Annexe bibliographique

J Stat Phys
DOI 10.1007/s10955-007-9464-0

Exact Results for Thermodynamics of the Hydrogen Plasma: Low-Temperature Expansions Beyond Saha Theory

A. Alastuey · V. Ballenegger · F. Cornu · Ph.A. Martin

Received: 20 April 2007 / Accepted: 6 November 2007

Abstract We study hydrogen in the Saha regime, within the physical picture in terms of a quantum proton-electron plasma. Long ago, Saha showed that, at sufficiently low densities and low temperatures, the system behaves almost as an ideal mixture made with hydrogen atoms in their groundstate, ionized protons and ionized electrons. More recently, that result has been rigorously proved in some scaling limit where both temperature and density vanish. In that Saha regime, we derive exact low-temperature expansions for the pressure and internal energy, where density ρ is rescaled in units of a temperature-dependent density ρ^* which controls the cross-over between full ionization ($\rho \ll \rho^*$) and full atomic recombination ($\rho \gg \rho^*$). Each term reduces to a function of ρ/ρ^* times temperature-dependent functions which decay exponentially fast when temperature T vanishes. Scaled expansions are ordered with respect to the corresponding decay rates. Leading terms do reduce to ideal contributions obtained within Saha theory. We consistently compute all corrections which are exponentially smaller by a factor $\exp(\beta E_H)$ at most, where E_H is the negative groundstate energy of a hydrogen atom and $\beta = 1/(k_B T)$. They include all effects arising from both the Coulomb potential and the quantum nature of the particles: excitations of atoms H, formation of molecules H_2, ions H_2^+ and H^-, thermal and pressure ionization, plasma polarization, screening, interactions between atoms and ionized charges, etc. Scaled low-temperature

---

A. Alastuey (✉)
Laboratoire de Physique, Université de Lyon, ENS Lyon, CNRS, 46 allée d'Italie, 69364 Lyon Cedex 07, France
e-mail: angel.alastuey@ens-lyon.fr

V. Ballenegger
Institut UTINAM, Université de Franche-Comté, CNRS, 16 route de Gray, 25030 Besançon Cedex, France

F. Cornu
Laboratoire de Physique Théorique, Université Paris-Sud, CNRS, Bâtiment 210, 91405 Orsay Cedex, France

Ph.A. Martin
Institut de Théorie des Phénomènes Physiques, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

expansions can be viewed as partial resummations of usual virial expansions up to arbitrary high orders in the density.

1 Introduction

Hydrogen is an important element, both at a theoretical level and for practical purposes. Here, we consider a non-relativistic quantum hydrogen plasma, made of protons and electrons with respective masses m_p and m_e, which interact *via* the familiar $1/r$-Coulomb potential (see Sect. 2.1). As far as thermodynamic properties of that system are concerned, an exact calculation at finite temperature T and finite density ρ, remains far beyond present human abilities. Nonetheless, by exploiting the exact knowledge of the spectrum of hydrogen atom and using Morita's method [51], Ebeling [23] first computed low-density expansions for pressure and free energy up to order ρ^2 at fixed non-zero temperature, in a closed analytical form (see also Ref. [39]). When ρ goes to zero, the system becomes fully ionized (see Ref. [43] for a rigorous proof). At order ρ^2, the recombination of a small fraction of charges into hydrogen atoms is exactly taken into account. Such low-density expansions have been more recently completed up to order $\rho^{5/2}$ [4–7]. Those results have been checked afterwards in Ref. [37], and their high-temperature form in the one-component case does coincide with that derived in Ref. [20]. In the opposite limit where ρ goes to infinity at zero temperature, the system behaves as a mixture of free Fermi gases, and expansions in inverse powers of ρ have been calculated (see Refs. [29] and [50] for the first calculations in the one-component case, and also Ref. [39] for similar works or extensions).

The previous exact asymptotic expansions are suitable for regimes where the system is almost fully ionized. The purpose of the present paper is to derive a similar expansion in the so-called Saha regime, where a non-vanishing fraction of charges is recombined into hydrogen atoms. That regime was introduced long ago [60] in the framework of the chemical picture. Assuming that the system is an ideal mixture of protons, electrons, and hydrogen atoms, its composition is then determined by applying the usual mass action law [26]. More recently, by starting from the physical description of the system in terms of a quantum plasma, it has been proved through successive works by Fefferman [27], Lieb et al. [18], Macris and Martin [45], that Saha approach is asymptotically exact in a scaling limit mixing the temperature and the chemical potential (see Sect. 2.2). As argued in Sect. 2.3, that limit defines quite diluted and low temperature conditions, namely the Saha regime, under which non-ideal contributions are small perturbations. In order to compute the corresponding contributions, we consider a formalism that combines the path integral representation of a quantum gas to familiar Mayer diagrammatics (see Sect. 2.4). Our key starting point in that framework is the so-called screened cluster expansion (SCE) of particle densities in terms of fugacities [8], which turns to be quite appropriate for studying recombined phases as illustrated in Refs. [10–12] (dielectric response of an atomic gas) or [9] (partial screening of van der Waals forces by free charges). The physical content of SCE is close to ideas first introduced by Rogers [54] for describing atomic or molecular recombination within the physical picture. In that approach, virial coefficients are numerically estimated within a priori modelizations, which incorporate quantum effects at short distances and classical Debye screening at large distances. The corresponding so-called ACTEX method has been developed through successive works [55, 56, 58, 59]. It has also been applied to hydrogen [57], with quite good results at low and moderate densities as described in Ref. [48]. Nevertheless, in the Saha regime, exact asymptotic expansions with analytical prescriptions for computing the successive terms, have not been derived within ACTEX method.

In Sect. 3, using the parametrization of chemical potential in terms of temperature introduced in Ref. [45], we show that every contribution in SCE of particle densities, decays exponentially fast when T goes to zero. Thanks to available inequalities for the spectrum of the considered Coulomb Hamiltonian, we extract the leading terms which do arise from free (ionized) protons and electrons, as well as from atoms H in their groundstate with energy $E_H = -me^4/(2\hbar^2)$ where m is the reduced mass $m = m_p m_e/(m_p + m_e)$. Next corrections are ordered with respect to their decay rate in the zero-temperature limit. They account for a large variety of physical effects: plasma polarization, formation of molecules H_2, ions H_2^+ and H^-, interactions between ionized charges and Hydrogen atoms. Such corrections are defined without any ambiguity or *a priori* modelizations, so they do not depend on any adjustable phenomenological parameter. In particular, SCE provides well-behaved expressions for the partition functions of a molecule H_2 or ions H_2^+ and H^- in the vacuum, which are the generalizations of quantum virial functions for the hydrogen atom [23] to more complex entities. Notice also that ionic contributions to charge neutrality or screening are consistently incorporated, as it should.

The equation of state (EOS) is derived by using thermodynamic identities in Sect. 4. This leads to our main result, namely scaled low-temperature (SLT) expansion of the pressure P around ideal Saha pressure

$$\beta P/\rho^* = \beta P_{\text{Saha}}/\rho^* + \sum_{k=1}^{\infty} \beta P_k/\rho^*, \tag{1.1}$$

considered as a function of the temperature and of the dimensionless density variable ρ/ρ^* where ρ is the electron number density (which is equal to the proton number density by neutrality). The temperature-dependent reference density ρ^* defined by

$$\rho^* = \frac{\exp(\beta E_H)}{2(2\pi\lambda_{pe}^2)^{3/2}} \quad \text{with } \lambda_{pe} = (\beta\hbar^2/m)^{1/2}, \tag{1.2}$$

determines the cross-over between full ionization for $\rho \ll \rho^*$, and full recombination for $\rho \gg \rho^*$ (see Sect. 2.2). Since E_H is negative, $E_H \simeq -13.6$ eV, ρ^* decays exponentially fast at low temperatures. In expansion (1.1), it is convenient to express the pressure in units of $k_B T \rho^*$ which turns out to be the natural reference pressure in the Saha regime. Then, each term in (1.1) is dimensionless. The first term is the usual Saha pressure expressed in terms of ρ/ρ^*

$$\beta P_{\text{Saha}}/\rho^* = \rho/\rho^* + (1 + 2\rho/\rho^*)^{1/2} - 1. \tag{1.3}$$

We see indeed that for $\rho \ll \rho^*$, the system becomes fully ionized ($\beta P_{\text{Saha}} \sim 2\rho$), whereas for $\rho \gg \rho^*$ all ionized charges recombine into neutral hydrogen atoms ($\beta P_{\text{Saha}} \sim \rho$). Each term in expansion (1.1) beyond that leading ideal contribution has the form of a non-linear function of ratio ρ/ρ^*, times a temperature-dependent function $h_k(\beta)$ (or possibly a polynomial in the $h_l(\beta)$, $l \le k$). The $h_k(\beta)$ decay exponentially fast when T vanishes and are ordered with respect to their decay rates $h_k(\beta) \sim e^{-\beta\delta_k}$, $0 < \delta_1 < \delta_2 < \cdots$. Hence the expansion (1.1) is organized as a series of exponential terms with increasingly faster exponential decay as $T \to 0$. The h_k-functions and their decay rates are governed by a balance between energy and entropy involving the ground-state energy $E^{(0)}_{N_p,N_e}$ of Coulomb Hamiltonian H_{N_p,N_e} for N_p protons and N_e electrons in mutual interaction. We determine the pressure in the Saha regime by computing exactly all terms in expansion (1.1) smaller than leading ideal contribution $\beta P_{\text{Saha}}/\rho^*$ of order 1 by exponentially decaying factors of maximum order $\exp(\beta E_H)$. We find

$$\begin{aligned}
&\beta P_k/\rho^* = (\text{function of } \rho/\rho^*) \times h_k(\beta), \quad h_k(\beta) \sim e^{-\beta\delta_k},\ k = 1,2,3,4, \\
&\beta P_5/\rho^* = (\text{function of } \rho/\rho^*) \times [h_1(\beta)]^2
\end{aligned} \tag{1.4}$$

with (δ_k in electronvolt units, $E_{2,1}^{(0)} = E_{H_2^+}$, $E_{1,2}^{(0)} = E_{H^-}$, $E_{2,2}^{(0)} = E_{H_2}$)

$$\begin{aligned}
\delta_1 &= |E_H|/2 \simeq 6.8, \\
\delta_2 &= |3E_H - E_{H_2}| \simeq 9.1, \\
\delta_3 &= 3|E_H|/4 \simeq 10.2, \\
\delta_4 &= |2E_H - E_{H_2^+}| \simeq 11.0.
\end{aligned} \tag{1.5}$$

The explicit forms of the density-dependent functions and of the $h_k(\beta)$ can be found in Sects. 4.1 and 4.2 together with a discussion of their interpretation and relative importance in different density and temperature regimes. In short, the h_k-functions incorporate various corrections to the ideal Saha term which can be described by

- $h_1(\beta)$: plasma polarization around ionized charges
- $h_2(\beta)$: formation of H_2 molecules and atom-atom interactions
- $h_3(\beta)$: atomic excitations and interactions between ionized charges
- $h_4(\beta)$: formation of H_2^+ and H^- ions, atom-charge interactions, and screening of atomic groundstate

The construction of SLT expansion (1.1), defined by taking the zero-temperature limit at fixed ratio ρ/ρ^*, is itself an important new result. It provides a non-trivial structure for the various corrections to ideal Saha pressure, which are properly ordered in that scaling limit. It turns out that keeping only the first correction $\beta P_1/\rho^*$, is equivalent to a modification of the Saha ionization rate which has been derived previously by several authors (see e.g. [41] and references quoted in [39]). To our knowledge, next terms $\beta P_k/\rho^*$ ($2 \leq k \leq 5$) are entirely new, and do not have counterparts in the literature. We provide their exact expressions, which involve suitably truncated few-body partition functions. Two-body truncated partition functions are merely related to quantum virial functions first introduced by Ebeling (see e.g. [39] and references quoted therein). Three- and four-body truncated partition functions are introduced and defined here for the first time. Previous terms ($2 \leq k \leq 5$) also account, beyond standard calculations, for interactions between recombined entities as well as screening effects. For instance, contributions of atom-atom interactions in $\beta P_2/\rho^*$ are evaluated without any a priori modelization, while screening of atomic groundstate embedded in $\beta P_4/\rho^*$ incorporates further corrections to the familiar Debye shift.

Corrections in SLT expansion (1.1) are ordered with respect to their decay rates when the temperature vanishes at fixed ratio ρ/ρ^*. The behavior of such corrections along a given low-temperature isotherm when the density is varied, displays some interesting physics. For very small densities $\rho \ll \rho^*$, all density-dependent functions in front of the $h_k's$ can be expanded in powers of ρ. Then, we retrieve the well-known virial expansion at fixed temperature up to order ρ^2 included (see Sect. 4.2). In particular, the leading correction of order $\rho^{3/2}$ is the familiar classical Debye term arising from the polarization contribution $\beta P_1/\rho^*$. When the density is increased, virial density-expansion can no longer be used, but SLT expansion still works and accounts for non-perturbative effects with respect to finite values of ρ/ρ^*. Up to moderate densities $\rho \simeq \rho^*$, $\beta P_1/\rho^*$ remains the leading correction to ideal Saha terms. Interestingly, that polarization contribution is reduced at higher densities $\rho \gg \rho^*$ because most protons and electrons are recombined into atoms H. Then, molecular contributions embedded in $\beta P_2/\rho^*$ provide the first correction to Saha pressure, since they also overcome contributions of atom-atom interactions, at least for a sufficiently low temperature isotherm. Ultimately, they are responsible for the breakdown of expansion (1.1) at too large densities. Our results clearly provide a better analytical knowledge of the thermodynamics in an extended part of the phase diagram, as illustrated by the validity domain drawn in Fig. 12 of Sect. 4.3. The SLT expansions can be viewed as infinite resummations of low-density expansions.

We emphasize again that the EOS (1.1) incorporates the screening effects in a coherent and consistent way for the whole range of densities $\rho \ll \rho^*$ (strongly ionized gas) and $\rho \gg \rho^*$ (recombined gas). When the interaction is Coulombic, one has to face the divergence of the sum of bound state contributions to the partition function of an isolated atom arising from the infinite number of Rydberg states. That important and well-known problem is usually dealt with the Planck-Larkin prescription to cut off states of energies E_n larger than $k_B T$ (see e.g. the discussion in [24]). In our implementation of the physical picture for the recombined phase, no divergence occurs since the partition function of the hydrogen atom appears naturally in a convergent truncated form, as a consequence of collective screening effects. Only that truncated partition function embedding both bound and ionized states is free from ambiguity. More comments about that point are offered in Sect. 3.2.

Collective screening effects also give raise to well-behaved partition functions for more complex entities, like ions H^- and H_2^+, or molecules H_2. Such partition functions are naturally defined according to a truncation procedure similar to that introduced for the atomic partition function. They also involve contributions from both bound and dissociated states. The molecular partition function accounts thus not only for molecular bound states, but also for diffusion states made with two protons and two electrons. The finiteness of few-body truncated partition functions is of course crucial in the analysis of their low-temperature behaviors, which are shown to be controlled by Boltzmann factors $\exp(-\beta E_H)$, $\exp(-\beta E_{H^-})$, $\exp(-\beta E_{H_2^+})$, $\exp(-\beta E_{H_2})$, associated with the corresponding recombined entities in their groundstate (as would trivially be expected in a system with short range forces [25, 35]). Contributions from excited or diffusion states are well-defined in those truncated partition functions, and they may be neglected when the temperature is low enough.

An exact treatment of screening in the many-body problem is also required to establish the correct classification of terms in the expansion (1.1) according to decaying exponentials. For instance, in addition to obvious contributions of atomic bound states, there are correction terms proportional to the inverse screening length κ (or powers of it), which is itself proportional to the square root of the density $\kappa \sim \sqrt{\rho}$. Since the latter is also exponentially small in the Saha regime (see Sect. 2.2), contributions of collective screening effects have to be compared to pure atomic terms, and may be predominant as exemplified by the first correction $\beta P_1/\rho^*$. Such systematic classification could not have been obtained without a unified theory which deals exactly with the interplay between screening effects and the other physical phenomena at stake (primarily the formation of atomic and molecular bound states).

Though SLT expansions are asymptotic, i.e. a priori valid in the zero-temperature limit, they can be used for quantitative purposes within a rather wide range of temperatures and densities. We have performed numerical calculations, for both the EOS and the internal energy. For T of the order a few thousand kelvins, the $h_k(\beta)$'s are not accurately reproduced by their simple low-temperature asymptotic forms: further contributions, which arise in particular from excited states of the recombined entities, must also be taken into account. Within a simple criterion on the order of magnitude of the various corrections to Saha pressure, we draw the validity domain of SLT expansion (1.1) in the plane (β, ρ) (see Fig. 12 in Sect. 4.3). That validity domain exemplifies the quantitative interest of our calculations, which can be applied to physical systems under observable conditions, like the Sun photosphere for instance. Furthermore, we have compared our findings to those of Militzer and Ceperley [48] obtained within Path Integral Monte Carlo (PIMC) simulations (PIMC methods have been implemented through successive works [15–17, 47, 49]). The agreement is satisfactory, as it should since PIMC results are computationally exact within statistical errors (see e.g. Ref. [40]). The detail of that comparison, as well as all our numerical results, will be presented in a forthcoming paper [1].

From a mathematical view point, all manifestations of screening stem from the screened potential introduced in Sect. 2.4 and studied in [13]. That potential can be viewed, in the quantum mechanical context, as the analogue of the classical Debye-Hückel potential. Because of its central role, we have devoted the long Appendix A: to a number of related properties which are used in our analysis. In Appendix B:, the low temperature behaviors of the truncated atomic, ionic and molecular partition functions mentioned above, are determined by methods using Green functions and path integral representations. In Appendix C:, we compute the contributions of interactions between atoms and ionized charges.

2 Saha Regime and Screened Cluster Expansion

2.1 Definition of the Model

The hydrogen plasma is a two-component system made of protons and electrons. In the present non-relativistic limit, the proton and the electron are viewed as quantum point particles with respective charges, masses, and spins, $e_p = e$ and $e_e = -e$, m_p and m_e, $\sigma_p = 1/2$ and $\sigma_e = 1/2$. The kinetic energy operator for each particle of species $\alpha = p, e$ with position $\mathbf{x}$ takes the Schrödinger form $-\hbar^2/(2m_\alpha)\Delta$ where Δ is the Laplacian with respect to $\mathbf{x}$. Two particles separated by a distance r interact via the instantaneous Coulomb potential $v(r) = 1/r$. The corresponding Coulomb Hamiltonian H_{N_p,N_e} for N_p protons and N_e electrons reads

$$H_{N_p,N_e} = -\sum_{i=1}^{N} \frac{\hbar^2}{2m_{\alpha_i}} \Delta_i + \frac{1}{2} \sum_{i \neq j} e_{\alpha_i} e_{\alpha_j} v(|\mathbf{x}_i - \mathbf{x}_j|) \tag{2.1}$$

where $N = N_p + N_e$ is the total number of particles. In (2.1), the subscript i is attached to protons for $i = 1, \ldots, N_p$ and to electrons for $i = N_p + 1, \ldots, N_p + N_e$, so the species index α_i reduces either to p or e while $\mathbf{x}_i$ denotes either the position $\mathbf{R}_i$ of the i-th proton or the position $\mathbf{r}_j$ of the j-th electron ($j = i - N_p$).

The system is enclosed in a box with volume Λ, in contact with a thermostat at temperature T and a reservoir of particles that fixes the chemical potentials equal to μ_p and μ_e for protons and electrons respectively. Its grand-partition function Ξ is

$$\Xi = \mathrm{Tr} \exp[-\beta(H_{N_p,N_e} - \mu_p N_p - \mu_e N_e)]. \tag{2.2}$$

In (2.2), the trace is taken over all states symmetrized according to the Fermionic nature of each species; the boundary conditions for the wave functions at the surface of the box can be chosen of the Dirichlet type. Lieb and Lebowitz [44] have proved that the thermodynamic limit ($\Lambda \to \infty$ at fixed β and μ_α) exists, thanks to Fermi statistics and screening. Indeed, the Fermionic statistics of at least one species implies the H-stability [21, 22]

$$H_{N_p,N_e} > -B(N_p + N_e), \quad B > 0 \tag{2.3}$$

that prevents the collapse of the system. On the other hand, screening ensures that it does not explode. In a fluid phase, the infinite system maintains local neutrality, i.e. the homogeneous local particle densities ρ_p and ρ_e for protons and electrons remain equal for any choice of the chemical potentials μ_α. In other words, the common particle density $\rho = \rho_p = \rho_e$, as well as all other bulk equilibrium quantities, depend on the sole combination

$$\mu = (\mu_p + \mu_e)/2, \tag{2.4}$$

and not on the difference $\nu = (\mu_e - \mu_p)/2$. In particular, in terms of the fugacities $z_\alpha = \exp(\beta\mu_\alpha)$, this means that both the density ρ and the pressure P are functions of only β and $z = (z_p z_e)^{1/2} = \exp(\beta\mu)$. Therefore individual chemical potentials μ_e, μ_p are not uniquely determined: we can choose their difference ν at will without changing the bulk densities. Among the possible choices, it is particularly convenient to set

$$\mu_p = \mu - \frac{3}{2}k_B T \ln \frac{\lambda_e}{\lambda_p}, \qquad \mu_e = \mu + \frac{3}{2}k_B T \ln \frac{\lambda_e}{\lambda_p} \tag{2.5}$$

where $\lambda_\alpha = (\beta\hbar^2/m_\alpha)^{1/2}$ is the thermal de Broglie wavelength of species α. This choice guarantees that Maxwell-Boltzmann densities of free (no interactions) proton and electron gases, respectively

$$\begin{aligned} \rho_p^{id} &= 2z_p/(2\pi\lambda_p^2)^{3/2}, \\ \rho_e^{id} &= 2z_e/(2\pi\lambda_e^2)^{3/2}, \end{aligned} \tag{2.6}$$

are identical, i.e.

$$\rho_p^{id} = \rho_e^{id} = 2z/(2\pi\lambda^2)^{3/2} \tag{2.7}$$

with $\lambda = (\lambda_p\lambda_e)^{1/2}$. The factors 2 in (2.6) accounts for spin degeneracy.

The enforced neutrality of the ideal mixtures is equivalent to the linear relation

$$\sum_\alpha e_\alpha z_\alpha/(2\pi\lambda_\alpha^2)^{3/2} = 0 \tag{2.8}$$

between the activities z_p and z_e, sometimes called the pseudo neutrality condition. That condition can be imposed without loss of generality when dealing with fugacity expansions in the grand canonical ensemble. As shown in Sect. 3, it considerably simplifies the analysis of diagrammatic series for the interacting system. If we consider other fugacities (z'_p, z'_e) which do not satisfy the pseudo neutrality condition, an infinite number of graphs contributes to any term with a given order in low-density expansions. The calculations of those terms then become rather cumbersome. Nevertheless, beyond that technical complication, their final expression would be identical to that derived by starting with the above fugacities satisfying both condition (2.8) and $z_p z_e = z'_p z'_e$, in agreement with Lieb and Lebowitz proof [44].

2.2 Rigorous Results at Low Density and Low Temperature

We briefly recall the Saha theory in its simplest form. From the elementary view point of the thermodynamic of ideal substances, equilibrium ionization phases can be considered in the so-called chemical picture [26] as mixtures of noninteracting gases of electrons, protons, and hydrogen atoms, with chemical potential μ_{at} obeying the law of chemical equilibrium $\mu_{at} = \mu_e + \mu_p$. According to (2.7) the corresponding densities of electrons and protons are

$$\rho_p^{id} = \rho_e^{id} = 2\left(\frac{(m_p m_e)^{1/2}}{2\pi\beta\hbar^2}\right)^{3/2} \exp(\beta\mu), \tag{2.9}$$

whereas the atomic density is

$$\rho_{at}^{id} = 4\left(\frac{M}{2\pi\beta\hbar^2}\right)^{3/2} \exp(-\beta(E_H - 2\mu)) \tag{2.10}$$

where $M = m_p + m_e$ is the atomic mass and the factor 4 is the number of spin states. Apart from the binding energy E_H of the Hydrogen atom, all other effects of the Coulomb interaction are disregarded, so the Saha EOS is that of a mixture of perfect gases

$$\beta P_{\text{Saha}} = \rho_p^{id} + \rho_e^{id} + \rho_{at}^{id}. \tag{2.11}$$

We see in (2.9) and (2.10) that, when $\mu = E_H$, all densities are of the same exponential order at low temperatures: this corresponds to the coexistence of ionized and atomic phases. It is appropriate to characterize the set of ionization equilibrium phases by a temperature-dependent chemical potential [45]

$$\mu = \mu(\beta) = E_H + k_B T \ln w \tag{2.12}$$

where w is a fixed parameter $0 < w < \infty$. As shown from (2.9) and (2.10), that parameter determines the relative proportion of atoms to ionized charges through

$$\frac{\rho_{at}^{id}}{\rho_{p,e}^{id}} = 2w\left(\frac{M}{m}\right)^{3/4} = \frac{\gamma}{2}, \tag{2.13}$$

where we have introduced the equivalent parameter $\gamma = 4(M/m)^{3/4}w$. According to the above definitions, the fugacity $z = \exp(\beta\mu)$ can be seen as parametrized by either w or γ at fixed temperature, i.e. $z = w\exp(\beta E_H)$ or

$$z = \left(\frac{m}{M}\right)^{3/4} \gamma \exp(\beta E_H)/4. \tag{2.14}$$

For further purposes, it is convenient to consider the temperature dependent reference density ρ^* defined by (1.2) in the Introduction. Then we can rewrite ideal densities as

$$\rho_p^{id} = \rho_e^{id} = \rho^*\gamma \tag{2.15}$$

and

$$\rho_{at}^{id} = \rho^*\frac{\gamma^2}{2}. \tag{2.16}$$

In terms of γ, the proton (or electron) density $\rho = \rho_p^{id} + \rho_{at}^{id} = \rho_e^{id} + \rho_{at}^{id}$ and the Saha pressure (2.11) respectively read

$$\rho = \rho^* \left(\gamma + \frac{\gamma^2}{2} \right) \tag{2.17}$$

and

$$\beta P_{\text{Saha}} = \rho^* \left(2\gamma + \frac{\gamma^2}{2} \right). \tag{2.18}$$

Inversion of relation (2.17) provides γ, and hence the chemical potential μ, as a function of the reduced density ρ/ρ^*. Substitution of that function in (2.18) finally yields the Saha EOS (1.3) for the dimensionless pressure written as a function of ρ/ρ^* (note that our density variable is half of the total number density). As said in the Introduction, ρ^* is the cross-over density between full ionization and atomic recombination.

The Saha picture has been rigorously justified from the statistical mechanics of the full interacting electron-proton gas in the following asymptotic sense. When the temperature goes to zero at fixed negative values of μ, the system obviously becomes highly dilute because all fugacities then vanish exponentially fast. If low temperatures favor recombination of electrons and protons into bound entities with negative ground state energies, on the contrary low densities favor dissociation. The chemical composition of the system will result of those two competing energy and entropy effects. That problem has been studied in a rigorous way by Fefferman [27], who proved the two following results using a refined version of the stability of matter (2.3) (see the discussion after (2.30) and Ref. [14] for a review). First, when $\beta \to \infty$ with $\mu < E_H$ (μ fixed), the pressure tends to that of an ideal mixture of protons and electrons with respective densities ρ_p^{id} and ρ_e^{id} (2.9) i.e. the system becomes fully ionized. Second, there exists some $\delta > 0$ such that, when $\beta \to \infty$ with $E_H < \mu < E_H + \delta$ (μ fixed), the pressure tends to that of an ideal gas of hydrogen atoms in their groundstate with density (2.10). In that case, there is full atomic recombination.

The previous discussion of the Saha EOS suggests that ionized protons, ionized electrons and Hydrogen atoms should coexist at $\mu = E_H$. This has been firmly settled by Lieb et al. [18] and also Macris and Martin [45] who proved that, when one introduces the temperature dependent chemical potential (2.12) and let $\beta \to \infty$, the EOS tends to that of an ideal mixture of protons, electrons, and Hydrogen-atoms in their ground state, namely

$$\beta P = (\rho_p^{id} + \rho_e^{id} + \rho_{at}^{id})[1 + O(\exp(-\beta\epsilon))] = \beta P_{\text{Saha}}[1 + O(\exp(-\beta\epsilon))] \tag{2.19}$$

for β large enough and $\epsilon > 0$. The original work of Fefferman provides a power-law bound $1/\beta$ to the error term; that bound was improved to an exponential one in [18]. Thus we see that all ideal densities vanish exponentially fast, while corrections to ideal terms in (2.19) decay exponentially faster. The mathematical methods used in [27] and [18] are adequate to obtain a rigorous control of the dominant term (the Saha pressure), but apparently not adapted to explicitly calculate the corrections. In this work, it is our purpose to develop tools that enable to systematically compute those corrections, by expanding the pressure beyond the Saha term in an exact way (see (1.1)). In order to characterize the Saha regime in our study of the interacting system, we shall still use the parametrization (2.14) of the fugacity z associated with the zero-temperature limit, in terms of the parameter γ.

2.3 Simple Physical Considerations about the Fugacity Expansion

Saha equation of state (2.19) can be recovered, at a heuristic level, from simple considerations on low fugacities series for the pressure. Those considerations will serve as a guide to the analytic estimations of various non-ideal contributions to the full EOS at finite temperature and density performed in next sections. Low fugacity series are easily inferred, at a formal level, from the identity

$$\beta P = \frac{\ln \Xi}{\Lambda}, \tag{2.20}$$

where the thermodynamic limit $\Lambda \to \infty$ is implicitly taken, as in the whole paper. They read

$$\beta P = \sum_{(N_p,N_e)\neq(0,0)} z_p^{N_p} z_e^{N_e} B_{N_p,N_e} \tag{2.21}$$

where Mayer coefficients B_{N_p,N_e} in (2.21) can be expressed as suitable traces,

$$B_{N_p,N_e} = \frac{1}{\Lambda}\mathrm{Tr}[\exp(-\beta H_{N_p,N_e})]_{\mathrm{Mayer}}. \tag{2.22}$$

The first Mayer operators $[\exp(-\beta H_{N_p,N_e})]_{\mathrm{Mayer}}$ read

$$\begin{aligned}
&[\exp(-\beta H_{1,0})]_{\mathrm{Mayer}} = \exp(-\beta H_{1,0}),\\
&[\exp(-\beta H_{0,1})]_{\mathrm{Mayer}} = \exp(-\beta H_{0,1}),\\
&[\exp(-\beta H_{1,1})]_{\mathrm{Mayer}} = \exp(-\beta H_{1,1}) - \exp(-\beta H_{1,0})\exp(-\beta H_{0,1}),\\
&[\exp(-\beta H_{2,0})]_{\mathrm{Mayer}} = \exp(-\beta H_{2,0}) - \frac{1}{2}\exp(-\beta H_{1,0})\exp(-\beta H_{1,0}),\\
&\ldots,
\end{aligned} \tag{2.23}$$

while a similar expression holds for any $[\exp(-\beta H_{N_p,N_e})]_{\mathrm{Mayer}}$

$$[\exp(-\beta H_{N_p,N_e})]_{\mathrm{Mayer}} = \exp(-\beta H_{N_p,N_e}) - \cdots. \tag{2.24}$$

In (2.24), terms left over reduce to a linear combination of products of Gibbs operators $\exp(-\beta H_{M_p,M_e})$ ($M_p \le N_p$, $M_e \le N_e$) associated with all possible partitions of N_p protons and N_e electrons. Traces (2.22) must be taken over Fermionic states which are products of anti-symmetrized states for each set of degrees of freedom associated with a Gibbs operator $\exp(-\beta H_{M_p,M_e})$. For instance, in space of positions and spins, $B_{2,0}$ reads

$$\begin{aligned}
B_{2,0} = \frac{1}{\Lambda}\int d\mathbf{R}_1 \int d\mathbf{R}_2 [&2\langle \mathbf{R}_1\mathbf{R}_2|\exp(-\beta H_{2,0})|\mathbf{R}_1\mathbf{R}_2\rangle\\
&- 2\langle \mathbf{R}_1|\exp(-\beta H_{1,0})|\mathbf{R}_1\rangle\langle \mathbf{R}_2|\exp(-\beta H_{1,0})|\mathbf{R}_2\rangle\\
&- \langle \mathbf{R}_2\mathbf{R}_1|\exp(-\beta H_{2,0})|\mathbf{R}_1\mathbf{R}_2\rangle].
\end{aligned} \tag{2.25}$$

For the term $\exp(-\beta H_{1,0})\exp(-\beta H_{1,0})$ subtracted in $[\exp(-\beta H_{2,0})]_{\mathrm{Mayer}}$, each Gibbs operator $\exp(-\beta H_{1,0})$ refers to a single proton, so no anti-symmetrization occurs and only diagonal matrix elements of $\exp(-\beta H_{1,0})$ appear in (2.25). Truncated Mayer operators can also be expressed in terms of Ursell operators [32–34].

Despite Mayer coefficients B_{N_p,N_e} diverge, leading contributions to the equation of state can be easily picked out in formal series (2.21), as follows. For ($N_p = 1, N_e = 0$) and ($N_p = 0, N_e = 1$), we obtain the simple exact expressions

$$B_{1,0} = \frac{2}{(2\pi\lambda_p^2)^{3/2}} \tag{2.26}$$

and

$$B_{0,1} = \frac{2}{(2\pi\lambda_e^2)^{3/2}}. \tag{2.27}$$

After multiplication by fugacity factors $\exp(\beta\mu_p)$ and $\exp(\beta\mu_e)$ respectively, we obtain the related contributions to pressure (2.21) which reduce, of course, to the ideal Maxwell-Boltzmann densities of ionized protons (ρ_p^{id}) and ionized electrons (ρ_e^{id}). For ($N_p = 1$, $N_e = 1$), it is reasonable to expect that hydrogen atoms with internal ground state energy E_H provide the leading low-temperature contribution which reads

$$\frac{4}{(2\pi\lambda_H^2)^{3/2}} \exp(-\beta E_H), \tag{2.28}$$

with $\lambda_H = (\beta\hbar^2/M)^{1/2}$ while factor 4 is due to spin degeneracy. The corresponding contribution to (2.21) is nothing but the ideal Maxwell-Boltzmann density ρ_{at}^{id} of Hydrogen atoms in their ground state. In the Saha regime, ideal densities of ionized protons (ρ_p^{id}), ionized electrons (ρ_e^{id}) and hydrogen atoms (ρ_{at}^{id}) are all found to be of the same order of magnitude $\exp(\beta E_H)$ disregarding powers of β, because of (2.12).

All other contributions to the EOS are expected to be small corrections to Saha pressure, as suggested by the following simple arguments and estimations. The Saha regime defines quite diluted conditions since ρ vanishes exponentially fast. Therefore, ionized charges and hydrogen atoms are expected to be weakly coupled and weakly degenerate. Let us introduce the various length and energy scales defined in Table 1, where we assume that each atom

Table 1 Length and energy scales in a quantum hydrogen plasma

Symbol	Value	Physical signification
Length		
a_B	$\hbar^2/(me^2)$	Bohr radius
$\lambda_{p,e,at}$	$\hbar(\beta/m_{p,e,at})^{1/2}$	de Broglie lengths
l_B	βe^2	Bjerrum length
a	$(3/(4\pi\rho))^{1/3}$	Mean interparticle distance
κ^{-1}	$(4\pi\beta e^2[\rho_p^{id} + \rho_e^{id}])^{-1/2}$	Debye screening length
l_Q	$\kappa^{-1}\vert\ln(\kappa\lambda)\vert$	Quantum screening distance
Energy		
ϵ_{H-H}	$e^2a_B^2/a^3$	Atom-atom interaction energy
ϵ_{H-c}	e^2a_B/a^2	Atom-charge interaction energy
ϵ_{c-c}	e^2/a	Charge-charge interaction energy
ϵ_{kin}	k_BT	Classical kinetic energy
E_H	$\vert E_H\vert = me^4/(2\hbar^2)$	Atom ground-state energy

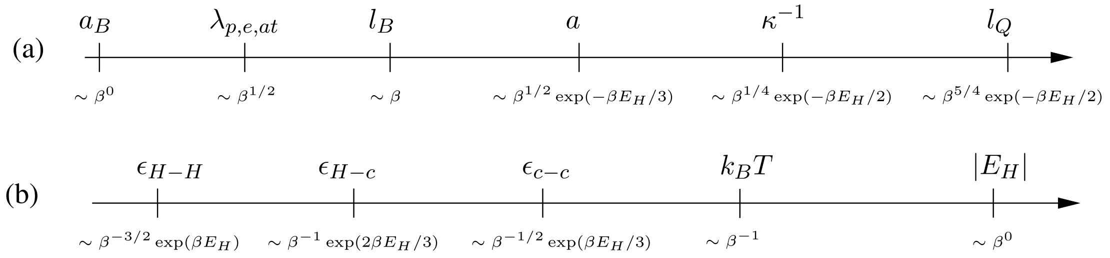


Fig. 1 Hierarchy of **a** length and **b** energy scales in the Saha regime

carries, roughly speaking, an instantaneous dipole of order ea_B, while the physical signification of l_Q is given in next Sect. 2.4. According to the hierarchies between those length and energy scales described in Fig. 1, both exchange and interaction contributions for ionized charges and hydrogen atoms should be exponentially smaller than above ideal terms. Similarly, we can estimate the contributions of complex entities which result from the quantum mechanical binding of N_e electrons and N_p protons, i.e. the existence of a bound state in the spectrum of H_{N_p,N_e} with negative ground state energy $E^{(0)}_{N_p,N_e}$. The ideal contribution of a given complex entity is easily extracted from B_{N_p,N_e}, and it is of order $\exp(-\beta E^{(0)}_{N_p,N_e})$. After multiplication by fugacity factor $\exp[\beta(\mu_p N_p + \mu_e N_e)]$, we find a contribution to pressure (2.21) which is of order

$$\exp[-\beta(E^{(0)}_{N_p,N_e} - (N_p + N_e - 1)E_H]\exp(\beta E_H), \tag{2.29}$$

where we have used parametrization (2.12) of the chemical potential. Ideal contribution (2.29) decays exponentially faster than $\exp(\beta E_H)$ for $(N_p, N_e) \neq (1,1), (1,0), (0,1)$, by virtue of inequality

$$E^{(0)}_{N_p,N_e} - (N_p + N_e - 1)E_H > 0, \quad (N_p, N_e) \neq (1,0), (0,1), (1,1) \tag{2.30}$$

which is a key ingredient in Fefferman's proof [27]. Although not yet proved, that inequality is satisfied by known complex entities [28] as illustrated below. Of course, and as for ionized charges and hydrogen atoms, exchange and interactions contributions for complex entities should be smaller than ideal ones.

Above heuristic arguments suggest that corrections to ideal Saha pressure (2.11) decay exponentially faster than leading terms when T vanishes, in agreement with the rigorous bound involved in (2.19). A precise evaluation of those corrections will be performed in Sect. 3 by using screened cluster expansion described in next Sect. 2.4. That method removes all long range Coulomb divergencies which plague Mayer coefficients B_{N_p,N_e}. It provides well-defined recipes for computing contributions from both interactions and complex entities. The simplest entities which appear are the molecule H_2 with groundstate energy $E^{(0)}_{2,2} = E_{H_2} \simeq -31.7$ eV, ion H_2^+ with $E^{(0)}_{2,1} = E_{H_2^+} \simeq -16.2$ eV, and ion H^- with $E^{(0)}_{1,2} = E_{H^-} \simeq -14.3$ eV. Notice that such groundstate energies do satisfy inequality 2.30. For complex entities made with five or more particles, we will assume inequality

$$E^{(0)}_{N_p,N_e} > (N_p + N_e - 2)E_H, \quad N_p + N_e \geq 5. \tag{2.31}$$

That inequality, more constraining than (2.30), is indeed satisfied by known stable complex entities. For instance, $E^{(0)}_{3,2} = E_{H_3^+} \simeq -36.5$ eV and $E^{(0)}_{2,3} = E_{H_2^-} \simeq -28.5$ eV are indeed larger than $3E_H \simeq -40.8$ eV. Previous groundstate energies are computed within the method

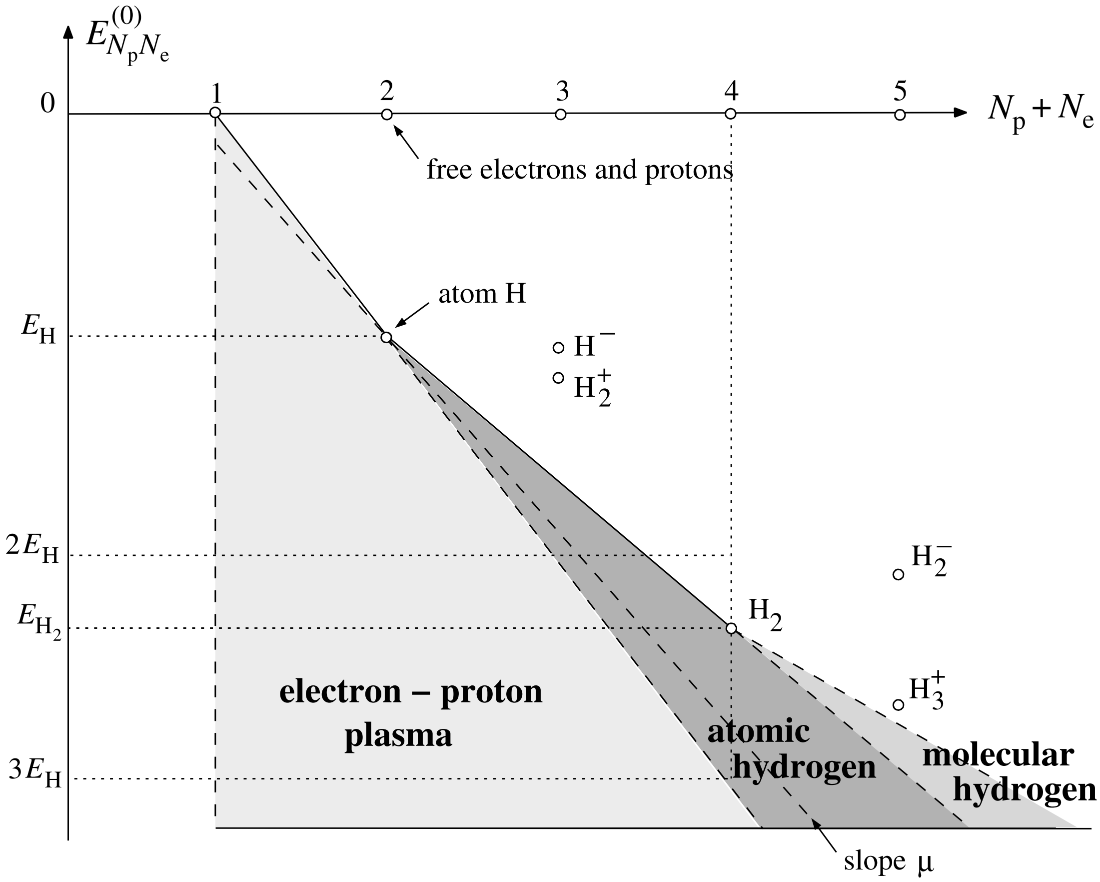


Fig. 2 Geometrical representation of inequalities (2.30). Consider a line of slope $\mu < 0$ which goes through the point associated with the hydrogen atom. If all the points $(N_p + N_e, E^{(0)}_{N_p,N_e})$ associated with other entities lie above that line, the inequalities (2.30) hold for that value of μ, and the system tends to a dilute atomic gas in the limit $\beta \to \infty$

described in [52]. The corresponding values are in excellent agreement with experiments and reported data in the literature. The resulting stability regimes of ionized, atomic and molecular phases are shown in Fig. 2.

2.4 Screened Cluster Expansion within Loop Formalism

Screened cluster expansions are devised within an auxiliary classical system of charged loops. As first shown by Ginibre [30], Ξ is identical to the grand-partition function of a classical system made with loops. That transformation starts with the expression of Ξ in space of positions and spins, and use of Feynman-Kac formula [38, 53, 61, 62],

$$\langle \mathbf{x}'_1 \cdots \mathbf{x}'_N | \exp(-\beta H_{N_p,N_e}) | \mathbf{x}_1 \cdots \mathbf{x}_N \rangle$$
$$= \prod_{i=1}^{N} \frac{\exp[-(\mathbf{x}'_i - \mathbf{x}_i)^2/(2\lambda^2_{\alpha_i})]}{(2\pi\lambda^2_{\alpha_i})^{3/2}} \int \prod_{i=1}^{N} \mathcal{D}(\boldsymbol{\xi}_i) \exp\Bigg[-\frac{\beta}{2}\sum_{i\neq j} e_{\alpha_i} e_{\alpha_j}$$
$$\times \int_0^1 \mathrm{d}s\, v(|(1-s)(\mathbf{x}_i - \mathbf{x}_j) + s(\mathbf{x}'_i - \mathbf{x}'_j) + \lambda_{\alpha_i}\boldsymbol{\xi}_i(s) - \lambda_{\alpha_j}\boldsymbol{\xi}_j(s)|)\Bigg]. \quad (2.32)$$

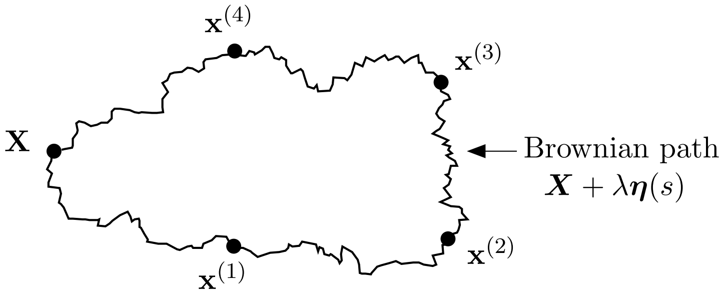


Fig. 3 A loop $\mathcal{L} = (\alpha, q, \mathbf{X}, \boldsymbol{\eta}(s))$ made up of 5 particles

In the r.h.s. of (2.32), functional integrations are performed over Brownian bridges $\boldsymbol{\xi}_i(s)$ ($\boldsymbol{\xi}_i(0) = \boldsymbol{\xi}_i(1) = 0$) with the normalized Gaussian measure $\mathcal{D}(\boldsymbol{\xi})$ defined by its covariance (see (2.34) with $q = 1$). Each Brownian bridge $\boldsymbol{\xi}_i(s)$ defines a path $(1-s)\mathbf{x}_i + s\mathbf{x}_i' + \lambda_{\alpha_i}\boldsymbol{\xi}_i(s)$ associated with a given particle. A loop $\mathcal{L}$ is then defined as the collection of open paths associated with particles exchanged in a given permutation cycle. This leads to the identity [14, 19, 46]

$$\Xi = \Xi_{loop} = \sum_{N=0}^{\infty} \frac{1}{N!} \int \prod_{i=1}^{N} \mathcal{D}(\mathcal{L}_i) z(\mathcal{L}_i) \prod_{i<j} \exp(-\beta V(\mathcal{L}_i, \mathcal{L}_j)), \tag{2.33}$$

where fugacity $z(\mathcal{L}_i)$ and two-body potential $V(\mathcal{L}_i, \mathcal{L}_j)$ are defined below.

