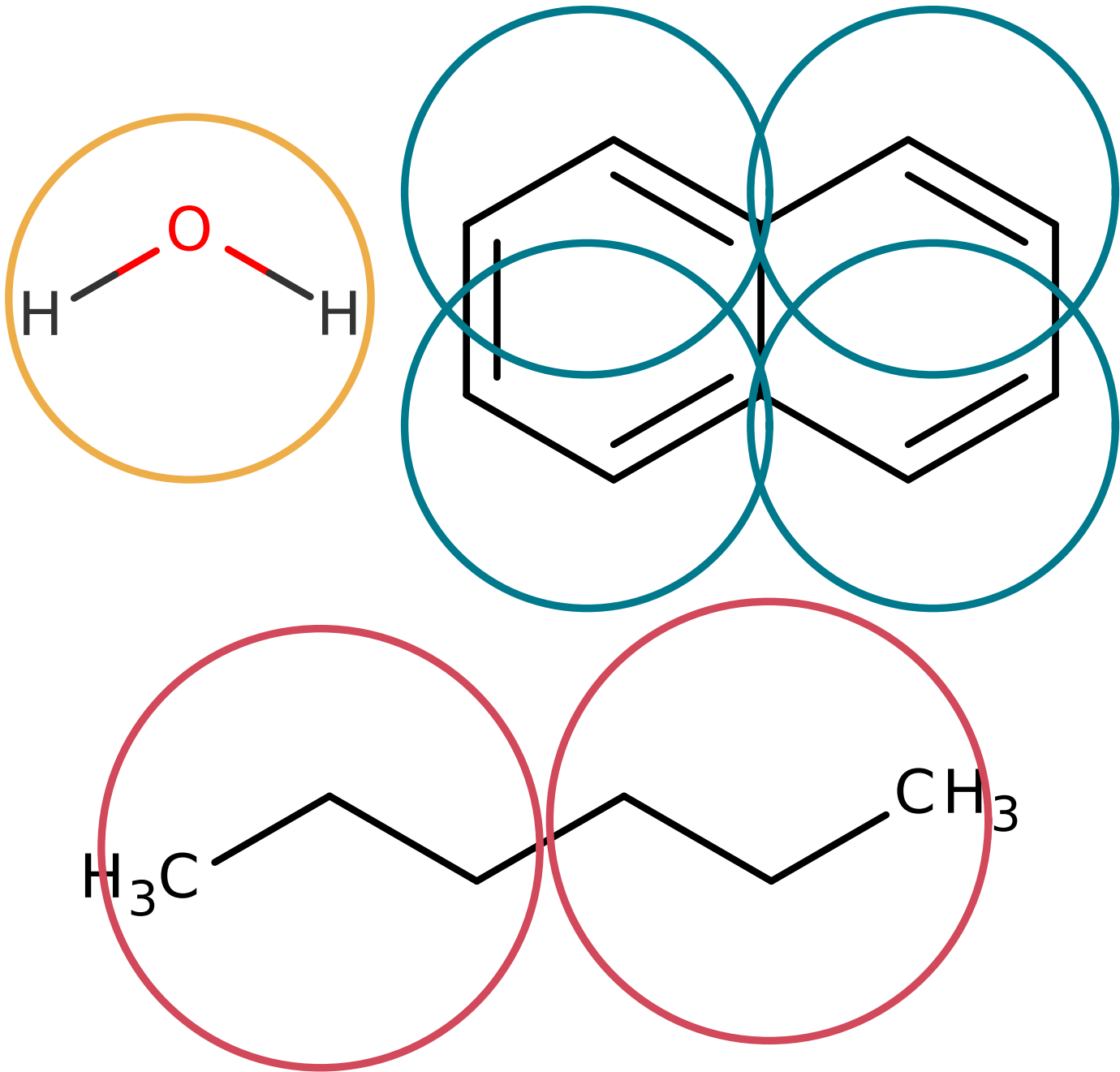
[45]M. Kohns, S. Reiser, M. Horsch, and H. Hasse, J. Chem. Phys. **144**, 084112 (2016).