A loop $\mathcal{L}$ is characterized by its position $\mathbf{X}$, species $\alpha = p, e$ and number q of exchanged particles, while its shape is defined by a closed Brownian path $\boldsymbol{\eta}(s)$ with $s \in [0, q]$ and $\boldsymbol{\eta}(0) = \boldsymbol{\eta}(q) = 0$. Genuine particle positions in matrix elements of $\exp(-\beta(H_{N_p,N_e})$ reduce to $\mathbf{x}^{(k)} = \mathbf{X} + \lambda_\alpha \boldsymbol{\eta}(k)$ with k integer, $k = 0, \ldots, q-1$ ($\mathbf{x}^{(0)} = \mathbf{x}^{(q)} = \mathbf{X}$) (see Fig. 3). Phase-space measure $\mathcal{D}(\mathcal{L})$ is the product of discrete summations over α and q, spatial integration over $\mathbf{X}$ and functional integration over $\boldsymbol{\eta}(s)$ with normalized Gaussian measure $\mathcal{D}(\boldsymbol{\eta})$ defined by its covariance

$$\int \mathcal{D}(\boldsymbol{\eta}) \eta_\mu(s) \eta_\nu(t) = \delta_{\mu\nu} q \inf(s/q, t/q)(1 - \sup(s/q, t/q)). \tag{2.34}$$

Fugacity $z(\mathcal{L})$ reads [14, 19]

$$\begin{aligned} z(\mathcal{L}) = (-1)^{q-1} \frac{2}{q} \frac{z_\alpha^q}{(2\pi q \lambda_\alpha^2)^{3/2}} \\ \times \exp\left[-\frac{\beta e^2}{2} \int_0^q \mathrm{d}s \int_0^q \mathrm{d}t (1 - \delta_{[s],[t]}) \tilde{\delta}(s-t) v(|\lambda_\alpha \boldsymbol{\eta}(s) - \lambda_\alpha \boldsymbol{\eta}(t)|) \right] \end{aligned} \tag{2.35}$$

where $\tilde{\delta}(s-t) = \sum_{n=-\infty}^{\infty} \delta(s-t-n)$ is Dirac comb, while $[s]$ ($[t]$) denotes the integer part of s (t). In (2.35), factor $(1 - \delta_{[s],[t]})$ avoids counting point particle self-energy contributions, while Dirac comb ensures that only loop elements with equal times (modulo an integer) interact, an essential feature specific to quantum mechanics. Eventually, two-body potential $V(\mathcal{L}_i, \mathcal{L}_j)$ reduces to

$$V(\mathcal{L}_i, \mathcal{L}_j) = e_{\alpha_i} e_{\alpha_j} \int_0^{q_i} \mathrm{d}s \int_0^{q_j} \mathrm{d}t \tilde{\delta}(s-t) v(|\mathbf{X}_i + \lambda_{\alpha_i} \boldsymbol{\eta}_i(s) - \mathbf{X}_j - \lambda_{\alpha_j} \boldsymbol{\eta}_j(t)|). \tag{2.36}$$

At large distances, V behaves as the Coulomb potential between point charges $q_i e_{\alpha_i}$ and $q_j e_{\alpha_j}$, i.e. $V(\mathcal{L}_i, \mathcal{L}_j) \sim q_i q_j e_{\alpha_i} e_{\alpha_j} / |\mathbf{X}_i - \mathbf{X}_j|$. Therefore, usual Mayer diagrammatics for

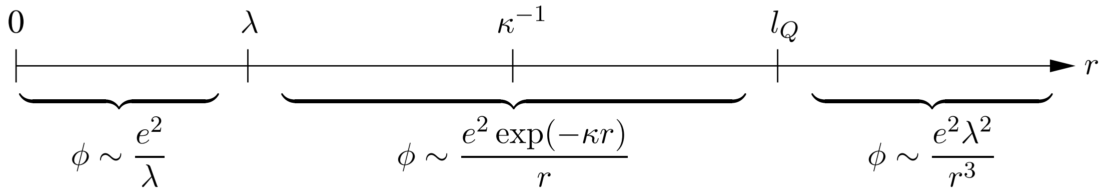


Fig. 4 Order of magnitude of effective potential ϕ at various distances. In region $\lambda < r < l_Q$, interactions are exponentially screened according to Debye potential. For $r > l_Q$, ϕ is dominated by (unscreened) multipolar interactions

loops are plagued with long-range divergences. As in the case of classical Coulomb fluids, they are removed by summing infinite chains built with V. This amounts to replace V by screened potential ϕ which can be viewed as the quantum analog of Debye potential [13]. The explicit formula for the Fourier transform of ϕ is recalled in Appendix A:. Its spatial behaviors, according to the hierarchy of scales displayed in Fig. 1, are roughly summarized in Fig. 4, where only orders of magnitude of ϕ are given (we set $r = |\mathbf{X}_i - \mathbf{X}_j|$ and omit all shape dependences which occur for $r < \lambda$ or $l_Q < r$). Notice that familiar exponential decay of ϕ breaks down at large distances $r \gg l_Q$. The asymptotic dipolar behavior of ϕ is sufficient for ensuring that every graph built with ϕ is finite [2, 3].

As detailed in [8], so-called screened cluster expansion for protonic density ρ_p follows from an exact transformation of formal Mayer diagrammatics for loop density $\rho(\mathcal{L}_a)$ which provides

$$\rho_p = \sum_G \frac{1}{S(G)} \int \mathcal{D}(C_a) Z_\phi^T(C_a) q_a \int \prod_{i=1}^{n} \mathcal{D}(C_i) Z_\phi^T(C_i) \Big[\prod \mathcal{F}_\phi\Big]_G \tag{2.37}$$

(a similar expression holds for ρ_e). In (2.37), bare potential V is replaced by screened potential ϕ. Graphs G are identical to usual Mayer graphs, where points are now particle clusters, except for some specific rules (arising from the replacement of V by ϕ) which are described below. Each cluster C_i $(i = 0, \ldots, n)$ contains $N_i^{(p)}$ protons and $N_i^{(e)}$ electrons. The internal state of a cluster $C(N_p, N_e)$ $(C \in \{C_i, i = 0, \ldots, n\})$ is determined by L_p and L_e loops $(\mathcal{L}_1^{(\alpha)}, \ldots, \mathcal{L}_{L_\alpha}^{(\alpha)})$ in which the N_p protons and N_e electrons are distributed (in root cluster $C_0 = C_a$, $\mathcal{L}_1^{(p)}$ is identified to $\mathcal{L}_a$ which contains the root proton). Integration within phase space measure $\mathcal{D}(C)$ reduces to the sum over all possible distributions of particles into loops combined with integrations over loop positions and shapes (with $\mathbf{X}_1^{(p)} = \mathbf{X}_a$ fixed at the origin for loop $\mathcal{L}_1^{(p)} = \mathcal{L}_a$). Statistical weight $Z_\phi^T(C)$ for a cluster $C(N_p, N_e)$ reads

$$Z_\phi^T(C) = \frac{\prod_{k=1}^{L_p} z_\phi(\mathcal{L}_k^{(p)}) \prod_{k=1}^{L_e} z_\phi(\mathcal{L}_k^{(e)})}{\prod_{q=1}^{N_p} n_p(q)! \prod_{q=1}^{N_e} n_e(q)!} \mathcal{B}_\phi^T(\{\mathcal{L}_k^{(\alpha)}\}), \tag{2.38}$$

where $n_\alpha(q)$ is the number of loops containing q particles of species α (for C_a, $n_p(q_a)!$ is replaced by $(n_p(q_a) - 1)!$). Weight $z_\phi(\mathcal{L})$ reduces to

$$z_\phi(\mathcal{L}) = z(\mathcal{L}) \exp[I_R(\mathcal{L})] \tag{2.39}$$

with ring sum $I_R(\mathcal{L})$ given by

$$I_R(\mathcal{L}) = \frac{1}{2} \int \mathcal{D}(\mathcal{L}_1) z(\mathcal{L}_1) \beta V(\mathcal{L}, \mathcal{L}_1) \beta \phi(\mathcal{L}_1, \mathcal{L}). \tag{2.40}$$

Truncated Mayer coefficient $\mathcal{B}^T_{\phi,N}$ is defined by a suitable truncation of usual Mayer coefficient $\mathcal{B}_{\phi,N}$ for N loops with pair interactions ϕ. This truncation ensures that $\mathcal{B}^T_{\phi,N}$ remains integrable over relative distances between loops when ϕ is replaced by V. First truncated Mayer coefficients are

$$\mathcal{B}^T_{\phi,1} = 1, \qquad \mathcal{B}^T_{\phi,2} = \exp(-\beta\phi) - 1 + \beta\phi - \frac{\beta^2\phi^2}{2!} + \frac{\beta^3\phi^3}{3!}, \qquad \ldots . \tag{2.41}$$

Bond $\mathcal{F}_\phi(C_i, C_j)$ can be either $-\beta\Phi$, $\beta^2\Phi^2/2!$, $-\beta^3\Phi^3/3!$, where total potential $\Phi(C_i, C_j)$ is the sum of pairwise interactions $\phi(\mathcal{L}, \mathcal{L}')$ over loops $\mathcal{L}$ and $\mathcal{L}'$ defining internal states of C_i and C_j respectively.

As for ordinary Mayer diagrams, two clusters are connected by at most one bond, and graph G is connected. Here, symmetry factor $S(G)$ is computed by permuting only clusters with identical numbers of protons and electrons. Moreover, for a cluster C different from C_a, the internal state of which is determined by a single loop $\mathcal{L}_1^{(\alpha)}$, when C is either, the intermediate cluster of a convolution $(-\beta\Phi) * (-\beta\Phi)$, or connected to the rest of the graph by a single bond $\beta^2\Phi^2/2!$, expression (2.38) of its statistical weight must be replaced by

$$Z^T_\phi(C) = z_\phi(\mathcal{L}_1^{(\alpha)}) - z(\mathcal{L}_1^{(\alpha)}). \tag{2.42}$$

Eventually, summation in (2.37) involves only graphs G which are no longer integrable over relative distances between clusters $\{C_i, i = 0, \ldots, n\}$ when ϕ is replaced by V. Screened cluster expansion for the pressure is inferred from use of (2.37) in thermodynamics identities, as described in Sect. 4.

3 Estimations of Ideal and Non-Ideal Contributions to Fugacity Expansions of Particle Densities

Now, we proceed to asymptotic estimations, in the Saha regime, of all contributions to $\rho = \rho_p$ in screened cluster expansion (2.37). Every contribution to ρ_p is expressed, similarly to (2.15) and (2.16), as ρ^* times a power of γ, and times a dimensionless temperature-dependent function. This provides a formal representation of ρ/ρ^* in powers of γ, where the coefficients depend only on temperature (see (4.2)). At low temperatures, every coefficient decays exponentially fast. In Sects. 3.1–3.6, we select all contributions which are smaller than leading terms (2.15) and (2.16) (divided by ρ^*) by exponentially decaying factors of maximum order $\exp(\beta E_H)$ ($\beta \to \infty$). In Sect. 3.7, we show that all other contributions decay faster by factors exponentially smaller than $\exp(\beta E_H)$. Beyond leading ideal contributions of ionized protons (2.15) and hydrogen atoms (2.16) (which are recovered in Sects. 3.1 and 3.2), we determine first corrections arising from their mutual interactions (Sects. 3.1, 3.4, 3.5 and 3.6). Such corrections are at most of order $(\rho^{id}_{p,e,at})^2$, so they are smaller than leading terms by exponential factor $\exp(\beta E_H)$. We also study ideal-like contributions of recombined entities, molecules H_2, ions H^- and H_2^+ (Sects. 3.3 and 3.6) which must be accounted for at that order. In the following, a graph with N_p protons and N_e electrons will be obviously denoted G_{N_p,N_e} (for $N_p + N_e > 1$, there are several graphs with identical particle numbers).

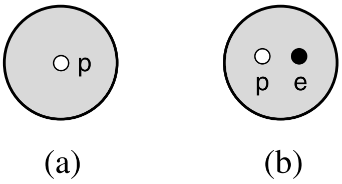


Fig. 5 Graphs representing simple entities: **a** ionized proton and **b** hydrogen atom

3.1 Ionized Proton and Plasma Polarization

An ionized proton appears in graph $G_{1,0}$ (see Fig. 5a) made with the sole root cluster C_a containing a single proton. The internal state of C_a is defined by the sole protonic loop $\mathcal{L}_a^{(p)}$ with $q_a = 1$. The contribution of $G_{1,0}$ to (2.37) then reads

$$\int \mathcal{D}(\boldsymbol{\xi}_a) z_\phi(\mathcal{L}_a^{(p)}) = \frac{2z_p}{(2\pi\lambda_p^2)^{3/2}} \int \mathcal{D}(\boldsymbol{\xi}_a) \exp(I_R(\mathcal{L}_a^{(p)})). \tag{3.1}$$

We stress that collective effects are embedded in ring sum $I_R(\mathcal{L}_a^{(p)})$. Thus, strictly speaking, $G_{1,0}$ describes an ionized proton dressed by the surrounding plasma of ionized protons and electrons. Within the present framework, that dressing mechanism accounts for the familiar plasma polarization induced by an immersed charge.

In the Saha regime, ring sum $I_R(\mathcal{L}_a^{(p)})$ can be evaluated by using the exact expression of ϕ (see Appendix A:). The corresponding asymptotic behavior can be easily recovered *via* the following simple estimation of convolution integral (2.40). At leading order, only terms $q_1 = 1$ ($\alpha_1 = p, e$) need to be retained into $\mathcal{D}(\mathcal{L}_1)$. Moreover the integration over position $\mathbf{X}_1$ is controlled by relative distances $|\mathbf{X}_1 - \mathbf{X}_a|$ of order κ^{-1}. At such distances, $\phi(\mathcal{L}_1, \mathcal{L}_a^{(p)})$ can be replaced by its Debye form, while $V(\mathcal{L}_a^{(p)}, \mathcal{L}_1)$ merely reduces to $ee_{\alpha_1}/|\mathbf{X}_1 - \mathbf{X}_a|$. This gives

$$I_R(\mathcal{L}_a^{(p)}) \sim \frac{\beta e^2 \kappa^2}{8\pi} \int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_1) \frac{\exp(-\kappa|\mathbf{X}_1 - \mathbf{X}_a|)}{|\mathbf{X}_1 - \mathbf{X}_a|^2} \sim \frac{\beta e^2 \kappa}{2}, \tag{3.2}$$

in perfect agreement with the detailed analysis of Appendix A:.

Since $\beta e^2 \kappa$ is small (see Fig. 1), dressing effects in (3.1) can be treated perturbatively by expanding $\exp(I_R(\mathcal{L}_a^{(p)}))$ in powers of I_R. The resulting leading contribution of $G_{1,0}$ reads

$$\frac{2z_p}{(2\pi\lambda_p^2)^{3/2}} \int \mathcal{D}(\boldsymbol{\xi}_a) = \frac{2z_p}{(2\pi\lambda_p^2)^{3/2}} = \rho_p^{id} = \rho^* \gamma, \tag{3.3}$$

where functional integration over $\boldsymbol{\xi}_a$ merely reduces to 1 by normalization of Gaussian measure $\mathcal{D}(\boldsymbol{\xi}_a)$. That leading term reduces to ideal Maxwell-Boltzmann density of ionized protons (2.15) (i.e. bare contribution $z_p B_{1,0}$ as it should).

Taking into account (3.2), we find that first correction to (3.3) is rewritten as

$$\frac{2z_p}{(2\pi\lambda_p^2)^{3/2}} \int \mathcal{D}(\boldsymbol{\xi}_a) I_R(\mathcal{L}_a^{(p)}) \sim \frac{2z_p}{(2\pi\lambda_p^2)^{3/2}} \beta e^2 \kappa \int \mathcal{D}(\boldsymbol{\xi}_a) = \rho_p^{id} \frac{\beta e^2 \kappa}{2}. \tag{3.4}$$

Contribution (3.4) involves a factor $z^{3/2}$, and hence a factor $\gamma^{3/2}$ times $\exp(3\beta E_H/2)$. One factor $\exp(\beta E_H)$ may be absorbed into the prefactor ρ^*. It remains a factor $\exp(\beta E_H/2)$, which multiplies the remaining part of the contribution. Therefore, we rewrite (3.4) as

$$\rho_p^{id} \frac{\beta e^2 \kappa}{2} = \rho^* \gamma^{3/2} S_{3/2}(1, 0) \exp(\beta E_H/2), \tag{3.5}$$

where we define screening function

$$S_{3/2}(1,0) = \frac{(\beta|E_H|)^{3/4}}{\pi^{1/4}}. \tag{3.6}$$

Index $3/2$ of screening function refers to the power of γ and $(1,0)$ to the single proton cluster. All subsequent contributions will be written according to the same prescription.

First correction (3.4) accounts for familiar plasma polarization induced by a single proton (i.e. cluster $(1,0)$) at lowest order. Its simple structure results from the almost classical and weakly coupled nature of the plasma mentioned in Sect. 2.3. Higher order collective corrections proportional to γ^p (with p integer or half-integer), can be rewritten similarly to (3.4) *via* the definition of screening functions $S_p(1,0)$ which depend only on β. For instance, next correction to (3.4) merely reduces to

$$\rho^* \gamma^2 S_2(1,0) \exp(\beta E_H), \tag{3.7}$$

with

$$S_2(1,0) = \frac{(\beta|E_H|)^{3/2}}{2\sqrt{\pi}} - \left(1 + \left(\frac{2m}{m_p}\right)^{1/2}\right) \frac{\beta|E_H|}{8}. \tag{3.8}$$

The first contribution in the r.h.s. of (3.8) arises from the quadratic term in the expansion of $\exp(I_R(\mathcal{L}_a^{(p)}))$, while the second one arises from the linear term where loop-shape dependence of $I_R(\mathcal{L}_a^{(p)})$ beyond classical form (3.3) is taken into account (see (A.11)). Further corrections are exponentially smaller than (3.7) as shown in Appendix A:.

The dressing mechanism associated with plasma polarization occurs for any particle in all other graphs G_{N_p,N_e}. At lowest order, every ring factor $\exp(I_R)$ can be replaced by 1, and first corrections are obtained by using (3.2).

3.2 Hydrogen Atom: Recombination and Dissociation Contributions

A hydrogen atom is expected to appear in graph $G_{1,1}$ made with single root cluster C_a (see Fig. 5b). The contribution of $G_{1,1}$ reads

$$\begin{aligned} &\int \mathcal{D}(\boldsymbol{\xi}_a) z_\phi(\mathcal{L}_a^{(p)}) \int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_1) z_\phi(\mathcal{L}_1^{(e)}) \mathcal{B}_\phi^T(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(e)}) \\ &= \frac{4 z_p z_e}{(2\pi\lambda^2)^3} \int \mathcal{D}(\boldsymbol{\xi}_a) \exp(I_R(a)) \int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_1) \exp(I_R(1)) \\ &\times \left[\exp(-\beta\phi(a,1)) - 1 + \beta\phi(a,1) - \frac{\beta^2\phi^2(a,1)}{2!} + \frac{\beta^3\phi^3(a,1)}{3!} \right] \end{aligned} \tag{3.9}$$

(with obvious simplified notations for the dependence of I_R and ϕ on loops $\mathcal{L}_a^{(p)}$ and $\mathcal{L}_1^{(e)}$). In (3.9), protonic loop $\mathcal{L}_a^{(p)}$ and electronic loop $\mathcal{L}_1^{(e)}$ contain one proton ($q_a = 1$) and one electron ($q_1 = 1$) respectively. Each of those particles are dressed like the ionized proton in $G_{1,0}$. Furthermore, their mutual interaction ϕ involves screening effects, which are also due to the surrounding plasma of ionized protons and electrons. At leading order, since $\phi(a,1)$ reduces to V at finite distances $|\mathbf{X}_1 - \mathbf{X}_a|$ (see Appendix A:), $\mathcal{B}_\phi^T(a,1)$ can be replaced by $\mathcal{B}^T(a,1)$ defined by (2.41) with V in place of ϕ. The resulting bare contribution of $G_{1,1}$, as well as first corrections due to collective effects, are successively estimated as follows.

3.2.1 Bare Contribution in the Vacuum

The bare contribution of $G_{1,1}$ reads

$$\frac{4z_p z_e}{(2\pi\lambda^2)^3}\int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_a)\mathcal{D}(\boldsymbol{\xi}_1) \\ \times \left[\exp(-\beta V(a,1)) - 1 + \beta V(a,1) - \frac{\beta^2 V^2(a,1)}{2!} + \frac{\beta^3 V^3(a,1)}{3!}\right]. \tag{3.10}$$

In (3.10), functional integrations over shapes $\boldsymbol{\xi}_a$ and $\boldsymbol{\xi}_1$ can be exactly rewritten in terms of matrix elements of suitable operators by applying backwards Feynman-Kac formula (2.32). For the exponential factor in $\mathcal{B}^T(a,1)$, we obviously obtain

$$\frac{1}{(2\pi\lambda_p^2)^{3/2}(2\pi\lambda_e^2)^{3/2}}\int \mathcal{D}(\boldsymbol{\xi}_a)\mathcal{D}(\boldsymbol{\xi}_1)\exp(-\beta V(a,1)) \\ = \langle \mathbf{R}_a\mathbf{r}_1|\exp(-\beta H_{1,1})|\mathbf{R}_a\mathbf{r}_1\rangle \tag{3.11}$$

with $\mathbf{R}_a = \mathbf{X}_a$ and $\mathbf{r}_1 = \mathbf{X}_1$. Functional integrations of powers of $V(a,1)$ in $\mathcal{B}^T(a,1)$, are related to the corresponding terms arising in Dyson expansion of $\langle \mathbf{R}_a\mathbf{r}_1|\exp(-\beta H_{1,1})|\mathbf{R}_a\mathbf{r}_1\rangle$ with respect to interaction part $V_{1,1}$ of $H_{1,1}$. Moreover, let us introduce position $\mathbf{R}^* = (m_p\mathbf{R}_a + m_e\mathbf{r}_1)/M$ of the atom mass center, and one-body Hamiltonian H_{pe} of relative particle with position $\mathbf{r}^* = \mathbf{r}_1 - \mathbf{R}_a$. Then, bare contribution (3.10) becomes

$$\rho^*\frac{\gamma^2}{8}Z(1,1)\exp(\beta E_H), \tag{3.12}$$

with

$$Z(1,1) = \frac{(2\pi\lambda_H^2)^{3/2}}{\Lambda}\mathrm{Tr}[\exp(-\beta H_{1,1})]^T_{\text{Mayer}} \\ = 4\int \mathrm{d}\mathbf{r}^*\langle\mathbf{r}^*|[\exp(-\beta H_{pe})]^T_{\text{Mayer}}|\mathbf{r}^*\rangle, \tag{3.13}$$

where $[\exp(-\beta H_{pe})]^T_{\text{Mayer}}$ stands for truncated Mayer operator

$$[\exp(-\beta H_{pe})]^T_{\text{Mayer}} \\ = \exp(-\beta H_{pe}) - \exp(-\beta K_{pe}) \\ + \int_0^\beta \mathrm{d}\tau_1 \exp[-(\beta-\tau_1)K_{pe}]V_{pe}\exp[-\tau_1 K_{pe}] \\ - \int_0^\beta \mathrm{d}\tau_1 \int_0^{\tau_1}\mathrm{d}\tau_2 \exp[-(\beta-\tau_1)K_{pe}]V_{pe}\exp[-(\tau_1-\tau_2)K_{pe}]V_{pe}\exp[-\tau_2 K_{pe}] \\ + \int_0^\beta \mathrm{d}\tau_1 \int_0^{\tau_1}\mathrm{d}\tau_2 \int_0^{\tau_2}\mathrm{d}\tau_3 \exp[-(\beta-\tau_1)K_{pe}]V_{pe}\exp[-(\tau_1-\tau_2)K_{pe}] \\ \times V_{pe}\exp[-(\tau_2-\tau_3)K_{pe}]V_{pe}\exp[-\tau_3 K_{pe}] \tag{3.14}$$

($K_{pe} = -\hbar^2\Delta/(2m)$ and $V_{pe} = -e^2/r$).

Partition function (3.13) is similar to the so-called direct quantum virial function first introduced by Ebeling [23] (see Sect. 4.1). It incorporates contributions from both bound states (recombination of proton and electron into an hydrogen atom) and diffusion states (dissociation of an hydrogen atom into ionized proton and electron). Contrary to the trace of $[\exp(-\beta H_{pe})]_{\mathrm{Mayer}}$, $Z(1,1)$ is finite because $\langle \mathbf{r}^*|[\exp(-\beta H_{pe})]^T_{\mathrm{Mayer}}|\mathbf{r}^*\rangle$ decays as $1/(r^*)^4$ at large distances. Though truncation in $[\exp(-\beta H_{pe})]^T_{\mathrm{Mayer}}$ can be traced back to collective screening effects, $Z(1,1)$ depends only on temperature, and no longer on density.

In order to estimate (3.13) at low temperatures, we can heuristically extend the very simple argument used in Section 2.3 for estimating $B_{1,1}$. For $r^* \sim a_B$, contribution of ground state $\psi_0(r^*)$ of H_{pe} to $\langle \mathbf{r}^*|\exp(-\beta H_{pe})|\mathbf{r}^*\rangle$ exponentially dominates all other contributions because of the finite gap between E_H and the rest of the spectrum. Moreover, truncated terms in $[\exp(-\beta H_{pe})]^T_{\mathrm{Mayer}}$, which are crucial for ensuring the finiteness of the trace, do not generate exponentially growing terms at low temperatures, because they only involve Gibbs operators associated with kinetic Hamiltonian K_{pe}. Therefore, the leading behavior of (3.13) when $\beta \to \infty$, obtained by replacing $\langle \mathbf{r}^*|[\exp(-\beta H_{pe})]^T|\mathbf{r}^*\rangle$ by $|\psi_0(\mathbf{r}^*)|^2 \exp(-\beta E_H)$, merely is

$$Z(1,1) \sim 4\exp(-\beta E_H). \tag{3.15}$$

Beyond the previous heuristic argument, we present in Appendix B: a non-perturbative derivation of (3.15), which is quite useful for further purposes (see Sects. 3.3 and 3.7) since it provides convincing low-temperature estimations of quantities similar to (3.13) involving three or more particles.

Eventually, according to formula (3.12), the leading bare contribution of $G_{1,1}$ reads

$$\rho^* \frac{\gamma^2}{2}, \tag{3.16}$$

which is nothing but ideal contribution (2.16) of hydrogen atoms in their groundstate. Beyond leading term (3.16), the rest of the bare contribution of $G_{1,1}$ can be rewritten as

$$\rho^* \frac{\gamma^2}{8} Z_{exc}(1,1)\exp(\beta E_H), \tag{3.17}$$

with $Z_{exc}(1,1) = Z(1,1) - 4\exp(-\beta E_H)$. At low temperatures, leading contribution to (3.17) arises from the first excited level ($E_H^{(1)} = E_H/4$) of the hydrogen atom and reads

$$2\rho^*\gamma^2 \exp\left(\frac{3\beta E_H}{4}\right). \tag{3.18}$$

It can be viewed as the ideal density of hydrogen atoms in their first excited state. As expected, that level is less populated than the ground state by exponentially decaying Boltzmann factor $\exp(3\beta E_H/4)$ associated with energy difference $E_H^{(1)} - E_H = -3E_H/4$ (apart from the trivial factor 4 arising from orbital degeneracy of the first excited state).

If the identification of atomic states contributions (like (3.16) or (3.18)) makes sense in the zero-temperature limit defining Saha regime, at finite temperatures the definition of an atomic part Z_H in $Z(1,1)$ is arbitrary, as it has been noticed for a long time (see Ref. [39] and references quoted therein). That ambiguity is related to the fact that contributions of bound states with $|E_H^{(p)}| \leq k_B T$ cannot be disentangled from that of diffusion states since they have the same order of magnitude. A possible definition of Z_H is a finite sum of terms analogous to (3.18) up to p_{max} such that $|E_H^{(p_{max})}| \simeq k_B T$: that procedure accounts for expected thermal ionization which prevents the existence of highly excited hydrogen atoms in so-called Rydberg states. As emphasized in Ref. [39], only the full contribution embedded in $Z(1,1)$, obviously independent of above arbitrariness, is relevant for thermodynamics. Notice that diffusion state contributions describe (unscreened) short-distance interactions between ionized proton and electron. Such contributions are similar to that involved in $G_{2,0}$ (see Sect. 3.3), and they are smaller than ideal contribution (3.16) by exponential factor $\exp(\beta E_H)$ apart from powers of β.

3.2.2 Collective Corrections

The first contributions of $G_{1,1}$ due to collective effects are obtained by expanding, in (3.9), ring factors $\exp(I_R(a))$ and $\exp(I_R(1))$ in powers of $I_R(a)$ and $I_R(1)$, and Mayer coefficient $\mathcal{B}_\phi^T(a,1)$ in powers of $(\phi - V)(a,1)$. At lowest order, $I_R(a)$ and $I_R(1)$ behave as $\beta e^2\kappa/2$, while $(\phi - V)(a,1)$ behaves as $e^2\kappa$ at distances $r < \beta e^2$. Therefore, first polarization corrections, which are smaller than leading bare contribution (3.16) by an extra factor $\beta e^2\kappa$, cancel out: an Hydrogen atom, which is a neutral entity, does not polarize its surrounding plasma at lowest order.

Collective corrections to (3.16) are then determined by the behavior of I_R and $(\phi - V)$ beyond the previous simple constants. In other words, the bare proton-electron Coulomb potential is modified, beyond the familiar Debye shift, by a coupling between quantum fluctuations of both particles and the surrounding plasma. That effect cannot be incorporated into an effective potential. The corresponding calculation, performed in Appendix A:, gives at lowest order,

$$\rho^* \gamma^3 S_3(1,1) \exp(2\beta E_H), \tag{3.19}$$

where screening function $S_3(1,1)$ for cluster $(1,1)$ is given by (A.16). Contribution (3.19) is exponentially smaller than (3.16) by factor $\exp(\beta E_H)$ and must be retained at that order, because $S_3(1,1)$ behaves as a power of β times $\exp(-\beta E_H)$ (see (A.17)). As shown in Appendix A:, higher order collective corrections decay exponentially faster than (3.19).

3.3 Other Complex Entities

3.3.1 Two-Proton Cluster

A two-proton cluster is described by graph $G_{2,0}$ made with single root cluster C_a (see Fig. 6a). There are two possible loop configurations for the internal state of root cluster C_a: either the two protons belong to two different loops $\mathcal{L}_a^{(p)}$ and $\mathcal{L}_1^{(p)}$, or they belong to a single loop $\mathcal{L}_a^{(p)}$. The corresponding contribution reads

$$\begin{aligned}
&\int \mathcal{D}(\boldsymbol{\xi}_a) z_\phi(\mathcal{L}_a^{(p)}) \int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_1) z_\phi(\mathcal{L}_1^{(p)}) \mathcal{B}_\phi^T(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p)}) + \int \mathcal{D}(\boldsymbol{\eta}_a) 2 z_\phi(\mathcal{L}_a^{(p)}) \\
&= \frac{4 z_p^2}{(2\pi\lambda_p^2)^3} \int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_a)\mathcal{D}(\boldsymbol{\xi}_1) \exp(I_R(a)) \exp(I_R(1)) \\
&\quad \times \left[\exp(-\beta\phi(a,1)) - 1 + \beta\phi(a,1) - \frac{\beta^2\phi^2(a,1)}{2!} + \frac{\beta^3\phi^3(a,1)}{3!}\right] \\
&\quad - \frac{2 z_p^2}{(4\pi\lambda_p^2)^{3/2}} \int \mathcal{D}(\boldsymbol{\eta}_a) \exp(I_R(a)) \exp(-\beta U(a)).
\end{aligned} \tag{3.20}$$

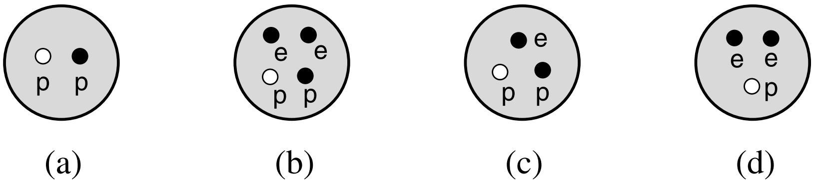


Fig. 6 Graphs representing various complex entities: **a** two-proton cluster; **b** molecule H_2; **c** ion H_2^+; **d** ion H^-

Like (3.9), (3.20) incorporates collective effects, i.e. dressing of each proton and screening of their mutual interactions.

At leading order, after applying backwards Feynman-Kac formula (2.32), we find that the bare contribution of Fig. 6a reduces to

$$\rho^* \frac{\gamma^2}{\sqrt{2}} \left(\frac{m}{m_p} \right)^{3/2} Z(2,0) \exp(\beta E_H), \tag{3.21}$$

with

$$\begin{aligned} Z(2,0) &= \frac{(\pi \lambda_p^2)^{3/2}}{\Lambda} \mathrm{Tr}[\exp(-\beta H_{2,0})]^T_{\mathrm{Mayer}} \\ &= \int \mathrm{d}\mathbf{r}^* \{ 2 \langle \mathbf{r}^* | [\exp(-\beta H_{pp})]^T_{\mathrm{Mayer}} | \mathbf{r}^* \rangle - \langle -\mathbf{r}^* | \exp(-\beta H_{pp}) | \mathbf{r}^* \rangle \}. \end{aligned} \tag{3.22}$$

In (3.22), H_{pp} is the one-body Hamiltonian of relative particle with position $\mathbf{r}^* = \mathbf{R}_1 - \mathbf{R}_a$ and mass $m_{pp} = m_p/2$, $H_{pp} = K_{pp} + V_{pp}$ with $K_{pp} = -\hbar^2 \Delta/(2m_{pp})$ and $V_{pp} = e^2/r$. Moreover, $[\exp(-\beta H_{pp})]^T_{\mathrm{Mayer}}$ is defined as (3.14) with K_{pp} and V_{pp} in place of K_{pe} and V_{pe} respectively. Like $Z(1,1)$, $Z(2,0)$ is also merely related to Ebeling quantum virial functions (see Sect. 4.1). Thanks to truncation in $[\exp(-\beta H_{pp})]^T_{\mathrm{Mayer}}$, the integral over $\mathbf{r}^*$ does converge contrary to the integral in (2.25) that formally defines $B_{2,0}$. Because of the continuous nature of the spectrum of H_{pp} which starts at zero, $Z(2,0)$ behaves as a power law at low temperatures. Contribution (3.21) then decays faster than ρ^* by exponential factor $\exp(\beta E_H)$ (discarding powers of β).

Collective corrections to (3.21) arise from expansions of ring factors and of Mayer coefficient in (3.20). At lowest order, we can use $I_R(a) \sim I_R(1) \sim \beta e^2 \kappa/2$ and $(\phi - V)(a,1) \sim -e^2\kappa$ $(r < \beta e^2)$ for $q_a = q_1 = 1$, while $I_R(a) \sim 2\beta e^2 \kappa$ for $q_a = 2$. Therefore, the first polarization correction to (3.21), which can be treated at a purely classical level, is smaller than ρ^* by factor $\exp(3\beta E_H/2)$.

3.3.2 Molecule H_2

Contribution of a molecule H_2 is embedded in graph $G_{2,2}$ made with the single root cluster C_a containing two protons and two electrons (Fig. 6b). Again, dressing of particles as well as screening of their mutual interactions can be treated perturbatively in the Saha regime. At leading order, the resulting bare contribution of $G_{2,2}$ is then transformed into

$$\rho^* \gamma^4 \frac{\sqrt{2}}{32} \left(\frac{m}{M} \right)^{3/2} Z(2,2) \exp(3\beta E_H), \tag{3.23}$$

with

$$\begin{aligned}Z(2,2) &= \frac{(2\pi\lambda_{H_2}^2)^{3/2}}{\Lambda}\,\mathrm{Tr}[\exp(-\beta H_{2,2})]^T_{\mathrm{Mayer}}\\ &= (2\pi\lambda_{H_2}^2)^{3/2}\int \mathrm{d}\mathbf{R}_1\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2\{4\langle\mathbf{R}_a\mathbf{R}_1\mathbf{r}_1\mathbf{r}_2|\exp(-\beta H_{2,2})|\mathbf{R}_a\mathbf{R}_1\mathbf{r}_1\mathbf{r}_2\rangle\\ &\quad - 2\langle\mathbf{R}_1\mathbf{R}_a\mathbf{r}_1\mathbf{r}_2|\exp(-\beta H_{2,2})|\mathbf{R}_a\mathbf{R}_1\mathbf{r}_1\mathbf{r}_2\rangle\\ &\quad - 2\langle\mathbf{R}_a\mathbf{R}_1\mathbf{r}_2\mathbf{r}_1|\exp(-\beta H_{2,2})|\mathbf{R}_a\mathbf{R}_1\mathbf{r}_1\mathbf{r}_2\rangle\\ &\quad + \langle\mathbf{R}_1\mathbf{R}_a\mathbf{r}_2\mathbf{r}_1|\exp(-\beta H_{2,2})|\mathbf{R}_a\mathbf{R}_1\mathbf{r}_1\mathbf{r}_2\rangle + \cdots\}\end{aligned} \tag{3.24}$$

($\lambda_{H_2} = (\beta\hbar^2/(2M))^{1/2}$). Like (3.14), truncated Mayer operator $[\exp(-\beta H_{2,2})]^T_{\mathrm{Mayer}}$ is defined as a suitable truncation of $[\exp(-\beta H_{2,2})]_{\mathrm{Mayer}}$ inherited from the structure of coefficients $\mathcal{B}^T_{\phi,N}$ ($N = 1, 2, 3, 4$). In addition to the terms already present in $[\exp(-\beta H_{2,2})]_{\mathrm{Mayer}}$, that truncation involves products of imaginary-time evolutions of interaction potentials between subsets of two protons and two electrons (for our purpose, it is not necessary to detail here all the numerous terms involved in that truncation). This ensures that $[\exp(-\beta H_{2,2})]^T_{\mathrm{Mayer}}$ has a finite trace contrary to $[\exp(-\beta H_{2,2})]_{\mathrm{Mayer}}$.

Similarly to (3.13), partition function (3.24) incorporates contributions from both recombination into molecules H_2, and dissociation (interactions at short distances between atoms H, ions H_2^+, ions H^-, ionized protons and ionized electrons). At low temperatures, the leading behavior of $Z(2,2)$ is determined by applying the method described in Appendix B:. A key ingredient is the discrete nature of the spectrum of $H_{2,2}$ (discarding the trivial contribution of the center of mass) near its infimum. Moreover, we assume quite weak bounds for three- and four-body Coulomb Green functions, inspired in part from their known exact two-body counterparts [36]. Then, we show that leading contribution to $Z(2,2)$ arises from the first four terms in the r.h.s of (3.24) evaluated for the ground state of molecule H_2 with energy $E_{H_2} = E^{(0)}_{2,2}$. Thus, despite truncated terms beyond matrix elements of $\exp(-\beta H_{2,2})$ not written explicitly in the r.h.s. of (3.24), are crucial for ensuring finiteness of $Z(2,2)$, they do not affect its leading low-temperature behavior which merely reads

$$Z(2,2) \sim \exp(-\beta E_{H_2}) \tag{3.25}$$

when $\beta \to \infty$. Since H_2 contains two protons, the resulting contribution (3.23) is twice ideal density $\rho^{id}_{H_2}$ of molecules H_2 in their para-groundstate where the two protons, as well as the two electrons, have opposite spin orientations, while the total angular momentum is zero.

First thermal corrections to (3.25) arise from molecular excited states. Contrarily to the atomic case, such states are not exactly known. However, according to the usual phenomenology, they are expected to be well described by para-states and ortho-states (the two protons have the same spin orientation) with non-zero angular momenta describing global rotations of the molecule [42]. Moreover, excited states with still higher energies can be associated with proton vibrations and ultimately electronic excitations [42].

Beyond above purely molecular terms, $Z(2,2)$ also incorporates short-range contributions which account for interactions between products of molecular dissociation, as well as the corresponding exchange effects. Similarly to the case of $Z(1,1)$ where atomic contributions are mixed to those of interactions between ionized-charges, the extraction of either a molecular part Z_{H_2} or an atom-atom contribution in $Z(2,2)$, remains arbitrary. Again, that arbitrariness does not cause any trouble for thermodynamics which depend only on the full contribution $Z(2,2)$.

Collective corrections to (3.23) embedded in $G_{2,2}$ can be studied as above (see Sect. 3.2). Like atom H, molecule H_2 is neutral so it does not polarize (at lowest order) the surrounding plasma. First collective corrections are then smaller than (3.25) by an extra factor $(\beta e^2\kappa)^2$ of order $\exp(\beta E_H)$. Therefore, they are smaller than ρ^* by a factor $\exp[\beta(4E_H - E_{H_2})]$, which is itself exponentially smaller than $\exp(\beta E_H)$ by virtue of inequality $3E_H < E_{H_2}$.

3.3.3 Ions H^- and H_2^+

Ions H_2^+ and H^- appear in graphs $G_{2,1}$ (Fig. 6c) and $G_{1,2}$ (Fig. 6d) respectively. The corresponding bare contributions are rewritten as

$$\rho^* \frac{\gamma^3}{16}\left(\frac{m_e(M+m_p)}{M^2}\right)^{3/2} Z(2,1)\exp(2\beta E_H) \tag{3.26}$$

and

$$\rho^* \frac{\gamma^3}{32}\left(\frac{m_p(M+m_e)}{M^2}\right)^{3/2} Z(1,2)\exp(2\beta E_H), \tag{3.27}$$

with

$$Z(2,1) = \frac{(2\pi\lambda_{H_2^+}^2)^{3/2}}{\Lambda}\,\mathrm{Tr}[\exp(-\beta H_{2,1})]^T_{\mathrm{Mayer}} \tag{3.28}$$

and

$$Z(1,2) = \frac{(2\pi\lambda_{H^-}^2)^{3/2}}{\Lambda}\,\mathrm{Tr}[\exp(-\beta H_{1,2})]^T_{\mathrm{Mayer}} \tag{3.29}$$

($\lambda_{H_2^+} = (\beta\hbar^2/(M+m_p))^{1/2}$ and $\lambda_{H^-} = (\beta\hbar^2/(M+m_e))^{1/2}$). Truncated Mayer operators $[\exp(-\beta H_{2,1})]^T_{\mathrm{Mayer}}$ and $[\exp(-\beta H_{1,2})]^T_{\mathrm{Mayer}}$ are defined similarly to $[\exp(-\beta H_{2,2})]^T_{\mathrm{Mayer}}$ and $[\exp(-\beta H_{1,1})]^T_{\mathrm{Mayer}}$. The low-temperature behaviors of (3.26) and (3.27) are determined by applying the method described in Appendix B:. As for (3.13) and (3.24), truncated terms beyond $\exp(-\beta H_{2,1})$ or $\exp(-\beta H_{1,2})$ do not contribute at leading order. Therefore, we find that (3.26) behaves as

$$\rho^* \frac{\gamma^3}{8}\left(\frac{m_e(M+m_p)}{M^2}\right)^{3/2} \exp[\beta(2E_H - E_{H_2^+})] = 2\rho^{id}_{H_2^+}, \tag{3.30}$$

where $\rho^{id}_{H_2^+}$ is the ideal density of ions H_2^+ in their groundstate with energy $E_{H_2^+} = E^{(0)}_{2,1}$, which is doubly degenerated because of electron spin. Similarly, we obtain leading behavior of (3.27), i.e.

$$\rho^* \frac{\gamma^3}{16}\left(\frac{m_p(M+m_e)}{M^2}\right)^{3/2} \exp[\beta(2E_H - E_{H^-})] = \rho^{id}_{H^-}, \tag{3.31}$$

where $\rho^{id}_{H^-}$ is the ideal density of ions H^- in their groundstate with energy $E_{H^-} = E^{(0)}_{1,2}$, which is doubly degenerated because of proton spin. Like (3.23), those ideal contributions decay exponentially faster than ρ^* in the Saha regime. Density effects embedded in $G_{2,1}$ and $G_{1,2}$ are similar to those encountered above for an ionized proton. They provide contributions which are smaller than ρ^* by factors $\exp(\beta(5E_H/2 - E_{H_2^+}))$ and $\exp(\beta(5E_H/2 - E_{H^-}))$, while such factors are themselves exponentially small compared to $\exp(\beta E_H)$ by virtue of inequalities $3E_H/2 < E_{H_2^+}$ and $3E_H/2 < E_{H^-}$ (see numerical values given in Sect. 2.3).

3.4 Interactions between Ionized Charges beyond Polarization Effects

Since the Saha regime is quite diluted and weakly coupled (see Sect. 2.3), leading contributions of screened interactions are embedded in the polarization mechanism described in Sect. 3.1 for a graph with a single particle. This provides well-known Debye correction (3.4). Beyond that mean-field contribution, next contributions of interactions between ionized charges arise from graphs involving two particles, namely $G_{1,1}$ and $G_{2,0}$ shown in Figs. 5b, 6a and 7a–c. As quoted above, graphs made with one cluster (Figs. 5b and 6a) involve contributions of unscreened interactions at short distances. Graphs made with two clusters, C_a (one proton) and C_1 (one proton or one electron), connected by a single bond $\mathcal{F}_\phi(C_a, C_1)$ which can be either $-\beta\Phi$ (Fig. 7a), $\beta^2\Phi^2/2!$ (Fig. 7b), or $-\beta^3\Phi^3/3!$ (Fig. 7c), account for large-distance screened contributions which are estimated as follows.

Graphs shown in Fig. 7a (with $\alpha = p, e$) provide contribution

$$
\begin{aligned}
&-\beta \int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_a)\mathcal{D}(\boldsymbol{\xi}_1) z_\phi(\mathcal{L}_a^{(p)})[z_\phi(\mathcal{L}_1^{(p)})\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p)}) + z_\phi(\mathcal{L}_1^{(e)})\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(e)})] \\
&\quad = -\frac{4\beta z_p}{(2\pi\lambda_p^2)^{3/2}} \int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_a)\mathcal{D}(\boldsymbol{\xi}_1) \exp(I_R(\mathcal{L}_a^{(p)})) \\
&\quad\quad \times \left[\frac{z_p}{(2\pi\lambda_p^2)^{3/2}} \exp(I_R(\mathcal{L}_1^{(p)}))\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p)}) \right. \\
&\quad\quad \left. + \frac{z_e}{(2\pi\lambda_e^2)^{3/2}} \exp(I_R(\mathcal{L}_1^{(e)}))\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(e)}) \right].
\end{aligned} \tag{3.32}
$$

The expansion of ring factors $\exp(I_R(\mathcal{L}_a^{(p)}))$ and $\exp(I_R(\mathcal{L}_1^{(p)}))$ provides a first contribution which vanishes by virtue of identity (A.3) derived in Appendix A:. The first non-vanishing contribution arises from linear terms $I_R(\mathcal{L}_1^{(p)})$ and $I_R(\mathcal{L}_1^{(e)})$ where loop-shape dependences beyond classical behavior (3.2) are included. At lowest order, ϕ can then be replaced by its classical Debye form, and the resulting leading contribution of Fig. 7a is

$$
-\rho^* \frac{\gamma^2}{2}[S_2(1,0) - S_2(0,1)]\exp(\beta E_H) \tag{3.33}
$$

with

$$
S_2(0,1) = \frac{(\beta|E_H|)^{3/2}}{2\sqrt{\pi}} - \left(1 + \left(\frac{2m}{m_e}\right)^{1/2}\right)\frac{\beta|E_H|}{8}, \tag{3.34}
$$

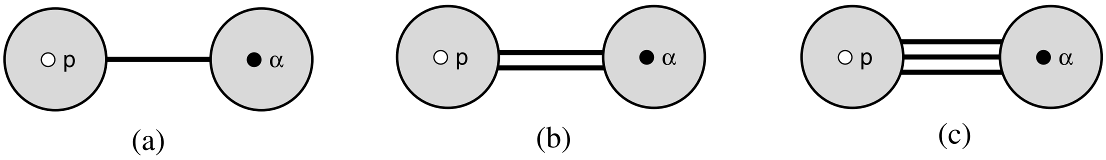


Fig. 7 Graphs describing screened interactions between one proton and one electron ($\alpha = e$), or between two ionized protons ($\alpha = p$)

which follows from (A.11). Further contributions decay exponentially faster than $\rho^* \exp(\beta E_H)$.

Because weight of cluster C_1 has specific form (2.42), contribution of Fig. 7b reads

$$\begin{aligned}
&\frac{\beta^2}{2} \int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_a)\mathcal{D}(\boldsymbol{\xi}_1) z_\phi(\mathcal{L}_a^{(p)})[(z_\phi(\mathcal{L}_1^{(p)}) - z(\mathcal{L}_1^{(p)}))\phi^2(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p)}) \\
&\qquad + (z_\phi(\mathcal{L}_1^{(e)}) - z(\mathcal{L}_1^{(e)}))\phi^2(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(e)})] \\
&\quad = \frac{2\beta^2 z_p}{(2\pi\lambda_p^2)^{3/2}} \int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_a)\mathcal{D}(\boldsymbol{\xi}_1) \exp(I_R(\mathcal{L}_a^{(p)})) \\
&\qquad \times \left[\frac{z_p}{(2\pi\lambda_p^2)^{3/2}} (\exp(I_R(\mathcal{L}_1^{(p)})) - 1)\phi^2(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p)}) \right. \\
&\qquad \left. + \frac{z_e}{(2\pi\lambda_e^2)^{3/2}} (\exp(I_R(\mathcal{L}_1^{(e)})) - 1)\phi^2(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(e)}) \right].
\end{aligned} \tag{3.35}$$

At lowest order, we can replace factors $(\exp(I_R(\mathcal{L}_1^{(p)})) - 1)$ and $(\exp(I_R(\mathcal{L}_1^{(e)})) - 1)$ by $\beta e^2 \kappa/2$ on one hand, and ϕ by its classical Debye form on another hand. This provides the leading contribution of (3.35)

$$\begin{aligned}
&(\rho_p^{id})^2 \frac{\beta^3 e^6}{2} \kappa \int \mathrm{d}\mathbf{X}_1 \frac{\exp(-2\kappa|\mathbf{X}_1 - \mathbf{X}_a|)}{|\mathbf{X}_1 - \mathbf{X}_a|^2} \\
&\quad = \pi(\rho_p^{id})^2 \beta^3 e^6 = \rho^* \gamma^2 [W(1,0|1,0) + W(1,0|0,1)] \exp(\beta E_H)
\end{aligned} \tag{3.36}$$

with

$$W(1,0|1,0) = W(1,0|0,1) = \frac{(\beta|E_H|)^{3/2}}{4\sqrt{\pi}}, \tag{3.37}$$

in agreement with asymptotic formula (A.6) derived in Appendix A:. Functions W can be interpreted as resulting from effective interactions between ionized charges generated by quadratic fluctuations of ϕ. Next corrections to (3.36) decay exponentially faster than $\rho^* \exp(\beta E_H)$, as inferred from (A.6) and (A.11).

Eventually, contribution of Fig. 7c is

$$\begin{aligned}
&\frac{-\beta^3}{3!} \int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_a)\mathcal{D}(\boldsymbol{\xi}_1) z_\phi(\mathcal{L}_a^{(p)}) \left[z_\phi(\mathcal{L}_1^{(p)})\phi^3(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p)}) + z_\phi(\mathcal{L}_1^{(e)})\phi^3(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(e)}) \right] \\
&\quad = -\frac{2\beta^3 z_p}{3(2\pi\lambda_p^2)^{3/2}} \int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_a)\mathcal{D}(\boldsymbol{\xi}_1) \exp(I_R(\mathcal{L}_a^{(p)})) \\
&\qquad \times \left[\frac{z_p}{(2\pi\lambda_p^2)^{3/2}} \exp(I_R(\mathcal{L}_1^{(p)}))\phi^3(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p)}) \right. \\
&\qquad \left. + \frac{z_e}{(2\pi\lambda_e^2)^{3/2}} \exp(I_R(\mathcal{L}_1^{(e)}))\phi^3(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(e)}) \right],
\end{aligned} \tag{3.38}$$

with $q_a = q_1 = 1$, $\boldsymbol{\eta}_a^{(p)} = \boldsymbol{\xi}_a$ and $\boldsymbol{\eta}_1^{(p,e)} = \boldsymbol{\xi}_1$. All collective effects can be omitted in (3.38) at lowest order, so leading contribution reads

$$-\rho^* \gamma^2 \frac{c_p (\beta |E_H|)^{3/2}}{12\pi^{3/2}} \exp(\beta E_H) \tag{3.39}$$

with numerical constant c_p given by (A.8). Next corrections to (3.39) decay exponentially faster than $\rho^* \exp(\beta E_H)$.

3.5 Interactions between an Atom and an Atom or an Ionized Charge

As argued in Sect. 2.3, atoms H are expected to be weakly coupled under Saha conditions, like ionized charges (see Sects. 3.1 and 3.4). Leading contributions of interactions between atoms and ionized charges should then involve either two atoms or a single one. As quoted in Sect. 3.3, short-range parts of those contributions are embedded in Figs. 6b–d made with a single cluster. Here, we consider other graphs made with two clusters which account for complementary parts including long-range effects.

3.5.1 Atom-Atom Interactions

Figures 8a–c describe interactions between two atoms. Contrary to the case of ionized charges, screening effects can now be omitted at leading order, because each atom is neutral. In other words, potential $\Phi(\mathcal{C}_a, \mathcal{C}_1)$ between clusters $\mathcal{C}_a$ and $\mathcal{C}_1$ can be replaced by its bare counterpart $V(\mathcal{C}_a, \mathcal{C}_1)$, which decays as a dipolar interaction (the corresponding $1/R^3$-decay is sufficient for ensuring integrability in Fig. 8a for symmetry reasons). Of course, in statistical weights defining internal states of $\mathcal{C}_a$ and $\mathcal{C}_1$, collective effects can be also ignored at leading order. Then contribution of Fig. 8a vanishes by symmetry, while the resulting bare contributions of Figs. 8b and 8c can be rewritten in terms of matrix elements of Gibbs operators by applying backwards Feynman-Kac formula (2.32). Leading contribution of Fig. 8b reads

$$\begin{aligned} z_p^2 z_e^2 \int \mathrm{d}\mathbf{R}_1 \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \Big\{ & 16 \langle \mathbf{R}_a \mathbf{R}_1 \mathbf{r}_1 \mathbf{r}_2 | \\ & \times \int_0^\beta \mathrm{d}\tau_1 \int_0^{\tau_1} \mathrm{d}\tau_2 \exp[-(\beta - \tau_1)(H_{at} + H_{at})] V_{at,at} \\ & \times \exp[-(\tau_1 - \tau_2)(H_{at} + H_{at})] V_{at,at} \\ & \times \exp[-\tau_2 (H_{at} + H_{at})] | \mathbf{R}_a \mathbf{R}_1 \mathbf{r}_1 \mathbf{r}_2 \rangle + \cdots \Big\}, \end{aligned} \tag{3.40}$$

where $H_{at} = H_{1,1}$ is the Hamiltonian of a single atom, while $V_{at,at}$ is the interaction potential between two atoms. Terms $\cdots$ in (3.40) have a structure analogous to those subtracted

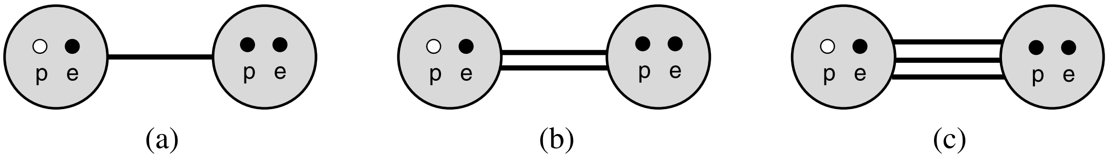


Fig. 8 Graphs accounting for interactions between two hydrogen atoms

from $\exp(-\beta H_{pe})$ in (3.14). The corresponding truncation, inherited from that in the $\mathcal{B}^T$'s, is analogous to that defining individual atomic partition functions: it ensures that spatial integration over $\mathbf{R}_1, \mathbf{r}_1, \mathbf{r}_2$ does converge. An expression similar to (3.40) can be obtained for Fig. 8c.

Full bare contribution of Figs. 8b and 8c takes the form (see Appendix C:)

$$\rho^* \gamma^4 W(1,1|1,1) \exp(3\beta E_H), \tag{3.41}$$

discarding terms which decay exponentially faster than $\rho^* \exp(\beta E_H)$. When $\beta \to \infty$, $W(1,1|1,1)$ behaves as

$$W(1,1|1,1) \sim \frac{c_{at,at}}{32\pi^{3/2}(\beta|E_H|)^{1/2}} \exp(-2\beta E_H) \tag{3.42}$$

where $c_{at,at}$ is the pure numerical coefficient (C.1). Function $W(1,1|1,1)$ accounts for unscreened interactions between two hydrogen atoms in their groundstate. Contributions from both short and large separations R are involved. In particular, contributions from familiar van der Waals interactions $U_{H-H}(R) = -A_{H-H}/R^6$ (with positive constant A_{H-H} computed from quantum perturbation theory at zero temperature [42]), do emerge through the large-distance ($R = |\mathbf{R}_2^* - \mathbf{R}_1^*| \gg \lambda_H$) behavior

$$\begin{aligned}
&\langle \mathbf{R}_a \mathbf{R}_1 \mathbf{r}_1 \mathbf{r}_2 | \int_0^\beta \mathrm{d}\tau_1 \int_0^{\tau_1} \mathrm{d}\tau_2 \exp[-(\beta-\tau_1)(H_{at}+H_{at})] V_{at,at} \\
&\quad \times \exp[-(\tau_1-\tau_2)(H_{at}+H_{at})] V_{at,at} \exp[-\tau_2(H_{at}+H_{at})] | \mathbf{R}_a \mathbf{R}_1 \mathbf{r}_1 \mathbf{r}_2 \rangle \\
&\sim -\frac{\exp(-2\beta E_H)}{(2\pi\lambda_H^2)^3} |\psi_0(r_1^*)|^2 |\psi_0(r_2^*)|^2 \beta U_{H-H}(|\mathbf{R}_1^* - \mathbf{R}_2^*|),
\end{aligned} \tag{3.43}$$

for spatial configurations $r_1^* \sim r_2^* \sim a_B$ and sufficiently low temperatures.

Collective corrections to (3.41) are exponentially smaller than its leading behavior. Notice that they arise from various effects: plasma polarization associated with ring factors $\exp(I_R)$, Debye exponential screening of interactions at scales κ^{-1}, and also modification of $1/R^6$-tails at distances larger than l_Q as detailed elsewhere [9].

3.5.2 Atom-Proton and Atom-Electron Interactions

Figs. 9a–f account for interactions between one atom H and a single ionized charge. Leading contribution of Fig. 9a (obtained by replacing ring factors by 1) vanishes by virtue of

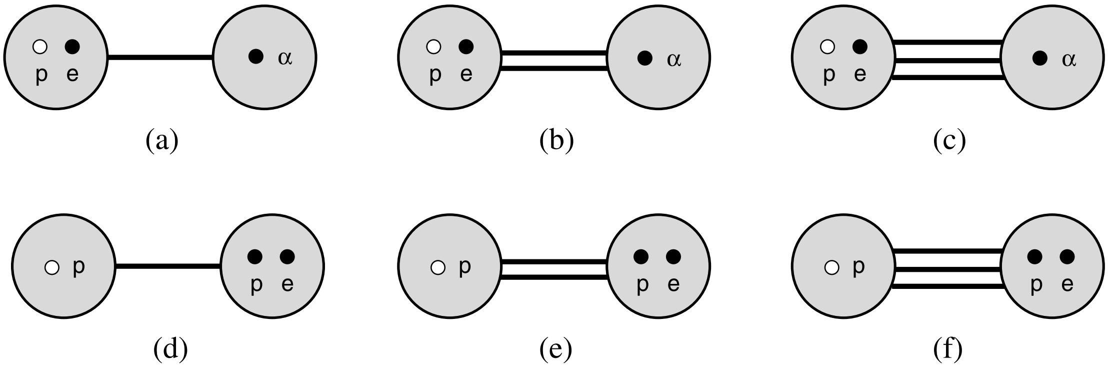


Fig. 9 Graphs accounting for interactions between a hydrogen atom and an ionized charge ($\alpha = p$ or e)

identity (A.3). Like Fig. 8a, bare contribution of Fig. 9d also vanishes for symmetry reasons. Therefore contributions of Figs. 9a and 9d decay faster than $\rho^* \exp(\beta E_H)$. Figs. 9e and 9f provide contributions obviously identical to that of Figs. 9b and 9c with $\alpha = p$. In Figs. 9b and 9c, all collective effects can be neglected at leading order, in particular $\phi(C_a, C_1)$ can be replaced by bare potential $V(C_a, C_1)$. Within that substitution, integrability at large distances R between C_a and C_1, is obviously ensured thanks to dipole-charge $1/R^2$ decay of $V(C_a, C_1)$. The resulting bare contributions of above graphs are rewritten in terms of matrix elements of Gibbs operators similarly to (3.40). For instance, bare contribution of Fig. 9b with $\alpha = p$ reads

$$z_p^2 z_e \int \mathrm{d}\mathbf{R}_1 \mathrm{d}\mathbf{r}_1 \{8\langle \mathbf{R}_a \mathbf{R}_1 \mathbf{r}_1 | \int_0^{\beta} \mathrm{d}\tau_1 \int_0^{\tau_1} \mathrm{d}\tau_2 \exp[-(\beta - \tau_1)(H_{at} + H_p)] V_{at,p} \\ \times \exp[-(\tau_1 - \tau_2)(H_{at} + H_p)] V_{at,p} \exp[-\tau_2 (H_{at} + H_p)] | \mathbf{R}_a \mathbf{R}_1 \mathbf{r}_1 \rangle + \cdots \} \quad (3.44)$$

where $H_p = H_{1,0}$ is the Hamiltonian of a single proton, while $V_{at,p}$ is the total interaction potential between an atom and a proton. Like in (3.40), terms $\cdots$ in (3.44) have a structure analogous to those subtracted from $\exp(-\beta H_{pe})$ in (3.14), which ensures spatial integrability over $\mathbf{R}_1$ and $\mathbf{r}_1$. Bare contributions of the other considered graphs can be expressed similarly to (3.44).

As shown in Appendix C:, the full bare contribution of Figs. 9a–f behaves as

$$\rho^* \gamma^3 [2W(1,1|1,0) + W(1,1|0,1)] \exp(2\beta E_H), \quad (3.45)$$

plus terms which decay exponentially faster than $\rho^* \exp(\beta E_H)$ when $\beta \to \infty$. Functions $W(1,1|1,0)$ and $W(1,1|0,1)$ account for unscreened interactions between an atom in its groundstate and an ionized charge. Their low-temperature behaviors are

$$\begin{aligned} W(1,1|1,0) &\sim \frac{c_{at,p}}{16\pi^{3/2}(\beta|E_H|)^{1/2}} \exp(-\beta E_H), \\ W(1,1|0,1) &\sim \frac{c_{at,e}}{16\pi^{3/2}(\beta|E_H|)^{1/2}} \exp(-\beta E_H) \end{aligned} \quad (3.46)$$

where $c_{at,\alpha}$ are pure numerical constants (C.3). Long-range contributions to $W(1,1|1,0)$ and $W(1,1|0,1)$ do reduce to that of the attractive interactions $U_{H-\alpha}(R) = -A_{H-\alpha}/R^4$ between an atom H and an ionized charge, with positive constant $A_{H-p} = A_{H-e}$ computed within quantum perturbation theory at zero temperature.

First collective corrections result from plasma polarization by the considered ionized charges, and they reduce to (3.45) multiplied by simple factor $\beta e^2 \kappa/2$ of order $\exp(\beta E_H/2)$. As for atom-atom interactions, part of further density corrections result from screening of atom-proton or atom-electron interactions at large distances.

3.6 Interactions between an Ionized Proton and Charged Clusters

Screened interactions between an ionized proton and charged clusters are embedded in any graph made with a root cluster C_a containing a single proton connected to a charged cluster C_1 *via* a bond $-\beta\Phi$. As shown below, such a graph provides a contribution which behaves, at leading order, as that of the part connected to C_a through C_1. Moreover, that mechanism enforces charge neutrality ($\rho_p = \rho_e$) by symmetrizing protonic and electronic contributions to SCE of ρ_p.

3.6.1 Two-Proton and Two-Electron Clusters

In Figs. 10a and 10b, C_1 contains either two protons or two electrons. Leading contribution of Fig. 10a arises from relative distances between clusters C_a and C_1 of order κ^{-1}, while relative distances between particles inside cluster C_1 are of order βe^2. For such configurations, $\Phi(C_a, C_1)$ can be replaced by its Debye classical form $-2\beta e^2 \exp(-\kappa X)/X$, where $\mathbf{X}$ is the relative distance between C_a and C_1 (cluster C_1 carries a total charge $2e$). At the same time, statistical weights Z_ϕ^T can be replaced by their bare forms. Then, integration over internal degrees of freedom of C_1 merely provides half contribution (3.21) of Fig. 6a made with a single root cluster identical to C_1: that factor $1/2$ arises from the combinatorics specific to root cluster of any graph (see comment after formula (2.38) and factor q_a in the corresponding contribution). Integration over internal degrees of freedom of C_a obviously provides ρ_p^{id}, while the remaining spatial integration over $\mathbf{X}$ reduces to

$$\int \mathrm{d}\mathbf{X}\left[-2\beta e^2 \frac{\exp(-\kappa X)}{X}\right] = -\frac{8\pi\beta e^2}{\kappa^2} = -\frac{1}{\rho_p^{id}}. \tag{3.47}$$

Eventually, leading contribution of Fig. 10a is

$$-\rho^* \frac{\gamma^2}{2\sqrt{2}}\left(\frac{m}{m_p}\right)^{3/2} Z(2,0)\exp(\beta E_H), \tag{3.48}$$

i.e. minus half bare contribution (3.21) of Fig. 6a. Next corrections to (3.47) decay exponentially faster than $\rho^* \exp(\beta E_H)$. A similar calculation provides leading contribution of Fig. 10b

$$\rho^* \frac{\gamma^2}{2\sqrt{2}}\left(\frac{m}{m_e}\right)^{3/2} Z(0,2)\exp(\beta E_H), \tag{3.49}$$

where we have used that C_1 carries a charge $-2e$. Next corrections to (3.49) also decay exponentially faster than $\rho^* \exp(\beta E_H)$.

3.6.2 Ions

Leading contributions of Figs. 10c and 10d can be treated as above. Taking into account that ion H_2^+ carries a charge e, we find for Fig. 10c

$$-\rho^* \frac{\gamma^3}{64}\left(\frac{m_e(M+m_p)}{M^2}\right)^{3/2} Z(2,1)\exp(2\beta E_H), \tag{3.50}$$

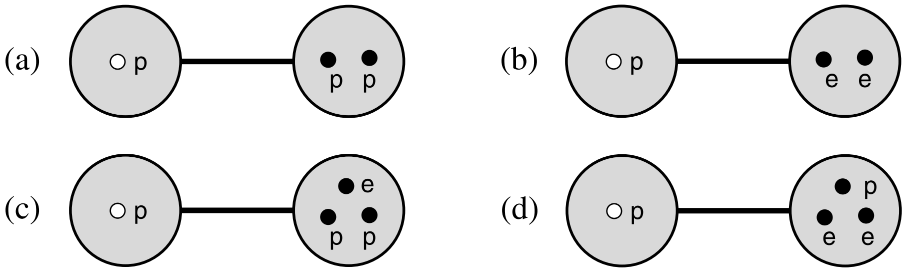


Fig. 10 Graphs accounting for interactions between an ionized proton and a charged cluster

i.e. minus one fourth bare contribution (3.26) of Fig. 6c. For Fig. 10d, no combinatorics factor 1/2 appears when integrating over internal degrees of freedom of C_1 because C_1 contains a single proton. Since C_1 carries a charge $-e$, leading contribution of Fig. 10d becomes

$$\rho^* \frac{\gamma^3}{64}\left(\frac{m_p(M+m_e)}{M^2}\right)^{3/2} Z(1,2)\exp(2\beta E_H), \tag{3.51}$$

i.e. half bare contribution (3.27) of Fig. 6d. Contributions (3.50) and (3.51) can be interpreted as the modification of density of ionized protons due to their coupling with ions H_2^+ and H^- respectively. As mentioned above, those contributions added to that of Figs. 6c and 6d provide a full contribution to ρ_p which is indeed identical to that relative to ρ_e. Thus, charge neutrality is indeed enforced by the structure of SCE (2.37).

Next corrections to (3.50) and (3.51) decay exponentially faster than $\rho^* \exp(\beta E_H)$, as well as all other non-ideal contributions of ions H_2^+ and H^-. Part of such contributions may be related to modifications of screening length, which are taken into account by summing suitable chain graphs (we have checked that this does provide the screening Debye length for a mixture of ionized charges and ions).

3.6.3 Other Charged Clusters

Eventually, Figs. 11a–d made with three clusters C_a, C_1 and C_2, also provide leading contributions of order $\rho^* \exp(\beta E_H)$ via the same mechanism as above. At leading order, $\Phi(C_a, C_1)$ can be replaced by its Debye classical form. Then, integrations over internal degrees of freedom of clusters C_1 and C_2, and over relative distance $\mathbf{X}_2 - \mathbf{X}_1$ between those clusters, are identical (apart from obvious substitutions $p \to e$) to those relative to Figs. 7b (for 11a), 7c (for 11b), 9e (for 11c) and 9f (for 11d). Using again identity (3.47) for integration over $\mathbf{X} = \mathbf{X}_1 - \mathbf{X}_a$, and (A.7) for the integral of ϕ^3, we obtain for Fig. 11b (C_1 is made with a single particle and carries a charge $\pm e$)

$$\rho^* \gamma^2 \frac{(c_p - c_e)(\beta|E_H|)^{3/2}}{24\pi^{3/2}} \exp(\beta E_H), \tag{3.52}$$

and for Figs. 11c and 11d

$$\rho^* \frac{\gamma^3}{2}[W(1,1|0,1) - W(1,1|1,0)]\exp(2\beta E_H). \tag{3.53}$$

Total leading contribution of Fig. 11a vanishes by charge neutrality constraint (2.8). Next corrections to (3.52) and (3.53) decay exponentially faster than $\rho^* \exp(\beta E_H)$.

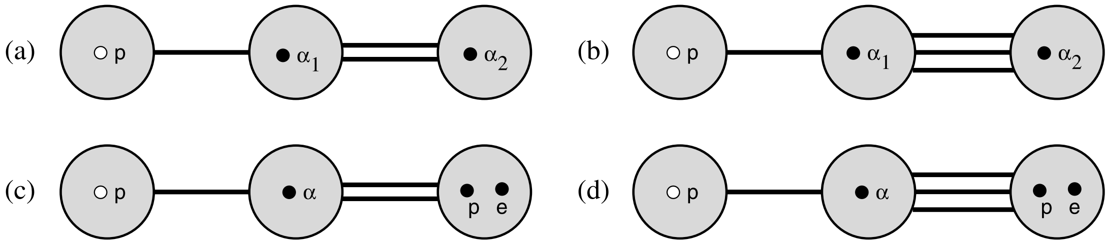


Fig. 11 Graphs of order $\rho^* \exp(\beta E_H)$ accounting for interactions between three clusters

3.7 Contributions with Arbitrary Particle Numbers

The evaluation of any contribution arising from a graph G_{N_p,N_e} made with at least three particles ($N_p + N_e \geq 3$), can be carried out by extending the methods described above for graphs with few particles. The outlines of the analysis are briefly sketched below. We first proceed to an estimation of the leading contribution. The behaviors of further collective corrections are discussed afterwards.

3.7.1 Leading Contributions

$*$ At leading order, we make the substitutions $\exp(I_R) \to 1$ and $\mathcal{B}_\phi^T \to \mathcal{B}^T$ in any weight $Z_\phi^T(C_i)$, and $(\exp(I_R) - 1) \to I_R$ in specific weight (2.42). Moreover, we explicit each $\mathcal{D}(C_i)$ in terms of spatial integrations over particle positions and of functional integrations over Brownian bridges.

$*$ Let consider two clusters C_i and C_j connected by a bond $\mathcal{F}_\phi(C_i, C_j)$. If one of them is electrically neutral, i.e. it contains the same number of protons and electrons, ϕ can be replaced by V in $\mathcal{F}_\phi(C_i, C_j)$. If both carry a net charge, ϕ must be replaced by its Debye classical form ϕ_D.

$*$ By virtue of Feynman-Kac formula, functional integrations over Brownian bridges reduce to matrix elements of either $\exp(-\beta H_{M_p,M_e})$, or interactions V evolved according to $\exp(-\tau H_{M_p,M_e})$ $(0 \leq \tau \leq \beta)$.

$*$ In graphs only made with bare bonds, integrations over positions of particles provide a function W accounting for bare interactions between clusters. If G_{N_p,N_e} contains a single cluster, such integrations give raise to partition function $Z(N_p, N_e)$.

$*$ When one or more bonds involve ϕ_D, positions of particles which belong to charged clusters connected by such bonds, are rewritten in terms of relative positions inside a given cluster and cluster position (arbitrarily defined as the position of a given particle). Let consider a charged cluster $C(M_p, M_e)$ $(M_p + M_e \neq 0)$, not connected to any neutral cluster. Integration over its position $\mathbf{X}$ can be disentangled from integrations over internal relative positions, since its internal weight $\mathcal{B}^T$ decays on a scale βe^2 much smaller than κ^{-1} which controls the decay of ϕ_D. Integration over its internal relative positions provide partition function $Z(M_p, M_e)$. Integration over $\mathbf{X}$ is performed by rescaling $\mathbf{X}$ in units of κ^{-1}. This provides multiplicative inverse powers of κ, with possible logarithmic terms $\ln(\kappa\lambda)$ arising from integrands built with ϕ_D^3.

$*$ According to above analysis and prescriptions, the leading contribution of G_{N_p,N_e} can be rewritten as (apart from a pure numerical coefficient which depends on ratio m_e/m_p)

$$\rho^* \gamma^{N_p+N_e-P/2} \exp[\beta(N_p + N_e - 1 - P/2)E_H] \prod Z \prod W \tag{3.54}$$

where each Z and each W depends only on temperature, while P is a positive integer. Term $\gamma^{-P/2}\exp(-P\beta E_H/2)$ arises from contribution $1/\kappa^P$, which is generated by both integrations over positions of charged clusters and specific weights I_R proportional to κ ($P = 0$ when G_{N_p,N_e} contains only neutral clusters).

$*$ The low-temperature behaviors of functions Z and W can be inferred from the methods exposed in Appendix B:. If there exists a bound state made with M_p protons and M_e electrons, partition function $Z(M_p, M_e)$ then behaves as

$$\exp(-\beta E^{(0)}_{M_p,M_e}), \tag{3.55}$$

apart from a multiplicative integer which accounts for groundstate degeneracy. In the other case, asymptotic behavior (3.55) has to be multiplied by some power of β. A given interaction function W behaves as the product of Boltzmann factors (3.55) associated with each interacting cluster times a power of β.

$*$ According to the above low-temperature behaviors of Z and W, leading contribution (3.54) of G_{N_p,N_e} reduces to $\rho^*\gamma^{N_p+N_e-P/2}$ times a power of β times

$$\exp[\beta(N_p+N_e-1-P/2)E_H]\prod\exp(-\beta E^{(0)}_{M_p,M_e}) \tag{3.56}$$

when $\beta\to\infty$. The precise form of factor (3.56) has been studied above for several graphs G_{N_p,N_e}. For all other graphs, we have checked that (3.56) is exponentially smaller than $\exp(\beta E_H)$. In particular, ideal contributions of complex entities made with more than four particles can be omitted at considered order. The analysis is achieved by using the known values of $E_{H_2^+}$, E_{H^-} and E_{H_2} given in Sect. 2.2, as well as inequality (2.31) for $M_p+M_e\geq 5$. Neutrality constraint (2.8) ensures the cancellation of the leading contributions of graphs which differ only by ending clusters made with either a single proton or a single electron: above statement then applies, strictly speaking, to the leading contribution of the sum of those graphs (for instance, see graphs $G_{1,1}$ and $G_{2,0}$ shown in Fig. 7a).

3.7.2 Collective Corrections

$*$ Collective corrections are obtained by expanding ring factors $\exp(I_R)$ in powers of I_R, and truncated Mayer coefficients $\mathcal{B}^T_\phi$ in powers of $(\phi-V)$. At the same time, both I_R and $(\phi-V)$ are expanded in positive integer powers of $\kappa\lambda$ as described in Appendix A:. Then, integration over cluster degrees of freedom involved in previous expansions, provide screening functions S. For a given graph G_{N_p,N_e}, the resulting corrections take the general form

$$\rho^*\gamma^{N_p+N_e-P/2+L/2}\exp[\beta(N_p+N_e-1-P/2+L/2)E_H]\prod Z\prod W\prod S \tag{3.57}$$

with L a positive integer.

$*$ The low-temperature behavior of S is analogous to those of Z and W, and it reduces to the product of a power of β times groundstate Boltzmann factors (3.55) associated with each cluster involved in S (for instance, see the calculation of $S_3(1,1)$ detailed in Appendix A:). Thus, and as expected from the weakly-coupled conditions enforced in the Saha regime, any correction (3.57) arising from G_{N_p,N_e} becomes exponentially smaller than its leading contribution (3.54) when $\beta\to\infty$. Collective corrections arising from graphs considered in Sects. 3.1–3.6 have been explicitly computed up to order $\rho^*\exp(\beta E_H)$ included. All other corrections, in particular those arising from other graphs, decay exponentially faster than $\rho^*\exp(\beta E_H)$.

4 Scaled Low-Temperature Expansions

According to the analysis of Sect. 3, we derive the structure of the asymptotic expansion of ρ/ρ^* (Sect. 4.1). Then, we proceed to the calculation of the pressure as a function of ρ (EOS), by using thermodynamical identities (Sect. 4.2). We derive the corresponding expansion around ideal Saha pressure (1.3), and the first four corrections are explicitly computed.

4.1 Structure of the Asymptotic Expansion of Particle Density

According to Sect. 3.7, every contribution arising from any graph G_{N_p,N_e} can be rewritten as ρ^* times γ^n times a temperature-dependent function. Power n is integer or half-integer, $n \geq 1$, while γ^n may be multiplied by integer powers of $\ln \gamma$ (it is not necessary to write explicitly such logarithmic terms since they do not play any role in the following). For a given n, there is a finite number of contributions proportional to γ^n, i.e. such that $N_p + N_e - P/2 + L/2 = n$. Their sum can be recast as

$$\rho^* \gamma^n g_n(\beta) \exp(\beta(n-1)E_H). \tag{4.1}$$

Functions $g_n(\beta)$ are expressed in terms of bare partition functions $Z(M_p, M_e)$ of clusters (M_p, M_e), bare interactions W between clusters, and screening functions S which may involve either a single or various clusters. Roughly speaking, the number of involved graphs, as well as the maximum total particle number $N_p + N_e$, increase with n.

Taking into account the results derived in Sect. 3, screened cluster expansion of common particle density $\rho = \rho_p = \rho_e$ can be formally rewritten as

$$\begin{aligned}\rho/\rho^* = {} & \gamma + \frac{\gamma^2}{2} + \gamma^{3/2} g_{3/2}(\beta) \exp(\beta E_H/2) + \gamma^2 g_{2,exc}(\beta) \exp(\beta E_H) \\ & + \sum_{n=5/2}^{\infty} \gamma^n g_n(\beta) \exp(\beta(n-1)E_H)\end{aligned} \tag{4.2}$$

where the sum runs over integer and half-integer values of n. In (4.2), we have extracted in $\gamma^2 g_2(\beta) \exp(\beta E_H)$ contribution (3.16) of atoms H in their groundstate, while the remaining part defines $g_{2,exc}(\beta)$. First two functions $g_{3/2}$ and $g_{2,exc}$ are

$$g_{3/2}(\beta) = S_{3/2}(1,0) \tag{4.3}$$

according to (3.4), and

$$\begin{aligned}g_{2,exc}(\beta) = {} & \frac{1}{2}[S_2(1,0) + S_2(0,1)] + W(1,0|1,0) + W(1,0|0,1) \\ & - \frac{(c_p + c_e)(\beta|E_H|)^{3/2}}{24\pi^{3/2}} \\ & + \frac{1}{8}\left[Z_{exc}(1,1) + \left(\frac{2m}{m_p}\right)^{3/2} Z(2,0) + \left(\frac{2m}{m_e}\right)^{3/2} Z(0,2)\right]\end{aligned} \tag{4.4}$$

by summing (3.7), (3.17), (3.21), (3.33), (3.36), (3.39), (3.48), (3.49) and (3.52). Notice that (4.4) is symmetric with respect to permutations of species indexes p and e in agreement with $\rho_p = \rho_e$.

We stress that, in the Saha regime, γ is a fixed parameter not necessarily small, while $\beta \to \infty$. Then, functions $g_{2,exc}(\beta) \exp(\beta E_H)$ and $g_n(\beta) \exp(\beta(n-1)E_H)$ with $n \geq 3/2$ and $n \neq 2$, decay exponentially fast. Thus, the whole sum over n in (4.2) can be reordered according to the corresponding decay rates. Each term $\gamma^n g_n(\beta) \exp(\beta(n-1)E_H)$ is then rewritten as $\gamma^n h_k(\beta)$ where $k = k(n)$ is some integer. Functions h_k decay exponentially fast, i.e. $h_k(\beta) \sim \exp(-\beta\delta_k)$ (apart from powers of β), with decay rates δ_k ranked as $0 < \delta_1 < \delta_2 < \cdots$: h_{k+1} decays exponentially faster than h_k when $\beta \to \infty$. According to the analytic results derived in Sects. 3.1–3.7 on the one hand, and to the numerical values of $E_H, E_{H^-}, E_{H_2^+}, E_{H_2}$ (see Sect. 2.3) on the other hand, we find

$$\begin{aligned} h_1(\beta) &= g_{3/2}(\beta)\exp(\beta E_H/2), & n_1 &= 3/2,\\ h_2(\beta) &= g_4(\beta)\exp(3\beta E_H), & n_2 &= 4,\\ h_3(\beta) &= g_{2,exc}(\beta)\exp(\beta E_H), & n_3 &= 2,\\ h_4(\beta) &= g_3(\beta)\exp(2\beta E_H), & n_4 &= 3. \end{aligned} \tag{4.5}$$

Their corresponding decay rates δ_k can be found in the table (1.5) given in the Introduction, while all higher-order functions $h_k(\beta)$ with $k \geq 5$ decay exponentially faster than $\exp(\beta E_H)$, i.e. their decay rates δ_k are larger than $|E_H| \simeq 13.6$. Notice that both $\gamma^{5/2} g_{5/2}(\beta)\exp(3\beta E_H/2)$ and $\gamma^{7/2} g_{7/2}(\beta)\exp(5\beta E_H/2)$ decay faster than $\exp(\beta E_H)$, so both $k(5/2)$ and $k(7/2)$ are strictly larger than 4. Within previous reordering, (4.2) becomes

$$\rho/\rho^* = \gamma + \frac{\gamma^2}{2} + \sum_{k=1}^{\infty} \gamma^{n_k} h_k(\beta). \tag{4.6}$$

At order $\exp(\beta E_H)$ included, all terms with $k \geq 5$ can be omitted in (4.6). Moreover, for the sake of consistency, it is sufficient to compute functions h_k with $1 \leq k \leq 4$ at the same order (beyond its leading behavior $\exp(-\beta\delta_k)$, h_k involves other exponentially small contributions). This gives

$$h_1(\beta) = \frac{(\beta|E_H|)^{3/4}}{\pi^{1/4}}\exp(\beta E_H/2), \tag{4.7}$$

$$h_2(\beta) = \frac{1}{64}\left(\frac{2m}{M}\right)^{3/2} Z(2,2)\exp(3\beta E_H) + W(1,1|1,1)\exp(3\beta E_H), \tag{4.8}$$

$$\begin{aligned} h_3(\beta) = {} & -\frac{1}{2} + \left[1 + \frac{1}{12}\ln\left(\frac{4m}{M}\right)\right]\frac{(\beta|E_H|)^{3/2}}{\pi^{1/2}}\exp(\beta E_H)\\ & + \frac{1}{8\pi^{1/2}}\left\{2Q(x_{pe}) + \left(\frac{2m}{m_p}\right)^{3/2}\left[Q(-x_{pp}) - \frac{1}{2}E(-x_{pp})\right]\right.\\ & \left. + \left(\frac{2m}{m_e}\right)^{3/2}\left[Q(-x_{ee}) - \frac{1}{2}E(-x_{ee})\right]\right\}\exp(\beta E_H), \end{aligned} \tag{4.9}$$

and

$$\begin{aligned} h_4(\beta) = {} & \frac{3}{64}\left[\left(\frac{m_e(M+m_p)}{M^2}\right)^{3/2} Z(2,1) + \left(\frac{m_p(M+m_e)}{M^2}\right)^{3/2} Z(1,2)\right]\exp(2\beta E_H)\\ & + S_3(1,1)\exp(2\beta E_H) + \frac{3}{2}[W(1,1|1,0) + W(1,1|0,1)]\exp(2\beta E_H). \end{aligned} \tag{4.10}$$

In (4.7) and (4.9), full contributions of respectively g_1 and $g_{2,exc}$, are kept, while analytic expressions (3.3), (3.8), (3.34), (3.37) and (3.39) have been used. Moreover, and according to formula (A.12) derived in Appendix A:, $Z_{exc}(1,1)$, $Z(2,0)$ and $Z(0,2)$ have been expressed in terms of Ebeling quantum virial functions Q (direct part) and E (exchange part) defined in Ref. [39], with arguments $x_{pe} = 2(\beta|E_H|)^{1/2}$, $x_{pp} = (2m_p/m)^{1/2}(\beta|E_H|)^{1/2}$ and $x_{ee} = (2m_e/m)^{1/2}(\beta|E_H|)^{1/2}$. Term $-1/2$ in $h_3(\beta)$ subtracts the ground-state contribution included in function $Q(x_{pe})$. In (4.8) and (4.10), contributions of g_4 and g_3 which decay exponentially faster than $\exp(\beta E_H)$ have been omitted. The resulting expression for h_2 is obtained by summing (3.23) and (3.41). Similarly, expression (4.10) for h_4 follows by summing (3.26), (3.27), (3.50), (3.51), (3.19), (3.45) and (3.53).

As a conclusion, it is useful to summarize the main features and ingredients of expansion (4.6). The h_k-functions are ordered, at sufficiently low temperatures, according to $|h_1(\beta)| > |h_2(\beta)| > |h_3(\beta)| > |h_4(\beta)| > \cdots$. They incorporate corrections to ideal Saha equation which arise from different physical phenomena, as listed in the Introduction. Explicit expressions for $h_1(\beta)$ and $h_3(\beta)$ are known, see (4.7) and Ref. [39], while $h_2(\beta)$ and $h_4(\beta)$ involve integrals associated with 3-body and 4-body problems which cannot be expressed in closed analytical forms. In $h_2(\beta)$, the internal partition function $Z(2,2)$ of a hydrogen molecule is defined in (3.24), and its low-temperature form is determined in Appendix B:. The function $W(1,1|1,1)$, which accounts for atom-atom interactions, is defined in (3.41), and its low-temperature form is computed in Appendix C:. In $h_4(\beta)$, the internal partition functions $Z(2,1)$ and $Z(1,2)$ of ions H_2^+ and H^- are defined in (3.28) and (3.29) respectively, while their asymptotic expressions at low temperatures are derived in Appendix B:. The interactions $W(1,1|1,0)$ and $W(1,1|0,1)$ between an atom and an ionized proton or electron, are defined in (3.45) and their low-temperature expressions are given in Appendix C:. Eventually, the screening function $S_3(1,1)$ of a hydrogen atom accounts for collective corrections to the bare proton-electron Coulomb potential beyond the familiar Debye shift, and it is given by formula (A.16) at low temperatures.

4.2 Equation of State

In order to compute the pressure, we consider identities

$$\begin{aligned} \rho_p &= z_p \frac{\partial \beta P}{\partial z_p}, \\ \rho_e &= z_e \frac{\partial \beta P}{\partial z_e}. \end{aligned} \tag{4.11}$$

Taking into account that P depends only on z and β, and using parametrization of z in terms of β and γ, $z = (m/M)^{3/4}\gamma \exp(\beta E_H)/4$, we rewrite such identities as

$$\frac{\partial \beta P}{\partial \gamma} = \frac{2\rho}{\gamma}, \tag{4.12}$$

where partial derivative of βP with respect to γ is taken at fixed β. After inserting expansion (4.6) of ρ into the r.h.s. of (4.12), a straightforward term by term integration with respect to γ provides

$$\beta P/\rho^* = 2\gamma + \frac{\gamma^2}{2} + \sum_{k=1}^{\infty} \frac{2\gamma^{n_k}}{n_k} h_k(\beta). \tag{4.13}$$

The required equation of state follows by inserting into (4.13) the expression of γ in terms of ρ obtained from the inversion of (4.6). That inversion can be performed perturbatively around the simple expression

$$\gamma_{\text{Saha}} = (1 + 2\rho/\rho^*)^{1/2} - 1 \tag{4.14}$$

obtained by retaining only the first two terms of (4.6). The resulting SLT expansion of the pressure takes the form (1.1) presented in the Introduction where $\beta P_{\text{Saha}}/\rho^*$ is given by (1.3). The general structure of $\beta P_k/\rho^*$ reduces to a function of ρ/ρ^* times a polynomial in the $h_l(\beta)$'s with $l \leq k$. Therefore, for a fixed ratio ρ/ρ^*, corrections $\beta P_k/\rho^*$ decay exponentially fast when $\beta \to \infty$. Moreover, each $\beta P_{k+1}/\rho^*$ decays faster than $\beta P_k/\rho^*$ for $k \geq 0$ (with $P_0 = P_{\text{Saha}}$). First corrections in (1.1) read

$$\beta P_1/\rho^* = \frac{[(1+2\rho/\rho^*)^{1/2} - 3][(1+2\rho/\rho^*)^{1/2} - 1]^{3/2}}{3(1+2\rho/\rho^*)^{1/2}} h_1(\beta), \tag{4.15}$$

$$\beta P_2/\rho^* = \frac{-[(1+2\rho/\rho^*)^{1/2} + 2][(1+2\rho/\rho^*)^{1/2} - 1]^4}{2(1+2\rho/\rho^*)^{1/2}} h_2(\beta), \tag{4.16}$$

$$\beta P_3/\rho^* = \frac{-[(1+2\rho/\rho^*)^{1/2} - 1]^2}{(1+2\rho/\rho^*)^{1/2}} h_3(\beta), \tag{4.17}$$

$$\beta P_4/\rho^* = \frac{-[(1+2\rho/\rho^*)^{1/2} + 3][(1+2\rho/\rho^*)^{1/2} - 1]^3}{3(1+2\rho/\rho^*)^{1/2}} h_4(\beta), \tag{4.18}$$

and

$$\beta P_5/\rho^* = \frac{[(1+2\rho/\rho^*)^{1/2} - \rho/\rho^*][(1+2\rho/\rho^*)^{1/2} - 1]^2}{(1+2\rho/\rho^*)^{3/2}} [h_1(\beta)]^2. \tag{4.19}$$

Next correction $\beta P_6/\rho^*$ decays faster than $\exp(\beta E_H)$.