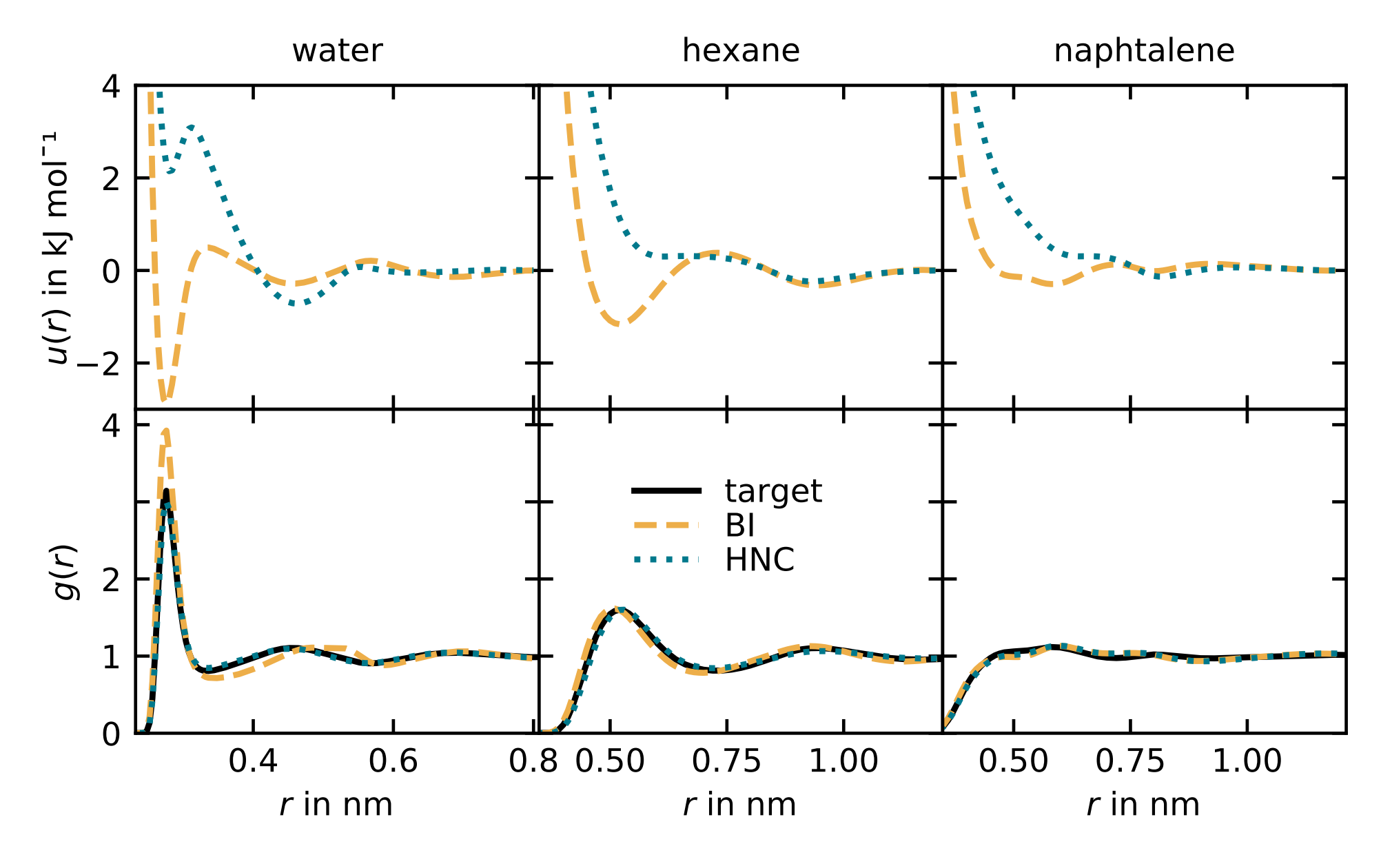
[46]P. T. Cummings and G. Stell, Mol. Phys. **46**, 383 (1982).

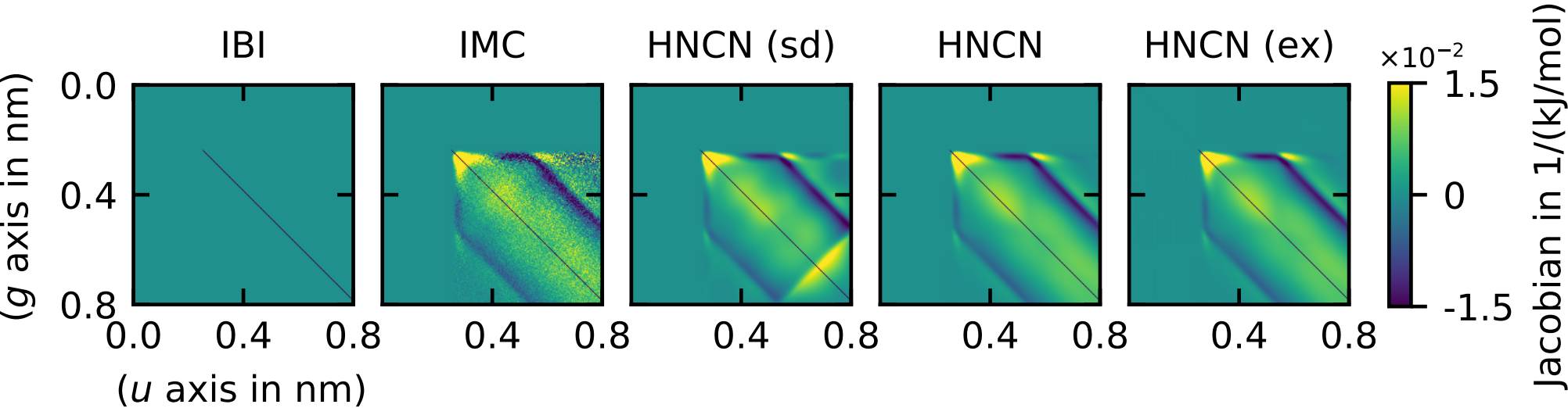
[47]M. Kohns, M. Horsch, and H. Hasse, J. Chem. Phys. **147**, 144108 (2017).