In previous corrections $\beta P_k/\rho^*$, functions $h_1(\beta)$ and $h_3(\beta)$ can be expressed in closed analytical forms according to (4.7) and (4.9) respectively. Similar analytical expressions for $h_2(\beta)$ and $h_4(\beta)$ are not available. Nevertheless, the low-temperature behaviors of those functions are exactly known, i.e.

$$h_2(\beta) \sim \frac{1}{64}\left(\frac{2m}{M}\right)^{3/2} \exp(\beta(3E_H - E_{H_2})) \tag{4.20}$$

and

$$h_4(\beta) \sim \frac{3}{64}\left(\frac{m_e(M+m_p)}{M^2}\right)^{3/2} \exp(\beta(2E_H - E_{H_2^+})) \tag{4.21}$$

when $\beta \to \infty$.

Eventually, the various terms in (1.1) display interesting behaviors with respect to ratio ρ/ρ^*, at fixed β sufficiently large:

- For ρ much smaller than ρ^*, each $\beta P_k/\rho^*$, as well as $\beta P_{\text{Saha}}/\rho^*$, can be expanded in powers of ρ/ρ^*. This leads to the virial expansion of βP in powers of ρ. Since all $\beta P_k/\rho^*$'s for $k \geq 6$ are at least of order $\rho^{5/2}$, the full contribution of terms with $k \leq 5$ in (1.1) provides the expansion of βP up to order ρ^2, i.e.

$$\beta P = 2\rho - \frac{2^{3/2}(2\pi)^{3/4}}{3}(\lambda_{pe})^{3/2}\exp(-\beta E_H/2)h_1(\beta)\rho^{3/2}$$

$$\begin{aligned}
&- (2\pi)^{3/2}(\lambda_{pe})^3 \exp(-\beta E_H)[2h_3(\beta) + 1 - 2(h_1(\beta))^2]\rho^2 + O(\rho^{5/2}) \\
&= 2\rho - \frac{(8\pi\beta e^2\rho)^{3/2}}{24\pi} - \frac{\pi}{\sqrt{2}}\bigg\{2\lambda_{pe}^3 Q(x_{pe}) + \lambda_{pp}^3\bigg[Q(-x_{pp}) - \frac{1}{2}E(-x_{pp})\bigg] \\
&\quad + \lambda_{ee}^3\bigg[Q(-x_{ee}) - \frac{1}{2}E(-x_{ee})\bigg]\bigg\}\rho^2 - \frac{\pi}{6}\ln\left(\frac{4m}{M}\right)\beta^3 e^6 \rho^2 + O(\rho^{5/2}) \qquad (4.22)
\end{aligned}$$

which does coincide with the well known expression derived previously by other methods [4, 23, 39] ($\lambda_{pe} = (\beta\hbar^2/m)^{1/2}$, $\lambda_{pp} = (\beta\hbar^2/m_{pp})^{1/2}$ and $\lambda_{ee} = (\beta\hbar^2/m_{ee})^{1/2}$). Notice that contribution of βP_4 is of order ρ^3, while that of βP_2 is of order ρ^4.

- For ρ of order ρ^*, leading term $\beta P_{\text{Saha}}/\rho^*$, as well as each correction $\beta P_k/\rho^*$, can be viewed as infinite resummations of terms with arbitrary high orders in the above low-density expansion. Such resummations account, in a non-perturbative way with respect to density, of recombination processes for any value of the ionization rate. The relative orders of magnitude of the various corrections to Saha pressure are mainly controlled by their decay rates δ_k. Therefore the larger correction indeed is βP_1, which results from plasma polarization around a given ionized charge, evaluated within Debye classical mean-field theory. That result is equivalent to the modified Saha condition which determines the ionization rate [39, 41].
- For ρ much larger than ρ^*, βP_{Saha} behaves as

$$\beta P_{\text{Saha}} \sim \rho, \qquad (4.23)$$

which illustrates the almost full atomic recombination of the plasma. The larger correction to Saha pressure is now βP_2 which behaves as

$$\beta P_2 \sim -2h_2(\beta)\rho^*\left(\frac{\rho}{\rho^*}\right)^2, \qquad (4.24)$$

so it overcomes βP_1 which grows only as $(\rho/\rho^*)^{3/4}$, as well as further corrections $\beta P_3 \sim (\rho/\rho^*)^{1/2}$, $\beta P_4 \sim (\rho/\rho^*)^{3/2}$ and $\beta P_5 \sim (\rho/\rho^*)^{1/2}$. Therefore molecular recombination prevails over plasma polarization. Of course, expansion (1.1) is no longer appropriate for too large values of ratio (ρ/ρ^*), since some corrections $\beta P_k/\rho^*$ become much larger than Saha pressure.

4.3 Numerical Estimations and Validity Domain of SLT Expansions

Quantitative estimations of corrections $\beta P_1/\rho^*$, $\beta P_3/\rho^*$ and $\beta P_5/\rho^*$ are easy, because functions $h_1(\beta)$ and $h_3(\beta)$ can be represented by simple analytical expressions at finite temperature. For functions $h_2(\beta)$ and $h_4(\beta)$ which involve 3 and 4-body contributions, no *explicit* finite-T representations are available beyond their low-temperature asymptotic forms determined in Appendices B and C. In order to obtain reliable values for those functions at moderate temperatures, we have used a simple approach in which important finite temperature effects (such as atomic vibrations and rotations) are phenomenologically taken into account. As mentioned in the Introduction, the corresponding numerical evaluations of the various corrections to Saha pressure (and internal energy), together with a comparison of our predicted isotherms with the results of PIMC simulations, will be presented in a forthcoming paper [1].

Here, we exhibit the validity domain of SLT expansion (1.1). A rigorous analysis of the convergence of that expansion is a tremendous mathematical task, much beyond the scope

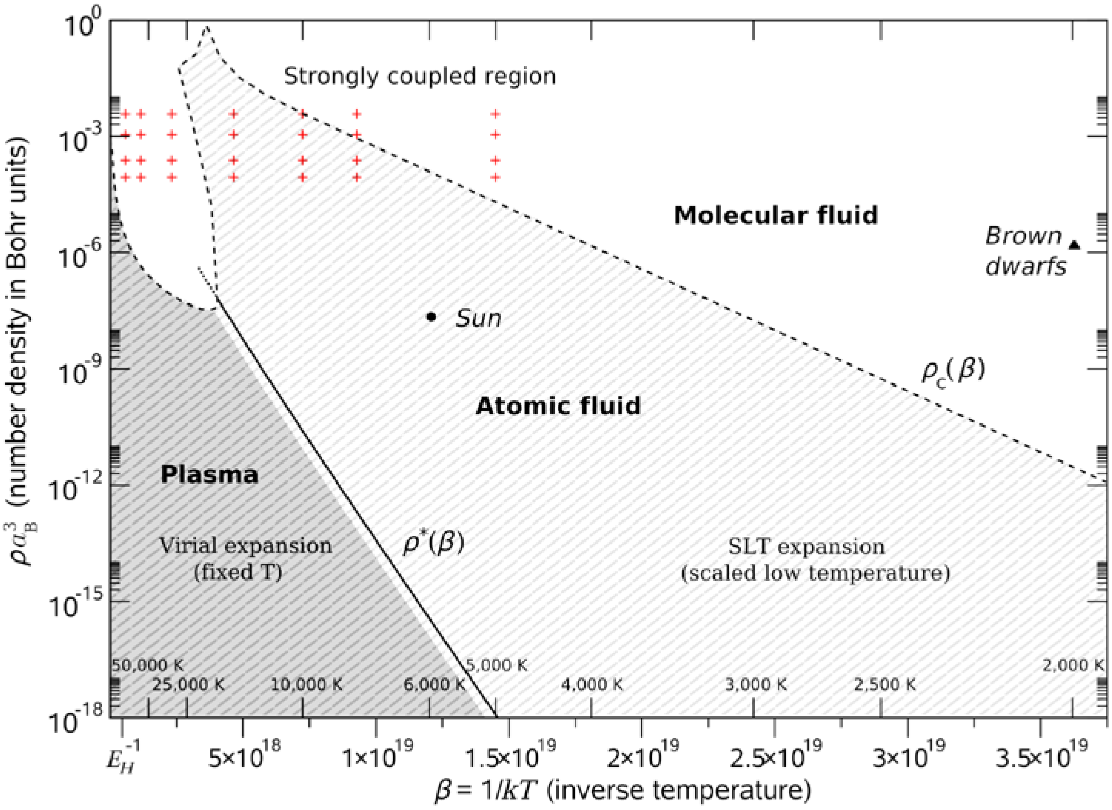


Fig. 12 Phase diagram showing the validity domains of SLT expansion (1.1) (*hatched region*) and of the virial expansion (*shaded region*). Atomic recombination density $\rho^*(\beta)$ (1.2) is a straight line at low temperatures in the $(\beta, \log\rho)$-plane. The validity domain is delimited at high densities, and for temperatures below 10000 K, by critical density $\rho_c(\beta)$ at which molecular recombination occurs. Crosses indicate state points where simulation results are available [48]. State points of astrophysical systems (Sun photosphere and Brown dwarfs) are also shown in the diagram

of the present work. We estimated a quite plausible validity domain, by employing the semi-empirical criterion $|P_k| < P_{\text{Saha}}/10$ for all five corrections ($1 \le k \le 5$): it covers the region hatched in the temperature-density plane shown in Fig. 12. The shaded region at low densities and high temperatures corresponds to the validity domain of the virial expansion (i.e. low-density expansion at fixed temperature) determined from a similar criterion. Obviously, SLT expansion improves widely upon virial expansion, by providing reliable results in the atomic phase, including the temperature and density ranges around $\rho^\star(\beta)$ which correspond to partially ionized hydrogen gases. In Fig. 12, we have also shown state points, symbolized by crosses, for which PIMC simulation results have been obtained [48]. It turns out that some of them lie within the validity domain of the SLT expansion. We have checked that our calculations, both for pressure and internal energy, are in agreement with PIMC results within statistical errors [1]. This confirms the reliability of SLT expansions in the domain inferred from the above semi-empirical criterion. Notice that such domain extends to rather high densities, up to $a/a_B = 6$ at 15000 K, which corresponds to a mean inter-particle distance of the order of twice the size of a hydrogen atom.

For temperatures below 10000 K, the validity domain is limited at high densities by molecular recombination which occurs around densities $\rho_c(\beta)$ shown as a dashed line in the phase diagram. That limitation is not intrinsic to the theory, and an SLT expansion applicable in the molecular regime can be derived as well. Such a generalization requires replacing the scaling (2.12) of the chemical potential by a similar scaling corresponding to a molecular regime (see Fig. 2), and performing the inversion $\mu = \mu(\rho)$ in the appropriate density range. Notice that if ρ is not too high above ρ_c, expansion (4.6) of particle density in terms of chemical potential should remain valid. Performing a non-perturbative inversion of the chemical potential in favor of the density, should then provide accurate thermodynamical functions which account not only for atomic recombination at $\rho \sim \rho^*$, but also for molecular recombination above ρ_c.

For temperatures above 10000 K, the borderline of the validity domain has a complicated shape determined by correction term P_3 which accounts for atomic excitations and for interactions between ionized electrons and protons. At high temperatures, typically above 30000 K, thermal ionization prevents recombination of protons and electrons into atoms, and our validity criterion is then equivalent to a weak coupling condition for ionized charges, i.e. coupling parameter $\varGamma = \beta e^2/a$ smaller than some value, which also determines the validity of the virial expansion.

Eventually, state points of two astrophysical systems of interest, Sun photosphere and a typical brown dwarf atmosphere, are also shown in Fig. 12. Our equation of state (1.1) clearly holds at the temperature and density of Sun photosphere. In order to be applicable to brown dwarf atmospheres, the SLT expansion would need to be generalized to the molecular regime, as discussed above.

Acknowledgements We are very indebted to Vincent ROBERT (Laboratoire de Chimie, ENS Lyon, CNRS) for his efficient and kind support. We thank him in particular for his spectroscopic calculations (ground state energies and excitations energies of a few compounds) which we used in Sect. 2.3 [52].

Appendix A: Screened Potential and Related Integrals

A.1 Expression and Behavior of ϕ

The Fourier transform of $\phi(\mathcal{L}_i, \mathcal{L}_j)$ with respect to $\mathbf{X}_i - \mathbf{X}_j$ reads

$$\tilde{\phi}(\mathbf{k}, \chi_i, \chi_j) = e_{\alpha_i} e_{\alpha_j} \int_0^{q_i} \mathrm{d}s \int_0^{q_j} \mathrm{d}t \exp[i\mathbf{k} \cdot (\lambda_{\alpha_i} \boldsymbol{\eta}_i(s) - \lambda_{\alpha_j} \boldsymbol{\eta}_j(t))] \times \sum_{n=-\infty}^{\infty} \frac{4\pi}{k^2 + \kappa^2(k,n)} \exp[-2n\pi i(s-t)], \tag{A.1}$$

where $\chi = (\alpha, q, \boldsymbol{\eta}(\cdot))$ denotes the loop internal degrees of freedom, while function $\kappa^2(k,n)$ is defined in Ref. [13]. Functions $\kappa^2(k,n)$ are analytical in k^2 near $k=0$, while $\kappa^2(0,n)=0$ for $n \neq 0$ and $\kappa^2(0,0) \neq 0$ is of order κ^2 . For large values of k, $\kappa^2(k,n)$ remains bounded by a constant independent of n (of order κ^2). For $k \sim \kappa$, $\kappa^2(k,n)$ for $n \neq 0$ is smaller than $\kappa^2(k,0)$ by a factor of order $\kappa^2\lambda^2$, while $\kappa^2(k,0)$ can be replaced by κ^2.

The behaviors of ϕ with respect to relative distance $r = |\mathbf{X}_i - \mathbf{X}_j|$ (roughly described in Fig. 4), can be readily derived from those of $\tilde{\phi}$ with respect to $\mathbf{k}$, as detailed in Ref. [13]. Here, we briefly summarize that analysis. For $k \gg \kappa$, each fraction $4\pi/(k^2 + \kappa^2(k,n))$ can be replaced by $4\pi/k^2$ in (A.1), so $\phi(\mathcal{L}_i, \mathcal{L}_j)$ behaves as $V(\mathcal{L}_i, \mathcal{L}_j)$ at short distances $r \ll \kappa^{-1}$. At distances $r \sim \kappa^{-1}$, we recover the Debye classical form $\phi_D(\mathcal{L}_i, \mathcal{L}_j) = q_i e_{\alpha_i} q_j e_{\alpha_j} \exp(-\kappa r)/r$, by noting that terms $n \neq 0$ in (A.1) provide contributions smaller than the one of $n=0$ by a factor of order $\kappa^2\lambda^2$. Eventually, terms $n \neq 0$ in (A.1) provide a singularity in the small-$\mathbf{k}$ expansion of $\tilde{\phi}(\mathbf{k}, \chi_i, \chi_j)$, which in turn induces a dipolar-like decay of $\phi(\mathcal{L}_i, \mathcal{L}_j)$ at large distances $r \gg l_Q$.

A.2 Integrals of Powers of ϕ

We consider $\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p,e)})$ where loops $\mathcal{L}_a^{(p)}$ and loop $\mathcal{L}_1^{(p,e)}$ contain, respectively, one proton ($q_a = 1$) and either one proton or one electron ($q_1 = 1$). According to the definition of $\tilde{\phi}(\mathbf{k}, \chi_a, \chi_1)$, the integral of $\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p,e)})$ over $\mathbf{X}_1$ and $\boldsymbol{\xi}_1 = \boldsymbol{\eta}_1$ is nothing but

$$\int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_1)\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p,e)}) = \int \mathcal{D}(\boldsymbol{\xi}_1)\tilde{\phi}(\mathbf{0}, \chi_a, \chi_1). \tag{A.2}$$

The r.h.s. of (A.2) is computed by taking the limit $\mathbf{k} \to \mathbf{0}$ of expression (A.1). The corresponding contribution of a term $n \neq 0$ is obtained by expanding phase factor $\exp[i\mathbf{k} \cdot (\lambda_p \boldsymbol{\xi}_a(s) - \lambda_{p,e}\boldsymbol{\xi}_1(t))]$ in powers of $\mathbf{k}$. Since all odd moments of measure $\mathcal{D}(\boldsymbol{\xi}_1)$ vanish, as well as $\int_0^1 \mathrm{d}t \exp(2n\pi i t)$, the first non-vanishing term in that expansion is at least of order k^3. It has to be multiplied by a factor of order $1/k^2$ which arises from fraction $4\pi/(k^2 + \kappa^2(k, n))$, so the resulting contribution to the r.h.s. of (A.2) vanishes. Therefore, the sole contribution arises from term $n = 0$, i.e.

$$\int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_1)\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p,e)}) = \pm\frac{4\pi e^2}{\kappa^2(0,0)}, \tag{A.3}$$

with a positive sign for $\mathcal{L}_1^{(p)}$ and a negative one for $\mathcal{L}_1^{(e)}$.

According to Fourier-Plancherel formula, the integral of $[\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p,e)})]^2$ over $\mathbf{X}_1$, $\boldsymbol{\xi}_a = \boldsymbol{\eta}_a$ and $\boldsymbol{\xi}_1 = \boldsymbol{\eta}_1$, is rewritten as

$$\begin{aligned}
&\int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_a)\mathcal{D}(\boldsymbol{\xi}_1)[\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p,e)})]^2 \\
&\quad= \frac{1}{(2\pi)^3}\int \mathrm{d}\mathbf{k} \int \mathcal{D}(\boldsymbol{\xi}_a)\mathcal{D}(\boldsymbol{\xi}_1)|\tilde{\phi}(\mathbf{k}, \chi_a, \chi_1)|^2 \\
&\quad= \frac{2e^4}{\pi}\int \mathrm{d}\mathbf{k} \int_0^1 \mathrm{d}s_1 \int_0^1 \mathrm{d}t_1 \int_0^1 \mathrm{d}s_2 \int_0^1 \mathrm{d}t_2 \\
&\qquad\times \sum_{n_1,n_2=-\infty}^{\infty} \frac{\exp[-2n_1\pi i(s_1 - t_1)]}{k^2 + \kappa^2(k, n_1)} \frac{\exp[-2n_2\pi i(s_2 - t_2)]}{k^2 + \kappa^2(k, n_2)} \\
&\qquad\times \int \mathcal{D}(\boldsymbol{\xi}_a)\exp[i\lambda_p \mathbf{k}\cdot\boldsymbol{\xi}_a(s_1 - s_2)] \int \mathcal{D}(\boldsymbol{\xi}_1)\exp[-i\lambda_{p,e}\mathbf{k}\cdot\boldsymbol{\xi}_1(t_1 - t_2)].
\end{aligned} \tag{A.4}$$

In the last equality of (A.4), we have used that the average over shape $\boldsymbol{\xi}$ of any function $f(\boldsymbol{\xi}_a(s_1) - \boldsymbol{\xi}_a(s_2))$ is identical to the average of $f(\boldsymbol{\xi}_a(s_1 - s_2))$, provided that $\boldsymbol{\xi}(s)$ for s outside $[0, 1]$ is defined as equal to $\boldsymbol{\xi}(s - [s])$ [13]. Within variable change $\mathbf{k} = \kappa\mathbf{q}$, we can replace $\kappa^2(\kappa q, n_i)$ by either κ^2 for $n_i = 0$, or 0 for $n_i \neq 0$, discarding terms which provide contributions smaller by a factor $(\kappa\lambda)^2$ at least. Summations over $n_i \neq 0$ are then performed according to identity $\sum_{n\neq 0}\exp[-2n\pi i(s - t)] = \delta(s - t) - 1$. Since measure $\mathcal{D}(\boldsymbol{\xi})$ is Gaussian with covariance (2.34) (for $q = 1$), we transform (A.4) into

$$\begin{aligned}
&\frac{8e^4}{\kappa}\int_0^1 \mathrm{d}s \int_0^1 \mathrm{d}t \int_0^\infty \mathrm{d}q \frac{q^2}{(q^2+1)^2}\exp[-\kappa^2\lambda_p^2 q^2 s(1-s)/2 - \kappa^2\lambda_{p,e}^2 q^2 t(1-t)/2] \\
&\quad+ \frac{8e^4}{\kappa}\int_0^1 \mathrm{d}s \int_0^\infty \mathrm{d}q \frac{1}{q^2}\{\exp[-\kappa^2(\lambda_p^2 + \lambda_{p,e}^2)q^2 s(1-s)/2] - 1\}
\end{aligned}$$

$$
\begin{aligned}
&-\frac{8e^4}{\kappa}\int_0^1 \mathrm{d}s \int_0^1 \mathrm{d}t \int_0^\infty \mathrm{d}q\, \frac{1}{q^2} \\
&\times \{\exp[-\kappa^2\lambda_p^2 q^2 s(1-s)/2 - \kappa^2\lambda_{p,e}^2 q^2 t(1-t)/2] - 1\} \qquad \text{(A.5)}
\end{aligned}
$$

discarding terms of order $\kappa^{-1}O((\kappa\lambda)^2)$. The integrals over q in (A.5) are computed in terms of elementary functions of arguments $[\kappa^2\lambda_p^2 s(1-s)/2 + \kappa^2\lambda_{p,e}^2 t(1-t)/2]^{1/2}$ and $[\kappa^2(\lambda_p^2 + \lambda_{p,e}^2)s(1-s)/2]^{1/2}$, which can be expanded in Taylor series since $\kappa\lambda$ is small. For the leading (order κ^{-1}) and first subleading (order $\kappa^{-1}O(\kappa\lambda)$) contributions, the remaining integrals over s and t are readily calculated (some complicated double integrals over s and t arising from respectively first and third terms in (A.5) cancel out). Eventually, we obtain

$$
\int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_a)\mathcal{D}(\boldsymbol{\xi}_1)[\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(\alpha)})]^2 = \frac{2\pi e^4}{\kappa}\left[1 - \frac{\sqrt{\pi}}{2\sqrt{2}}\kappa\lambda_{p\alpha} + O((\kappa\lambda)^2)\right]. \qquad \text{(A.6)}
$$

The integral of $[\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(\alpha)})]^3$ over $\mathbf{X}_1$, $\boldsymbol{\xi}_a$ and $\boldsymbol{\xi}_1$, can be evaluated within similar techniques and tricks. Discarding terms of order $O(\kappa\lambda)$, its leading behavior reduces to a constant times $\ln(\kappa\lambda_{p\alpha})$ plus another constant. When the two integrals corresponding respectively to $\alpha = p$ and $\alpha = e$ are summed together, logarithmic terms in κ cancel out. Therefore, we obtain

$$
\int \mathrm{d}\mathbf{X}_1 \int \mathcal{D}(\boldsymbol{\xi}_a) \int \mathcal{D}(\boldsymbol{\xi}_1)[(\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p)}))^3 + (\phi(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(e)}))^3] = c_p e^6 + O(\kappa\lambda) \qquad \text{(A.7)}
$$

where c_p is the constant

$$
\begin{aligned}
c_p = \frac{2}{\pi^3}\int_0^1 \mathrm{d}s \int_0^s \mathrm{d}t \int \mathrm{d}\mathbf{q}_1 \int \mathrm{d}\mathbf{q}_2 &\frac{1}{q_1^2 q_2^2 |\mathbf{q}_1 + \mathbf{q}_2|^2} \\
\times \{&\exp[-(q_1^2 s(1-s) + q_2^2 t(1-t) + 2\mathbf{q}_1\cdot\mathbf{q}_2 t(1-s))] \\
-&\exp[-(q_1^2 s(1-s) + q_2^2 t(1-t) + 2\mathbf{q}_1\cdot\mathbf{q}_2 t(1-s))m_p/(2m)]\} \qquad \text{(A.8)}
\end{aligned}
$$

entirely determined by ratio m_p/m. As it should, leading contribution $c_p e^6$ in (A.7) is nothing but the value of the considered integral with bare potential V in place of ϕ (that bare integral does converge thanks to the $1/|\mathbf{X}_1|^4$-decay of $[V(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(p)})]^3 + [V(\mathcal{L}_a^{(p)}, \mathcal{L}_1^{(e)})]^3$). When the root proton is replaced by a root electron ($\mathcal{L}_a^{(p)} \rightarrow \mathcal{L}_a^{(e)}$), the resulting integral behaves similarly to (A.7) where constant c_e is given by (A.8) with m_e in place of m_p.

A.3 Behavior of I_R

We consider a loop $\mathcal{L}$ containing a single particle of species α. Convolution formula (2.40) for $I_R(\mathcal{L})$ is first transformed according to Fourier-Plancherel identity, in which $\tilde{\phi}(\mathbf{k}, \chi_a, \chi_1)$ is replaced by (A.1). Discarding terms smaller by a factor $O((\kappa\lambda)^2)$, only the contributions of loops $\mathcal{L}_1^{(p,e)}$ associated with a single proton or a single electron, are retained. Moreover, at the same order, after making variable change $\mathbf{k} = \kappa\mathbf{q}$, we can replace $\kappa^2(\kappa q, n)$ by either κ^2 for $n = 0$, or 0 for $n \neq 0$. Using again identity $\sum_{n\neq 0}\exp[-2n\pi i(s-t)] = \delta(s-t) - 1$, we then obtain

$$
I_R(\mathcal{L}) = \frac{\beta e^2\kappa}{2} + \frac{\beta e^2\kappa}{4\pi^2}\sum_\gamma \int_0^1 \mathrm{d}s \int_0^1 \mathrm{d}s_1 \int_0^1 \mathrm{d}t_1 \int \mathcal{D}(\boldsymbol{\xi}_1) \int \mathrm{d}\mathbf{q} \frac{1}{q^2(q^2+1)}
$$

$$
\begin{aligned}
&\times [\exp(i\kappa \mathbf{q}\cdot(\lambda_\alpha \boldsymbol{\xi}(s) - \lambda_\gamma \boldsymbol{\xi}_1(s) + \lambda_\gamma \boldsymbol{\xi}_1(s_1) - \lambda_\alpha \boldsymbol{\xi}(t_1))) - 1] \\
&+ \frac{\beta e^2 \kappa}{4\pi^2} \sum_\gamma \int_0^1 \mathrm{d}s \int_0^1 \mathrm{d}s_1 \int_0^1 \mathrm{d}t_1 \int \mathcal{D}(\boldsymbol{\xi}_1) \int \mathrm{d}\mathbf{q} \frac{1}{q^4} \\
&\times \exp(i\kappa \mathbf{q}\cdot(\lambda_\alpha \boldsymbol{\xi}(s) - \lambda_\gamma \boldsymbol{\xi}_1(s) + \lambda_\gamma \boldsymbol{\xi}_1(s_1))) \\
&\times [\exp(-i\kappa \mathbf{q}\cdot \lambda_\alpha \boldsymbol{\xi}(s_1)) - \exp(-i\kappa \mathbf{q}\cdot \lambda_\alpha \boldsymbol{\xi}(t_1))] + \beta e^2 \kappa O((\kappa\lambda)^2). \qquad \text{(A.9)}
\end{aligned}
$$

The leading behavior of $I_R(\mathcal{L})$ reduces to the first term in the r.h.s. of (A.9). In the second term of (A.9), we can first perform the integration over $\mathbf{q}$ thanks to Cauchy's theorem. The resulting elementary functions of the argument $\kappa|\lambda_\alpha \boldsymbol{\xi}(s) - \lambda_\gamma \boldsymbol{\xi}_1(s) + \lambda_\gamma \boldsymbol{\xi}_1(s_1) - \lambda_\alpha \boldsymbol{\xi}(t_1)|$ are then expanded in Taylor series since $\kappa\lambda$ is small. The remaining integrations over times and shape $\boldsymbol{\xi}_1$ provide a contribution of order $\beta e^2 \kappa O(\kappa\lambda)$ which depends on $\boldsymbol{\xi}$. The third term in the r.h.s. of (A.9) has the same order and a similar shape-dependence, as shown by variable changes $\mathbf{q} = \kappa|\lambda_\alpha \boldsymbol{\xi}(s) - \lambda_\gamma \boldsymbol{\xi}_1(s) + \lambda_\gamma \boldsymbol{\xi}_1(s_1) - \lambda_\alpha \boldsymbol{\xi}(t)|\mathbf{u}$ with $t = s_1$ or $t = t_1$ (the integral over $\mathbf{q}$ is splitted as the sum of two integrals by adding and subtracting 1 to $[\exp(-i\kappa \mathbf{q}\cdot \lambda_\alpha \boldsymbol{\xi}(s_1)) - \exp(-i\kappa \mathbf{q}\cdot \lambda_\alpha \boldsymbol{\xi}(t_1))]$).

The integration of $I_R(\mathcal{L})$ over shape $\boldsymbol{\xi}$ readily follows from (A.9). Now in the second and third terms of (A.9), it is convenient to first perform integration over shapes $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_1$, using the previous trick relative to differences $\boldsymbol{\xi}(s) - \boldsymbol{\xi}(t_1)$ and $\boldsymbol{\xi}_1(s) - \boldsymbol{\xi}_1(s_1)$, as well as the Gaussian nature of measures $\mathcal{D}(\boldsymbol{\xi})$ and $\mathcal{D}(\boldsymbol{\xi}_1)$. This leads to

$$
\begin{aligned}
\int \mathcal{D}(\boldsymbol{\xi}) I_R(\mathcal{L}) = \frac{\beta e^2 \kappa}{2} &+ \frac{\beta e^2 \kappa}{\pi} \sum_\gamma \int_0^1 \mathrm{d}s_1 \int_0^1 \mathrm{d}t_1 \int_0^\infty \mathrm{d}q \frac{1}{(q^2+1)} \\
&\times \{\exp[-\kappa^2 \lambda_\gamma^2 q^2 s_1(1-s_1)/2 - \kappa^2 \lambda_\alpha^2 q^2 t_1(1-t_1)/2] - 1\} \\
&+ \frac{\beta e^2 \kappa}{\pi} \sum_\gamma \int_0^1 \mathrm{d}s_1 \int_0^\infty \mathrm{d}q \frac{1}{q^2} \{\exp[-\kappa^2(\lambda_\alpha^2 + \lambda_\gamma^2) q^2 s_1(1-s_1)/2] - 1\} \\
&- \frac{\beta e^2 \kappa}{\pi} \sum_\gamma \int_0^1 \mathrm{d}s_1 \int_0^1 \mathrm{d}t_1 \int_0^\infty \mathrm{d}q \frac{1}{q^2} \\
&\times \{\exp[-\kappa^2 \lambda_\gamma^2 q^2 s_1(1-s_1)/2 - \kappa^2 \lambda_\alpha^2 q^2 t_1(1-t_1)/2] - 1\} \\
&+ \beta e^2 \kappa O((\kappa\lambda)^2). \qquad \text{(A.10)}
\end{aligned}
$$

The integrals over q in the second, third and fourth terms of (A.10) are computed, similarly to that arising in (A.5), in terms of elementary functions which are afterwards expanded in powers of $\kappa\lambda$. Contributions of second and fourth terms with order $\beta e^2 \kappa O(\kappa\lambda)$ cancel out, so it remains

$$
\int \mathcal{D}(\boldsymbol{\xi}) I_R(\mathcal{L}) = \frac{\beta e^2 \kappa}{2} \left[1 - \frac{\sqrt{\pi}}{8\sqrt{2}} \sum_\gamma \kappa \lambda_{\alpha\gamma} + O((\kappa\lambda)^2) \right]. \qquad \text{(A.11)}
$$

A.4 Truncated Integrals of Powers of V

Quantum virial functions $Q(\pm x_{\alpha\gamma})$ are defined [39] through a truncation similar to that arising in $\langle \mathbf{r}|[\exp(-\beta H_{\alpha\gamma})]^T_{\text{Mayer}}|\mathbf{r}\rangle$, where matrix elements of time-evolved operators $V_{\alpha\gamma}$ and $[V_{\alpha\gamma}]^2$ are replaced by $\beta e_\alpha e_\gamma / r$ and $\beta^2 e^4 / r^2$ respectively, while $[V_{\alpha\gamma}]^3$-term is omitted.

Within such truncation, convergence at large distances is ensured by taking the limit $R \to \infty$ of the corresponding spatial integral inside a sphere with radius R plus logarithmic counter terms [39]. When partition functions $Z_{exc}(1,1)$, $Z(2,0)$ and $Z(0,2)$ are expressed in terms of the Q's and E's, the integrals arising from $V_{\alpha\gamma} - e_\alpha e_\gamma/r$, $V_{\alpha\gamma}^2 - e^4/r^2$ and $[V_{\alpha\gamma}]^3$ are computed within previous methods applied to similar integrals of powers of ϕ. The sum of contributions due to $V_{\alpha\gamma} - e_\alpha e_\gamma/r$ vanishes by virtue of identity $\lambda_{pp}^2 + \lambda_{ee}^2 - 2\lambda_{pe}^2 = 0$. We then find

$$\begin{aligned}
&\frac{1}{8}\left[Z_{exc}(1,1) + \left(\frac{2m}{m_p}\right)^{3/2} Z(2,0) + \left(\frac{2m}{m_e}\right)^{3/2} Z(0,2)\right] \\
&\quad = \frac{1}{8\pi^{1/2}}\left\{2Q(x_{pe}) + \left(\frac{2m}{m_p}\right)^{3/2}\left[Q(-x_{pp}) - \frac{1}{2}E(-x_{pp})\right]\right. \\
&\qquad \left. + \left(\frac{2m}{m_e}\right)^{3/2}\left[Q(-x_{ee}) - \frac{1}{2}E(-x_{ee})\right]\right\} \\
&\qquad - \frac{1}{2}\exp(-\beta E_H) + \frac{\beta^2 e^4}{32\lambda_{pe}^3}(2\lambda_{pe} + \lambda_{pp} + \lambda_{ee}) + \frac{\beta^3 e^6}{24(2\pi\lambda_{pe}^2)^{3/2}}(c_p + c_e) \\
&\qquad + \frac{\beta^3 e^6}{12(2\pi)^{1/2}\lambda_{pe}^3}\ln(\lambda_{pp}\lambda_{ee}/\lambda_{pe}^2).
\end{aligned} \tag{A.12}$$

In the r.h.s of (A.12), the first additional term to the linear combination of the Q's and E's merely arises from the substraction of groundstate contribution in $Z_{exc}(1,1)$, while the last one is due to the logarithmic counter terms introduced in the definitions of the Q's.

A.5 Integral Mixing ϕ, I_R and V

At lowest order, effects of atom polarization are entirely embedded in the integral

$$\begin{aligned}
&\frac{4z^2}{(2\pi\lambda^2)^3\Lambda}\int \mathrm{d}\mathbf{R}\mathrm{d}\mathbf{r}\int \mathcal{D}(\boldsymbol{\xi}_a)\mathcal{D}(\boldsymbol{\xi}_1)\mathcal{B}^T(a,1) \\
&\quad \times [I_R(\mathcal{L}_a^p) + I_R(\mathcal{L}_1^e) - \beta(\phi(\mathcal{L}_a^p,\mathcal{L}_1^e) - V(\mathcal{L}_a^p,\mathcal{L}_1^e))]
\end{aligned} \tag{A.13}$$

where we set $\mathbf{R} = \mathbf{R}_a$ and $\mathbf{r} = \mathbf{r}_1$. Similarly to the case of bare integral (3.10), the presence of Boltzmann factor $\exp(-\beta V(\mathcal{L}_a^p,\mathcal{L}_1^e))$ in $\mathcal{B}^T(a,1)$ implies that leading contributions in (A.13) arise from configurations where loop sizes are at most of order λ, while relative proton-electron distance $|\mathbf{r}-\mathbf{R}|$ is of order the extension a_B of the atom groundstate. The I_R's are then expanded with respect to small parameter $\kappa\lambda$ as above, while a similar expansion is derived for $(\phi - V)$ by starting from a convolution relation analogous to (2.40) and by noting that κa_B is also a small parameter. This provides

$$\begin{aligned}
&I_R(\mathcal{L}_a^p) + I_R(\mathcal{L}_1^e) - \beta(\phi(\mathcal{L}_a^p,\mathcal{L}_1^e) - V(\mathcal{L}_a^p,\mathcal{L}_1^e)) \\
&\quad = \frac{\beta e^2\kappa^2}{4}\int_0^1 \mathrm{d}s\int_0^1 \mathrm{d}t[2|\mathbf{r} + \lambda_e\boldsymbol{\xi}_1(t) - \mathbf{R} - \lambda_p\boldsymbol{\xi}_a(s)| - \lambda_e|\boldsymbol{\xi}_1(t) - \boldsymbol{\xi}_1(s)| \\
&\qquad - \lambda_p|\boldsymbol{\xi}_a(t) - \boldsymbol{\xi}_a(s)|] + \beta e^2\kappa O((\kappa\lambda)^2).
\end{aligned} \tag{A.14}$$

In (A.14), terms proportional to $\beta e^2\kappa$ cancel out, as well as terms proportional to $\beta e^2\kappa^2|\mathbf{r} - \mathbf{R}|$ which do not depend on loop shapes.

Using (A.14) in (A.13), the functional integrations over loop shapes can be rewritten in terms of matrix elements of Gibbs operators. For instance, the integral associated with the first term $\exp(-\beta V(\mathcal{L}_a^p, \mathcal{L}_1^e))$ in truncated Mayer coefficient $\mathcal{B}^T(a,1)$ can be rewritten as

$$\begin{aligned}
&\frac{2z^2e^2\kappa^2}{\beta\Lambda}\int \mathrm{d}\mathbf{R}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{R}_1\mathrm{d}\mathbf{R}_2\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2\\
&\quad\times\int_0^\beta \mathrm{d}\tau_1\int_0^{\tau_1}\mathrm{d}\tau_2\langle\mathbf{R}\mathbf{r}|\exp[-(\beta-\tau_1)H_{1,1}]|\mathbf{R}_1\mathbf{r}_1\rangle\\
&\quad\times\langle\mathbf{R}_1\mathbf{r}_1|\exp[-(\tau_1-\tau_2)H_{1,1}]|\mathbf{R}_2\mathbf{r}_2\rangle\langle\mathbf{R}_2\mathbf{r}_2|\exp[-\tau_2H_{1,1}]|\mathbf{R}\mathbf{r}\rangle\\
&\quad\times[2|\mathbf{r}_2-\mathbf{R}_1|-|\mathbf{r}_2-\mathbf{r}_1|-|\mathbf{R}_2-\mathbf{R}_1|]
\end{aligned}\tag{A.15}$$

discarding terms of order $\beta e^2\kappa O((\kappa\lambda)^2)$. Next subtracted terms in $\mathcal{B}^T(a,1)$ are rewritten similarly to (A.15) where imaginary-time evolution operators are associated with purely kinetic Hamiltonian $H_{1,0}+H_{0,1}$. At low temperatures, such terms become exponentially smaller than (A.15), the behavior of which is controlled by atomic groundstate contributions. That behavior is determined by starting with decomposition $H_{1,1}=-\hbar^2\Delta_{\mathbf{R}^*}/(2M)+H_{pe}$. An eigenstate of $H_{1,1}$ reduces then to the product of a plane wave $\exp(i\mathbf{K}\cdot\mathbf{R}^*)/\Lambda^{1/2}$ for position $\mathbf{R}^*=(m_p\mathbf{R}+m_e\mathbf{r})/M$ of the atom mass center, times an internal wave function $\psi_p(\mathbf{r}^*)$ for relative position $\mathbf{r}^*=\mathbf{r}-\mathbf{R}$, while its energy reads $\hbar^2\mathbf{K}^2/(2M)+E_H^{(p)}$. For bound states, $p\to(n,l,m)$ where n is the usual quantum principal number ($E_H^{(p)}=E_H/n^2, 1\le n$), l is the orbital number ($0\le l\le n-1$) and m is the azimuthal number $-l\le m\le l$ ($0\to(1,0,0)$ denotes the ground state); for diffusive states, $p\to(k,l,m)$ where k parametrizes the positive energy $E_H^{(p)}=\hbar^2k^2/(2m)$ while l and m are again the orbital and azimuthal numbers with $0\le l$ (the precise forms of the corresponding ψ_p's are given in Ref. [42] for instance). After changing proton and electron positions in favor of their mass center and relative particle counterparts in (A.15), the matrix elements are evaluated by suitable insertions of closure relation for the basis made with previous eigenstates. The resulting integrals over τ_1 and τ_2 are readily performed for each set of involved eigenstates. According to the scaling prescriptions defined in Sect. 3, integral (A.13) is then rewritten as (3.19) plus terms which decay exponentially faster than $\rho^*\exp(\beta E_H)$, while screening function $S_3(1,1)$ reads

$$\begin{aligned}
S_3(1,1)=&\frac{\sqrt{2}(\beta|E_H|)^{1/2}}{64\pi^5}\exp(-\beta E_H)\Bigg\{4\int \mathrm{d}\mathbf{K}\mathrm{d}\mathbf{Q}\frac{\sinh(\mathbf{K}\cdot\mathbf{Q})}{\mathbf{K}\cdot\mathbf{Q}}\exp(-(K^2+Q^2)/2)\\
&\times\int \mathrm{d}\mathbf{X}\mathrm{d}\mathbf{r}_1^*\mathrm{d}\mathbf{r}_2^*\exp(-2i\mathbf{K}\cdot\mathbf{X})|\psi_0(\mathbf{r}_1^*)|^2|\psi_0(\mathbf{r}_2^*)|^2\\
&\times\left[2\left|\left(\frac{2\beta|E_H|m}{M}\right)^{1/2}\mathbf{X}+\frac{m_e}{M}\mathbf{r}_1^*+\frac{m_p}{M}\mathbf{r}_2^*\right|\right.\\
&-\left|\left(\frac{2\beta|E_H|m}{M}\right)^{1/2}\mathbf{X}-\frac{m_p}{M}\mathbf{r}_1^*+\frac{m_p}{M}\mathbf{r}_2^*\right|\\
&\left.-\left|\left(\frac{2\beta|E_H|m}{M}\right)^{1/2}\mathbf{X}+\frac{m_e}{M}\mathbf{r}_1^*-\frac{m_e}{M}\mathbf{r}_2^*\right|\right]
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{p\neq 0}\int \mathrm{d}\mathbf{K}\mathrm{d}\mathbf{Q}\frac{\exp(-K^2/2)}{\beta(E_H^{(p)}-E_H)+Q^2/2-K^2/2}\\
&\times\int \mathrm{d}\mathbf{X}\mathrm{d}\mathbf{r}_1^*\mathrm{d}\mathbf{r}_2^*\exp(i(\mathbf{K}-\mathbf{Q})\cdot\mathbf{X})\\
&\times\overline{\psi}_0(\mathbf{r}_1^*)\psi_0(\mathbf{r}_2^*)\psi_p(\mathbf{r}_1^*)\overline{\psi}_p(\mathbf{r}_2^*)\Bigg[2\left|\left(\frac{2\beta|E_H|m}{M}\right)^{1/2}\mathbf{X}+\frac{m_e}{M}\mathbf{r}_1^*+\frac{m_p}{M}\mathbf{r}_2^*\right|\\
&-\left|\left(\frac{2\beta|E_H|m}{M}\right)^{1/2}\mathbf{X}-\frac{m_p}{M}\mathbf{r}_1^*+\frac{m_p}{M}\mathbf{r}_2^*\right|\\
&-\left|\left(\frac{2\beta|E_H|m}{M}\right)^{1/2}\mathbf{X}+\frac{m_e}{M}\mathbf{r}_1^*-\frac{m_e}{M}\mathbf{r}_2^*\right|\Bigg]\Bigg\}.
\end{aligned}
\tag{A.16}
$$

In (A.16), all integration variables are dimensionless, in particular $\mathbf{r}_i^*$ is expressed in units of a_B. Moreover, the sum over diffusive states must be understood as an integral over k from 0 to ∞ and a discrete sum over l and m. When $\beta\to\infty$, $S_3(1,1)$ behaves as

$$
S_3(1,1)\sim\frac{c_{at}}{8\pi^{3/2}(\beta|E_H|)^{1/2}}\exp(-\beta E_H)
\tag{A.17}
$$

with pure numerical constant

$$
\begin{aligned}
c_{at}=&\sum_{p\neq 0}\frac{|E_H|}{E_H^{(p)}-E_H}\int \mathrm{d}\mathbf{Y}\mathrm{d}\mathbf{r}_1^*\mathrm{d}\mathbf{r}_2^*\frac{\exp(-|\mathbf{Y}|)}{|\mathbf{Y}|}\overline{\psi}_0(\mathbf{r}_1^*)\psi_0(\mathbf{r}_2^*)\psi_p(\mathbf{r}_1^*)\overline{\psi}_p(\mathbf{r}_2^*)\\
&\times\Bigg[2\left|\left(\frac{m|E_H|}{M(E_H^{(p)}-E_H)}\right)^{1/2}\mathbf{Y}+\frac{m_e}{M}\mathbf{r}_1^*+\frac{m_p}{M}\mathbf{r}_2^*\right|\\
&-\left|\left(\frac{m|E_H|}{M(E_H^{(p)}-E_H)}\right)^{1/2}\mathbf{Y}-\frac{m_p}{M}\mathbf{r}_1^*+\frac{m_p}{M}\mathbf{r}_2^*\right|\\
&-\left|\left(\frac{m|E_H|}{M(E_H^{(p)}-E_H)}\right)^{1/2}\mathbf{Y}+\frac{m_e}{M}\mathbf{r}_1^*-\frac{m_e}{M}\mathbf{r}_2^*\right|\Bigg].
\end{aligned}
\tag{A.18}
$$

Notice that only the second term ($\sum_{p\neq 0}\cdots$) in (A.16) contributes to asymptotic behavior (A.17) (the first term provide contributions smaller by a factor $\ln(\beta|E_H|)/(\beta|E_H|)$ as a consequence of the spherical symmetry of groundstate wavefunction $\psi_0(\mathbf{r}^*)=\psi_0(r^*)$). An accurate simplified expression for c_{at} can be derived by setting $m/M=m_e/M=0$ and $m_p/M=1$ in agreement with numerical value of ratio $m_e/m_p\simeq 1/1850$, i.e.

$$
c_{at}\simeq -4\pi\sum_{p\neq 0}\frac{|E_H|}{E_H^{(p)}-E_H}\int \mathrm{d}\mathbf{r}_1^*\mathrm{d}\mathbf{r}_2^*\overline{\psi}_0(\mathbf{r}_1^*)\psi_0(\mathbf{r}_2^*)\psi_p(\mathbf{r}_1^*)\overline{\psi}_p(\mathbf{r}_2^*)|\mathbf{r}_1^*-\mathbf{r}_2^*|
\tag{A.19}
$$

which provides $c_{at}\simeq 10.1$.

Appendix B: Low-Temperature Behavior of Few-Body Partition Functions

B.1 Behavior of $Z(1,1)$

Two-body proton-electron partition function $Z(1,1)$ reads

$$Z(1,1) = 4\int \mathrm{d}\mathbf{x}\langle\mathbf{x}|[\exp(-\beta H_{pe})]^{T}_{\text{Mayer}}|\mathbf{x}\rangle, \tag{B.1}$$

where reduced Hamiltonian H_{pe} describes a single particle with mass m submitted to attractive Coulomb potential $V_{pe}(\mathbf{x}) = -e^2/|\mathbf{x}|$. Here, for the sake of notational convenience, we set $\mathbf{x} = \mathbf{r}^*$, while z will denote a complex number. The integral over $\mathbf{x}$ can be splitted into two parts, $|\mathbf{x}| < \beta e^2$ and $\mathbf{x} > \beta e^2$, the contributions of which are determined as follows.

For $|\mathbf{x}| < \beta e^2$, truncated Mayer operator $[\exp(-\beta H_{pe})]^{T}_{\text{Mayer}}$ is replaced by its definition (3.14) in (B.1). The contribution of second term in (3.14), as well as those of next terms involving imaginary-time evolutions of V_{pe} with kinetic Hamiltonian K_{pe}, are readily computed in terms of elementary functions by exploiting the Gaussian nature of matrix elements of $\exp(-\beta K_{pe})$. Such contributions are bounded by a power of β. The contribution of first term in (3.14) is analyzed by introducing the Green function $\widehat{G}(\mathbf{x},\mathbf{y};z)$ defined as the matrix elements of resolvent $(z + H_{pe})^{-1}$. That function is the solution of partial differential equation

$$\left(-\frac{\hbar^2}{2m}\Delta_{\mathbf{x}} - \frac{e^2}{|\mathbf{x}|} + z\right)\widehat{G}(\mathbf{x},\mathbf{y};z) = \delta(\mathbf{x}-\mathbf{y}), \tag{B.2}$$

with suitable boundary conditions [36]. Its exact expression reads [36]

$$\begin{aligned}\widehat{G}(\mathbf{x},\mathbf{y};z) = {}& \frac{m\Gamma(1-i\nu)}{2\pi\hbar^2|\mathbf{x}-\mathbf{y}|}[W_{i\nu,1/2}(-ikd_+)\dot{M}_{i\nu,1/2}(-ikd_-) \\ & - \dot{W}_{i\nu,1/2}(-ikd_+)M_{i\nu,1/2}(-ikd_-)],\end{aligned} \tag{B.3}$$

with $k = (-2mz/\hbar^2)^{1/2}$ ($\Im(k) > 0$), $\nu = 1/(ka_B)$, $d_+ = |\mathbf{x}| + |\mathbf{y}| + |\mathbf{x}-\mathbf{y}|$ and $d_- = |\mathbf{x}| + |\mathbf{y}| - |\mathbf{x}-\mathbf{y}|$, while $\Gamma(u)$ is the familiar gamma-function and $W_{i\nu,1/2}(u)$, $M_{i\nu,1/2}(u)$ are Whittaker functions [31] ($\dot{W}_{i\nu,1/2}(u) = \partial W_{i\nu,1/2}(u)/\partial u$, $\dot{M}_{i\nu,1/2}(u) = \partial M_{i\nu,1/2}(u)/\partial u$). Green function $\widehat{G}(\mathbf{x},\mathbf{y};z)$ is analytical in z in the whole complex plane, except on the negative real axis ($\Re(z) \le 0$, $\Im(z) = 0$) which is a cut line, and also at $z = z_n = -E_H/n^2$ ($n \ge 1$) which are simple poles (of course, such singularities are controlled by the spectrum of H_{pe}). When $\mathbf{x} \to \mathbf{y}$, $\widehat{G}(\mathbf{x},\mathbf{y};z)$ diverges as $m/(2\pi\hbar^2|\mathbf{x}-\mathbf{y}|)$, as shown by expanding Whittaker functions for arguments close to $-2ik|\mathbf{x}|$. That $1/|\mathbf{x}-\mathbf{y}|$-behavior, is also displayed by free Green function $\widehat{G}_0(\mathbf{x},\mathbf{y};z) = m\exp(ik|\mathbf{x}-\mathbf{y}|)/(2\pi\hbar^2|\mathbf{x}-\mathbf{y}|)$, and it can be interpreted by quoting that (B.2) reduces to Poisson equation for $|\mathbf{x}-\mathbf{y}|$ small. It is convenient to define

$$\begin{aligned}\widehat{H}(\mathbf{x},\mathbf{y};z) &= \widehat{G}(\mathbf{x},\mathbf{y};z) - \widehat{G}_0(\mathbf{x},\mathbf{y};z) + \int \mathrm{d}\mathbf{r}\widehat{G}_0(\mathbf{x},\mathbf{r};z)V_{pe}(|\mathbf{r}|)\widehat{G}_0(\mathbf{r},\mathbf{y};z) \\ &= \int \mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2\widehat{G}_0(\mathbf{x},\mathbf{r}_1;z)V_{pe}(|\mathbf{r}_1|)\widehat{G}_0(\mathbf{r}_1,\mathbf{r}_2;z)V_{pe}(|\mathbf{r}_2|)\widehat{G}(\mathbf{r}_2,\mathbf{y};z)\end{aligned} \tag{B.4}$$

where the second equality follows from a standard identity for interacting and free Green functions. That function $\widehat{H}(\mathbf{x},\mathbf{y};z)$ has the same analytical properties as $\widehat{G}(\mathbf{x},\mathbf{y};z)$. Using above expansions of Whittaker functions, we find that $\widehat{H}(\mathbf{x},\mathbf{x};z)$ reads

$$\widehat{H}(\mathbf{x},\mathbf{x};z) = \frac{-imk\Gamma(1-i\nu)}{2\pi\hbar^2}\Big[2\dot{W}_{i\nu,1/2}\dot{M}_{i\nu,1/2} - \ddot{W}_{i\nu,1/2}M_{i\nu,1/2} - W_{i\nu,1/2}\ddot{M}_{i\nu,1/2}\Big](-2ik|\mathbf{x}|) - \frac{imk}{2\pi\hbar^2} - \frac{m^2e^2}{4\pi^2\hbar^4}\int \mathrm{d}\mathbf{r}\frac{\exp(2ik|\mathbf{x}-\mathbf{r}|)}{r|\mathbf{x}-\mathbf{r}|^2}. \tag{B.5}$$

Notice that, contrarily to $\widehat{G}-\widehat{G}_0$ which diverges as $-2m^2e^2\ln(|k||\mathbf{x}|)/(\pi\hbar^4)$ when $\mathbf{x}\to\mathbf{0}$, $\widehat{H}(\mathbf{x},\mathbf{x};z)$ remains finite at $\mathbf{x}=\mathbf{0}$ thanks to the addition of the integral of $\widehat{G}_0V_{pe}\widehat{G}_0$ in (B.4).

Since $\widehat{G}(\mathbf{x},\mathbf{y};z)$ is nothing but the Laplace transform with respect to β of density matrix $\langle\mathbf{x}|\exp(-\beta H_{pe})|\mathbf{y}\rangle$, the standard inversion formula provides

$$\langle\mathbf{x}|\exp(-\beta H_{pe})|\mathbf{y}\rangle = \frac{1}{2i\pi}\int_{D_\sigma}\mathrm{d}z\exp(\beta z)\widehat{G}(\mathbf{x},\mathbf{y};z) \tag{B.6}$$

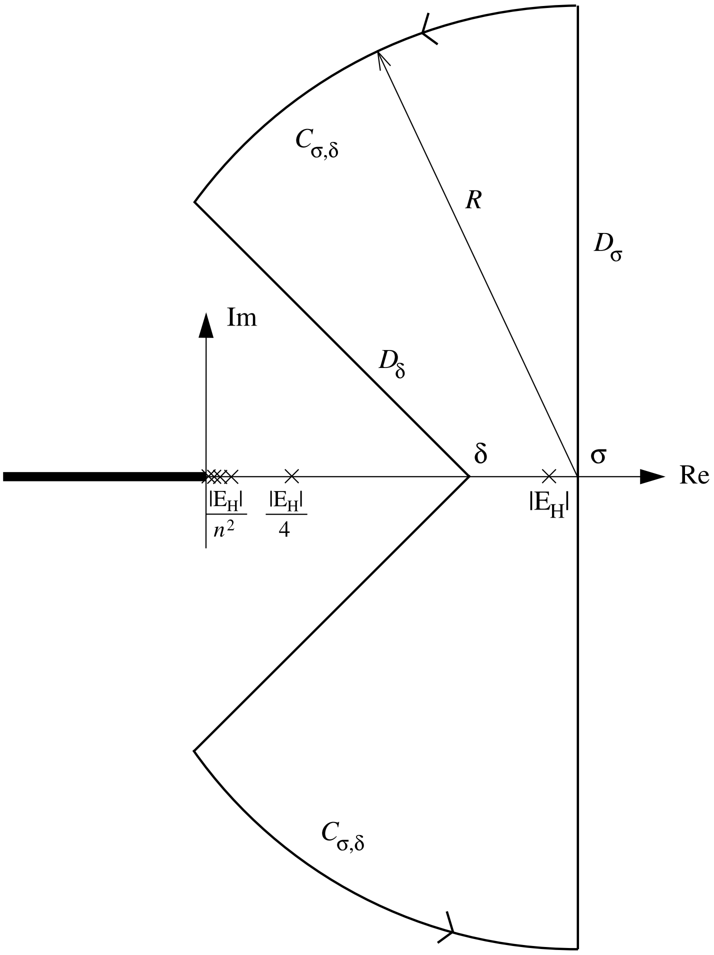
where D_σ is any straight line parallel to imaginary axis and which cuts real axis at $\sigma > -E_H$. Inserting decomposition of $\widehat{G}$ in terms of $\widehat{G}_0$, V_{pe} and $\widehat{H}$ into (B.6), we find that the contribution of terms involving $\widehat{G}_0$ give raise to free density matrix $\langle\mathbf{x}|\exp(-\beta K_{pe})|\mathbf{y}\rangle = \exp(-|\mathbf{x}-\mathbf{y}|^2/(2\lambda_{pe}^2))/(2\pi\lambda_{pe}^2)^{3/2}$. The resulting expression of $\langle\mathbf{x}|\exp(-\beta H_{pe})|\mathbf{y}\rangle$ can be specified to diagonal elements $\mathbf{x}=\mathbf{y}$ because $\widehat{H}(\mathbf{x},\mathbf{x};z)$ is finite. The integral of $\widehat{H}(\mathbf{x},\mathbf{x};z)\exp(\beta z)$ along D_σ is then transformed by applying Cauchy's theorem with contour $C_{\sigma,\delta}$ shown in Fig. 13 ($-E_H/4 < \delta < -E_H$). Function $\widehat{H}(\mathbf{x},\mathbf{x};z)\exp(\beta z)$ is analytical inside $C_{\sigma,\delta}$ except at $z=z_1=-E_H$ which is a simple pole with residue $|\psi_0(\mathbf{x})|^2\exp(-\beta E_H)$. Moreover, it satisfies Jordan's lemma on the circular parts of $C_{\sigma,\delta}$ which connect D_δ to D_σ, so the corresponding parts of the contour integral vanish when the radius is sent to infinity. This provides

$$\begin{aligned}&\langle\mathbf{x}|\exp(-\beta H_{pe})|\mathbf{x}\rangle\\ &\quad= |\psi_0(\mathbf{x})|^2\exp(-\beta E_H) - \frac{1}{2i\pi}\int_{D_\delta}\mathrm{d}z\exp(\beta z)\widehat{H}(\mathbf{x},\mathbf{x};z)\\ &\qquad+\frac{1}{(2\pi\lambda_{pe}^2)^{3/2}}\left[1+\frac{\beta e^2}{|\mathbf{x}|}\int_0^1\mathrm{d}s\,\Phi(|\mathbf{x}|/(\sqrt{2}s(1-s)\lambda_{pe}))\right]\end{aligned} \tag{B.7}$$

where Φ is the familiar Error function [31]. Last term in (B.7) is bounded by a power of β. Along D_δ, index $i\nu$ of Whittaker functions follows a closed curve in complex plane which cuts real axis at non-strictly positive integers. At the same time, their argument $u=-2ik|\mathbf{x}|$ remains inside sector $-3\pi/8 < \arg(u) < 3\pi/8$. Consequently, $|\widehat{H}(\mathbf{x},\mathbf{x};z)|$ remains bounded by a constant when z runs along D_δ. The modulus of the corresponding integral in (B.7) is then bounded by a power of β times $\exp(\beta\delta)$. Taking into account above power-law bounds for the contributions of truncated terms in $[\exp(-\beta H_{pe})]^T_{\text{Mayer}}$, and noting that $\psi_0(\mathbf{x})$ decays exponentially fast for $|\mathbf{x}| > \beta e^2$, we eventually obtain

$$\int_{|\mathbf{x}|<\beta e^2}\mathrm{d}\mathbf{x}\langle\mathbf{x}|[\exp(-\beta H_{pe})]^T_{\text{Mayer}}|\mathbf{x}\rangle = \exp(-\beta E_H), \tag{B.8}$$

Fig. 13 Path in complex plane



discarding additional terms which are exponentially smaller than $\exp(-\beta E_H)$ when $\beta \to \infty$.

For $|\mathbf{x}| > \beta e^2$, it is convenient to use Feynman-Kac expression of $\langle \mathbf{x}|[\exp(-\beta \times H_{pe})]^T_{\text{Mayer}}|\mathbf{x}\rangle$. Within the variable change $\mathbf{x} = \beta e^2 \mathbf{v}$, the corresponding integral is rewritten as

$$\begin{aligned}
&\int_{|\mathbf{x}|>\beta e^2} \mathrm{d}\mathbf{x} \langle \mathbf{x}|[\exp(-\beta H_{pe})]^T_{\text{Mayer}}|\mathbf{x}\rangle \\
&\quad = \frac{1}{(2\pi)^{3/2}} \left(\frac{\beta e^2}{\lambda_{pe}}\right)^3 \int_{|\mathbf{v}|>1} \mathrm{d}\mathbf{v} \int \mathcal{D}(\boldsymbol{\xi}) \\
&\quad \times \left[\exp\left(\int_0^1 \mathrm{d}s \left|\mathbf{v} + \frac{\lambda_{pe}}{\beta e^2}\boldsymbol{\xi}(s)\right|^{-1}\right) - 1 - \int_0^1 \mathrm{d}s \left|\mathbf{v} + \frac{\lambda_{pe}}{\beta e^2}\boldsymbol{\xi}(s)\right|^{-1}\right. \\
&\quad \left. - \frac{1}{2}\left(\int_0^1 \mathrm{d}s \left|\mathbf{v} + \frac{\lambda_{pe}}{\beta e^2}\boldsymbol{\xi}(s)\right|^{-1}\right)^2 - \frac{1}{6}\left(\int_0^1 \mathrm{d}s \left|\mathbf{v} + \frac{\lambda_{pe}}{\beta e^2}\boldsymbol{\xi}(s)\right|^{-1}\right)^3\right]. \qquad \text{(B.9)}
\end{aligned}$$

When $\beta \to \infty$, ratio $\lambda_{pe}/\beta e^2$ vanishes. Then, potential $|\mathbf{v} + \frac{\lambda_{pe}}{\beta e^2}\boldsymbol{\xi}(s)|^{-1}$ can be merely replaced by $|\mathbf{v}|$ because the mean extension of $\boldsymbol{\xi}(s)$ is of order 1. The corresponding functional integral over $\boldsymbol{\xi}(s)$ reduces to 1 by normalization of Wiener measure $\mathcal{D}(\boldsymbol{\xi})$, and the remaining integral over $\mathbf{v}$ is a pure number. Thus, the integral in the l.h.s. of (B.9) behaves as a power of β, and it becomes exponentially smaller than (B.8) when $\beta \to \infty$. This implies

$$Z(1,1) = 4\exp(-\beta E_H) \tag{B.10}$$

discarding terms which are exponentially smaller when $\beta \to \infty$. Next corrections to that leading behavior arising from excited atomic states, can be readily obtained within a similar approach by adjusting the position δ between two successive poles, i.e. $z_{n+1} < \delta < z_n$ with $n > 1$.

B.2 Behavior of $Z(1,2)$

The low-temperature behavior of $Z(1,2)$, which involves one proton and two electrons, is determined by a straightforward extension of previous methods applied to $Z(1,1)$. If an exact expression of three-body Green function is not available, its main properties of interest can be guessed, under rather weak assumptions, by exploiting relations and analogies with its free counterpart, as well as known results on the spectrum of $H_{1,2}$. Such properties appear to be quite natural extensions of the exact behaviors observed in the two-body case.

Let $\mathbf{R} = (m_p\mathbf{R}_a + m_e\mathbf{r}_1 + m_e\mathbf{r}_2)/M_{H^-}$ be the position of center of mass ($M_{H^-} = m_p + 2m_e$), while $\mathbf{x}_1 = (\mathbf{r}_1 - \mathbf{r}_2)/\sqrt{2}$ and $\mathbf{x}_2 = (m_p/M_{H^-})^{1/2}(\mathbf{r}_1 + \mathbf{r}_2 - 2\mathbf{R}_a)/\sqrt{2}$ are the reduced variables. After expressing $H_{1,2}$ in terms of those variables, we find that three-body partition function $Z(1,2)$ can be rewritten as

$$\begin{aligned} Z(1,2) = 2\int \mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2\{&2\langle \mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{1,2})|\mathbf{x}_1\mathbf{x}_2\rangle \\ &- \langle -\mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{1,2})|\mathbf{x}_1\mathbf{x}_2\rangle + \cdots\} \end{aligned} \tag{B.11}$$

with reduced Hamiltonian

$$\begin{aligned} H^*_{1,2} = &-\frac{\hbar^2}{2m_e}\Delta_{\mathbf{x}_1} - \frac{\hbar^2}{2m_e}\Delta_{\mathbf{x}_2} + \frac{e^2}{\sqrt{2}|\mathbf{x}_1|} - \frac{\sqrt{2}e^2}{|\mathbf{x}_1 + (M_{H^-}/m_p)^{1/2}\mathbf{x}_2|} \\ &- \frac{\sqrt{2}e^2}{|\mathbf{x}_1 - (M_{H^-}/m_p)^{1/2}\mathbf{x}_2|}. \end{aligned} \tag{B.12}$$

Off-diagonal matrix elements in the r.h.s. of (B.11) are associated with the exchange of the electrons. First potential term in (B.12) describes Coulomb repulsion between those electrons, while the second and third ones account for Coulomb attractions between the proton and each electron. In the double integral involved in (B.11), we make a partition of space integration into three domains $\Omega^{(i)}$ ($i = 0, 1, 2$), such that i, and only i, sides of triangle $(\mathbf{0}, \mathbf{x}_1, \mathbf{x}_2)$ are smaller than βe^2 in $\Omega^{(i)}$.

For $\mathbf{x}_1, \mathbf{x}_2$ inside $\Omega^{(2)}$, we express matrix elements of $\exp(-\beta H^*_{1,2})$ as inverse Laplace transforms of Green function $\widehat{G}(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}_1, \mathbf{y}_2; z)$ solution of

$$(H^*_{1,2} + z)\widehat{G}(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}_1, \mathbf{y}_2; z) = \delta(\mathbf{x}_1 - \mathbf{y}_1)\delta(\mathbf{x}_2 - \mathbf{y}_2) \tag{B.13}$$

with suitable boundary conditions. That Green function is analytical in z in the whole complex plane, except on a part of real axis with $\Re(z) < -E_{H^-}$, while $z_1 = -E_{H^-}$ is a simple (isolated) pole with residue $\psi_0(\mathbf{x}_1, \mathbf{x}_2)\overline{\psi}_0(\mathbf{y}_1, \mathbf{y}_2)$ (ψ_0 is the groundstate wavefunction of

$H^*_{1,2}$ with energy E_{H^-}). For a given z not close to its singularities, $\widehat{G}(\mathbf{x}_1,\mathbf{x}_2;\mathbf{y}_1,\mathbf{y}_2;z)$ displays a position-dependence analogous to that of free Green function $\widehat{G}_0(\mathbf{x}_1,\mathbf{x}_2;\mathbf{y}_1,\mathbf{y}_2;z)$ solution of Helmholtz equation in six dimensions. In particular, for $\mathbf{x}_i$ close to $\mathbf{y}_i$, potential terms in (B.13) can be omitted and $\widehat{G}(\mathbf{x}_1,\mathbf{x}_2;\mathbf{y}_1,\mathbf{y}_2;z)$ also diverges as $1/[(\mathbf{x}_1-\mathbf{y}_1)^2+(\mathbf{x}_2-\mathbf{y}_2)^2]^2$, i.e. the Coulomb potential in six dimensions. In order to handle such divergences for diagonal matrix elements, we consider identity

$$\begin{aligned}\widehat{G}&(\mathbf{x}_1,\mathbf{x}_2;\mathbf{y}_1,\mathbf{y}_2;z)\\&=\widehat{G}_0(\mathbf{x}_1,\mathbf{x}_2;\mathbf{y}_1,\mathbf{y}_2;z)\\&\quad-\int \mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2\widehat{G}_0(\mathbf{x}_1,\mathbf{x}_2;\mathbf{r}_1,\mathbf{r}_2;z)V^*_{1,2}(\mathbf{r}_1,\mathbf{r}_2)\widehat{G}(\mathbf{r}_1,\mathbf{r}_2;\mathbf{y}_1,\mathbf{y}_2;z)\end{aligned} \tag{B.14}$$

where $V^*_{1,2}$ is the potential part of $H^*_{1,2}$. Successive iterations of formula (B.14) generate the perturbative expansion of $\widehat{G}(\mathbf{x}_1,\mathbf{x}_2;\mathbf{y}_1,\mathbf{y}_2;z)$ in powers of $V^*_{1,2}$. We define $\widehat{H}(\mathbf{x}_1,\mathbf{x}_2;\mathbf{y}_1,\mathbf{y}_2;z)$ as $\widehat{G}(\mathbf{x}_1,\mathbf{x}_2;\mathbf{y}_1,\mathbf{y}_2;z)$ minus the first six terms (up to order $(V^*_{1,2})^5$ included) of that expansion. Similarly to (B.4), $\widehat{H}$ has the same analytical properties as $\widehat{G}$, while $\widehat{H}(\mathbf{x}_1,\mathbf{x}_2;\mathbf{x}_1,\mathbf{x}_2;z)$ is now finite. After inserting the expression of $\widehat{G}$ in terms of $\widehat{G}_0$, $V^*_{1,2}$ and $\widehat{H}$ into the inversion formula for diagonal matrix element of $\exp(-\beta H^*_{1,2})$, we find that terms built with $\widehat{G}_0$ and $V^*_{1,2}$ are bounded by powers of β. For dealing with the contribution of $\widehat{H}$, we introduce a closed contour $C_{\sigma,\delta}$ in complex plane similar to that shown in Fig. 13, with $z_2<\delta<z_1$ and $-z_2$ the first singularity of $\widehat{H}$ below z_1. Along that contour, z stays at a finite distance at any singularity of $\widehat{H}$. Also, for $|z|$ large, in the expression of $\widehat{H}$ as a spatial integral of $\widehat{G}_0V^*_{1,2}\widehat{G}_0\cdots V^*_{1,2}\widehat{G}$ (similar to that involved in (B.4)), $\widehat{G}$ can be replaced by $\widehat{G}_0$. Therefore, $\widehat{H}$ goes to zero as a power of k, as shown by variable changes $\mathbf{r}_i\to|k|\mathbf{r}_i$ (integrals over the $\mathbf{r}_i$'s do converge thanks to exponentially decaying factors $\exp(ik|\mathbf{r}_i-\mathbf{r}_j|)$ arising from the $\widehat{G}_0$'s). Hence, we conclude that $|\widehat{H}(\mathbf{x}_1,\mathbf{x}_2;\mathbf{x}_1,\mathbf{x}_2;z)|$ remains bounded along D_δ, while Jordan's lemma applies to the integral of $\widehat{H}(\mathbf{x}_1,\mathbf{x}_2;\mathbf{x}_1,\mathbf{x}_2;z)\exp(\beta z)$ upon the circular parts of $C_{\sigma,\delta}$. This provides

$$2\langle\mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{1,2})|\mathbf{x}_1\mathbf{x}_2\rangle=2|\psi_0(\mathbf{x}_1,\mathbf{x}_2)|^2\exp(-\beta E_{H^-}) \tag{B.15}$$

discarding terms which are exponentially smaller. Within similar methods, we can also estimate the off-diagonal matrix element $\langle-\mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{1,2})|\mathbf{x}_1\mathbf{x}_2\rangle$. Contributions of terms involving $\widehat{G}_0$'s and $V^*_{1,2}$'s are readily bounded by powers of β by rescaling positions in units of λ_e. Contribution of $\widehat{H}$ is treated as above since $\widehat{H}(-\mathbf{x}_1,\mathbf{x}_2;\mathbf{x}_1,\mathbf{x}_2;z)$ is bounded along D_δ and decays sufficiently fast for $|z|$ large. After using $\psi_0(-\mathbf{x}_1,\mathbf{x}_2)=\psi_0(\mathbf{x}_1,\mathbf{x}_2)$ ($H^*_{1,2}$ is invariant under transformation $\mathbf{x}_1\to-\mathbf{x}_1$ at fixed $\mathbf{x}_2$), we find

$$\langle-\mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{1,2})|\mathbf{x}_1\mathbf{x}_2\rangle=|\psi_0(\mathbf{x}_1,\mathbf{x}_2)|^2\exp(-\beta E_{H^-}), \tag{B.16}$$

discarding terms which are exponentially smaller. Next terms $\cdots$ in the r.h.s. of (B.11), which arise from truncation in $[\exp(-\beta H_{1,2})]^T_{\text{Mayer}}$, can be also estimated by similar techniques. For instance, term

$$\int_0^\beta \mathrm{d}\tau\exp[-(\beta-\tau)(H_{1,1}+H_{0,1})]V_{at,e}\exp[-\tau(H_{1,1}+H_{0,1})] \tag{B.17}$$

provides a contribution which can be rewritten as the inverse Laplace transform of

$$\int d\mathbf{y}_1 d\mathbf{y}_2 \langle \mathbf{x}_1\mathbf{x}_2|(z + H^*_{pe,e})^{-1}|\mathbf{y}_1\mathbf{y}_2\rangle V_{at,e}(\mathbf{y}_1, \mathbf{y}_2)\langle \mathbf{y}_1\mathbf{y}_2|(z + H^*_{pe,e})^{-1}|\mathbf{x}_1\mathbf{x}_2\rangle \tag{B.18}$$

with

$$H^*_{pe,e} = -\frac{\hbar^2}{2m_e}\Delta_{\mathbf{x}_1} - \frac{\hbar^2}{2m_e}\Delta_{\mathbf{x}_2} - \frac{\sqrt{2}e^2}{|\mathbf{x}_1 + (M_{H^-}/m_p)^{1/2}\mathbf{x}_2|} \tag{B.19}$$

and

$$V_{at,e}(\mathbf{x}_1, \mathbf{x}_2) = \frac{e^2}{\sqrt{2}|\mathbf{x}_1|} - \frac{\sqrt{2}e^2}{|\mathbf{x}_1 - (M_{H^-}/m_p)^{1/2}\mathbf{x}_2|}. \tag{B.20}$$

Green function defined as matrix elements of $(z + H^*_{pe,e})^{-1}$, is analytical with respect to z in the whole complex plane, except on part $\Re(z) \leq -E_H$ of the real axis. In the Laplace inversion formula, we introduce a contour analogous to that of Fig. 13 with $\delta > -E_H$. We also define a regular part of $\langle \mathbf{x}_1\mathbf{x}_2|(z + H^*_{pe,e})^{-1}|\mathbf{y}_1\mathbf{y}_2\rangle$ which remains finite at $\mathbf{x}_i = \mathbf{y}_i$. When z follows D_δ, $\Im(k)$ remains larger than a given positive constant, so previous regular part is bounded by an exponentially decaying function of $[(\mathbf{x}_1 - \mathbf{y}_1)^2 + (\mathbf{x}_2 - \mathbf{y}_2)^2]^{1/2}$ (for large separations of the arguments, Coulomb potential terms vanish so Green functions behave as their free counterparts which decay exponentially on a scale $(\Im(k))^{-1}$). This implies that (B.18) remains bounded by a constant along D_δ. Contribution of (B.17) is then found to be bounded by a power of β times $\exp(\beta\delta)$, with δ arbitrarily close to $-E_H$. Since $-E_H < -E_{H^-}$, that contribution is exponentially smaller than (B.15) and (B.16). Contributions of all the other truncated terms in $[\exp(-\beta H_{1,2})]^T_{\text{Mayer}}$ behave similarly because groundstate energies of Hamiltonians $(H_{1,0} + H_{0,2})$ and $(H_{1,0} + H_{0,1} + H_{0,1})$ (which both vanish) are strictly larger than E_{H^-}. Since volume of $\Omega^{(2)}$ is bounded by a constant times $(\beta e^2)^6$ on one hand, while $\psi_0(\mathbf{x}_1, \mathbf{x}_2)$ decays exponentially fast for $|\mathbf{x}_i|$ large on another hand, we eventually obtain

$$\begin{aligned}\int_{\Omega^{(2)}} & d\mathbf{x}_1 d\mathbf{x}_2 \{2\langle \mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{1,2})|\mathbf{x}_1\mathbf{x}_2\rangle - \langle -\mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{1,2})|\mathbf{x}_1\mathbf{x}_2\rangle + \cdots\} \\ &= \exp(-\beta E_{H^-})\end{aligned} \tag{B.21}$$

discarding terms which are exponentially smaller.

For $\mathbf{x}_1, \mathbf{x}_2$ inside $\Omega^{(1)}$, two of three distances $|\mathbf{x}_1|$, $|\mathbf{x}_1 - \mathbf{x}_2|$, $|\mathbf{x}_2|$ are larger than βe^2. For instance, we may have both $|\mathbf{x}_1|$ and $|\mathbf{x}_1 - \mathbf{x}_2|$ larger than βe^2, while $|\mathbf{x}_2|$ is smaller than βe^2. For such configurations, both distances $|\mathbf{x}_1|$ and $|\mathbf{x}_1 - (M_{H^-}/m_p)^{1/2}\mathbf{x}_2|$ are larger than βe^2. In the Feynman-Kac formula for $\langle \mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{1,2})|\mathbf{x}_1\mathbf{x}_2\rangle$, potentials $\int_0^1 ds e^2/\sqrt{2}|\mathbf{x}_1 + \lambda_e\boldsymbol{\xi}_1(s)|$ and $-\int_0^1 ds\sqrt{2}e^2/|\mathbf{x}_1 + \lambda_e\boldsymbol{\xi}_1(s) - (M_{H^-}/m_p)^{1/2}(\mathbf{x}_2 + \lambda_e\boldsymbol{\xi}_2(s))|$ can then be replaced by their classical counterparts $e^2/\sqrt{2}|\mathbf{x}_1|$ and $-\sqrt{2}e^2/|\mathbf{x}_1 - (M_{H^-}/m_p)^{1/2}\mathbf{x}_2|$ respectively, because $\lambda_e/\beta e^2$ goes to zero when β diverges. Thus, $\langle \mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{1,2})|\mathbf{x}_1\mathbf{x}_2\rangle$ behaves as

$$\langle \mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{pe,e})|\mathbf{x}_1\mathbf{x}_2\rangle \exp(-\beta V_{at,e}(\mathbf{x}_1, \mathbf{x}_2)) \tag{B.22}$$

at leading order, where $H^*_{pe,e}$ and $V_{at,e}(\mathbf{x}_1, \mathbf{x}_2)$ are given by (B.19) and (B.20) respectively. A similar estimation holds for truncated terms in $[\exp(-\beta H_{1,2})]^T_{\text{Mayer}}$ built with powers of

$V_{at,e}$, so the corresponding full contribution integrated upon considered configurations behaves as

$$2\int_{|\mathbf{x}_1|,|\mathbf{x}_1-\mathbf{x}_2|>\beta e^2,|\mathbf{x}_2|<\beta e^2} d\mathbf{x}_1 d\mathbf{x}_2 \langle \mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{pe,e})|\mathbf{x}_1\mathbf{x}_2\rangle \times\Bigg[\exp(-\beta V_{at,e}(\mathbf{x}_1,\mathbf{x}_2))+\beta V_{at,e}(\mathbf{x}_1,\mathbf{x}_2)-\frac{1}{2}(\beta V_{at,e}(\mathbf{x}_1,\mathbf{x}_2))^2 + \frac{1}{6}(\beta V_{at,e}(\mathbf{x}_1,\mathbf{x}_2))^3\Bigg]. \tag{B.23}$$

According to above properties of Green functions, $\langle \mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{pe,e})|\mathbf{x}_1\mathbf{x}_2\rangle$ is bounded by some power of β times $\exp(\beta\delta)$ with $-E_H<\delta<-E_{H^-}$. Therefore, integral (B.23) is also bounded by a power of β times $\exp(\beta\delta)$, because purely classical integral

$$\int_{|\mathbf{x}_1|,|\mathbf{x}_1-\mathbf{x}_2|>\beta e^2,|\mathbf{x}_2|<\beta e^2} d\mathbf{x}_1 d\mathbf{x}_2 \Bigg[\exp(-\beta V_{at,e}(\mathbf{x}_1,\mathbf{x}_2))+\beta V_{at,e}(\mathbf{x}_1,\mathbf{x}_2) -\frac{1}{2}(\beta V_{at,e}(\mathbf{x}_1,\mathbf{x}_2))^2+\frac{1}{6}(\beta V_{at,e}(\mathbf{x}_1,\mathbf{x}_2))^3\Bigg] \tag{B.24}$$

is proportional to $(\beta e^2)^6$ as shown by rescaling $\mathbf{x}_i$ in units of βe^2. A similar analysis applies to the contributions of the other terms in $[\exp(-\beta H_{1,2})]^T_{\text{Mayer}}$. When off-diagonal matrix elements are involved, we use bounds inferred from properties of Green functions, which decay exponentially fast with respect to relative distances between different arguments. For instance, $|\langle -\mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{1,2})|\mathbf{x}_1\mathbf{x}_2\rangle|$ is bounded by a constant times $\exp(-\beta E_{H^-})\exp(-c|\mathbf{x}_1|/a_B)$ with $c>0$, so

$$-\int_{|\mathbf{x}_1|,|\mathbf{x}_1-\mathbf{x}_2|>\beta e^2,|\mathbf{x}_2|<\beta e^2} d\mathbf{x}_1 d\mathbf{x}_2 \langle -\mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{1,2})|\mathbf{x}_1\mathbf{x}_2\rangle \tag{B.25}$$

decays exponentially faster than $\exp(-\beta E_{H^-})$. Previous analysis can be also repeated for the other configurations belonging to $\Omega^{(1)}$, i.e. $\{|\mathbf{x}_2|,|\mathbf{x}_1-\mathbf{x}_2|>\beta e^2, |\mathbf{x}_1|<\beta e^2\}$ and $\{|\mathbf{x}_1|,|\mathbf{x}_2|>\beta e^2, |\mathbf{x}_1-\mathbf{x}_2|<\beta e^2\}$. We eventually find that

$$\int_{\Omega^{(1)}} d\mathbf{x}_1 d\mathbf{x}_2 \{2\langle \mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{1,2})|\mathbf{x}_1\mathbf{x}_2\rangle - \langle -\mathbf{x}_1\mathbf{x}_2|\exp(-\beta H^*_{1,2})|\mathbf{x}_1\mathbf{x}_2\rangle + \cdots\} \tag{B.26}$$

decays exponentially faster than $\exp(-\beta E_{H^-})$.

For $\mathbf{x}_1,\mathbf{x}_2$ inside $\Omega^{(0)}$, all distances $|\mathbf{x}_1|$, $|\mathbf{x}_1-\mathbf{x}_2|$, and $|\mathbf{x}_2|$ are larger than βe^2. For diagonal matrix elements, potential parts can be treated classically at leading order, as immediately seen from Feynman-Kac formula by noting that $\lambda_e/\beta e^2$ vanishes. Such matrix elements then behave as their free counterparts times classical Boltzmann factors. The corresponding full contribution integrated upon $\Omega^{(0)}$ is shown to be proportional to $(\beta e^2/\lambda_e)^6$, as shown by variable changes $\mathbf{x}_i=\beta e^2\mathbf{v}_i$. The contribution of remaining terms with off-diagonal matrix elements decays exponentially faster than $\exp(-\beta E_{H^-})$, thanks to the existence of bounds which are proportional to $\exp(-\beta E_{H^-})$ and decay exponentially fast for large separations (over a finite length scale proportional to a_B). Thus, and like (B.26), contribution of $\Omega^{(0)}$ to $Z(1,2)$ also decays exponentially faster than $\exp(-\beta E_{H^-})$, so we eventually obtain

$$Z(1,2) = 2\exp(-\beta E_{H^-}) \tag{B.27}$$

discarding terms which are exponentially smaller when $\beta \to \infty$.

B.3 Behaviors of $Z(2,1)$, $Z(2,2)$, ...

The low-temperature behaviors of $Z(2,1)$ and $Z(2,2)$ can be also determined by previous methods introduced for studying $Z(1,1)$ and $Z(1,2)$. Again, analytic properties of Green functions associated with resolvents $(z + H^*_{2,1})^{-1}$ and $(z + H^*_{2,2})^{-1}$ play a crucial role in the derivations. Such functions are analytical in the whole complex plane, except on a part of real axis with $\Re(z) < -E_{H_2^+}$ or $\Re(z) < -E_{H_2}$, while $z_1 = -E_{H^-}$ or $z_1 = -E_{H_2}$ is a simple (isolated) pole. Along integration contours analogous to that described in Fig. 13, they can be bounded as above, within quite plausible arguments based on the properties of their free counterparts, solutions of Helmholtz equation in six or nine dimensions. Integration space upon reduced positions is splitted into several parts according to the values of relative distances compared to βe^2. The parts inside which all relative distances are smaller than βe^2, provide the leading contributions, i.e.

$$Z(2,1) = 2\exp(-\beta E_{H_2^+}) \tag{B.28}$$

and

$$Z(2,2) = \exp(-\beta E_{H_2}) \tag{B.29}$$

discarding terms which are exponentially smaller when $\beta \to \infty$.

The analysis can be applied to any partition function $Z(N_p, N_e)$. However, when the infimum of reduced Hamiltonian $H^*_{N_p,N_e}$ is not separated from the rest of the spectrum (i.e. in the corresponding groundstate, the N_p protons and the N_e electrons are not binded together), the first singularity (with the largest real part) $z_1 = -E^{(0)}_{N_p,N_e}$ of Green function associated with $(z + H^*_{N_p,N_e})^{-1}$, is a branching point which is not isolated from other singularities. Then, contour $C_{\sigma,\delta}$ cuts real axis at $\delta > z_1$, so the previous methods show that $|Z(N_p, N_e)|$ is bounded by a power of β times $\exp(\beta\delta)$. A more detailed analysis of the behavior of Green function for z close to z_1 is then required for determining the precise leading behavior of $Z(N_p, N_e)$. Nevertheless, since previous bound is valid for δ arbitrarily close to z_1, it is quite reasonable to assume that $Z(N_p, N_e)$ then behaves as a power of β times $\exp(-\beta E^{(0)}_{N_p,N_e})$ (such a behavior is indeed observed for $Z(2,0)$ with $E^{(0)}_{2,0} = 0$).