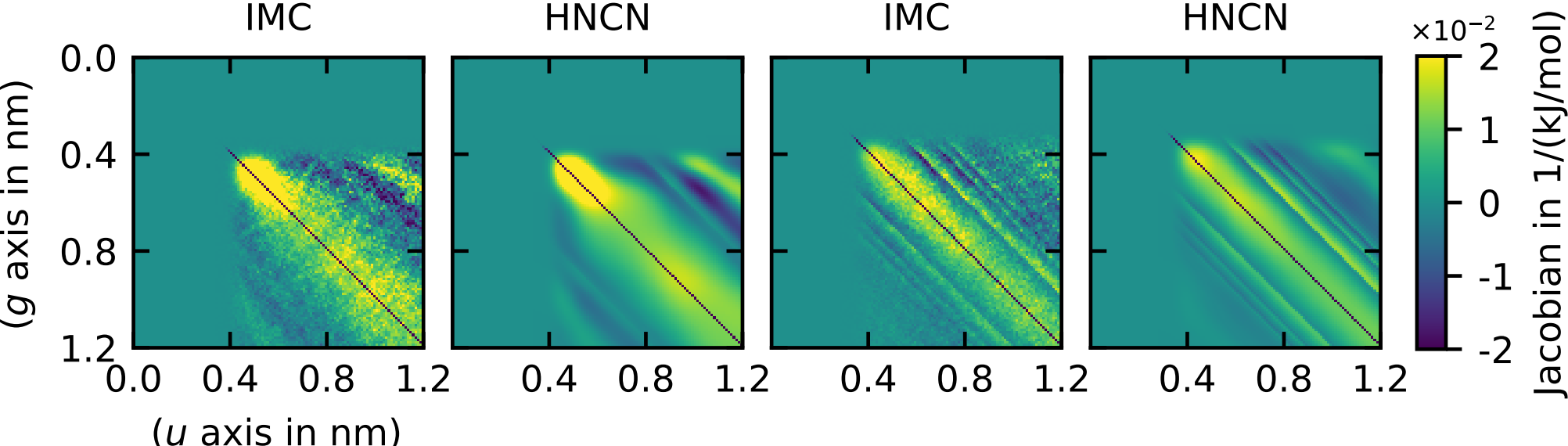
[48]M. Jochum, D. Andrienko, K. Kremer, and C. Peter, J. Chem. Phys. **137**, 064102 (2012).