Appendix C: Leading Contributions of Interactions between Atoms and Ionized Charges

C.1 Expression of $W(1,1|1,1)$

The low-temperature behavior of bare contributions of Figs. 8b and 8c, is determined along similar lines as that of polarization contribution (A.13). The integrals of interest are again expressed in terms of the atom mass centers and of the reduced variables. Matrix elements are also evaluated *via* insertions of the closure relation for a suitable basis. Each eigenstate in that basis, is the product of plane waves describing mass center motions, times atomic internal wavefunctions. The resulting integrals over times τ_i are readily performed for each set of involved eigenstates. According to the scaling prescriptions defined in Sect. 3, full bare contribution of Figs. 8b and 8c is then rewritten as (3.41) plus terms which decay exponentially faster than $\rho^* \exp(\beta E_H)$. Like $Z(1,1)$ or $S_3(1,1)$, function $W(1,1|1,1)$ is determined by atomic groundstate contributions, and further contributions of excited states might be bounded by methods similar to that exposed in Appendix B:. Similarly to expression (A.16) for $S_3(1,1)$, $W(1,1|1,1)$ reduces to the product $\exp(-2\beta E_H)$ of atomic groundstate Boltzmann factors, times a function of $\beta|E_H|$ which remains bounded by a power law at low temperatures. Its asymptotic form when $\beta \to \infty$ is given by (3.42), where $c_{at,at}$ is the pure numerical coefficient

$$\begin{aligned}
c_{at,at} &= \frac{2}{\pi}\left\{\frac{M}{m}\int \mathrm{d}\mathbf{K}\frac{|D^{(at,at)}_{00,00}(\mathbf{K})|^2}{K^6}\right.\\
&\quad\left. + \sum_{(p_1,p_2)\neq(0,0)}\int \mathrm{d}\mathbf{K}\frac{2|E_H|}{(E_H^{(p_1)}+E_H^{(p_2)}-2E_H+\hbar^2K^2/(Ma_B^2))}\frac{|D^{(at,at)}_{0p_1,0p_2}(\mathbf{K})|^2}{K^4}\right\}\\
&\quad\times\frac{-1}{\pi^3}\left\{\left(\frac{M}{m}\right)^2\int \mathrm{d}\mathbf{K}\mathrm{d}\mathbf{Q}\frac{D^{(at,at)}_{00,00}(\mathbf{K})D^{(at,at)}_{00,00}(\mathbf{Q}-\mathbf{K})D^{(at,at)}_{00,00}(-\mathbf{Q})}{K^4Q^4|\mathbf{K}-\mathbf{Q}|^2}\right.\\
&\quad + \sum_{(p_1,p_2,p_3,p_4)\neq(0,0,0,0)}\int \mathrm{d}\mathbf{K}\mathrm{d}\mathbf{Q}\frac{2|E_H|}{(E_H^{(p_1)}+E_H^{(p_2)}-2E_H+\hbar^2K^2/(Ma_B^2))}\\
&\quad\times\frac{2|E_H|}{(E_H^{(p_3)}+E_H^{(p_4)}-2E_H+\hbar^2Q^2/(Ma_B^2))}\\
&\quad\left.\times\frac{D^{(at,at)}_{0p_1,0p_2}(\mathbf{K})D^{(at,at)}_{p_1p_3,p_2p_4}(\mathbf{Q}-\mathbf{K})D^{(at,at)}_{p_30,p_40}(-\mathbf{Q})}{K^2Q^2|\mathbf{K}-\mathbf{Q}|^2}\right\}. \qquad (C.1)
\end{aligned}$$

In (C.1), $\mathbf{K}$ and $\mathbf{Q}$ are dimensionless (units a_B^{-1}) wavenumbers, while function $D^{(at,at)}_{0p_1,0p_2}(\mathbf{K})$ reduces to

$$\begin{aligned}
D^{(at,at)}_{0p_1,0p_2}(\mathbf{K}) &= \langle\psi_0|\exp\left(-i\frac{m}{m_e}\mathbf{K}\cdot\mathbf{r}^*\right)|\psi_{p_1}\rangle\langle\psi_0|\exp\left(i\frac{m}{m_e}\mathbf{K}\cdot\mathbf{r}^*\right)|\psi_{p_2}\rangle\\
&\quad+\langle\psi_0|\exp\left(i\frac{m}{m_p}\mathbf{K}\cdot\mathbf{r}^*\right)|\psi_{p_1}\rangle\langle\psi_0|\exp\left(-i\frac{m}{m_p}\mathbf{K}\cdot\mathbf{r}^*\right)|\psi_{p_2}\rangle\\
&\quad-\langle\psi_0|\exp\left(-i\frac{m}{m_e}\mathbf{K}\cdot\mathbf{r}^*\right)|\psi_{p_1}\rangle\langle\psi_0|\exp\left(-i\frac{m}{m_p}\mathbf{K}\cdot\mathbf{r}^*\right)|\psi_{p_2}\rangle\\
&\quad-\langle\psi_0|\exp\left(i\frac{m}{m_p}\mathbf{K}\cdot\mathbf{r}^*\right)|\psi_{p_1}\rangle\langle\psi_0|\exp\left(i\frac{m}{m_e}\mathbf{K}\cdot\mathbf{r}^*\right)|\psi_{p_2}\rangle. \qquad (C.2)
\end{aligned}$$

C.2 Expressions of $W(1,1|1,0)$ and $W(1,1|0,1)$

A straightforward extension of previous methods provides the low-temperature behaviors of bare contributions of Figs. 9a–f. Using again the scaling prescriptions defined in Sect. 3, the corresponding full bare contribution is then rewritten as (3.45) plus terms which decay exponentially faster than $\rho^* \exp(\beta E_H)$. Functions $W(1,1|1,0)$ and $W(1,1|0,1)$ are also determined by atomic groundstate contributions. They behave as $\exp(-\beta E_H)$ times functions of $\beta|E_H|$ which remain bounded by power laws. Their asymptotic forms when $\beta \to \infty$ are given by 3.46, where pure numerical constants reduce to

$$\begin{aligned}
c^{(2)}_{at,\alpha} &= \frac{2}{\pi}\left\{\frac{2m_\alpha M}{m(m_\alpha+M)}\int \mathrm{d}\mathbf{K}\frac{|D^{(at,\alpha)}_{00}(\mathbf{K})|^2}{K^6}\right.\\
&\quad + \sum_{p_1\neq 0}\int \mathrm{d}\mathbf{K}\frac{2|E_H|}{(E_H^{(p_1)}-E_H+\hbar^2K^2/(2Ma_B^2)+\hbar^2K^2/(2m_\alpha a_B^2))}\left.\frac{|D^{(at,\alpha)}_{0p_1}(\mathbf{K})|^2}{K^4}\right\}\\
&\quad \times \frac{-1}{\pi^3}\left\{\left(\frac{2m_\alpha M}{m(m_\alpha+M)}\right)^2\int \mathrm{d}\mathbf{K}\mathrm{d}\mathbf{Q}\frac{D^{(at,\alpha)}_{00}(\mathbf{K})D^{(at,\alpha)}_{00}(\mathbf{Q}-\mathbf{K})D^{(at,\alpha)}_{00}(-\mathbf{Q})}{K^4Q^4|\mathbf{K}-\mathbf{Q}|^2}\right.\\
&\quad + \sum_{(p_1,p_2)\neq(0,0)}\mathrm{d}\mathbf{K}\mathrm{d}\mathbf{Q}\frac{2|E_H|}{(E_H^{(p_1)}-E_H+\hbar^2K^2/(2Ma_B^2)+\hbar^2K^2/(2m_\alpha a_B^2))}\\
&\quad \times \frac{2|E_H|}{(E_H^{(p_2)}-E_H+\hbar^2Q^2/(2Ma_B^2)+\hbar^2Q^2/(2m_\alpha a_B^2))}\\
&\quad \times \left.\frac{D^{(at,\alpha)}_{0p_1}(\mathbf{K})D^{(at,\alpha)}_{p_1p_2}(\mathbf{Q}-\mathbf{K})D^{(at,\alpha)}_{p_20}(-\mathbf{Q})}{K^2Q^2|\mathbf{K}-\mathbf{Q}|^2}\right\},
\end{aligned} \tag{C.3}$$

with

$$\begin{aligned}
D^{(at,p)}_{0p_1}(\mathbf{K}) &= -D^{(at,e)}_{0p_1}(\mathbf{K})\\
&= \langle\psi_0|\exp\left(-i\frac{m}{m_e}\mathbf{K}\cdot\mathbf{r}^*\right)|\psi_{p_1}\rangle - \langle\psi_0|\exp\left(i\frac{m}{m_p}\mathbf{K}\cdot\mathbf{r}^*\right)|\psi_{p_1}\rangle.
\end{aligned} \tag{C.4}$$

References

1. Alastuey, A., Ballenegger, V., Cornu, F.: Low temperature isotherms of the hydrogen plasma: comparisons between analytical and Monte Carlo results. ENS Lyon preprint
2. Alastuey, A., Martin, Ph.A.: Absence of exponential clustering for static quantum correlations and time-displaced correlations in charged fluids. Eur. Phys. Lett. **6**, 385–390 (1988)
3. Alastuey, A., Martin, Ph.A.: Absence of exponential clustering in quantum Coulomb fluids. Phys. Rev. A **40**, 6485–6520 (1989)
4. Alastuey, A., Perez, A.: Virial expansion of the equation of state of a quantum plasma. Europhys. Lett. **20**, 19–24 (1992)
5. Alastuey, A., Perez, A.: Virial expansion for quantum plasmas: Fermi–Bose statistics. Phys. Rev. E **53**, 5714–5728 (1996)
6. Alastuey, A., Cornu, F., Perez, A.: Virial expansion for quantum plasmas: diagrammatic resummations. Phys. Rev. E **49**, 1077–1093 (1994)
7. Alastuey, A., Cornu, F., Perez, A.: Virial expansion for quantum plasmas: Maxwell–Boltzmann statistics. Phys. Rev. E **51**, 1725–1744 (1995)
8. Alastuey, A., Ballenegger, V., Cornu, F., Martin, Ph.A.: Screened cluster expansions for partially ionized gases. J. Stat. Phys. **113**, 455–503 (2003)
9. Alastuey, A., Cornu, F., Martin, Ph.A.: Van der Waals forces in presence of free charges: an exact derivation from equilibrium quantum correlations. J. Chem. Phys. **127**, 054506 (2007)
10. Ballenegger, V.: Étude des phénomènes d'écran et de polarisation dans un plasma quantique par la méthode des graphes de Mayer. PhD thesis, École Polytechnique et Fédérale de Lausanne (2002)
11. Ballenegger, V., Martin, Ph.A.: Quantum Coulomb systems: some exact results in the atomic limit. Physica A **306**, 59–67 (2002)
12. Ballenegger, V., Martin, Ph.A.: Dielectric versus conductive behavior in quantum gases: Exact results for the Hydrogen plasma. Physica A **328**, 97–144 (2003)

13. Ballenegger, V., Martin, Ph.A., Alastuey, A.: Quantum Mayer graphs for Coulomb systems and the analog of the Debye potential. J. Stat. Phys. **108**, 169–211 (2002)
14. Brydges, D.C., Martin, Ph.A.: Coulomb systems at low density: a review. J. Stat. Phys. **96**, 1163–1330 (1999)
15. Ceperley, D.M.: Fermion nodes. J. Stat. Phys. **63**, 1237–1267 (1991)
16. Ceperley, D.M.: Path integrals in the theory of condensed helium. Rev. Mod. Phys. **67**, 279–355 (1995)
17. Ceperley, D.M.: In: Binder, E.K., Ciccotti, G. (eds.) Monte Carlo and Molecular Dynamics of Condensed Matter Systems. Editrice Compositori, Bologna (1996)
18. Conlon, J.G., Lieb, E.H., Yau, H.T.: The Coulomb gas at low temperature and low density. Commun. Math. Phys. **125**, 153–180 (1989)
19. Cornu, F.: Correlations in quantum plasmas, I: resummations in Mayer-like diagrammatics. Phys. Rev. E **53**, 4562 (1996)
20. DeWitt, H.E., Schlanges, M., Sakakura, A.Y., Kraeft, W.D.: Low density expansion of the equation of state for a quantum electron gas. Phys. Lett. A **197**, 326–329 (1995)
21. Dyson, F., Lenard, A.: Stability of matter I. J. Math. Phys. **8**, 423–434 (1967)
22. Dyson, F., Lenard, A.: Stability of matter II. J. Math. Phys. **9**, 698–711 (1968)
23. Ebeling, W.: Ann. Phys. Leipz. **19**, 104 (1967)
24. Ebeling, W.: Coulomb interaction and ionization equilibrium in partially ionized plasmas. Physica **63**, 293 (1969)
25. Ebeling, W.: Statistical derivation of the mass-action law for interacting gases and plasmas. Physica **73**, 573–584 (1974)
26. Ebeling, W., Kraeft, W.D., Kremp, D.: Theory of Bound States and Ionization Equilibrium in Plasmas and Solids. Akademie, Berlin (1976)
27. Fefferman, C.: The atomic and molecular nature of matter. Rev. Math. Iberoam. **1**, 1–44 (1985)
28. Fefferman, C.: The n-body problem in quantum mechanics. Commun. Pure Appl. Math. **39**(S1), 67–109 (1986)
29. Gell-Mann, M., Brueckner, K.A.: Correlation energy of an electron gas at high density. Phys. Rev. **106**, 364–368 (1957)
30. Ginibre, J.: Some applications of functional integration in statistical mechanics. In: DeWitt, C., Stora, R. (eds.) Statistical Mechanics and Quantum Field Theory. Gordon and Breach, les Houches (1971)
31. Gradsteyn, I.S., Ryzhik, I.M.: In: Jeffrey, A. (ed.) Tables of Integrals, Series and Products. Academic Press, London (1994)
32. Gruter, P., Laloe, F.: Ursell operators in statistical physics I. J. Phys. I. (France) **5**, 1255 (1995)
33. Gruter, P., Laloe, F.: Ursell operators in statistical physics II. J. Phys. I. (France) **5**, 181 (1995)
34. Gruter, P., Laloe, F.: Ursell operators in statistical physics III. J. Phys. I. (France) **7**, 485 (1997)
35. Hill, L.T.: Statistical Mechanics. McGraw-Hill, New York (1956)
36. Hostler, L.: Coulomb's Green functions and the Furry's approximation. J. Math. Phys. **5**, 591 (1964)
37. Kahlbaum, T.: The quantum-diffraction term in the free energy for Coulomb plasma and the effective-potential approach. J. Phys. IV **10**(P5), 455 (2000)
38. Kleinert, H.: Path Integrals in Quantum Mechanics, Statistics, Polymer Physics and Financial Markets. World Scientific, Singapore (2004)
39. Kraeft, W.D., Kremp, D., Ebeling, W., Ropke, G.: Quantum Statistics of Charged Particle Systems. Plenum, New York (1986)
40. Krauth, W.: Statistical Mechanics: Algorithms and Computations. Oxford University Press, Oxford (2006)
41. Kremp, D., Kraeft, W.D., Lambert, A.J.M.D.: Equation of state and ionization equilibrium for non-ideal plasmas. Physica A **127**, 72–86 (1984)
42. Landau, L.D., Lifchitz, E.M.: Quantum Mechanics, 3rd edn. Course of Theoretical Physics, vol. 3. Pergamon, Oxford (1977)
43. Lebowitz, J., Pena, R.: Low density form of the free energy of real matter. J. Chem. Phys. **59**, 1362–1364 (1973)
44. Lieb, E.H., Lebowitz, J.: The constitution of matter: existence of thermodynamics for systems composed of electrons and nuclei. Adv. Math. **9**, 316–398 (1972)
45. Macris, N., Martin, Ph.A.: Ionization equilibrium in the proton-electron gas. J. Stat. Phys. **60**, 619–637 (1990)
46. Martin, Ph.A.: Quantum Mayer graphs: applications to Bose and Coulomb gases. Acta Phys. Pol. B **34**, 3629 (2003)
47. Militzer, B., Ceperley, D.M.: Path integral Monte Carlo calculation of the deuterium hugoniot. Phys. Rev. Lett. **85**, 1890–1893 (2000)
48. Militzer, B., Ceperley, D.M.: Path integral Monte Carlo simulation of the low-density hydrogen plasma. Phys. Rev. E **63**, 066404 (2001)

49. Militzer, B., Pollock, E.L.: Variational density matrix method for warm, condensed matter: application to dense hydrogen. Phys. Rev. E **61**, 3470–3482 (2000)
50. Montroll, E.W., Ward, J.C.: Quantum statistics of interacting particles: General theory and some remarks on properties of an electron gas. Phys. Fluid **1**, 55 (1958)
51. Morita, T.: Equation of state of high temperature plasma. Prog. Theor. Phys. **22**, 757 (1959)
52. All spectroscopic calculations were performed by V. Robert, who used a coupled cluster (CCSD(T)) approach including an extended basis set for hydrogen atoms $(4s3p2d1f)$
53. Roepstorff, G.: Path Integral Approach to Quantum Physics. Springer, Berlin (1994)
54. Rogers, F.J.: Statistical mechanics of Coulomb gases of arbitrary charge. Phys. Rev. A **10**, 2441 (1974)
55. Rogers, F.J.: Equation of state of dense, partially degenerate, reacting plasmas. Phys. Rev. A **24**, 1531 (1981)
56. Rogers, F.J.: Occupation numbers for reacting plasmas—the role of the Planck–Larkin function. Astrophys. J. **310**, 723 (1986)
57. Rogers, F.J.: A distribution function approach for effective occupation numbers and the equation of state of hydrogen plasmas. Astrophys. J. **352**, 689 (1990)
58. Rogers, F.J.: In: Chabrier, G., Schatzman, E. (eds.) The Equation of State in Astrophysics. Cambridge University Press, New York (1994)
59. Rogers, F.J., Young, D.A.: Validation of the activity expansion method with ultrahigh pressure shock equations of state. Phys. Rev. E **56**, 5876 (1997)
60. Saha, M.: Philos. Mag. **40**, 472 (1920)
61. Schulman, L.S.: Techniques and Applications of Path Integrals. Wiley, New York (1981)
62. Simon, B.: Functional Integration and Quantum Physics. Academic, New York (1979)

PHYSICAL REVIEW E **86**, 066402 (2012)

Thermodynamics of atomic and ionized hydrogen: Analytical results versus equation-of-state tables and Monte Carlo data

A. Alastuey
Laboratoire de Physique, ENS Lyon, UMR CNRS 5672, 46 allée d'Italie, 69364 Lyon Cedex 07, France

V. Ballenegger
Institut UTINAM, Université de Franche-Comté, UMR CNRS 6213, 16, route de Gray, 25030 Besançon cedex France

(Received 8 August 2012; published 6 December 2012)

We compute thermodynamical properties of a low-density hydrogen gas within the physical picture, in which the system is described as a quantum electron-proton plasma interacting via the Coulomb potential. Our calculations are done using the exact scaled low-temperature (SLT) expansion, which provides a rigorous extension of the well-known virial expansion—valid in the fully ionized phase—into the Saha regime where the system is partially or fully recombined into hydrogen atoms. After recalling the SLT expansion of the pressure [A. Alastuey *et al.*, J. Stat. Phys. **130**, 1119 (2008)], we obtain the SLT expansions of the chemical potential and of the internal energy, up to order $\exp(-|E_H|/kT)$ included ($E_H \simeq -13.6$ eV). Those truncated expansions describe the first five nonideal corrections to the ideal Saha law. They account exactly, up to the considered order, for all effects of interactions and thermal excitations, including the formation of bound states (atom H, ions H^- and ${H_2}^+$, molecule $H_2, \ldots$) and atom-charge and atom-atom interactions. Among the five leading corrections, three are easy to evaluate, while the remaining ones involve well-defined internal partition functions for the molecule H_2 and ions H^- and ${H_2}^+$, for which no closed-form analytical formula exist currently. We provide accurate low-temperature approximations for those partition functions by using known values of rotational and vibrational energies. We compare then the predictions of the SLT expansion, for the pressure and the internal energy, with, on the one hand, the equation-of-state tables obtained within the opacity program at Livermore (OPAL) and, on the other hand, data of path integral quantum Monte Carlo (PIMC) simulations. In general, a good agreement is found. At low densities, the simple analytical SLT formulas reproduce the values of the OPAL tables up to the last digit in a large range of temperatures, while at higher densities ($\rho \sim 10^{-2}$ g/cm^3), some discrepancies among the SLT, OPAL, and PIMC results are observed.

DOI: 10.1103/PhysRevE.86.066402 PACS number(s): 52.25.Kn, 67.10.Fj

I. INTRODUCTION

As the lightest and most simple element, hydrogen is important both theoretically and for practical applications. It is also the most abundant element in the universe, and a precise knowledge of its thermodynamical properties is needed by astrophysicists over a wide range of pressures and temperatures. In that context, the derivation of accurate tables for thermodynamical functions is quite useful. This motivated the celebrated opacity program at Livermore (OPAL), which, in addition, provides tabulations of the opacity as a function of temperature and density, a key ingredient for astrophysical diagnosis.

The OPAL equation-of-state tables [1] have been derived from the activity expansion (ACTEX) method, first introduced in Ref. [2] and implemented through successive papers [3]. That approach is built within the physical picture, where hydrogen is described in terms of a quantum plasma made with protons and electrons interacting *via* the $1/r$ Coulomb potential. For a given set of thermodynamical parameters, one proceeds to suitable estimations of the expected relevant contributions in the activity expansions determined by simple physical arguments. This allows one to account for complex phenomena arising from the formation of chemical species and their interactions. The resulting OPAL tables are very reasonably accurate over a wide range of temperatures and densities, as checked through comparisons to quantum Monte Carlo simulations [4] and to high-pressure shock experiments [5].

Aside from the OPAL tables, exact asymptotic expansions can be used to provide reliable numerical data. It turns out that such an expansion, the so-called scaled low-temperature (SLT) expansion, has been recently derived in the Saha regime [6], where hydrogen reduces to a dilute partially ionized atomic gas. That regime is of particular astrophysical interest since it is observed, for instance, in the Sun interior. The main purpose of that paper is to derive, from the SLT expansion, simple and very precise estimations of the contributions of all the mechanisms at work in the Saha regime to any thermodynamical function. Our calculations avoid approximations introduced in the ACTEX approach, and they are written in terms of tractable analytic formulas which are quite easy to handle for determining the quantities of interest at any temperature and any density. No interpolation has to be performed as in other purely numerical tables like OPAL. The corresponding high-accuracy and thermodynamically consistent calculations should be quite useful for various applications, in particular the interpretation of recent seismology measurements in the Sun [7].

In the physical picture, the equation of state is studied by applying methods of quantum statistical mechanics to Coulombic matter. Various analytical methods have been developed for this purpose, such as effective potential methods [8–11], many-body perturbation theory [12,13], and Mayer diagrams in the polymer representation of the quantum system [14–16]. Numerical techniques have also been elaborated, in particular, density functional theory molecular dynamics [17–19] and path-integral Monte Carlo (PIMC) simulations [20]. In the present work, we use a suitable extension [21]—needed for dealing with a partially recombined phase—of the quantum Mayer diagrams method introduced previously to derive in particular the virial expansion of the equation of state up to order $\rho^{5/2}$ in the density, both in the absence [14,22,23] and presence [24] of a magnetic field. This framework avoids the problems associated with the more widely used chemical approach [25–29] in which bound states (atoms H, molecules H_2, ions H^-, ${H_2}^+$, ...) are treated as preformed constituents that are assumed to interact via some given effective potentials [30,31] with the ionized charges and between themselves. The SLT expansion [6] solves the difficult problem of dealing consistently and exactly with screening and bound states in the Saha regime within the physical picture. Effect of atom-atom interactions and screened interactions between ionized charges and atoms appear, for instance, in our calculations, without introducing any intermediate modelization, as a consequence of the basic Coulombic interactions between the electrons and protons. They are embedded in functions $h_2(\beta)$ and $h_4(\beta)$ defined in Sec. II D. In the SLT expansion, the internal partition functions of all bound entities are finite thanks to a systematic account of collective screening effects. That expansion exhibits no missing term, nor double counting, for instance, between contributions associated to a hydrogen molecule or to two hydrogen atoms, despite the atoms may form a molecule at short distances.

The Saha regime corresponds to low-temperature and sufficiently low densities, so the most abundant chemical species are ionized protons, ionized electrons, and hydrogen atoms in their ground state. The corresponding thermal ionization equilibrium $\mathrm{H} \rightleftarrows e + p$ is well described in first approximation by the mass-action law for ideal mixtures [32]

$$\frac{\rho_{\mathrm{at}}^{\mathrm{id}}}{\rho_p^{\mathrm{id}}\rho_e^{\mathrm{id}}} = \left(\frac{2\pi\hbar^2\beta}{m}\right)^{3/2} e^{-\beta E_{\mathrm{H}}}, \qquad (1)$$

which relates the number density $\rho_{\mathrm{at}}^{\mathrm{id}}$ of hydrogen atoms in their ground state with energy $E_{\mathrm{H}} = -me^4/(2\hbar^2) \simeq -13.6\,\mathrm{eV}$ to the number densities ρ_p^{id} and ρ_e^{id} of ionized protons and electrons with $\rho_p^{\mathrm{id}} = \rho_e^{\mathrm{id}}$ because of charge neutrality. In ionization equation (1), β is the inverse temperature, while m_p and m_e are the proton and electron masses, and $m = m_p m_e/(m_p + m_e)$ is the mass of the reduced particle. All ideal densities can be computed in terms of the sole total electron or proton density $\rho = \rho_e^{\mathrm{id}} + \rho_{\mathrm{at}}^{\mathrm{id}} = \rho_{\mathrm{p}}^{\mathrm{id}} + \rho_{\mathrm{at}}^{\mathrm{id}}$ and of the temperature-dependent density

$$\rho^* = \frac{\exp(\beta E_{\mathrm{H}})}{2\left(2\pi\lambda_{\mathrm{pe}}^2\right)^{3/2}} \quad \text{with} \quad \lambda_{pe} = (\beta\hbar^2/m)^{1/2}, \qquad (2)$$

which naturally emerges in ionization equation (1). The resulting Saha equation of state (EOS) follows from adding the partial pressures of the three ideal gases in the mixture, and it reads

$$\beta P_{\mathrm{Saha}} = \rho + \rho^*[\sqrt{1 + 2\rho/\rho^*} - 1]\,. \qquad (3)$$

Temperature-dependent density ρ^* controls the crossover between full ionization and full recombination, as illustrated by the respective behaviors $\beta P_{\mathrm{Saha}} \sim 2\rho$ for $\rho \ll \rho^*$ and $\beta P_{\mathrm{Saha}} \sim \rho$ for $\rho \gg \rho^*$. As recalled in Sec. II, within the physical picture, the Saha predictions have been proved to be asymptotically valid in a suitable scaling limit in the grand-canonical ensemble [33], where temperature T is decreased while chemical potentials μ_p and μ_e go to E_{H} with a linear dependence in T. The corresponding density decreases exponentially fast with T, in order to keep the same energy-entropy balance and, hence, the same ionization degree. The identification of that scaling limit opened up the possibility to construct systematic expansions beyond Saha theory [6]. The structure of the corresponding SLT expansion of the density in terms of the chemical potential is described in Sec. II. The successive terms depend on temperature-dependent functions $h_k(\beta)$ which decay exponentially fast when $T \to 0$ with increasing decay rates. Their physical content is discussed in relation with the formation of chemical species, interaction, and screening effects.

In Sec. III, starting from the SLT expansion of density and using standard thermodynamical identities, we derive the SLT expansions of chemical potential, pressure, and internal energy. By construction, all expressions are thermodynamically consistent, and similar expressions for other thermodynamical quantities can be easily derived along similar lines. In those SLT expansions, beyond the leading terms given by Saha theory, each correction reduces to an algebraic function of ratio ρ/ρ^* times a temperature-dependent function which decays exponentially fast when $T \to 0$. We give the expressions of all corrections up to order $\exp(\beta E_{\mathrm{H}})$ included. Such corrections account for various phenomena such as plasma polarization, thermal atomic excitations, shift of the atomic energy levels, formation of hydrogen molecules H_2 and ions H^- and ${H_2}^+$, and interactions between ionized charges and atoms.

As usual for asymptotic expansions, and aside from the question of convergence in a strict mathematical sense, the truncation of SLT expansions can be reasonably expected to provide reliable quantitative informations on thermodynamics. Here, since the characteristic energy scale $|E_{\mathrm{H}}|$ involved in SLT expansions is rather large, the corresponding calculations should be reliable up to temperatures of the order 10^4 K for which the condition $kT \ll |E_{\mathrm{H}}|$ is indeed fulfilled. Furthermore, we stress that, though the SLT expansion is built by considering a low-density and low-temperature scaling, it can provide actually accurate predictions in a rather large range of densities and temperatures that cover the fully ionized, partially ionized, and atomic phases of the hydrogen gas. Indeed, the SLT expansion reduces by construction to the standard virial expansion when $\rho \ll \rho^*$ at fixed T [6]. Thus, the SLT formulas remain valid in the fully ionized regime where one may have $T > T_{\mathrm{Rydberg}} = |E_{\mathrm{H}}|/k = 157\,801$ K, as long as the density is not too high, namely the coupling parameter $\Gamma = \beta e^2/a$ must remain small.

If we keep all corrections to Saha leading terms up to order $\exp(\beta E_{\mathrm{H}})$ included, as provided by the SLT formulas, the knowledge of the first four functions $h_1(\beta)$, $h_2(\beta)$, $h_3(\beta)$, and $h_4(\beta)$ is required. If functions $h_1(\beta)$ and $h_3(\beta)$ are explicitly known in closed elementary forms and can be calculated exactly at any temperature, no similar formulas for functions $h_2(\beta)$ and $h_4(\beta)$ are available since analytical results on the three- and four-body quantum problem are very scarce. In Sec. IV, we propose simple approximations of those functions which account for their exact low-temperature forms, on the one hand, and incorporate the usual reliable descriptions of the spectra of ions H^-, $H_2{}^+$ and of molecule H_2, on the other hand. Those approximations are sufficient for computing thermodynamical properties of a partially ionized hydrogen gas for temperatures up to about 30 000 K. More refined calculations of functions $h_2(\beta)$ and $h_4(\beta)$ would be required at higher temperatures, in particular for state points where recombination into hydrogen molecules or ions give a significant contribution.

In Sec. V, within previous simple representations of the $h_k(\beta)$'s, we study the importance of the various nonideal corrections to Saha pressure and internal energy along various isotherms and isochores. The predictions of our analytical SLT formulas are compared to the OPAL tables which, up to now, are expected to provide the most reliable numerical data in the considered regimes. A very good agreement is found at low densities, for all temperatures, if one corrects the OPAL tables by using the exact ground-state energy of the hydrogen atom with the reduced mass m in place of m_e [34]. When the density is increased, our predictions differ somewhat from those of the OPAL tables. We compare also our predictions to data of quantum Monte Carlo simulations [4]. These comparisons, together with a simple semiempirical criterion, allow us to determine the validity domain of the SLT expansion in the temperature-density plane (see Fig. 12). Final comments and possible extensions are given in Sec. VI.

II. THE SCALED LOW-TEMPERATURE EXPANSION

A. The Saha regime in the grand-canonical ensemble

Within the physical picture, a hydrogen gas is viewed as a system of quantum point particles which are either protons or electrons, interacting via the instantaneous Coulomb potential $v(r) = 1/r$. Protons and electrons have respective charges, masses, and spins, $e_p = e$ and $e_e = -e, m_p$ and $m_e, \sigma_p = \sigma_e = 1/2$. In the present nonrelativistic limit, the corresponding Hamiltonian for $N = N_p + N_e$ particles reads

$$H_{N_p,N_e} = -\sum_{i=1}^{N} \frac{\hbar^2}{2m_{\alpha_i}}\Delta_i + \frac{1}{2}\sum_{i\neq j} e_{\alpha_i}e_{\alpha_j}v(|\mathbf{x}_i - \mathbf{x}_j|), \quad (4)$$

where $\alpha_i = p,e$ is the species of the ith particle and Δ_i is the Laplacian with respect to its position $\mathbf{x}_i$. The system is enclosed in a box with volume Λ, in contact with a thermostat at temperature T and a reservoir of particles that fixes the chemical potentials equal to μ_p and μ_e for protons and electrons, respectively. Because the infinite system maintains local neutrality $\rho_p = \rho_e$ in any fluid phase, the bulk equilibrium quantities depend in fact solely on the mean,

$$\mu = (\mu_p + \mu_e)/2, \quad (5)$$

while the difference $\nu = (\mu_e - \mu_p)/2$ is not relevant, as rigorously proved in Ref. [35]. Consequently, the common particle density $\rho = \rho_p = \rho_e$ depends only on T and μ.

In the present framework, the EOS (3) has been proved to become exact in some limit introduced by Macris and Martin, who extended Fefferman's work on the atomic phase of the hydrogen plasma [36] to a partially ionized phase [33]. In that limit, the temperature T goes to zero while the average chemical potential μ of protons and electrons approaches the value E_H with a definite slope [33]. More precisely, let γ be the dimensionless parameter defined through the parametrization

$$\mu = E_H + kT\{\ln(\gamma) + \ln[(m/M)^{3/4}/4]\} \quad (6)$$

with $M = m_p + m_e$. The state of the system then is equivalently defined by either the usual set (T,μ) of thermodynamical parameters in the grand-canonical ensemble or the set (T,γ), since both sets are univocally related. As proved in Ref. [33], in the limit $T \to 0$ at fixed γ, density ρ and pressure P behave as $(c > 0)$

$$\rho = \left(\rho_p^{\text{id}} + \rho_{\text{at}}^{\text{id}}\right)[1 + O(e^{-c\beta})] = \left(\rho_e^{\text{id}} + \rho_{\text{at}}^{\text{id}}\right)[1 + O(e^{-c\beta})] \quad (7)$$

and

$$\beta P = \left(\rho_p^{\text{id}} + \rho_e^{\text{id}} + \rho_{\text{at}}^{\text{id}}\right)[1 + O(e^{-c\beta})], \quad (8)$$

where ideal densities reduce to

$$\rho_p^{\text{id}} = \rho_e^{\text{id}} = \rho^*\gamma \quad (9)$$

and

$$\rho_{\text{at}}^{\text{id}} = \rho^*\frac{\gamma^2}{2}. \quad (10)$$

Notice that ideal densities (9) and (10) do satisfy the Saha ionization equation (1) for the total proton/electron density $\rho = \rho^*\gamma(1 + \gamma/2)$. Moreover, the leading contribution to pressure in formula (8) indeed describes an ideal mixture of free protons, free electrons, and hydrogen atoms in their ground state, which can be rewritten in the form Eq. (3). Thus, discarding exponentially vanishing terms embedded in $O(e^{-c\beta})$, Saha predictions are rigorously recovered in the scaling limit of Macris and Martin $T \to 0$ at fixed γ. The parameter γ may be fixed at an arbitrary positive value; it controls the density and also the ionization ratio since $\rho_p^{\text{id}}/\rho = 1/(1 + \gamma/2)$. Contrary to the zero-temperature limit at fixed chemical potential used in the atomic and molecular limit theorem [15,36,37] (see also Ref. [38] for further physical considerations around that limit), we consider γ fixed and a chemical potential that varies as $T \to 0$ according to Eq. (6) so the e-p plasma tends in the limit $T \to 0$ to a partially ionized hydrogen gas with a well-defined ionization ratio.

The Saha regime corresponds to quite diluted conditions, since the densities of ionized particles and of atoms vanish exponentially fast when $\beta \to \infty$, with a rate determined by the ground-state energy of the hydrogen atom $E_H \simeq -13.6$ eV. The low-temperature condition, namely $kT \ll |E_H|$, ensures that atoms can form, while they maintain their individuality thanks to $a \gg a_B$, where $a = (3/(4\pi\rho))^{1/3}$ is the mean interparticle distance and $a_B = \hbar^2/(2me^2)$ is the Bohr radius. Because of the high dilution, the system is both weakly coupled and weakly degenerate. In particular, the ionized charges are almost classical, and the corresponding screening length reduces to its Debye expression $\kappa^{-1} = [4\pi\beta e^2(\rho_p^{\text{id}} + \rho_e^{\text{id}})]^{-1/2}$.

According to the above rigorous derivation, corrections to Saha theory decay exponentially fast in the scaled limit $T \to 0$ at fixed γ. Nevertheless, they cannot be explicitly computed within the corresponding mathematical techniques, so one has to use different tools as described further.

B. About the interplay between recombination and screening

In the Saha regime, in addition to atoms, recombination processes lead to the formation of molecules H_2, ions H^- and $H_2{}^+$, and also more complex entities like $H_2{}^-$, $H_3{}^+$, H_3, and so on. A controlled analysis of the corresponding contributions is the central problem for deriving systematic corrections to Saha theory. The well-known difficulty lies in a suitable account of the individual contribution of a single chemical species at finite temperature, which is free from the divergences arising from Rydberg states. In the literature, in general, that problem has been tackled within phenomenological prescriptions, in particular those leading to the so-called Brillouin-Planck-Larkin formula for atomic contributions [39,40]. The physical idea underlying that phenomenological approach is that the divergent contributions of Rydberg states are in fact screened by the free charges present in the system. Accordingly, the estimation of contributions from recombined entities cannot be disentangled from that of screened interactions between ionized charges.

A systematic procedure for dealing simultaneously with recombination and screening has been constructed through the combination of path integral and diagrammatical methods [21]. This provided some kind of cluster representation for equilibrium quantities in the grand-canonical ensemble. In the quite diluted Saha regime, the statistical weight of a given cluster made with N_p protons and N_e electrons involves the cluster partition function

$$Z(N_p,N_e) = 2\pi\lambda^2_{N_p,N_e} \lim_{\Lambda\to\infty} \frac{1}{\Lambda}\mathrm{Tr}[\exp(-\beta H_{N_p,N_e})]^T_{\mathrm{Mayer}}, \tag{11}$$

which is a truncated trace of Gibbs operator $\exp(-\beta H_{N_p,N_e})$ built with bare Coulomb Hamiltonians, while $\lambda_{N_p,N_e} = (\beta\hbar^2/[N_p m_p + N_e m_e])^{1/2}$ is the thermal de Broglie wavelength of the cluster. The trace in Eq. (11) converges despite the long range of the Coulomb interaction, thanks to a systematic truncation procedure which accounts for the screening by ionized charges [21]. Roughly speaking, that procedure amounts to subtract and add counterterms to the genuine Gibbs operators, which involve nontraceable operators built with the Coulomb potential. The Gibbs operator and the subtracted counterterms give rise to the finite partition function (11) which depends only on T and no longer on γ. The added counterterms are recombined together with other divergent contributions *via* chain resummations, which ultimately provide finite contributions involving the screening length κ^{-1} associated with the ionized charges present in the medium.

Remarkably, the familiar chemical species naturally emerge from cluster partition function $Z(N_p,N_e)$, which is intrinsic to the considered cluster in the vacuum. A given chemical species made with N_p protons and N_e electrons is associated with bound states of bare Hamiltonian H_{N_p,N_e}. In the zero-temperature limit, it provides the leading contribution to $Z(N_p,N_e)$ which behaves as

$$\exp\left(-\beta E^{(0)}_{N_p,N_e}\right) \tag{12}$$

apart from possible integer degeneracy factors and where $E^{(0)}_{N_p,N_e}$ is the groundstate energy of Hamiltonian H_{N_p,N_e}. At finite temperatures, $Z(N_p,N_e)$ involves not only contributions from thermally excited bound states but also contributions from diffusive states describing the dissociation of the considered chemical species.

Cluster partition functions $Z(N_p,N_e)$ can be viewed as generalizations of Ebeling virial coefficients [9] introduced for dealing with contributions from two-particle clusters. The contributions of interactions between chemical species can be expressed also in terms of cluster functions similar to $Z(N_p,N_e)$, so all contributions related to the formation of complex entities are properly taken into account. We stress that, as far as thermodynamical properties are concerned, only the full contribution of $Z(N_p,N_e)$ and of its related screened counterterms makes an unambiguous sense. The considered formalism [21] avoids both arbitrary and uncontrolled definitions of internal partition functions for chemical species, which are key ingredients in phenomenological chemical approaches.[1]

C. Systematic corrections to Saha theory

Within the combination of path integral and diagrammatical methods evoked above [21], systematic corrections to Saha theory have been explicitly computed in Ref. [6]. A pedagogical summary of both rather long papers is given in Ref. [41]. Here, leaving aside the tedious technical details involved in the derivation, we can guess and explain the mathematical structure of the corresponding expansion through simple arguments based on the considerations exposed just earlier.

In the so-called screened cluster representation of particle density ρ [21], any contribution reduces to a graph made with particle clusters connected by screened bonds. The statistical weight of a cluster made with N_p protons and N_e electrons, reduces to, roughly speaking, the product $\exp[\beta(N_p+N_e)\mu]$ of the particle fugacity factors times the cluster partition function $Z(N_p,N_e)$ and times some dressing factor which accounts for collective polarization effects. The integration over the relative distances between particle clusters generate powers of the Debye screening wave number κ, while dressing factors can be also expanded in powers of κ. Since $\exp(\beta\mu)$ is proportional to both γ and $\exp(-\beta|E_\mathrm{H}|)$, while κ is proportional to both $\gamma^{1/2}$ and $\exp(-\beta|E_\mathrm{H}|/2)$, any contribution reduces to some integer or half-integer power of γ times a temperature-dependent function. The low-temperature behavior of that function results from the competition between factors varying exponentially fast, namely positive powers of $\exp(-\beta|E_\mathrm{H}|)$ arising from fugacity factors, positive or negative powers of $\exp(-\beta|E_\mathrm{H}|/2)$ arising from screened interactions and polarization effects, and exploding Boltzmann factors $\exp(-\beta E^{(0)}_{N_p,N_e})$ arising from the contribution of bound entities with ground-state energy

---

[1] Of course, one might express the trace defining $Z(N_p,N_e)$ over the complete basis made with the eigenstates of H_{N_p,N_e}. This would provide a convergent infinite sum of bound-state contributions, which might be identified as an internal partition function. Nevertheless, extending the analysis carried out in Ref. [42], one can rewrite $Z(N_p,N_e)$ as a finite contribution plus another truncated partition function, which would ultimately provide a different internal part. For instance, various atomic partition functions can be extracted from Ebeling virial coefficient [42], including the famous Planck-Larkin-Brillouin expression.

$E^{(0)}_{N_p,N_e} < 0$ in cluster partition functions. In the Saha regime, namely $T \to 0$ with γ fixed, the order of a given contribution is determined by subtle inequalities involving $E^{(0)}_{N_p,N_e}$ and E_{H}. The leading contributions to density ρ are easily identified as arising from graphs made with single clusters carrying either one particle (one proton or one electron) or two particles (one proton and one electron), and they are of order $\exp(-\beta|E_{\mathrm{H}}|)$. All the other contributions decay exponentially faster, in agreement with the rigorous estimation of order $O(e^{-c\beta})$ for the full deviation to Saha theory. Accordingly, the corresponding SLT expansion for the dimensionless density ρ/ρ^* takes the following mathematical form Ref. [6]:

$$\rho/\rho^* = \gamma + \frac{\gamma^2}{2} + \sum_{k=1}^{\infty} \gamma^{n_k} h_k(\beta). \tag{13}$$

In Eq. (13), the leading first two terms are the ideal contributions predicted by Saha theory, while the sum accounts for the corrections. In each correction with order k, power n_k is integer or half-integer while γ^{n_k} may be multiplied by logarithmic terms. Furthermore, function $h_k(\beta)$ decays exponentially fast in the zero-temperature limit, $h_k(\beta) \sim \exp(-\beta\delta_k)$ when $\beta \to \infty$, except for possible multiplicative powers of β. We stress that expansion (13) is not ordered with respect to powers of γ, i.e., the n_k's do not necessarily increase with k, but it is ordered with respect to increasing decay rates, $0 < \delta_1 < \delta_2 < \ldots$, of functions $h_k(\beta)$. In other words, instead of γ, which is kept fixed here, the small parameter is built with the temperature which is sent to zero. That small parameter may be identified with $\exp(-\beta|E_{\mathrm{H}}|)$, so the leading low-temperature behavior of each correction of order k reduces to some positive real power $\delta_k/|E_{\mathrm{H}}|$ of that parameter. Notice that each function $h_k(\beta)$ does not reduce to its leading low-temperature form in general, but it also involves contributions which decay exponentially faster than $\exp(-\beta\delta_k)$.

The SLT expansion (13) provides an exact relationship between the density and the chemical potential [recall definition (6) of γ] that is very useful in the Saha regime because the series converges then rapidly and can be safely truncated. We comment on the mathematical form and the physical content of the first four corrections $h_k(\beta)$ in the next subsection. It will then be shown in Sec. III how Eq. (13) can be used to compute explicitly in the Saha regime any thermodynamical quantity as a function of the natural physical variables ρ and T.

D. First corrections and their physical content

The first four functions $h_k(\beta)$ as well as the corresponding n_k's, $k = 1,2,3,4$, are explicitly computed in Ref. [6], where it is also shown that all other h_k's with $k \geqslant 5$ decay faster than $\exp(\beta E_{\mathrm{H}})$, i.e., $\delta_k > |E_{\mathrm{H}}|$ for $k \geqslant 5$. Thus, if we truncate expansion (13) up to order $\exp(\beta E_{\mathrm{H}})$ included, it is consistent to only retain contributions which are at most of that order in the first four h_k's. In the following, we recall the corresponding expressions and we discuss their physical content.

1. *Term* $k = 1$*: plasma polarization around ionized charges*

That correction arises from a single cluster with one proton (electron) where dressing many-body effects on its statistical weight are computed at leading order. This provides the fugacity factor $\exp(\beta\mu)$ multiplied by the Debye screening factor κ. Accordingly, we find $n_1 = 1 + 1/2 = 3/2$ and $\delta_1 = -(E_{\mathrm{H}} + E_{\mathrm{H}}/2 - E_{\mathrm{H}}) = |E_{\mathrm{H}}|/2$, once ρ has been expressed in units of temperature-dependent density $\rho^* \sim \exp(-\beta|E_{\mathrm{H}}|)$. The precise form of function $h_1(\beta)$ reads

$$h_1(\beta) = \frac{(\beta|E_{\mathrm{H}}|)^{3/4}}{\pi^{1/4}} \exp(\beta E_{\mathrm{H}}/2), \tag{14}$$

since $Z(1,0) = Z(0,1) = 2$ for a single proton or a single electron, for which no truncation occurs, namely $[\exp(-\beta H_{1,0})]^T_{\mathrm{Mayer}} = \exp(-\beta H_{1,0})$ and $[\exp(-\beta H_{0,1})]^T_{\mathrm{Mayer}} = \exp(-\beta H_{0,1})$.

The present correction accounts for the familiar polarization of the plasma surrounding an ionized charge. In the literature, that mechanism was taken into account for the first time in Ref. [43] through a suitable modification of Saha ionization equilibrium (1). We have checked that the corresponding correction to Saha theory can be exactly recovered by keeping only the first term $k = 1$ in the SLT expansion (13).

2. *Term* $k = 2$*: formation of molecules and atom-atom interactions*

That correction arises from a single cluster made with two protons and two electrons and from two interacting neutral clusters where each of them is made with one proton and one electron. At leading order, dressing collective effects in statistical weights can be neglected, while screening of interactions between neutral clusters can be also omitted since the corresponding bare interactions are integrable. Then, power n_2 is merely determined by the product of four fugacity factors $\exp(\beta\mu)$, which provides $n_2 = 4$. Function $h_2(\beta)$ reduces to

$$h_2(\beta) = \frac{1}{64}\left(\frac{2m}{M}\right)^{3/2} Z(2,2)\exp(3\beta E_{\mathrm{H}}) + W(1,1|1,1)\exp(3\beta E_{\mathrm{H}}), \tag{15}$$

where $W(1,1|1,1)$ is a suitable trace analogous to expression (11) which now involves two Gibbs operators $\exp(-\beta H_{1,1})$ associated with two proton-electron pairs, as well as their bare Coulomb interactions. Factors $\exp(3\beta E_{\mathrm{H}})$ arise from the product of the four fugacity factors $\exp(\beta\mu)$ and the rewriting of ρ in units of ρ^*. In the zero-temperature limit, the leading contribution in expression (15) is that of the molecular ground state in $Z(2,2)$ with energy $E_{\mathrm{H}_2} = E^{(0)}_{2,2}$. Notice that, because of inequality $3E_{\mathrm{H}} < E_{\mathrm{H}_2} < 2E_{\mathrm{H}}$, function $h_2(\beta)$ indeed decays exponentially fast with the rate $\delta_2 = |3E_{\mathrm{H}} - E_{\mathrm{H}_2}| \simeq 9.1$ eV. That inequality ensures that molecules H_2 are very scarce in the Saha regime compared to atoms H, despite that they are more stable energetically, Moreover, δ_2 is indeed larger than $\delta_1 = |E_{\mathrm{H}}|/2 \simeq 6.8$ eV.

Molecular contributions are embedded in $Z(2,2)$, which is indeed finite thanks to the truncation procedure inherited from screening, as well as the short-range part of atom-atom interactions. Long-range atom-atom interactions, including familiar van der Waals interactions, appear in $W(1,1|1,1)$. Notice that the screened counterterms related to the truncations involved here provide contributions to expansion (13) which decay faster than $\exp(\beta E_{\mathrm{H}})$, and they arise in terms with

$k \geqslant 5$. We stress that both molecular formation and atom-atom interactions are properly taken into account, without any *a priori* modelizations like in usual chemical approaches. Here the corresponding contributions are expressed in terms of the associated few-body Coulomb Hamiltonians, and they naturally emerge through the fundamental quantum mechanisms at work. In particular, the quantum mechanical operators involved in both $Z(2,2)$ and $W(1,1|1,1)$ automatically and correctly take care of the unavoidable mixing between the contributions from two interacting atoms, on the one hand, and from a single molecule, on the other hand.

3. Term k = 3: atomic excitations and charge-charge interactions

That correction arises from single clusters made with either one or two particles. Contributions of two-particle clusters are controlled by the product of two fugacity factors $\exp(\beta\mu)$ and of two-body cluster partition functions $Z(2,0)$, $Z(0,2)$, and $Z(1,1)$, while contributions from the one-particle clusters reduce to one fugacity factor $\exp(\beta\mu)$ multiplied by a factor κ^2 which accounts for polarization effects beyond the Debye mean-field result of order κ. This leads to $n_3 = 1 + 1 = 1 + 2 \times 1/2 = 2$ and

$$
\begin{aligned}
h_3(\beta) = {} & -\frac{1}{2} + \left[1 + \frac{1}{12}\ln\left(\frac{4m}{M}\right)\right]\frac{(\beta|E_{\mathrm{H}}|)^{3/2}}{\pi^{1/2}}\exp(\beta E_{\mathrm{H}}) \\
& + \frac{1}{8\pi^{1/2}}\Bigg\{2Q(x_{pe}) + \left(\frac{2m}{m_p}\right)^{3/2} \\
& \times \left[Q(-x_{pp}) - \frac{1}{2}E(-x_{pp})\right] + \left(\frac{2m}{m_e}\right)^{3/2}[Q(-x_{ee})] \\
& - \frac{1}{2}E(-x_{ee})]\Bigg\}\exp(\beta E_{\mathrm{H}}),
\end{aligned} \tag{16}
$$

where the two-particle partition functions, $Z(2,0)$, $Z(0,2)$ and $Z(1,1)$ have been rewritten in terms of Ebeling's functions $Q(x)$ and $E(x)$ [9] with $x_{pe} = 2(\beta|E_{\mathrm{H}}|)^{1/2}$, $x_{pp} = (2m_p/m)^{1/2}(\beta|E_{\mathrm{H}}|)^{1/2}$, and $x_{ee} = (2m_e/m)^{1/2}(\beta|E_{\mathrm{H}}|)^{1/2}$. The ground-state contribution has been extracted from $Z(1,1)$ and it provides the leading atomic contribution $\gamma^2/2$ in Eq. (13). Consequently, the leading low-temperature behavior of $h_3(\beta)$ arises from the contribution to $Q(x_{pe})$ of the first excited state of an atom H with energy $E_{\mathrm{H}}/4$, so δ_3 reduces to $\delta_3 = -E_{\mathrm{H}}/4 + E_{\mathrm{H}} = -3E_{\mathrm{H}}/4$. That decay rate $\delta_3 \simeq 10.2$ eV is indeed larger than $\delta_2 \simeq 9.1$ eV.