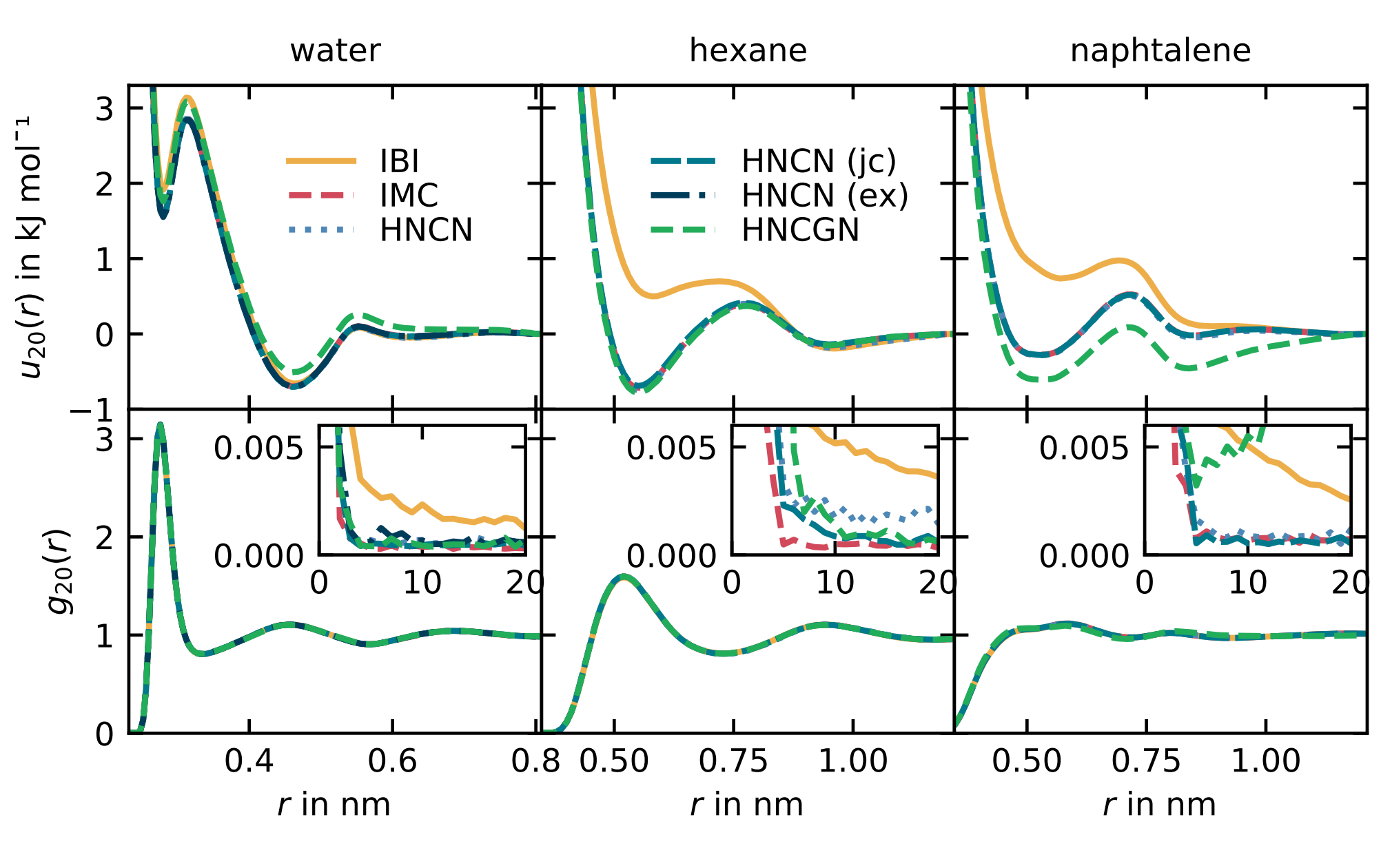

0.06
0.04
0.02
0.00
−0.02
−0.04
0.0
0.5
1.0
1.5
r in nm
h†
h†ext, 0
h†ext, 5

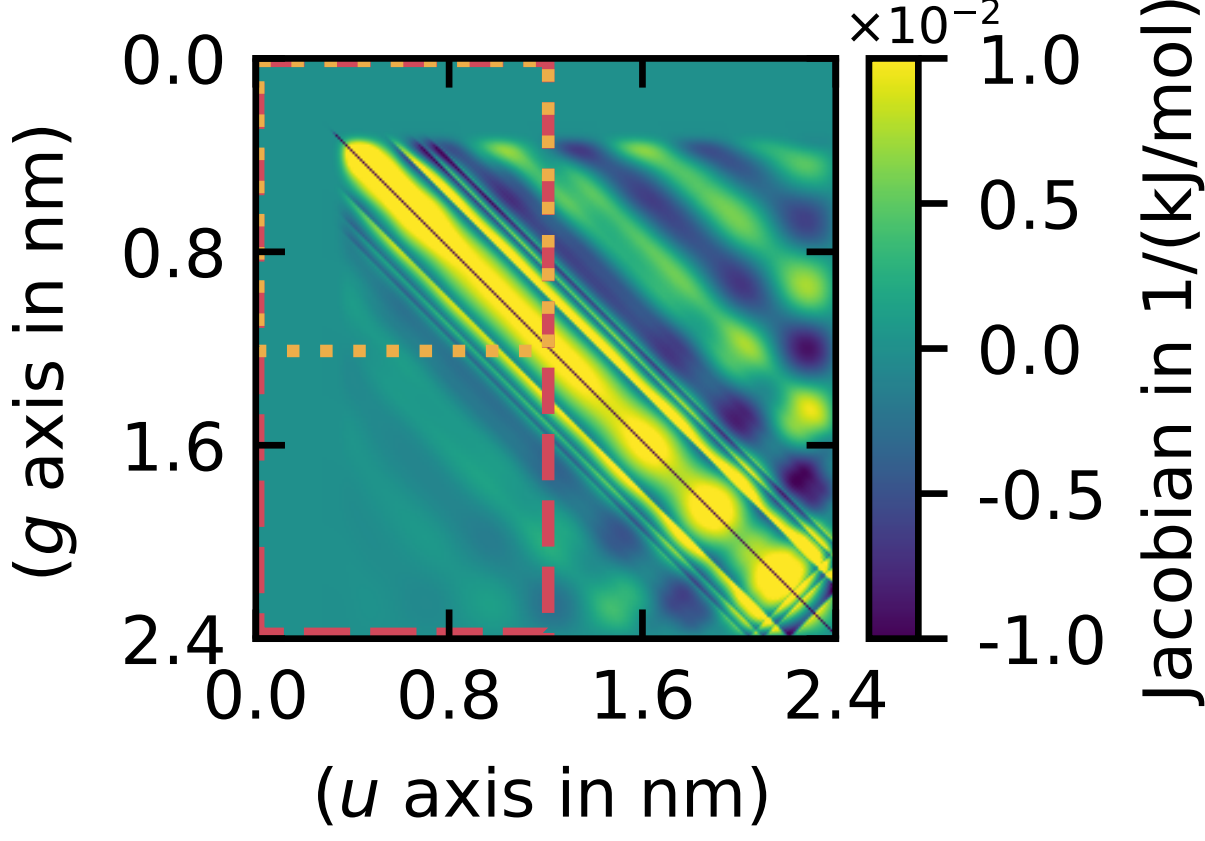

O
H
H
H3C
CH3