The present correction involves contributions of the atomic excited states, as well as of interactions between two ionized charges. The screened long-range part of such interactions are precisely the counterterms related to the truncations ensuring the finiteness of $Z(2,0)$, $Z(0,2)$, and $Z(1,1)$ or, equivalently, of the Ebeling function Q.

4. Term k = 4: formation of ions and atom-charge interactions

That correction arises from single three-particle clusters, a two-particle cluster interacting with a one-particle cluster, and a single two-particle cluster dressed by many-body effects. All contributions provide the same power $n_4 = 3$ of γ, as resulting from either the product of three fugacity factors $\exp(\beta\mu)$ or the product of two fugacity factors $\exp(\beta\mu)$ times a factor κ^2 arising from polarization effects, namely $n_4 = 1 + 1 + 1 = 1 + 1 + 2 \times 1/2 = 3$. The corresponding function $h_4(\beta)$ reads

$$
\begin{aligned}
h_4(\beta) = {} & \frac{3}{64}\Bigg\{\left[\frac{m_e(M+m_p)}{M^2}\right]^{3/2} Z(2,1) + \left[\frac{m_p(M+m_e)}{M^2}\right]^{3/2} \\
& \times Z(1,2)\Bigg\}\exp(2\beta E_{\mathrm{H}}) + S_3(1,1)\exp(2\beta E_{\mathrm{H}}) \\
& + \frac{3}{2}[W(1,1|1,0) + W(1,1|0,1)]\exp(2\beta E_{\mathrm{H}}).
\end{aligned} \tag{17}
$$

The leading low-temperature behavior of $h_4(\beta)$ arises from the ground-state contribution of ion $\mathrm{H_2}^+$ in $Z(2,1)$, so $\delta_4 = E_{\mathrm{H_2}^+} - 2E_{\mathrm{H}} \simeq 11.0$ eV, which is indeed larger than $\delta_3 \simeq 10.2$ eV.

Several phenomena contribute to the present correction. First, formation of ions $\mathrm{H_2}^+$ and H^- are embedded in partition functions $Z(2,1)$ and $Z(1,2)$, respectively. Second, contributions of bare interactions between an atom H and a single ionized charge are described by the functions $W(1,1|1,0)$ and $W(1,1|0,1)$. Third, the function $S_3(1,1)$ accounts for modifications of the atomic ground state due to the polarization of the surrounding plasma, beyond the familiar Debye shift.

III. THERMODYNAMICAL FUNCTIONS

A. Chemical potential as a function of density

In physical systems, the natural thermodynamical parameters are the temperature and the density. Thus, it is quite useful to invert the SLT expansion (13), namely to determine $\gamma(\rho,T)$. Then, by using standard thermodynamical identities, we are able to compute consistently all thermodynamical quantities as functions of T and ρ. In the present low-temperature limit, the inversion can be performed in a perturbative way as follows. First, if we neglect all the exponentially small corrections embedded in the h_k's, the density reduces to its Saha expression,

$$
\rho/\rho^* = \gamma + \frac{\gamma^2}{2}. \tag{18}
$$

The inversion of that relation, which amounts here to solving a simple second-order equation for γ, gives

$$
\gamma_{\mathrm{Saha}}(\rho,T) = \gamma_{\mathrm{S}}(\xi) = \sqrt{1+2\xi} - 1 \quad \text{with} \quad \xi = \rho/\rho^*, \tag{19}
$$

which is the leading form of $\gamma(\rho,T)$ in the Saha regime. The inversion of the full relation (13) is then achieved by writing $\gamma(\rho,T) = \gamma_{\mathrm{Saha}}(\rho,T)$ plus a small correction which is treated perturbatively. This leads to

$$
\gamma(\rho,T) = \sqrt{1+2\rho/\rho^*} - 1 + \sum_{k=1}^{\infty} a_k(\rho/\rho^*)g_k(\beta), \tag{20}
$$

where a_k depends only on $\xi = \rho/\rho^*$, while g_k depends only on temperature. The a_k's can be determined in terms of $\gamma_{\mathrm{S}}(\xi)$ and of its derivatives with respect to ξ. Each g_k reduces to a polynomial in the h_l's with $1 \leqslant l \leqslant k$. Thus, the g_k's decay exponentially fast when $\beta \to \infty$, and they are ordered with respect to increasing decay rates. The first five terms in Eq. (20) are

$$a_1 = -\frac{\gamma_S^{3/2}}{1+\gamma_S}; \quad a_2 = -\frac{\gamma_S^4}{1+\gamma_S}; \quad a_3 = -\frac{\gamma_S^2}{1+\gamma_S}; \quad a_4 = -\frac{\gamma_S^3}{1+\gamma_S}; \quad a_5 = \frac{\gamma_S^2(2\gamma_S+3)}{2(1+\gamma_S)^3} \tag{21}$$

$$g_1 = h_1; \quad g_2 = h_2; \quad g_3 = h_3; \quad g_4 = h_4; \quad g_5 = h_1^2, \tag{22}$$

where γ_S is given by the simple algebraic function (19) of ρ/ρ^*. All g_k's with $k \geqslant 6$ decay exponentially faster than $\exp(\beta E_H)$.

Similarly to SLT expansion (13), the small parameter in series (20) is the temperature or, equivalently, the exponentially small factor $\exp(-\beta|E_H|)$. Now, the fixed parameter is the ratio $\xi = \rho/\rho^*$ which can take arbitrary values. The corresponding expansion of chemical potential μ straightforwardly follows by inserting series (20) into relation (6). Other thermodynamical functions can be expanded in a similar way by using thermodynamical identities, as explained further for pressure and internal energy.

B. Pressure

We start from the standard relation in the grand-canonical ensemble which expresses the density as the derivative of the pressure with respect to the fugacity $z = \exp(\beta\mu)$. According to the variable change defined through relation (6), that identity can be rewritten as

$$\rho = \frac{\gamma}{2}\frac{\partial \beta P}{\partial \gamma}, \tag{23}$$

where the partial derivative is taken at fixed β. Replacing ρ by its SLT expansion (13) into the right-hand side of identity (23), we easily obtain

$$\beta P/\rho^* = 2\gamma + \frac{\gamma^2}{2} + \sum_{k=1}^{\infty} \frac{2\gamma^{n_k}}{n_k} h_k(\beta), \tag{24}$$

where we have that ρ^* depends only on β, while $\beta P/\rho^*$ vanishes for infinite dilution, namely for $\gamma = 0$. Replacing each factor γ by the SLT inverted series (20), we recast expression (24) as the SLT expansion of the pressure at fixed ratio ρ/ρ^*, namely

$$\beta P/\rho^* = \beta P_{\text{Saha}}/\rho^* + \sum_{k=1}^{\infty} \beta P_k/\rho^*. \tag{25}$$

The leading term is nothing but the well-known Saha pressure (3) in units of ρ^*. The general structure of the k^{th} correction reads

$$\beta P_k/\rho^* = b_k(\rho/\rho^*)\alpha_k(\beta), \tag{26}$$

where α_k is a polynomial in the $h_l(\beta)$'s with $l \leqslant k$. Therefore, for a fixed ratio ρ/ρ^*, corrections $\beta P_k/\rho^*$ decay exponentially fast when $\beta \to \infty$, while the corresponding decay rates increase with k. The functions $[b_k(\rho/\rho^*),\alpha_k(\beta)]$ involved in the first five corrections read

$$b_1 = \frac{\gamma_S^{3/2}(\gamma_S-2)}{3(1+\gamma_S)}; \quad b_2 = -\frac{\gamma_S^4(\gamma_S+3)}{2(1+\gamma_S)}; \quad b_3 = -\frac{\gamma_S^2}{1+\gamma_S}; \quad b_4 = -\frac{\gamma_S^3(\gamma_S+4)}{3(1+\gamma_S)}; \quad b_5 = \frac{\gamma_S^2(2-\gamma_S^2)}{2(1+\gamma_S)^3} \tag{27}$$

$$\alpha_1 = h_1; \quad \alpha_2 = h_2; \quad \alpha_3 = h_3; \quad \alpha_4 = h_4; \quad \alpha_5 = h_1^2, \tag{28}$$

while next correction $\beta P_6/\rho^*$ decays faster than $\exp(\beta E_H)$.

C. Internal energy

In the grand-canonical ensemble and for a finite volume Λ, we set $\langle N\rangle = (\langle N_p\rangle + \langle N_e\rangle)/2$ for the average common number of protons and electrons. We then define, in the thermodynamic limit, the internal energy per particles pair $u = \lim_{TL} U/\langle N\rangle$. Standard thermodynamical identities provide the relation

$$u = \frac{\partial}{\partial\beta}\left(2\beta\mu - \frac{\beta P}{\rho}\right), \tag{29}$$

where the partial derivative with respect to β is taken at fixed density ρ. Inserting into identity (29) the expression (6) of the chemical potential in terms of γ and β, we find

$$u = \left(2 - \frac{\beta P}{\rho}\right)E_H + \frac{3P}{2\rho} + 2\frac{\partial}{\partial\beta}\ln\gamma - \frac{\rho^*}{\rho}\frac{\partial}{\partial\beta}\frac{\beta P}{\rho^*}, \tag{30}$$

where we have also used

$$\frac{\partial}{\partial\beta}\rho^* = (E_H - 3kT/2)\rho^*, \tag{31}$$

inferred from definition (2) of ρ^*. The insertion of expansions (20) and (25) of γ and $\beta P/\rho^*$ into relation (30) provide the corresponding SLT expansion of u. The partial derivatives with respect to β give rise to functions $h_k'(\beta) = dh_k/d\beta$, which decay exponentially fast at low temperatures with the same decay rates as the $h_k(\beta)$'s. Moreover, since ρ^* depends on β, coefficients $a_k(\rho/\rho^*)$ and $b_k(\rho/\rho^*)$ also provide contributions to the partial derivatives with respect to β at fixed ρ. After straightforward algebraic calculations, we eventually obtain

$$u = u_{\text{Saha}} + \sum_{k=1}^{\infty} u_k \tag{32}$$

with

$$u_{\text{Saha}} = (1+\gamma_S\xi^{-1})3kT/2 + (1-\gamma_S\xi^{-1})E_H, \tag{33}$$

and

$$u_1 = -\gamma_S^{1/2}(1+\gamma_S)^{-1}[(1+\gamma_S\xi^{-1})kT + 2(1-\gamma_S\xi^{-1})E_H] \times h_1(\beta), \tag{34}$$

$$u_2 = -\gamma_S^4\xi^{-1}h_2'(\beta)/2 + \gamma_S^3(1+\gamma_S)^{-1}(1+\gamma_S\xi^{-1}/2) \times \left(E_H - \tfrac{3}{2}kT\right)h_2(\beta), \tag{35}$$

$$u_3 = -\gamma_S^2\xi^{-1}h_3'(\beta) + \gamma_S^2\xi^{-1}(1+\gamma_S)^{-1}\left(E_H - \tfrac{3}{2}kT\right)h_3(\beta),$$

$$u_4 = -2\gamma_S^3\xi^{-1}h_4'(\beta)/3 + 2(1+\gamma_S\xi^{-1})\gamma_S^2(1+\gamma_S)^{-1} \times \left(E_H - \tfrac{3}{2}kT\right)h_4(\beta)/3, \tag{37}$$

$$u_5 = (1+\gamma_S)^{-3}\left(3\xi\, kT + \gamma_S^2(3\gamma_S+4)(2+\gamma_S)^{-1}\, E_H\right) \times [h_1(\beta)]^2. \tag{38}$$

Expression (33) of Saha internal energy can be easily interpreted as follows. Let $x_p^{\text{id}} = \rho_p^{\text{id}}/\rho$, $x_e^{\text{id}} = \rho_e^{\text{id}}/\rho$, and $x_{\text{at}}^{\text{id}} = \rho_{\text{at}}^{\text{id}}/\rho$ be the respective molar fractions of ionized protons, ionized electrons, and hydrogen atoms. We can rewrite formula (33) as

$$u_{\text{Saha}} = \left(x_p^{\text{id}} + x_e^{\text{id}}\right) 3kT/2 + x_{\text{at}}^{\text{id}}(E_{\text{H}} + 3kT/2), \tag{39}$$

so each ionized charge does provide a classical kinetic contribution $3kT/2$, while each atom does provide the kinetic contribution $3kT/2$ associated with motion of its mass center plus the ground-state energy E_{H}, as it should. Notice that the low-temperature behavior of corrections (35) and (37) can be similarly interpreted in terms of the contributions of molecules H_2 and ions H^-, ${H_2}^+$ respectively.

Other equilibrium quantities can also be derived from previous expansions *via* thermodynamical identities, like the specific heats, or the isentropic compressibility which determines the sound speed. Such derivations are, of course, consistent within the general framework of thermodynamics: A given quantity can be computed following different routes which all lead to the same expression.

IV. REPRESENTATIONS OF TEMPERATURE-DEPENDENT FUNCTIONS

For practical applications of SLT expansions, we need simple representations of functions $h_k(\beta)$ at finite temperatures. Though asymptotic low-temperature behaviors of such functions are exactly known for $k = 1,2,3,4$, explicit analytic expressions at finite temperature are available only for $h_1(\beta)$ and $h_3(\beta)$, thanks to our exact knowledge of the whole spectrum of two-body Coulomb Hamiltonians (see Sec. IV A). For functions $h_2(\beta)$ and $h_4(\beta)$, we construct simple approximations which are expected to be sufficiently accurate for temperatures up to 30 000 K (see Sec. IV B). Eventually, we provide the corresponding numerical tables and plots for all those functions.

A. Analytical expressions for one- and two-body functions

Function $h_1(\beta)$ is given by the simple formula (14) which reduces to an elementary function of dimensionless parameter βE_{H}. Function $h_3(\beta)$ is given by expression (16) in terms of Ebeling virial functions $Q(x)$ and $E(x)$, which have been widely studied in the literature. In particular, entire series expansions in powers of x have been derived [44,45],

$$\begin{aligned} Q(x) &= -\frac{1}{6}x - \frac{\sqrt{\pi}}{8}x^2 - \frac{1}{6}\left(\frac{C}{2} + \ln 3 - \frac{1}{2}\right)x^3 + \sum_{n=4}^{\infty} q_n x^n, \\ q_n &= \sqrt{\pi}\,\frac{\zeta(n-2)}{2^n \Gamma\left(\frac{n}{2}+1\right)}, \end{aligned} \tag{40}$$

where $C = 0.57721\ldots$ is the Euler-Mascheroni constant and $\zeta(s)$ is Riemann's function, while

$$\begin{aligned} E(x) &= \frac{1}{2} + \frac{\sqrt{\pi}\ln 2}{4}x^2 + \frac{\pi^2}{72}x^3 + \sum_{n=4}^{\infty} e_n x^n , \\ e_n &= \sqrt{\pi}(1 - 2^{2-n})\frac{\zeta(n-1)}{2^n \Gamma\left(\frac{n}{2}+1\right)} \end{aligned} \tag{41}$$

with x replaced by $x_{pe} = 2(\beta|E_{\text{H}}|)^{1/2}$, $-x_{pp} = -(2m_p/m)^{1/2}(\beta|E_{\text{H}}|)^{1/2}$, or $-x_{ee} = -(2m_e/m)^{1/2}(\beta|E_{\text{H}}|)^{1/2}$. Since x is proportional to $1/\sqrt{T}$, such series can be viewed as high-temperature expansions. A controversy has arisen recently on Eq. (40), with Kraeft [46] and Kremp, Schlanges, and Kraeft [13] maintaining that no linear term $-x/6$ should be present. Starting from the definition of function $Q(x)$ see, for instance, Eq. (7.1) in Ref. [23], one can calculate quite easily its high-temperature behavior by using the Feynman-Kac representation [47] of the density matrix

$$\langle \boldsymbol{r}|e^{-\beta h}|\boldsymbol{r}\rangle = \frac{1}{(2\pi\lambda^2)^{3/2}}\int D(\boldsymbol{\xi})\, e^{\beta e^2 \int_0^1 ds\, v(|\boldsymbol{r}+\lambda\boldsymbol{\xi}(s)|)}, \tag{42}$$

where h is the Hamiltonian of a particle of mass m in the attractive Coulomb potential $-e^2/r$ and $\boldsymbol{\xi}(s)$ is a Brownian bridge distributed according to the Wiener measure $D_W(\boldsymbol{\xi})$. At high temperatures, the exponential can be linearized, and a straightforward calculation confirms that the leading behavior is indeed given by Ebeling's result, $-x/6$. Asymptotic large-x, i.e., low-temperature, expansions read

$$\begin{aligned} Q(x) = {} & 2\sqrt{\pi}\left[\sum_{n=1}^{\infty} n^2 \left(e^{x^2/(4n^2)} - 1 - \frac{x^2}{4n^2}\right) - \frac{x^2}{8}\right] \\ & - \frac{x^3}{6}\left(\ln x + 2C + \ln 3 - \frac{11}{6}\right) - \frac{x}{12} \\ & - \frac{1}{60x} + O\left(\frac{1}{x^3}\right) \end{aligned} \tag{43}$$

for $x > 0$ (attractive case) and

$$\begin{aligned} Q(x) = {} & -\frac{x^3}{6}\left(\ln|x| + 2C + \ln 3 - \frac{11}{6}\right) - \frac{x}{12} \\ & - \frac{1}{60x} + O\left(\frac{1}{x^3}\right) \end{aligned} \tag{44}$$

for $x < 0$ (repulsive case). When $x \to -\infty$, a semiclassical calculation [48] shows that exchange function $E(x)$ decays exponentially fast as

$$E(x) \simeq \frac{4}{\sqrt{3\pi}}|x| \exp\left[-\frac{3}{2}\left(\pi^2 \frac{x^2}{2}\right)^{1/3}\right]. \tag{45}$$

As far as numerical calculations are concerned, high-temperature series (40) and (41) are quite useful because their radius of convergence is infinite. When x becomes very large, calculations using large-x expansions (43)–(45) are, of course, faster. Those low-temperature expansions can be used in fact to compute $h_3(T)$ from formula (16), up to the Rydberg temperature $|E_{\text{H}}|/k \simeq 157\,800$ K since $|x_{\text{ab}}|$, for ab $= ep$, pp, ee, remains larger than 1. Figure 1 shows a plot of functions $Q(x)$ and $E(x)$.

Notice that the first sum in the right-hand side of expression (43) is nothing but the so-called Brillouin-Planck-Larkin partition function

$$Z_{\text{BPL}}(T) = \sum_{n=1}^{\infty} n^2 \left(e^{-\beta E_{\text{H}}/n^2} - 1 + \frac{\beta E_{\text{H}}}{n^2}\right), \tag{46}$$

which was introduced in the 1930s on the basis of heuristic arguments as discussed in Refs. [49,50]. It turns out that $Q(x_{pe}) = Q(T)$ may be approximated by $Z_{\text{BPL}}(T)$ if the temperature is not too high [49,51]. Notice that the leading asymptotic behavior of $Q(T)$ and $Z_{\text{BPL}}(T)$ are identical as

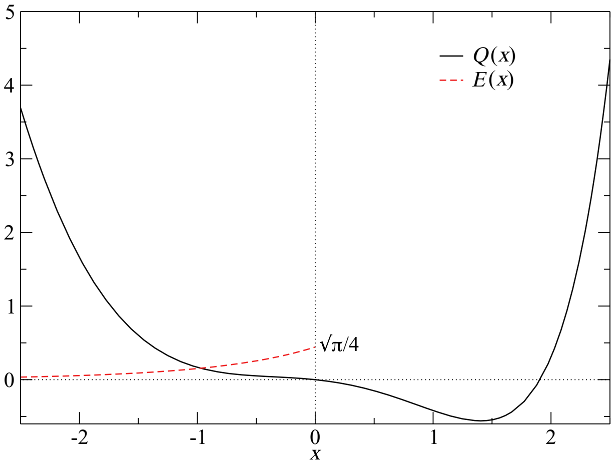


FIG. 1. (Color online) Plot of Ebeling direct and exchange functions $Q(x)$ (solid line) and $E(x)$ (dashed line).

$T \to 0$, whereas one has $Q(T) \sim 1/\sqrt{T}$ versus $Z_{\rm BPL}(T) \sim 1/T^2$ at high temperatures. The quality of that approximation, and of two even simpler approximations, is shown in Fig. 2. The relative error when approximating $Q(T)$ by $Z_{\rm BPL}(T)$ is less than 3% for $T \leqslant 25\,000$ K, while it reaches 30% at 50 000 K. Truncating the sum in $Z_{\rm BPL}(T)$ at $n = 1$ provides actually a better approximation that exhibits an accuracy of 8% at 50 000 K (see Fig. 2). The latter approximation is quite successful because contributions from diffusive states and from terms associated with screening effects in $Q(T)$, more or less cancel out the contributions from excited bound states at the considered temperatures. We stress that our further calculations do not use those approximations but rather the exact expressions for $Q(x)$ recalled earlier.

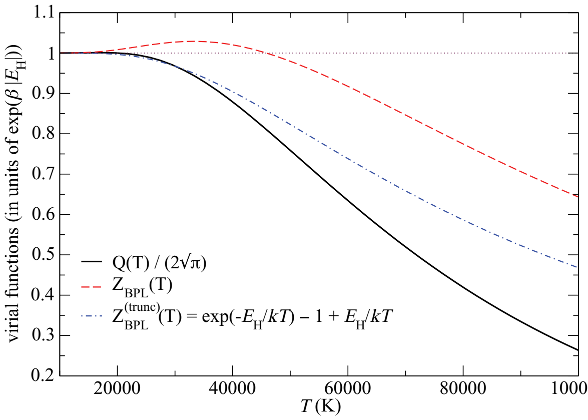


FIG. 2. (Color online) Comparison of the Brillouin-Planck-Larkin partition function (dashed line) to the exact virial function $Q(x_{pe}(T))$ (solid line) for temperatures up to 100 000 K. Truncating the sum in $Z_{\rm BPL}$ after the first term yields a quite good approximation (dot-dashed line). The dotted line (constant value 1) corresponds to keeping only the ground-state contribution $\exp(-E_{\rm H}/kT)$.

B. Simple approximations for three- and four-body functions

Because of our rather poor knowledge of the whole spectrum of three- and four-body Coulomb Hamiltonians, closed analytical expressions for $h_2(\beta)$ and $h_4(\beta)$ cannot be derived at the moment. So here we proceed to simple estimations of those functions, inspired by the above considerations on $Z(1,1)$, which should work reasonably well in the temperature range considered here.

1. Case of $h_2(\beta)$

The relative importance of the contributions of the various operators involved in $[\exp(-\beta H_{2,2})]^T_{\rm Mayer}$, the trace of which defines partition function $Z(2,2)$, can be readily estimated by using the ground-state energies of Hamiltonians $H_{2,2}$, $H_{2,1}$, $H_{1,2}$, $H_{1,1}$, $H_{1,0}$, and $H_{0,1}$. Up to 30 000 K, it is sufficient to retain only contributions from $\exp(-\beta H_{2,2})$ and from subtracted operators involving $\exp(-\beta H_{1,1})\exp(-\beta H_{1,1})$. All other combinations of Gibbs operators associated with products of molecular dissociation differing from two atoms, like $\exp(-\beta H_{1,0})\exp(-\beta H_{1,2})$ associated with $(p,{\rm H}^-)$ or $\exp(-\beta H_{0,1})\exp(-\beta H_{2,1})$ associated with $(e,{\rm H_2}^+)$, can be safely neglected. After adding to $Z(2,2)$ the contribution of $W(1,1|1,1)$ in the expression (15) of $h_2(\beta)$, we find that terms which involve imaginary-time evolutions of $V_{\rm at,at}$, $V^2_{\rm at,at}$, and $V^3_{\rm at,at}$ cancel out. This provides a simple estimation of $h_2(\beta)$ as arising entirely from operator $[\exp(-\beta H_{2,2}) - \exp(-\beta H_{1,1})\exp(-\beta H_{1,1})]$.

Similarly to the case of $Z(1,1)$, in the considered temperature range, the main contributions from $[\exp(-\beta H_{2,2}) - \exp(-\beta H_{1,1})\exp(-\beta H_{1,1})]$ can be reasonably expected to arise from the lowest-energy molecular bound states. In addition, we describe such states within the familiar picture where the electrons are in their ground state, while global rotations and vibrations of the molecule are taken into account within a rigid-rotator model and a harmonic oscillator, respectively. This leads to the approximation

$$h_2(\beta) \simeq \frac{\sqrt{2}m^{3/2}}{32M^{3/2}} Z_{\rm H_2} \exp(3\beta E_{\rm H}), \tag{47}$$

where the molecular partition function $Z_{\rm H_2}$ factorizes into [52]

$$Z_{\rm H_2} = \exp\left(-\beta E_{\rm H_2}\right) Z^{\rm (rot)}_{\rm H_2} Z^{\rm (vib)}_{\rm H_2} \tag{48}$$

with the vibrational part

$$Z^{\rm (vib)}_{\rm H_2} = \frac{1}{1 - \exp\left(-\beta\epsilon^{\rm (vib)}_{\rm H_2}\right)} \tag{49}$$

and the rotational part

$$Z^{\rm (rot)}_{\rm H_2} = \left[\sum_{l=0}^{\infty}(4l+1)\exp\left(-2l(2l+1)\beta\epsilon^{\rm (rot)}_{\rm H_2}\right) + 3\sum_{l=0}^{\infty}(4l+3)\exp\left(-(2l+1)(2l+2)\beta\epsilon^{\rm (rot)}_{\rm H_2}\right)\right]. \tag{50}$$

In formula (50), the first sum runs over rotational states of parahydrogen and the second sum over rotational states of orthohydrogen. The energy quanta $\epsilon^{\rm (vib)}_{\rm H_2} = kT^{\rm (vib)}_{\rm H_2}$ and $\epsilon^{\rm (rot)}_{\rm H_2} = kT^{\rm (rot)}_{\rm H_2}$ associated to proton vibrations and global rotations are

TABLE I. Spectroscopic data of some hydrogen bound states: ground-state energy $E^{(0)}$ (in atomic units) and rotational and vibrational temperatures (in K). The values include neither relativistic nor radiative corrections.

	H	H^-	H_2	${H_2}^+$	${H_2}^-$	${H_3}^+$
$E^{(0)}$	-0.5	-0.527733147	-1.164663172	-0.597139063	-1.048274	-1.323 ± 0.002
$T^{(\mathrm{rot})}$	/	/	85.26	41.87	50.42	(not used[a])
$T^{(\mathrm{vib})}$			5986.98	3150.78	2228.32	
Ref.		[b]	[c]	[d]	[e]	[f]

[a] For the rovibrational partition function of H_3^+, see L. Neale and J. Tennyson, *Astrophys. J.* **454**, L169 (1995).

[b] *Handbook of Atomic, Molecular, and Optical Physics*, edted by G. Drake (Springer, 2006).

[c] L. Wolniewicz, J. Chem. Phys. **99**, 1851 (1993).

[d] J. Ph. Karr and L. Hilico, J. Phys. B: At. Mol. Opt. Phys. **39**, 2095 (2006).

[e] Calculated by V. Robert using a coupled cluster [CCSD(T)] approach including an extended basis set for hydrogen atoms ($4s3p2d1f$).

[f] P. C. Cosby and H. Helm, Chem. Phys. Lett. **152**, 71 (1988).

listed in Table I. We recall that the rotational partition function reduces to the classical result

$$Z_{H_2}^{(\mathrm{rot})} \sim \frac{2T}{T_{H_2}^{(\mathrm{rot})}}, \tag{51}$$

when $T \gg T_{H_2}^{(\mathrm{rot})}$. Notice that more precise descriptions of the lowest-energy excited states of H_2, which do not neglect rotation-vibration coupling as the rigid rotor approximation (48), are available in the literature [53] but are not required for our purpose.

Approximation (47) can be viewed as a suitable extrapolation of the the exact low-temperature behavior of $h_2(\beta)$, which includes both leading and sub-leading terms. Up to 30 000 K, contributions of electronic excitations can be omitted because the corresponding energy gaps are of order 10 eV at least. Moreover, and similarly to the case of $Z(1,1)$, either diffusive states like those associated with the dissociation of the molecule into two atoms or substracted terms involved in the truncated operator $[\exp(-\beta H_{2,2})]_{\mathrm{Mayer}}^{T}$ like $\exp(-\beta H_{1,1})\exp(-\beta H_{1,1})$ provide contributions which can be safely neglected in that relatively low-temperature range. As detailed in Appendix, this has been checked within a simplified model which is often used for describing the energy levels of the molecule H_2.

2. Case of $h_4(\beta)$

Function $h_4(\beta)$, defined by formula (17), can be approximated within a construction similar to the above derivation for $h_2(\beta)$. This provides

$$\begin{aligned} h_4(\beta) \simeq {} & \frac{3m_e^{3/2}(M+m_p)^{3/2}}{64M^3} Z_{{H_2}^+} \exp(2\beta E_{\mathrm{H}}) \\ & + \frac{3m_p^{3/2}(M+m_e)^{3/2}}{64M^3} Z_{H^-} \exp(2\beta E_{\mathrm{H}}) \\ & + \frac{c_{\mathrm{at}}}{8\pi^{3/2}(\beta|E_{\mathrm{H}}|)^{1/2}} \exp(\beta E_{\mathrm{H}}), \end{aligned} \tag{52}$$

which is the analog of expression (47) for $h_2(\beta)$. Here, internal partition functions for ions ${H_2}^+$ and H^-, which arise from $Z(2,1)$ and $Z(1,2)$, read

$$Z_{{H_2}^+} = 2\exp(-\beta E_{{H_2}^+}) Z_{{H_2}^+}^{(\mathrm{rot})} Z_{{H_2}^+}^{(\mathrm{vib})} \tag{53}$$

and

$$Z_{H^-} = 2\exp(-\beta E_{H^-}), \tag{54}$$

where the required spectroscopic data are given in Table I, while the contribution of $S_3(1,1)\exp(2\beta E_{\mathrm{H}})$ has been replaced by its low-temperature form with constant $c_{\mathrm{at}} \simeq 10.065$ [6].

Similarly to formula (47), approximate expression (52) incorporates both leading and subleading contibutions to the low-temperature representation of $h_4(\beta)$. A similar accuracy for that approximation can be reasonable expected up to a few thousands kelvins, where ionic ground-state contributions dominate. However, when temperature is increased up to 30 000 K, approximation (52) becomes surely less accurate than its counterpart (47) for $h_2(\beta)$. Indeed, the binding energies of ions ${H_2}^+$ and H^- are of order 2.6 and 0.7 eV, respectively, so contributions of all terms which arise from diffusive states, truncations defining $Z(2,1)$ and $Z(1,2)$, or interactions involved in $W(1,1|1,0)$ and $W(1,1|0,1)$ can no longer be neglected for $T > 10\,000$ K. Nevertheless, the accuracy of approximation (52) should be sufficient for our purpose, because contributions to thermodynamics associated with $h_4(\beta)$ remain quite small in the considered density-temperature range.

C. Numerical values and plots

Using the exact representations for $h_1(\beta)$ and $h_3(\beta)$, as well as the approximate forms of $h_2(\beta)$ and $h_4(\beta)$, we can compute those functions easily and quickly. The corresponding plots, in logarithmic units, are shown in Fig. 3, while numerical values at some specific temperatures are given in Table II. Functions $h_2(\beta)$ and $h_4(\beta)$ are computed only up to 30 000 K, because their approximate expressions introduced in Sec. IV B are expected to become inaccurate at higher temperatures.

At low temperatures, all functions are positive and increase monotonously with T up to 30 000 K. Function h_3 changes sign near 90 000 K, while a change of sign in functions h_2 and h_4 might also happen at temperatures above 30 000 K. In function h_4, the contributions arising from ions ${H_2}^+$ and H^- have the same order of magnitude and are much larger than the third contribution in Eq. (52) which accounts for a shift in the atomic ground state. For $T < 30\,000$ K, all functions remain quite small in relation with their exponential decay in the zero-temperature limit. The hierarchy $h_1(\beta) \gg h_2(\beta) \gg h_3(\beta) \gg$

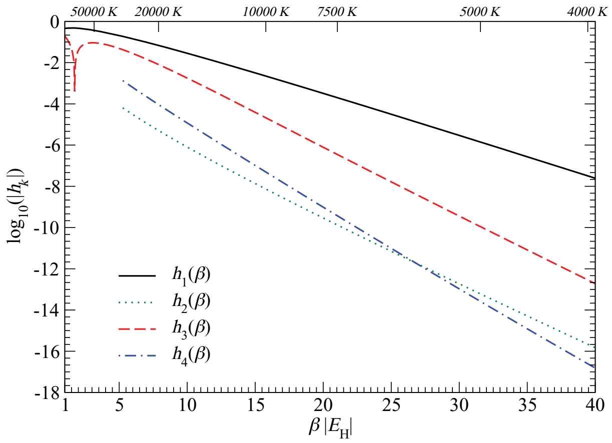


FIG. 3. (Color online) Plot of functions $h_k(\beta)$ for $k = 1,2,3,4$. Note that $\beta|E_H| = T_{\text{Rydberg}}/T$.

$h_4(\beta)$ is satisfied at very low temperature, in agreement with the ordering of the corresponding low-temperature decay rates $\delta_1 < \delta_2 < \delta_3 < \delta_4$. When the temperature is increased, the relative differences between those functions are reduced. In particular, h_3 overcomes h_2 near 1000 K, and h_4 overcomes h_2 near 5900 K.

The ordering of the h_k functions is connected to the relative importance of the corresponding corrections to the ideal Saha equation in the SLT limit $T \to 0$. At finite temperature and density, the ordering of the various corrections can differ from the low-temperature ordering, not only due to the temperature dependence of the h_k functions but also because those corrections involve also functions of ratio ρ/ρ^* which depends on both temperature and density as shown by formula (26). For instance, the corrections associated with h_2 describing molecular formation and atom-atom interactions, will dominate in a sufficiently dense atomic phase, despite the fact that function h_2 is much smaller than h_1 (see Sec. V A).

V. COMPARISONS TO OPAL TABLES AND PIMC DATA

We compute numerically the various corrections to the ideal Saha equation of state that appear in the SLT expansion of the pressure [Eq. (25)] and of the internal energy [Eq. (32)] and compare the predictions to the most accurate current tabulations of those thermodynamical functions. The present calculations correct[2] and complement the initial results for the pressure published in Ref. [54].

TABLE II. Numerical values of functions $h_k(\beta)$ ($k = 1,2,3,4$) at different temperatures.

T (K)	$h_1(\beta)$	$h_2(\beta)$	$h_3(\beta)$	$h_4(\beta)$
2000	1.46×10^{-16}	2.89×10^{-28}	3.99×10^{-26}	1.00×10^{-31}
6000	1.70×10^{-5}	2.73×10^{-12}	6.08×10^{-9}	3.01×10^{-12}
10 000	2.23×10^{-3}	7.40×10^{-9}	2.11×10^{-5}	4.94×10^{-8}
20 000	6.84×10^{-2}	5.12×10^{-6}	8.09×10^{-3}	9.26×10^{-5}
30 000	1.88×10^{-1}	6.41×10^{-5}	4.24×10^{-2}	1.35×10^{-3}

A. General properties of the isotherms

We start by studying general properties of isotherms that follow from the structure of the terms in SLT expansion (25). Along a given isotherm, the various corrections to Saha pressure depend on the ratio ρ/ρ^* where the crossover density ρ^* is kept fixed. Because of the nonlinear dependence in ρ/ρ^* of coefficients $b_k(\rho/\rho^*)$, their relative importance changes drastically from low densities $\rho \ll \rho^*$ to high densities $\rho \gg \rho^*$. For $\rho \ll \rho^*$, each $b_k(\rho/\rho^*)$ can be expanded in powers of ρ/ρ^*, as well as Saha pressure (3) itself. This leads to the well-known virial expansion of βP in powers of ρ at fixed T [9,23,44,55,56], as shown in Ref. [54]. The corresponding leading term entirely arises from Saha pressure and describes full ionization of the plasma. First correction of order $\rho^{3/2}$, entirely provided by βP_1, is the familiar Debye contribution for a classical plasma of ionized protons and electrons with density ρ. Saha pressure (3), βP_3, and βP_5 contribute to a second correction of order ρ^2, which accounts for atomic recombination as well as two-body interactions between ionized charges. Ionic contributions embedded in βP_4 are of order ρ^3, while molecular ones embedded in βP_2 are of order ρ^4, in agreement with the numbers of protons and electrons involved in the ions $H_2{}^+$ and H^- and molecule H_2.

For ρ of order ρ^*, beyond the leading contribution of Saha pressure (3), the ranking of the various corrections is essentially that of temperature-dependent functions $h_k(\beta)$. However, notice that the *a priori* first correction βP_1 vanishes at $\rho = 4\rho^*$ and becomes positive for $\rho > 4\rho^*$. Therefore, βP_1 reduces to the familiar Debye contribution $\beta P_{\text{Debye}} = -\kappa^3/(24\pi)$ only in the fully ionized region $\rho \ll \rho^*$, while its structure begins to differ in the atomic region $\rho > \rho^*$ as a subtle consequence of recombination processes. In the region close to $\rho = 4\rho^*$, the first correction to Saha pressure is then given by βP_3 as illustrated below.

For $\rho \gg \rho^*$, Saha pressure reduces to the ideal pressure of an atomic gas with density ρ since almost all charges are recombined. According to the large-ρ/ρ^* behavior of coefficients $b_k(\rho/\rho^*)$, the first correction to Saha pressure is now given by βP_2 which describes both molecular recombination and atom-atom interactions. Since it increases as ρ^2, that correction would overcome Saha pressure at sufficiently high densities. In fact, and as discussed further, this provides an upper density for the validity of the SLT expansion. Notice that while βP_4 also increases faster than $\beta P_{\text{Saha}} \sim \rho$, i.e., as $\rho^{3/2}$, all other considered corrections increase more slowly than ρ. This can be easily interpreted by noting that such corrections, namely βP_1, βP_3, and βP_5, are related to the presence of ionized protons and electrons which tend to disappear at high densities.

The previous analysis of the behavior of the pressure along a given isotherm is summarized in Table III. Similar results hold

---

[2] In Ref. [54], formula (9) contains a typo and the full SLT curve in Fig. 2 of Ref. [54] is incorrect due to an ill-placed parenthesis when computing the correction $k = 3$.

TABLE III. First correction to ideal Saha pressure along a low-temperature isotherm.

Density	First correction	Physical origin
$\rho \ll \rho^*$	βP_1	Debye plasma polarization
$\rho \simeq \rho^*$	βP_3	e-p interactions and excited H atoms
$\rho \gg \rho^*$	βP_2	H_2 molecules and H-H interactions

for the behavior of the internal energy per particle given by its SLT expansion (32) together with the expressions (33)–(38). In particular, at low densities, we find that the first correction to the classical thermal energy $3kT$ of the fully ionized plasma is indeed the Debye contribution which arises entirely from u_1. For densities $\rho \gg \rho^*$, u_2 becomes the leading correction to $u_{\rm Saha}$, since it increases as ρ while u_4 grows only as $\rho^{1/2}$, u_3 tends to a constant, and both u_1 and u_5 vanish. If we replace $h_2'(\beta)$ by its low-temperature form $(3E_{\rm H} - E_{\rm H_2})h_2(\beta)$, while we introduce the molar fraction of molecules H_2 defined by $x^{\rm id}_{\rm H_2} = \rho^{\rm id}_{\rm H_2}/\rho$, then the sum of leading and subleading contributions in Eq. (32) can be rewritten as

$$\left(x_{\rm at} + x^{\rm id}_{\rm H_2}\right)\frac{3kT}{2} + x_{\rm at}E_{\rm H} + x^{\rm id}_{\rm H_2}E_{\rm H_2}, \tag{55}$$

where the molar fraction $x_{\rm at}$ of atoms H accounts for the partial recombination of protons and electrons into molecules H_2, i.e., $x_{\rm at} = x^{\rm id}_{\rm at} - 2x^{\rm id}_{\rm H_2}$. Expression (55) shows that, for $kT \ll |E_{\rm H}|$ and $\rho \gg \rho^*$, the system reduces to an ideal mixture, made of a small fraction of molecules H_2 diluted in a gas of atoms H, in their molecular and atomic ground states, respectively.

B. Isotherms at a few thousands kelvins

Because of the relatively large value of temperature scale $|E_{\rm H}|/k \simeq 150\,000$ K and of the occurrence of exponentially decaying factors, crossover density ρ^* is extremely small below a few thousand kelvins. For instance, at $T = 300$ K, we find $\rho^* \simeq 3.4 \times 10^{-204}{\rm m}^{-3}$, which corresponds to tremendously diluted conditions that are not physically accessible. This means that a stable partially ionized atomic phase, which exists when ρ is of the order of ρ^*, cannot be realized in practice for hydrogen at such low temperatures. For instance, under the standard conditions of the Earth's atmosphere, density $\rho \simeq 10^{26}{\rm m}^{-3}$ is so large with respect to ρ^* that SLT expansion breaks down, due to correction βP_2 being too large, in agreement with the emergence of molecules H_2 as the most important species. Even at $T = 2000$ K, $\rho^* = 6.1 \times 10^{-9}{\rm m}^{-3}$ still corresponds to quite diluted conditions. Interesting physical systems with similar temperatures in the range 1000K $< T <$ 2000K are the atmospheres of brown dwarfs. The corresponding densities lie in the range $10^{25}{\rm m}^{-3} < \rho < 10^{27}{\rm m}^{-3}$, so they are too large compared to the corresponding ρ^*'s to use SLT expansions. In fact, as for Earth's atmosphere, hydrogen is essentially recombined into molecules. Thus, we will not consider isotherms below $T =$ 2000 K within SLT expansions, because the corresponding density ranges of validity do not correspond to known physical systems of interest. We note that the standard virial expansion cannot provide as well any useful information on the hydrogen gas when $T < 2000$ K, for the very same reason that the fully ionized phase is not thermodynamically stable at low temperatures for physically accessible densities.

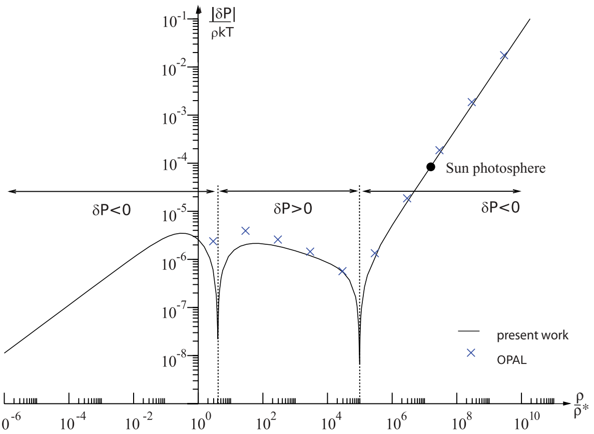


FIG. 4. (Color online) Logarithmic plot of deviations to Saha pressure for pure hydrogen along isotherm $T = 6000$ K ($\rho^* = 2.12 \times 10^{15}$ m^{-3}). Crosses correspond to tabulated points of the OPAL equation of state [1] (with corrected ground-state energy [34]).

When T increases up to a few thousands kelvins, the atomic crossover density ρ^* reaches higher values which are encountered in some systems. For fixing ideas, we consider isotherm $T = 6000$ K, which is typical of the Sun's photosphere and for which $\rho^* \simeq 2.12 \times 10^{15}{\rm m}^{-3}$. In Fig. 4, we plot deviation $\delta P = P - P_{\rm Saha}$ along that isotherm in the range $10^9{\rm m}^{-3} < \rho < 10^{27}{\rm m}^{-3}$, where δP does not exceed a few percentages of $P_{\rm Saha}$. When $\rho/\rho^* \leqslant 10^5$, the dominant contribution is due to the polarization of the plasma around ionized charges, embedded in correction P_1. For $\rho/\rho^* \leqslant 10^{-2}$, that contribution is negative and reduces to the familiar Debye expression $-\kappa^3/(24\pi)$ that appears in the virial expansion. At $\rho/\rho^* = 4$, correction P_1 changes sign, as seen on expression (27) for coefficient $b_1(\rho/\rho^*)$. The plasma-polarization correction thus is not given at high densities by the Debye formula with a modified Debye length computed with the density of ionized charges, as it could naively be expected in a phenomenological approach. At densities $\rho/\rho^* \geqslant 10^5$, molecular contributions embedded in term P_2 become the most important correction, as expected at high densities. Since the formation of molecules reduces the pressure, δP then becomes negative again. When $\rho/\rho^* \geqslant 10^{11}$, the SLT expansion fails to converge because molecular recombination can no longer be treated perturbatively.

Contributions of the first excited atomic state, embedded in correction P_3, and contributions of the ions $H_2{}^+$ and H^- embedded in P_4, are essentially negligible along the whole isotherm. For completeness, we mention that it is only near the special density $4\rho^*$ that P_3 turns out to provide the dominant correction, while P_4 is the dominant correction near the density $10^5\rho^*$ where corrections P_1 and P_2 compensate each other and where P_3 is negligible since there are almost no ionized charges. At these special densities, the deviation to the Saha pressure is about 10^{-9}.

We can compare our results to those of the OPAL tables [1], which are shown as crosses in Fig. 4. A very good agreement is found for densities $\rho/\rho^* > 10^4$. At lower densities, the

TABLE IV. Pressure at typical temperature and density of Sun photosphere.