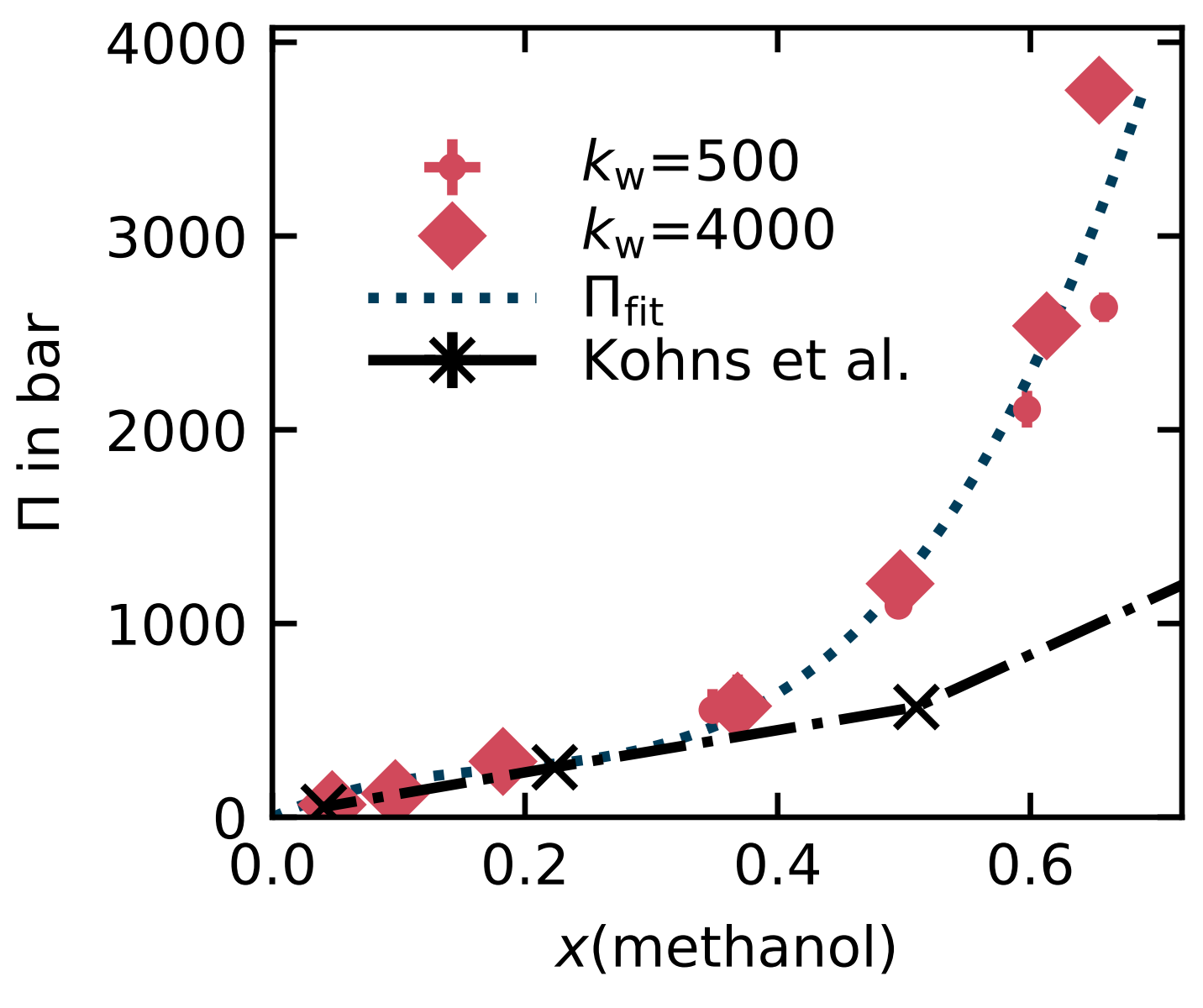

water
hexane
naphtalene
u(r) in kJ mol⁻¹
g(r)
target
BI
HNC
r in nm
r in nm
r in nm
4
2
0
−2
4
2
1
0
0.4
0.6
0.8
0.50
0.75
1.00
0.50
0.75
1.00

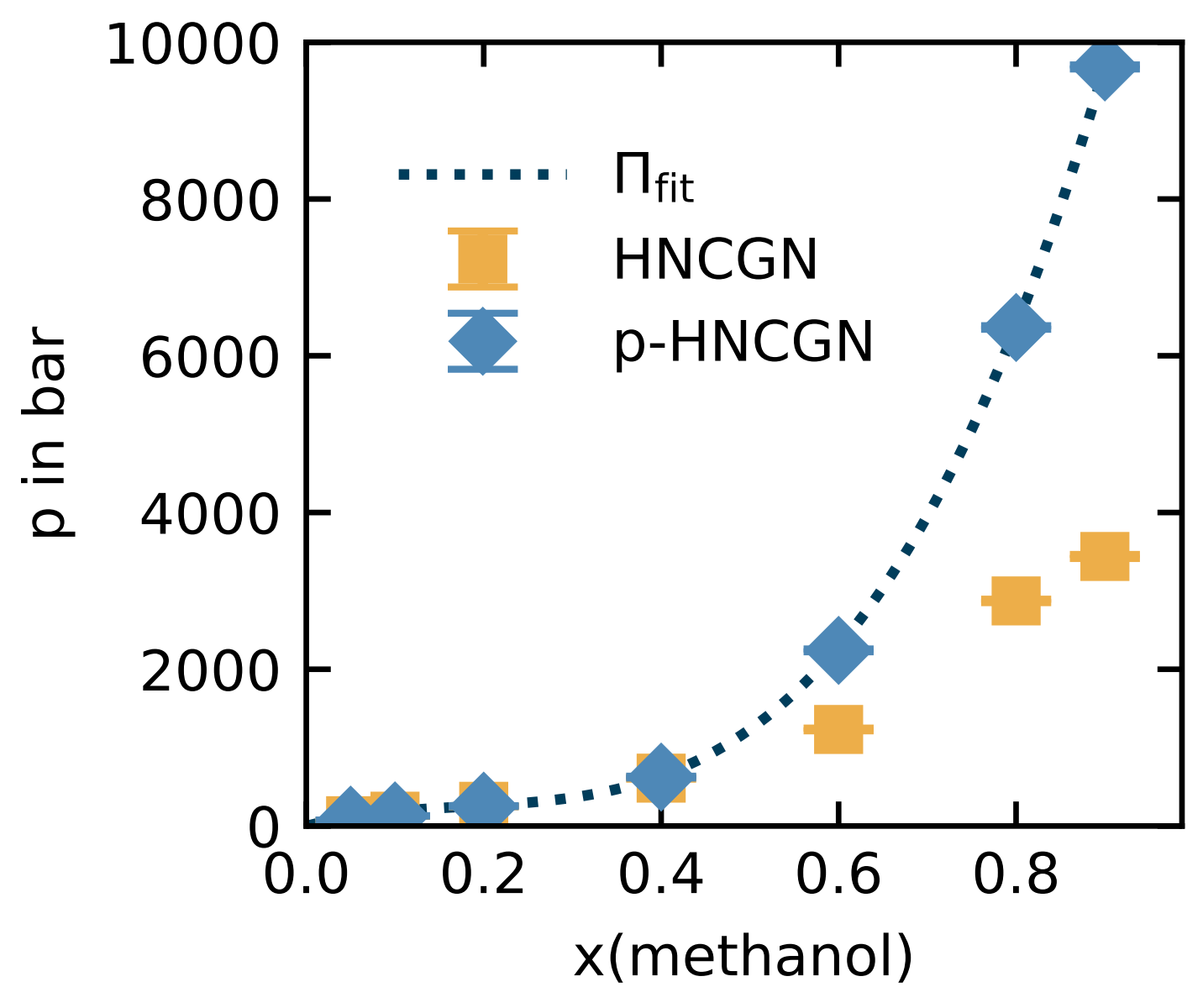

IBI
IMC
HNCN (sd)
HNCN
HNCN (ex)
0.0
0.4
0.8
(g axis in nm)
(u axis in nm)
×10⁻²
1.5
0
-1.5
Jacobian in 1/(kJ/mol)

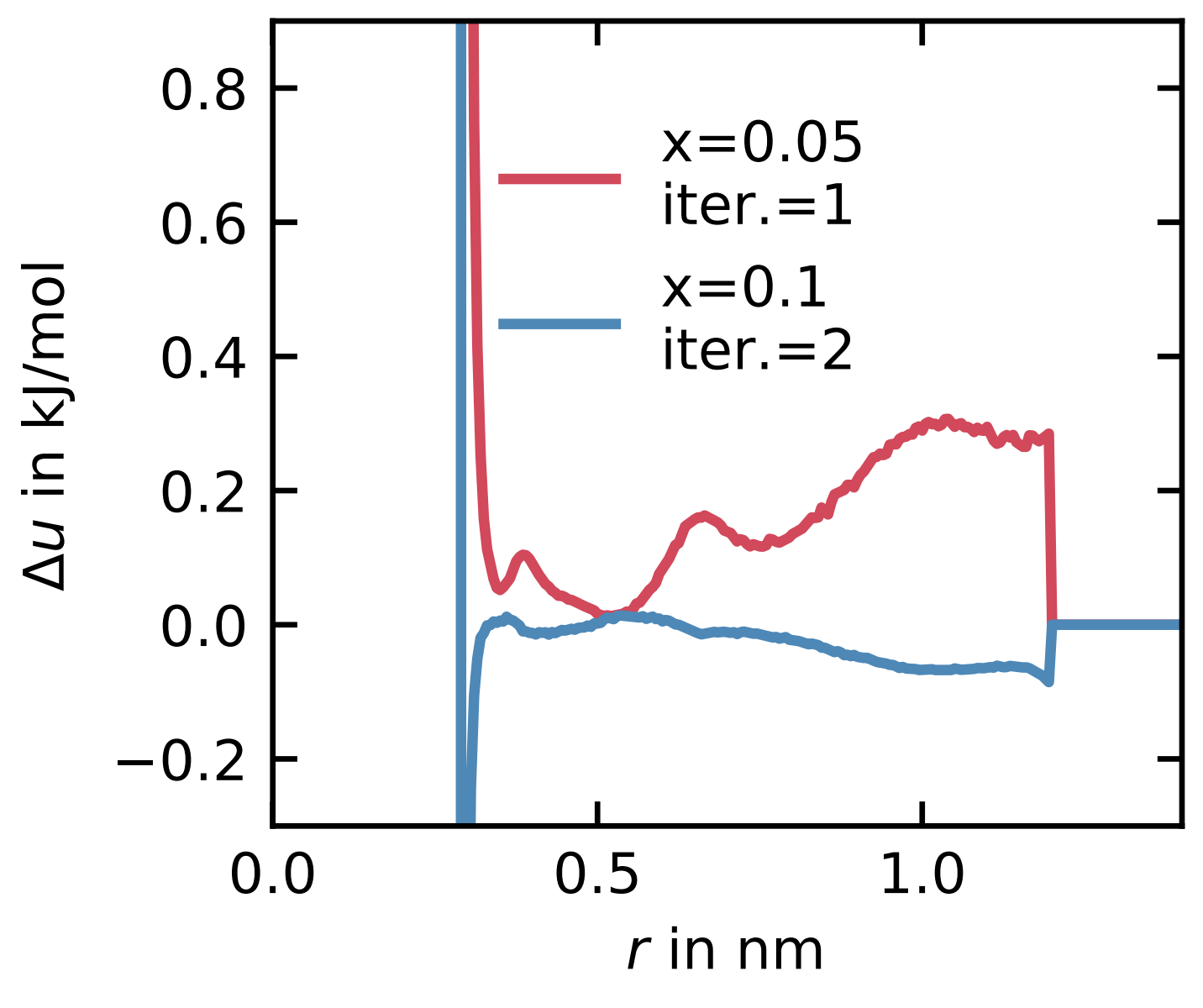

IMC
HNCN
IMC
HNCN
×10⁻²
2
1
0
-1
-2
Jacobian in 1/(kJ/mol)
0.0
0.4
0.8
1.2
(g axis in nm)
0.0
0.4
0.8
1.2
0.4
0.8
1.2
0.4
0.8
1.2
0.4
0.8
1.2
(u axis in nm)

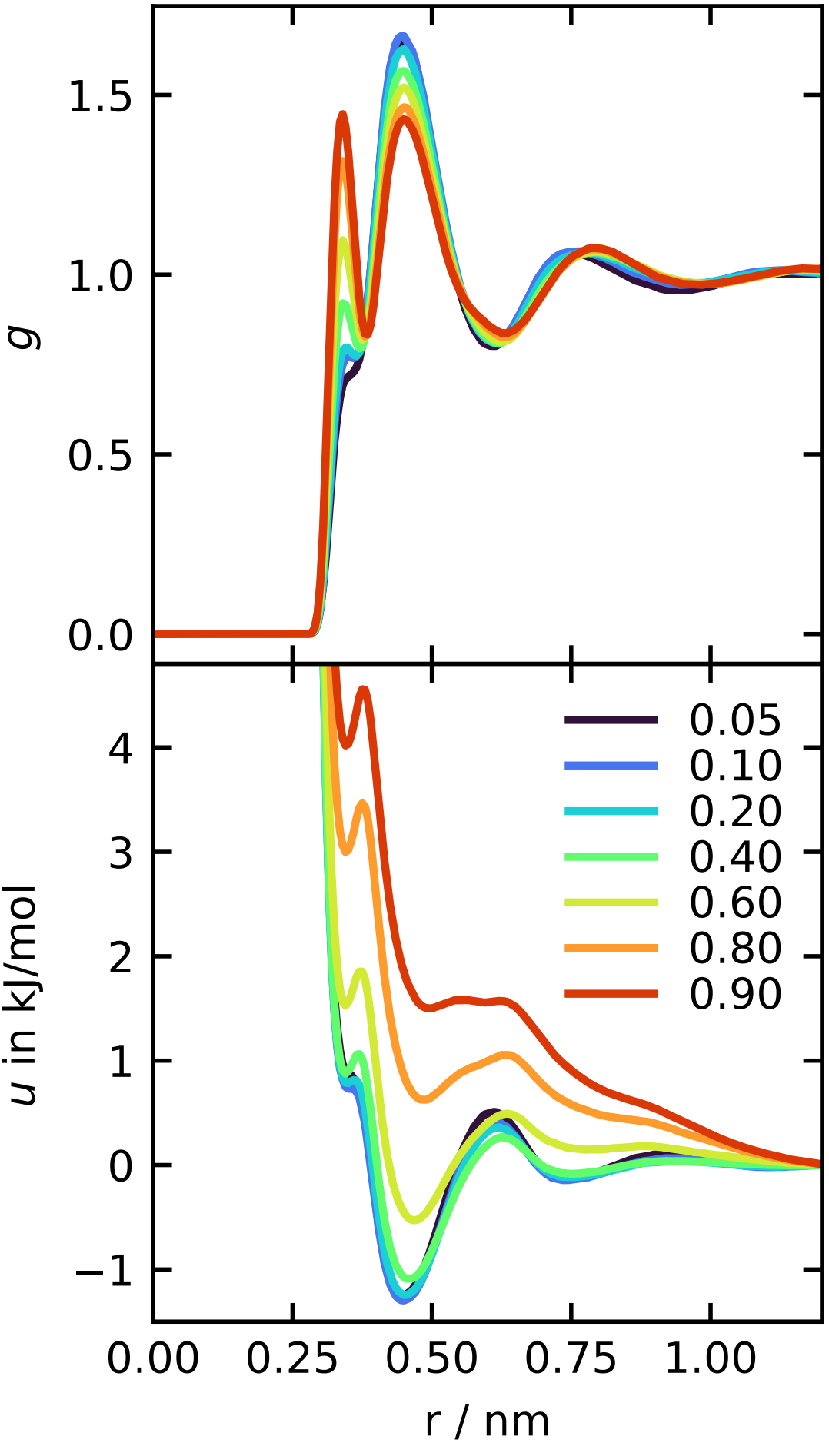

water
hexane
naphtalene
IBI
IMC
HNCN
HNCN (jc)
HNCN (ex)
HNCGN
$u_{20}(r)$ in kJ mol^{-1}
$g_{20}(r)$
r in nm
r in nm
r in nm


0.0
0.8
1.6
2.4
(g axis in nm)
0.0
0.8
1.6
2.4
(u axis in nm)
×10−2
1.0
0.5
0.0
-0.5
-1.0
Jacobian in 1/(kJ/mol)


4000
3000
2000
1000
0
Π in bar
0.0
0.2
0.4
0.6
x(methanol)
k_w=500
k_w=4000
Π_fit
Kohns et al.


10000
8000
6000
4000
2000
0
p in bar
0.0
0.2
0.4
0.6
0.8
x(methanol)
Π_fit
HNCGN
p-HNCGN


0.8
0.6
0.4
0.2
0.0
−0.2
Δu in kJ/mol
x=0.05
iter.=1
x=0.1
iter.=2
0.0
0.5
1.0
r in nm


g
1.5
1.0
0.5
0.0
u in kJ/mol
4
3
2
1
0
−1
0.05
0.10
0.20
0.40
0.60
0.80
0.90
0.00
0.25
0.50
0.75
1.00
r / nm