	Pressure
$\beta P_{\mathrm{Saha}}/\rho$	$1 + 1.70 \times 10^{-4}$
$\beta P_1/\rho$	1.04×10^{-7}
$\beta P_2/\rho$	-3.79×10^{-4}
$\beta P_3/\rho$	-1.03×10^{-12}
$\beta P_4/\rho$	-2.36×10^{-8}
$\beta P_5/\rho$	-2.44×10^{-14}

discrepancies are certainly due to the fact that extracting a deviation to Saha pressure from the OPAL tables is a difficult task when the deviation is of the order of 10^{-6}. Indeed, the values in the OPAL tables are given with at most six digits, and a slight difference in the values of the fundamental constants, like in the ground-state energy E_{H}, can induce a small variation of the Saha pressure P_{Saha} that is comparable to the deviation $\delta P = P - P_{\mathrm{Saha}}$ itself.

A change of sign of δP is observed in the OPAL EOS around density $\rho/\rho^* \simeq 10^5$ at 6 000 K, in agreement with our SLT EOS: This sign change is induced by negative contributions in P_2 associated with molecular recombination overcoming the plasma polarization correction P_1. Notice that the nontrivial variations of δP with two sign changes are brought to light by simple physical interpretations within our approach. At the point with density $\rho \simeq 1.47 \times 10^{23}\mathrm{m}^{-3}$, typical of the Sun's photosphere shown in Fig. 4, electrons and protons are almost fully recombined into hydrogen atoms since $\rho \gg \rho^*$, and the various corrections to Saha pressure, which is itself close to $P_{\mathrm{at}}^{(\mathrm{id})}$, are given in Table IV. The full pressure is below $P_{\mathrm{at}}^{(\mathrm{id})}$ because of molecular recombination. Moreover, contributions of ions $\mathrm{H_2}^+$ and H^- are still smaller than the positive polarization contribution due to the ionized protons and electrons despite the fact that their dilution is quite large. Such subtle effects cannot be anticipated nor accurately described with phenomenological approaches.

C. Isotherms between ten and thirty thousands kelvins

We consider various isotherms above $T = 10\,000$ K up to $T = 30\,000$ K. The corresponding δP's are plotted in Fig. 5, while the respective values of ρ^* are given in Table V. The OPAL values of δP are shown, moreover, on the plots with symbols. As discovered in Ref. [34], the OPAL tables were computed using the value $E_{\mathrm{H}} = 1\mathrm{Ry} \simeq -13.60569$ eV, corresponding to an infinitely heavy nucleus, instead of the correct value, $E_{\mathrm{H}} = -me^4/(2\hbar^2) \simeq -13.59829$ eV. That inaccuracy in E_{H} induces variations of the Saha pressure that can be larger by an order of magnitude than the deviation δP itself for state points in the crossover region between the ionized and atomic phases. The OPAL deviations δP shown in all figures of the present paper were determined by subtracting from the OPAL values the ideal pressure P_{Saha} computed with $E_{\mathrm{H}} = 1$ Ry.

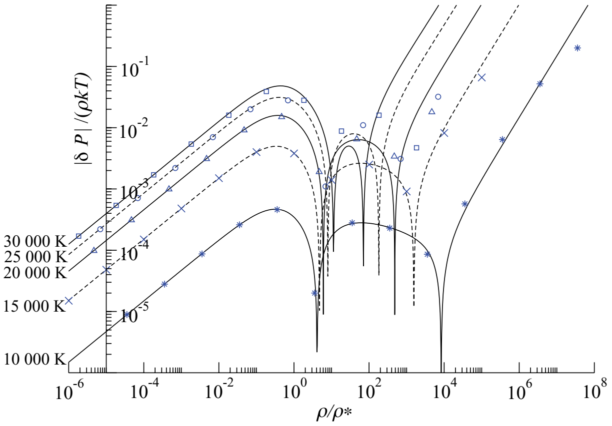


FIG. 5. (Color online) Logarithmic plots of deviations to Saha pressure along isotherms from 10 000 K up to 30 000 K according to the SLT EOS (solid and dashed lines) and to the tabulated OPAL EOS (symbols).

At the low temperatures 10 000 K and 15 000 K, we find excellent agreement between our analytical SLT EOS and the tabulated OPAL EOS in the considered density range. A small discrepancy is observed only for the last highest-density point on each of these two curves. This discrepancy at high density (massic density 0.1 g/cm^3) and low temperatures is due to the fact that the system is close to being in a molecular phase.

At low densities, $\rho/\rho^* < 1$, the agreement between both equations of state is also excellent, even at rather high temperatures. We note that the deviation δP is dominated in that region by the Debye plasma polarization correction P_1. Although that correction changes sign at $\rho/\rho^* = 4$, deviation δP changes sign, for isotherms with $T > 10\,000$ K, at a density ρ/ρ^* slightly higher than 4 because of negative contributions arising from correction P_3, which describes excited atoms and charge-charge interactions.

We observe, in Fig. 5, discrepancies in region $\rho/\rho^* \simeq 10^2$ and $T \geqslant 20\,000$ K. Those discrepancies are due to the negative corrections P_2 and P_4, which shrink more strongly, in our calculations, the region where deviation δP is positive than in the OPAL EOS. We note that those corrections may be somewhat overestimated since we computed in the present work an approximation to functions h_2 and h_4 in which only molecular and ionic bound-state contributions are kept. The discarded truncation terms in h_2 can provide positive contributions at high temperatures, which describe in particular atom-atom interactions. An accurate calculation of functions h_2 and h_4 at temperatures higher than 20 000 K is needed to provide fully reliable results at such temperatures and densities. Tests have shown that the OPAL deviations δP around $\rho/\rho^* \simeq 10^2$ are in fact fully explained by retaining solely the Debye plasma polarization effect P_1.

TABLE V. Atomic recombination density ρ^* in m^{-3} and in $\mathrm{g/cm}^3$ at various temperatures.

T (K)	$\rho^*(\mathrm{m}^{-3})$	$\rho^*(\mathrm{g/cm}^3)$
10 000	1.69×10^{20}	2.83×10^{-10}
15 000	5.98×10^{22}	1.00×10^{-7}
20 000	1.28×10^{24}	2.14×10^{-6}
25 000	8.65×10^{24}	1.54×10^{-5}
30 000	3.26×10^{25}	5.45×10^{-5}

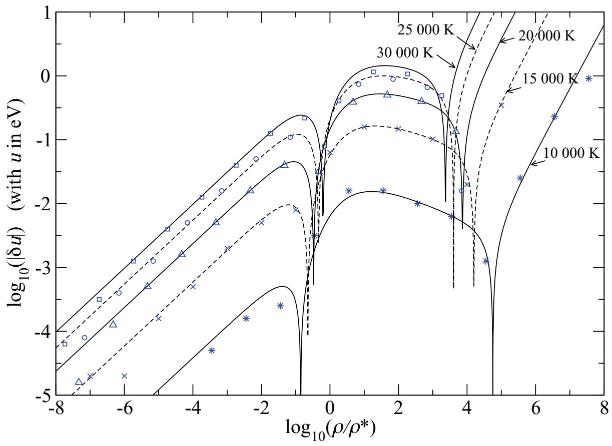


FIG. 6. (Color online) Deviations to Saha internal energy for isotherms between 10 000 K and 30 000 K according to the SLT EOS (solid and dashed lines). Points of the (corrected) OPAL EOS are shown by symbols (up to density 0.1 g/cm^3).

Deviations $\delta u = u - u_{\text{Saha}}$ of the internal energy per proton are shown along several isotherms in Fig. 6. The variations of δu at temperatures up to 10 000 K are fully controlled by the two terms u_1 and u_2. At low densities ($\rho/\rho^* \ll 1$), the plasma polarization term u_1 is dominant and negative. That correction changes sign when the condition

$$\frac{\rho}{\rho^*} = 4\epsilon \frac{1+\epsilon}{(1-\epsilon)^2}, \qquad \epsilon = \frac{kT}{2|E_H|} \tag{56}$$

is met, as can be seen from Eq. (34). At high densities ($\rho/\rho^* \simeq 10^4$), δu becomes negative again because the term u_2 becomes dominant and the formation of molecules indeed lowers the energy. When $T > 10\,000$ K, term u_3 comes into the game and has the effect of enlarging the domain where the deviations δu are positive, similarly to the case of the pressure deviations δP.

D. Low-density isochores

A plot of pressure deviations δP along two low-density isochores is displayed in Fig. 7 for temperatures between 2000 K and 100 000 K. The predictions of the OPAL tables, which are available for many temperature points, are also shown in those plots. At very low densities (isochore 10^{-8} g/cm^3), our calculations agree very well with the OPAL tables. When the temperature is high, the system is fully ionized and the dominant correction to Saha pressure arises from the Debye plasma polarization correction P_1, which behaves as $-\kappa^3/(24\pi) \propto T^{-3/2}$ when $T \to \infty$. On decreasing the temperature, correction P_1, and, hence, also δP, changes sign when the condition $\rho/\rho^* = 4$ is met. The deviation δP displays a second change of sign at a lower temperature, because the correction P_2 becomes dominant due to the formation of hydrogen molecules in the system. That correction P_2 grows quickly when the temperature is further lowered, and a point is reached where P_2 is no longer a small correction. This signals the formation of the molecular phase, which is outside the scope of the present calculations. Since the pressure of the ideal molecular gas is $\rho kT/2$ and that of the ideal atomic gas is ρkT, the pressure deviation $|P - P_{\text{Saha}}|/(\rho kT)$ should tend to 0.5 at low temperatures and low densities, as is indeed observed for the OPAL deviations in Fig. 7.

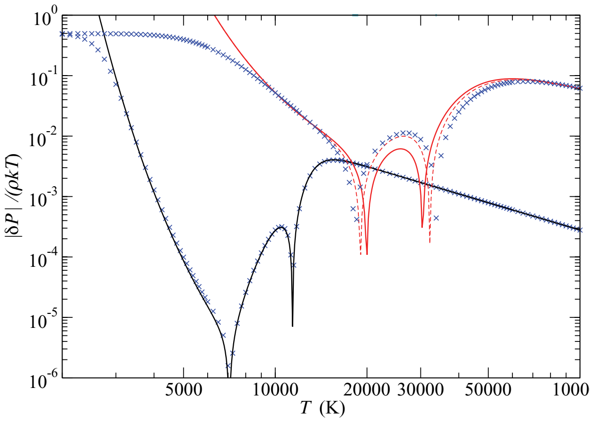


FIG. 7. (Color online) Deviations to Saha pressure along isochores 10^{-8} g/cm^3 (black line) and 10^{-3} g/cm^3 [red (light gray) line] according to the SLT EOS. Crosses correspond to values of the (corrected) OPAL EOS. The red (light gray) dashed line shows the effect of neglecting term P_4 in the SLT EOS.

Along isochore 10^{-3} g/cm^3 shown in Fig. 7, we can observe some discrepancies between our predictions for the deviations δP and those of the OPAL tables. When $T < 10\,000$ K, the differences are due to the formation of the molecular phase. In region 20 000K $< T <$ 30 000 K, deviation δP is positive and is somewhat larger in the OPAL tables than predicted by the SLT EOS. In that region, deviation δP is the result of the sum of the first four SLT terms, with P_2 and P_4 partially compensating P_1 and P_3. While the latter two terms are known exactly, the former two terms are currently estimated in our calculations by keeping only molecular and ionic bound states (see Sec. IV B). Not surprisingly, accurate calculations of the corresponding functions h_2 and h_4 in the temperature range 20 000–70 000 K, are required for a fully reliable description. When $T \geqslant 80\,000$ K, the SLT and OPAL predictions coincide. At such high temperatures, the deviation δP is due to the sum of the SLT terms P_1, P_3, and P_5 associated with ionized protons and electrons, and it reduces to the predictions of the standard virial expansion.

A low-density isochore of the internal energy is shown in Fig. 8, where the SLT EOS is compared to other equations of state. At high temperatures, the system is fully ionized and the variations of the internal energy per proton are mainly controlled by the average thermal kinetic energy $3kT/2$ of the particles (dot-dashed line). When the temperature is reduced to 10 000 K, a sharp drop of the internal energy is observed due to the formation of the atomic phase. Reducing further the temperature, a second drop in the potential energy occurs when the molecular phase is formed. Since the SLT EOS accounts perturbatively for all deviations to the ideal Saha equation of state that describes the transition between the ionized and the atomic phase, its validity domain is limited at low temperature by the formation of the molecular phase, which is a nonperturbative effect (see deviations at low temperatures in

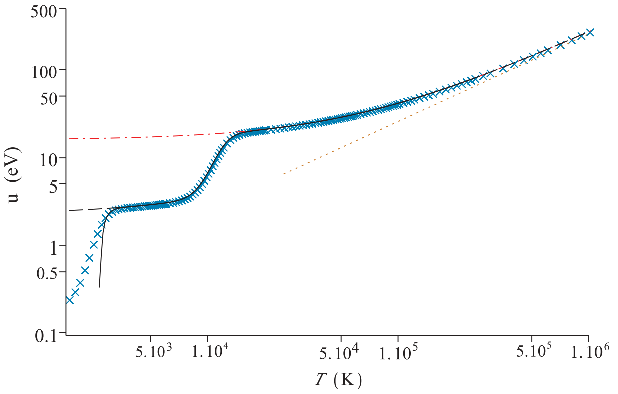


FIG. 8. (Color online) Internal energy per proton along isochore 10^{-8} g/cm^3 according to the SLT EOS (black line), OPAL EOS (crosses), and Saha EOS (dashed line). For the difference between the two former and the latter curves, see Fig. 9. All energies are shifted upwards by $|E_{\mathrm{H}_2}|/2$. The plot shows also curves $|E_{\mathrm{H}_2}|/2 + 3kT$ (red dot-dashed line) and $3kT$ (dotted line).

Fig. 8). On the scale of Fig. 8, the deviations $\delta u = u - u_{\mathrm{Saha}}$ between the SLT EOS or the OPAL EOS, and the ideal Saha values are almost indiscernible. Those deviations are shown in Fig. 9 for two isochores, 10^{-8} and 10^{-3} g/cm^3. The agreement on deviations δu between our analytical calculations and the OPAL tables is impressively good. As in the case of the pressure isotherms (see Fig. 5), δu changes sign twice: once in the ionized-atomic transition, when condition (56) is met, and once in the atomic-molecular crossover region. The only significant disagreement between the SLT and OPAL EOS along the very low density isochore (10^{-8} g/cm^3) occurs at low temperatures due to the formation of the molecular phase, where the OPAL deviation δu tends to the difference in energy per proton between the atomic and the molecular phase [$\log_{10}(E_{\mathrm{H}_2}/2 - E_{\mathrm{H}}) \simeq 0.35$]. Along isochore 10^{-3} g/cm^3, rather small differences are observed only near the two sign changes of δu, similarly to the deviations δP along the same isochore. Term u_5 contributes sensitively to δu only when $T \geqslant 30\,000$ K.

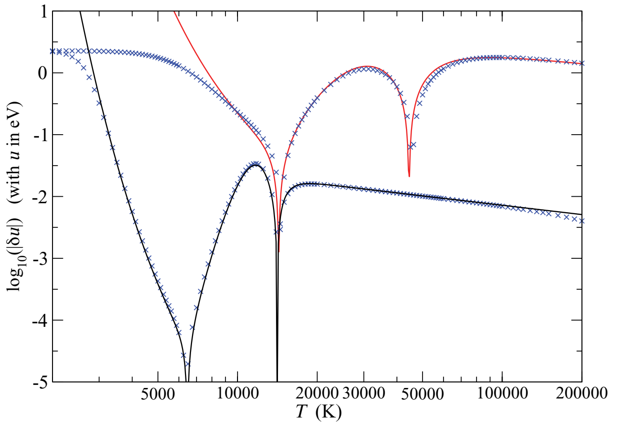


FIG. 9. (Color online) Deviations to Saha internal energy for isochores 10^{-8} g/cm^3 (black line) and 10^{-3} g/cm^3 [red (light gray) line] according to the SLT EOS. Crosses denote tabulated points of the (corrected) OPAL EOS.

E. Comparison to PIMC data

Results of quantum path integral Monte Carlo (PIMC) simulations of a dilute e-p gas are available at six densities between 10^{-3} g/cm^3 and 0.15 g/cm^3 and eight temperatures between 5000 K and 250 000 K [4] (see state points in Fig. 12). Very low densities, such as 10^{-6} g/cm^3 as in the Sun photosphere, are not within reach of PIMC simulations because sufficient statistics cannot be collected in very diluted conditions. Fortunately, the SLT expansion converges quickly at such low densities, and it has been shown in the previous section that the predictions of the SLT EOS coincide in this regime with those of the (corrected) OPAL tables, if we exclude the molecular phase which is outside the validity domain of the SLT expansion.

Comparison of our results with the PIMC data is instructive along the moderate density isochore 10^{-3} g/cm^3; see Fig. 10. The decrease of the pressure as the temperature is lowered is shown in Fig. 10(a), with P varying from $2\rho kT$ at high temperature (fully ionized gas) to ρkT (atomic gas at around 15 000 K), and eventually down to $\rho kT/2$ (molecular phase). A small plateau corresponding to the atomic phase can also be identified in Fig. 10(b) for the internal energy, though it is much less visible than on isochore 10^{-8} g/cm^3 (see Fig. 8). The deviations δP and δu from the ideal Saha values are shown in Figs. 10(a′) and 10(b′). The various sign changes of these deviations predicted by the SLT formulas can indeed be observed in the simulations. The uncertainties of the simulation data at temperatures $T \leqslant 15\,000$ K are quite large, especially in the case of the pressure. Notice that $\delta P/(\rho kT)$ tends trivially to $-1/2$ in a dilute molecular phase, so the simulation data do not contain much information in that regime. In the crossover region between the atomic phase and the fully ionized phase, the agreement on δP and δu between the simulation data and the SLT and OPAL equations of state is rather good. In the temperature range 10 000–25 000 K, the predictions of the SLT EOS are within the error bars of the PIMC results. At very high temperatures, $T > 100\,000$ K, the PIMC results for the energy deviation δu do not agree fully with the SLT nor the OPAL EOS. The uncertainties in the simulation results are maybe underestimated. When $T = 62\,500$ K, and to a lesser extent when $T = 31\,250$ K, the OPAL EOS agrees slightly better with the PIMC data than the SLT EOS. This might be due to our inaccurate description of functions h_2 and h_4 for such temperatures.

In Fig. 11, we compare our results along a denser isochore at massic density 0.0125 g/cm^3 (for which $r_s = a/a_{\mathrm{B}} = 6$), where significant differences between the predictions of the three approaches (SLT EOS, OPAL EOS, and PIMC simulations) can be observed, especially in the case of the pressure deviations. Disregarding the molecular phase at low temperatures, deviations δP and δu remain small (less than 10%) along this isochore, so the SLT expansion should still apply. Since the OPAL EOS agrees quite well with the PIMC data for the energy deviations δu and the OPAL EOS is thermodynamically consistent, the disagreement on the OPAL pressures with the PIMC data when $T \leqslant 30\,000$ K, i.e., in the

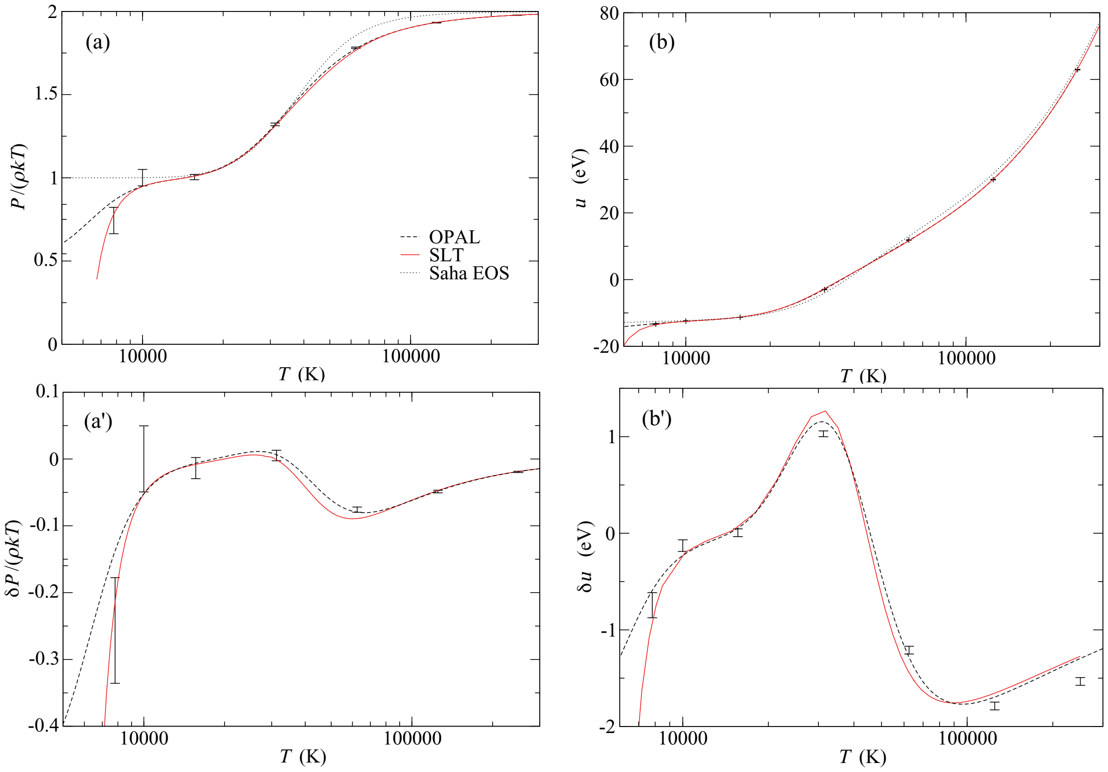


FIG. 10. (Color online) Pressure (a) and internal energy (b) per electron-proton pair as a function of temperature along isochore 10^{-3} g/cm^3, according to SLT EOS [red (light gray) line], OPAL EOS (dashed line), and ideal Saha EOS (dotted line). Points with error bars are simulation results of Ref. [4]. Plots (a′) and (b′) show deviations to the ideal Saha values along the isochore.

crossover region between the atomic and the molecular phase, indicates probably that the errors on the PIMC pressures are underestimated. Some disagreement between the OPAL (and SLT) equations of state and the PIMC results is also observed at 31 250 K and at high temperatures for $T \geqslant 100\,000$ K. As for the above lower-density isochore, some discrepancies between the SLT EOS and the other data might be due to inaccuracies in the calculation of functions $h_2(T)$ and $h_4(T)$ for temperatures above 30 000 K.

F. Validity domain

As exemplified in the previous sections, truncation of the SLT expansion (25) at $k = 5$ gives accurate results as long as the corresponding deviations $\delta P = \sum_{k=1}^{5} P_k$ or $\delta u = \sum_{k=1}^{5} u_k$ remain small compared to their ideal Saha value. The validity of the SLT expansion is limited at high densities along an isotherm, respectively at low temperatures along an isochore, by the formation of the molecular phase, where βP_2 then becomes larger than $P_{\rm Saha}$ itself. We can estimate the borderline of the validity domain by introducing the empirical criterion $|\delta P(T_{\rm v},\rho)| = P_{\rm Saha}/10$. At high densities, or low temperatures, $\rho \gg \rho^*$ so P_2 becomes the leading correction in δP. Criterion $|P_2|/P_{\rm Saha} = 0.1$ then gives

$$\rho_{\rm c}(T) = \frac{\rho^*(T)}{20|h_2(T)|} \qquad (57)$$

as a borderline for the validity domain in the (ρ,T) plane. If the temperature is low, function $h_2(T)$ behaves as

$$h_2(T) \sim \frac{1}{64}\left(\frac{2m}{M}\right)^{3/2} \exp[(3E_{\rm H} - E_{\rm H_2})/(kT)], \qquad (58)$$

and $\rho_c(T)$ reduces to a straight line in the $(\log\rho,\ \beta)$ plane (see Fig. 12). The curve $\rho_c(T)$ defines quite precisely the borderline of the validity domain in a large part of the phase diagram, as checked by the comparison with the data of the PIMC simulations. In particular, this is illustrated by the plots in Sec. IV E and the state points denoted by crosses in Fig. 12. At a given density, the SLT expansion converges only for temperatures that are sufficiently high to avoid the formation of the molecular phase. Some of the lower-temperature state points in the PIMC simulations are, for instance, outside of the validity domain of the SLT expansion, while others, like those to the left of line $\rho_c(T)$ in Fig. 12, are within the validity domain. When the density increases, the minimum temperature required for the SLT expansion to converge also increases. At density 0.1 g/cm^3, the temperature must be higher than 30 000 K. Since the exact behavior of $h_2(T)$ above 30 000 K is currently not available, the precise position of the borderline of the validity domain in the region of high densities and temperatures is yet not fully known.

The crossover density $\rho^*(T)$ [Eq. (2)] between the fully ionized and the atomic phase is also shown in Fig 12. That line gives essentially the borderline of the validity domain

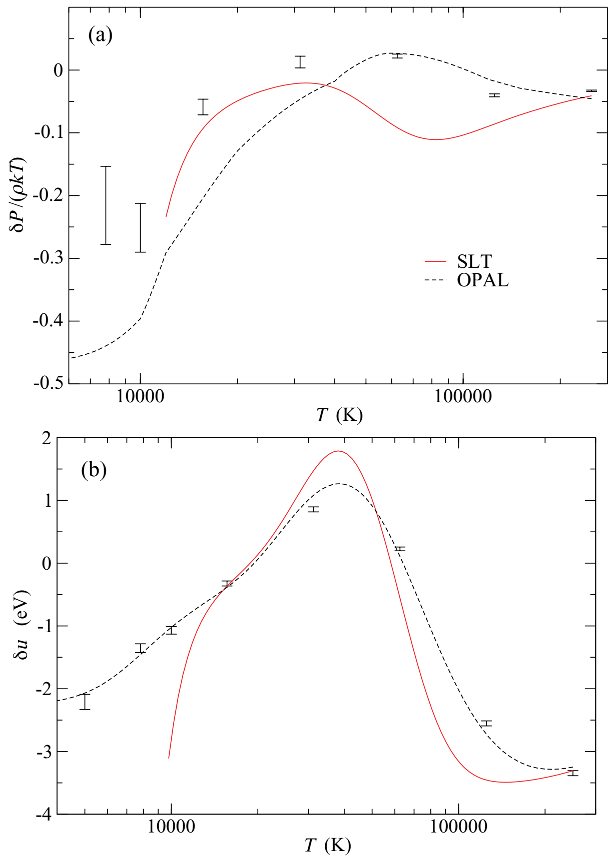


FIG. 11. (Color online) Deviations (with respect to ideal Saha values) of pressure (a) and internal energy per electron-proton pair (b) as a function of temperature along isochore 0.0125 g/cm^3, according to the SLT EOS [red (light gray) line] and OPAL EOS (dashed line). Points are simulation results of Ref. [4].

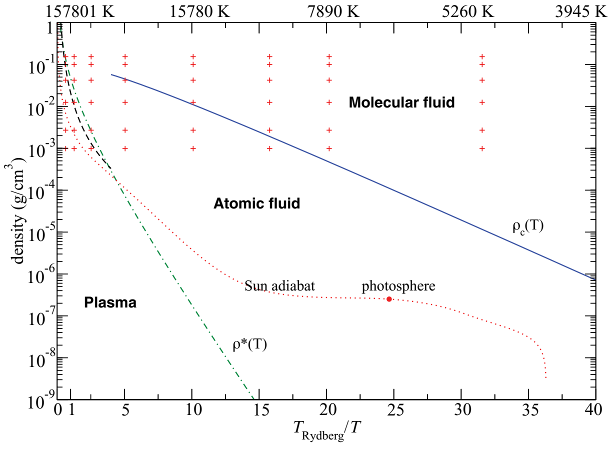


FIG. 12. (Color online) Phase diagram of a pure hydrogen gas at low densities. The crosses denote state points where simulation results are available [4]. The crossover density $\rho^*(T)$ between the plasma and the atomic phase (dot-dashed line) gives essentially the borderline of the validity domain of the standard virial expansion, which applies only in the plasma phase. The dashed line corresponds to a plasma coupling parameter $\Gamma = 0.5$. The SLT expansion is valid in both the plasma and atomic phases, up to the solid (blue) line $\rho_c(T)$ [Eq. (57)], which locates the crossover to the molecular phase. The state point of the Sun photosphere and the track of the Sun adiabat (dotted line) are also shown.

of the standard virial expansion, which holds only in the weakly coupled ($\Gamma \ll 1$) fully ionized phase. The dashed line in Fig. 12 shows state points where the coupling parameter $\Gamma = \beta^2/a$ is equal to 0.5. That line lies quite close to $\rho^*(T)$. In the narrow strip at high densities and temperatures between the lines $\Gamma = 0.5$ and $\rho^*(T)$, the SLT expansion is expected to converge only slowly because Γ is close to 1 ($\Gamma \leqslant 0.7$ in that region).

Figure 12 shows that the SLT expansion provides an accurate analytical knowledge of the thermodynamics of the quantum e-p gas in a rather large range of densities and temperatures that includes the fully ionized phase ($\rho \ll \rho^*$), the partially ionized phase ($\rho \sim \rho^*$), and the atomic phase ($\rho^* \ll \rho < \rho_c$).

VI. CONCLUSIONS

In this work, we extend the exact analytical knowledge on the thermodynamics of hydrogen at low densities by deriving the first five terms in the SLT expansion of the internal energy [Eq. (32)], a result that complements the corresponding expansions for the pressure [Eq. (25)] and the chemical potential [Eq. (20)]. Similar expansions for any thermodynamical quantity can be easily derived without any loss of thermodynamic consistency. We performed also extensive numerical calculations of isotherms and isochores and compared in detail the predictions of the SLT formulas with the numerical OPAL EOS and data of PIMC simulations. Our analytical SLT formulas for the deviations to the ideal Saha law can be evaluated numerically very easily and quickly. As applying the SLT EOS is straightforward, no extensive tabulation, with the associated loss of accuracy due to interpolation, is required. A plot and a tabulation of the functions $h_k(\beta)$ ($k = 1,2,3,4$) have been provided as guide to help the user in applying numerically our formulas.

We emphasize the following points:

(1) The exact SLT expansion overcomes the restriction $\rho \ll \rho^*(T)$ of the standard virial expansion and has a validity domain that extends up to the density $\rho_c(T)$ in the atomic phase (see Fig. 12).

(2) At low densities ($\rho < 10^{-5}$ g/cm^3), excellent agreement is found between the predictions of the SLT EOS for the pressure and the internal energy and the values in the OPAL tables for pure hydrogen. Notice that if the density is very low, the OPAL tables need to be corrected as explained in Ref. [50]. As several important ingredients in the OPAL EOS, which is available only in the form of precomputed tables, are unknown, attempts have been made to emulate this equation of state [57]. It is very satisfactory to see that the OPAL EOS for pure hydrogen can be fully reproduced at low densities by our simple analytical formulas. Furthermore, the physical content of the various corrections of interest is enlightening, as well as the subtle cancellations between some contributions. No PIMC simulation results are available at such low densities because the statistics becomes poor.

(3) At higher densities, some small discrepancies can be observed among the SLT EOS, the OPAL EOS, and the simulation results. We plotted the pressure deviations $\delta P = P - P_{\text{Saha}}$ and the energy deviations $\delta u = u - u_{\text{Saha}}$ to the ideal Saha law to extract the variations of those thermodynamical quantities that are due to nonideality. Those deviations δP and δu show two sign changes, which are also seen in the simulation data and which are fully explained by our analytical formulas. For densities up to 10^{-3} g/cm^3 and temperatures up to 30 000 K, the SLT EOS and the OPAL EOS are both within the error bars of the PIMC simulations.

(4) At densities around 10^{-2} g/cm^3 and above, and for temperatures above 30 000 K, there are some discrepancies between the deviations, especially δP, calculated from the PIMC data and those of the SLT and OPAL EOS. In that range of temperatures, our present approximate expressions for $h_2(\beta)$ and $h_4(\beta)$ are not reliable. Better estimations of those functions, based on a numerical evaluation of path integral formulas for the internal partition functions $Z(2,2)$, $Z(2,1)$, and $Z(1,2)$, would provide an interesting improvement of our calculations, in particular in the region $\rho \geqslant 10^{-3}$ g/cm^3 and $T \geqslant 30\,000$ K.

As shown in Ref. [34], expansion (13) of particle density in terms of chemical potential should remain valid in the molecular regime, i.e., for $\rho \geqslant \rho_c$, provided the density is not too high. The accurate knowledge of $Z(2,2)$, $Z(2,1)$, and $Z(1,2)$ within numerical path integration, together with a numerical inversion of Eq. (13), should, therefore, provide a precise description of the crossover transition from the atomic gas to the molecular gas, much beyond the level of accuracy of current calculations, including PIMC simulations [34]. The accurate calculations of $Z(2,2)$, $Z(2,1)$, and $Z(1,2)$ via numerical path integrations would include in particular the contributions of atom-atom interactions. Such contributions are not easy to determine within simple modelizations because of the difficulty, intrinsic to quantum mechanics, in separating them from purely molecular contributions.[3] The accurate knowledge of $Z(2,2)$, $Z(2,1)$, and $Z(1,2)$ might also be useful for improving chemical approaches, like the Saumon-Chabrier theory [58], the MHD model [25], or the SAHA-S model [59].

The track of the Sun adiabat [60] stays well within the validity domain of the SLT expansion. The presents results are, therefore, of interest for astrophysics, where a very accurate EOS is needed, for instance, to interpret recent seismology measurements in the Sun [7,60]. The SLT expansion of other thermodynamical properties, such the adiabatic exponent and the sound speed, can be derived along similar lines. For real applications to helioseismology, the present calculations must be generalized to the case of a hydrogen-helium mixture within similar tools, an *a priori* feasible task. The contributions of other heavier elements might be determined within simple ideal approximations since their dilution is very high. Notice that relativistic effects associated with the electrons should be also incorporated as discussed in Refs. [61,62].

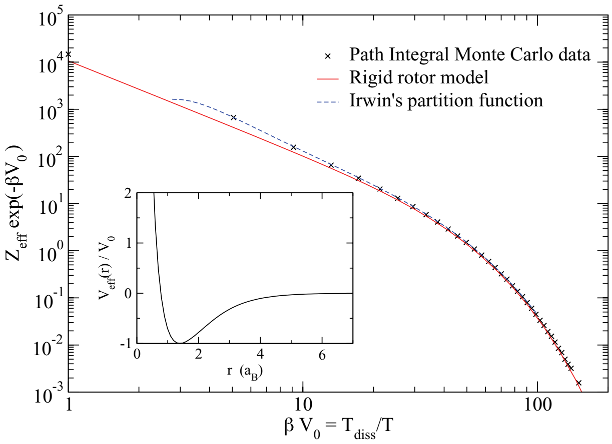


FIG. 13. (Color online) The truncated trace $Z_{\text{eff}} = \text{Tr}[\exp(-\beta H_{\text{eff}}) - \exp(-\beta H_0)]$ for a quantum particle in the effective proton-proton potential of the H_2 molecule (shown in the inset), as obtained from a numerically exact path integral Monte Carlo calculation (crosses), from the rigid rotor approximation (A4) (solid line) and from Irwin's partition function [53] (dashed curve). The dissociation temperature of the H_2 molecule is $T_{\text{diss}} = V_0/k = 55\,459$ K.

APPENDIX: A SIMPLIFIED MODEL

Let us consider a simple model which is often used for describing the H_2 molecule [63]. In that model, the two protons separated by a distance R interact *via* a potential $V_{\text{eff}}(R)$ which is inferred from the electronic ground-state wave function for that fixed protonic configuration. The potential $V_{\text{eff}}(R)$ is repulsive at short distances, attractive at large distances, and minimum at $R = R_0$ with $V_{\text{eff}}(R_0) < 0$. A plot of that potential, obtained by fitting the data of Ref. [63] to the formula

$$V_{\text{eff}}(R) = (1 + aR + bR^2 + cR^3)\frac{\exp(-R/d_A)}{R} - \frac{6.5}{d_B^6 + R^6} \tag{A1}$$

is shown in the inset of Fig. 13, while the values of the fitted coefficients are given in Table VI. Equation (A1) is written in atomic units: lengths are measured in units of the Bohr radius and energies in units of the Hartree energy (1 E_h = 2 Ry). The zero of energy corresponds here to the state with two hydrogen atoms infinitely far apart. The minimum of the potential is located at $R_0 = 1.3924$ (in atomic units) with value

TABLE VI. Parameters of potential (A1) that best fit the data of Ref. [63] for the hydrogen molecule.

Coefficient	Value (in a.u.)
a	−0.52906
b	−0.60479
c	−0.37294
d_A	0.64160
d_B	2.64591

---

[3] Notice that in Ref. [54], we proposed a simple approximation for those contributions, which needs to be improved for tackling regimes with $T > 30\,000$ K.

$V(R_0) = -V_0 = -0.175629$ hartree. The dissociation energy of the hydrogen molecule is $T_{\text{diss}} = V_0/k = 55\,459$ K.

For our purpose, it is sufficient to consider that the protons are spinless. A proper account of the spins would lead to the usual coupling between spin and position variables, which in turn induce a different counting of ortho- and parahydrogen contributions; see, for instance, formula (50).

The Hamiltonian of the relative particle with mass $m^* = m_{\text{p}}/2$ submitted to $V_{\text{eff}}(R)$ reads

$$H_{\text{eff}} = -\frac{\hbar^2}{2m^*}\Delta_{\mathbf{R}} + V_{\text{eff}}(R). \tag{A2}$$

Within that simplified model, the analog of the contribution of $[\exp(-\beta H_{2,2}) - \exp(-\beta H_{1,1})\exp(-\beta H_{1,1})]$ to $h_2(\beta)$ is the truncated trace

$$Z_{\text{eff}} = \text{Tr}[\exp(-\beta H_{\text{eff}}) - \exp(-\beta H_0)], \tag{A3}$$

where H_0 is the kinetic part of H_{eff}.

Let $E_0 = E_{\text{H}_2} - 2E_{\text{H}} \simeq -0.94188\, V_0$ be the ground-state energy of H_{eff}. Instead of determining exactly all bound-state energies of H_{eff} by solving the corresponding radial Schrodinger equation for various values of orbital number l, we consider the usual rigid-rotor and harmonic well approximations for describing global rotations and vibrations. Then, on the one hand, the approximation analoguous to Eq. (47) for $h_2(\beta)$ becomes

$$Z_{\text{eff}} \simeq \exp(-\beta E_0)\frac{1}{1-\exp\left(-\beta\epsilon_{\text{eff}}^{(\text{vib})}\right)}\sum_{l=0}^{\infty}(2l+1) \times \exp\left[-l(l+1)\beta\epsilon_{\text{eff}}^{(\text{rot})}\right], \tag{A4}$$

where $\epsilon_{\text{eff}}^{(\text{rot})} = kT_{\text{eff}}^{(\text{rot})}$ and $\epsilon_{\text{eff}}^{(\text{vib})} = kT_{\text{eff}}^{(\text{vib})}$ are the rotational and vibrational quanta which can be determined respectively from R_0 and from the shape of $V_{\text{eff}}(R)$ around its minimum at $R = R_0$: $T_{\text{eff}}^{(\text{rot})} = 88.7$ K and $T_{\text{eff}}^{(\text{vib})} = 6524$ K. On the other hand, we have performed a numerical calculation of Z_{eff} within path integral Monte Carlo methods applied to its Feynman-Kac path integral representation [47],

$$Z_{\text{eff}} = \frac{1}{(2\pi(\lambda^*)^2)^{3/2}}\int d\mathbf{R}\int \mathcal{D}_{\text{W}}(\boldsymbol{\xi}) \times \left[\exp\left[-\beta\int_0^1 ds\, V_{\text{eff}}(\mathbf{R}+\lambda^*\boldsymbol{\xi}(s))\right]-1\right], \tag{A5}$$

where $\boldsymbol{\xi}(s)$ is a Brownian bridge such that $\boldsymbol{\xi}(1) = \boldsymbol{\xi}(0) = 0$, $\mathcal{D}_{\text{W}}(\boldsymbol{\xi})$ is the normalized Wiener measure, and $\lambda^* = (\beta\hbar^2/m^*)^{1/2}$. At low temperatures ($T \ll T_{\text{diss}}$), the truncated trace Z_{eff} is dominated by the contributions arising from bound states with negative energies since the corresponding contributions grow exponentially fast when $\beta \to \infty$. Figure 13 shows that the approximation (A4) (which amounts to keeping only bound-state contributions evaluated within the rigid rotor model) represents quite well the Monte Carlo values for Z_{eff} up to $T \simeq 30\,000$K. Thus, we can reasonably expect a similar accuracy for approximation (47). We note that Irwin's partition function [53], which accounts for rotational-vibration coupling in molecule H_2 by summing explicitly on calculated rovibrational energy levels, provides a better fit to the Monte Carlo data for this simplified model, but Irwin's partition function is available only up to 16 000 K. At high temperatures (30 000 K and above), a precise evaluation of the contributions besides those of the bound states, i.e., contributions arising from diffusive states as well as those due to the truncation terms in $h_2(\beta)$, becomes mandatory.

[1] F. J. Rogers and A. Nayfonov, Astrophys. J. **576** 1064 (2002); F. J. Rogers, F. J. Swenson, and C. A. Iglesias, *ibid.* **456**, 902 (1996).

[2] F. J. Rogers, Phys. Rev. A **10**, 2441 (1974).

[3] F. J. Rogers, Phys. Rev. A **24**, 1531 (1981); Astrophys. J. **310**, 723 (1986); in *The Equation of State in Astrophysics*, edited by G. Chabrier and E. Schatzman (Cambridge University Press, New York, 1994).

[4] B. Militzer and D. M. Ceperley, Phys. Rev. E **63**, 066404 (2001).

[5] F. J. Rogers and D. A. Young, Phys. Rev. E **56**, 5876 (1997).

[6] A. Alastuey, V. Ballenegger, F. Cornu, and Ph. A. Martin, J. Stat. Phys. **130**, 1119 (2008).

[7] W. Däppen, J. Phys. A: Math. Gen. **39**, 4441 (2006).

[8] T. Morita, Prog. Theor. Phys. **22**, 757 (1959).

[9] W. Ebeling, Ann. Phys. Leipzig **19**, 104 (1967).

[10] C. Deutsch, Phys. Lett. A **60**, 317 (1977).

[11] H. Minoo, M. M. Gombert, and C. Deutsch, Phys. Rev. A **23**, 924 (1981).

[12] H. E. DeWitt, M. Schlanges, A. Y. Sakakura, and W. D. Kraeft, Phys. Lett. A **197**, 326 (1995).

[13] D. Kremp, M. Schlanges, and W.-D. Kraeft, *Quantum Statistics of Nonideal Plasmas* (Springer, Berlin, 2005).

[14] A. Alastuey, F. Cornu, and A. Perez, Phys. Rev. E **49**, 1077 (1994).

[15] D. C. Brydges and Ph. A. Martin, J. Stat. Phys. **96**, 1163 (1999).

[16] V. Ballenegger, Ph. A. Martin, and A. Alastuey, J. Stat. Phys. **108**, 169 (2002).

[17] F. Perrot and M. W. C. Dharma-wardana, Phys. Rev. E **52**, 5352 (1995).

[18] T. J. Lenosky, S. R. Bickham, J. D. Kress, and L. A. Collins, Phys. Rev. B **61**, 1 (2000).

[19] W. Lorenzen, B. Holst, and R. Redmer, Phys. Rev. Lett. **102**, 115701 (2009).

[20] C. Pierleoni, D. M. Ceperley, B. Bernu, and W. R. Magro, Phys. Rev. Lett. **73**, 2145 (1994).

[21] A. Alastuey, V. Ballenegger, F. Cornu, and Ph. A. Martin, J. Stat. Phys. **113**, 455 (2003).

[22] A. Alastuey, F. Cornu, and A. Perez, Phys. Rev. E **51**, 1725 (1995).

[23] A. Alastuey and A. Perez, Phys. Rev. E **53**, 5714 (1996).

[24] F. Cornu, Phys. Rev. E **58**, 5268 (1998); **58**, 5293 (1998); **58**, 5322 (1998).

[25] D. Mihalas, W. Däppen, and D. Hummer, Astrophys. J. **331**, 815 (1988).

[26] A. Y. Potekhin, Phys. Plasm. **3**, 4156 (1996).

[27] G. Massacrier, A. Y. Potekhin, and G. Chabrier, Phys. Rev. E **84**, 056406 (2011).

[28] M. R. Zaghloul, Phys. Rev. E **69**, 026702 (2004).

[29] Y. V. Arkhipov, F. B. Baimbetov, and A. E. Davletov, Phys. Rev. E **83**, 016405 (2011).
[30] A. V. Filinov, V. O. Golubnychiy, M. Bonitz, W. Ebeling, and J. W. Dufty, Phys. Rev. E **70**, 046411 (2004).
[31] Y. A. Omarbakiyeva, C. Fortmann, T. S. Ramazanov, and G. Röpke, Phys. Rev. E **82**, 026407 (2010).
[32] M. Saha, Philos. Mag. **40**, 472 (1920).
[33] N. Macris and Ph. A. Martin, J. Stat. Phys. **60**, 619 (1990).
[34] A. Alastuey and V. Ballenegger, Contrib. Plasma Phys. **52**, 95 (2012).
[35] E. H. Lieb and J. Lebowitz, Adv. Math. **9**, 316 (1972).
[36] C. Fefferman, Rev. Math. Iberoamericana **1**, 1 (1985).
[37] J. G. Conlon, E. H. Lieb, and H. T. Yau, Commun. Math. Phys. **125**, 153 (1989).
[38] I. Iosilevskiy, Contrib. Plasma Phys. **49**, 713 (2009).
[39] L. Brillouin, *Les Statistiques Quantiques et leurs Applications* (Presse Univ. de France, Paris, 1930).
[40] A. I. Larkin, Sov. Phys. JETP **11**, 1363 (1960).
[41] A. Alastuey and V. Ballenegger, J. Phys. A: Math. Theor. **42**, 214031 (2009).
[42] D. Bollé, Ann. Phys. (NY) **121**, 131 (1979).
[43] D. Kremp, W. D. Kraeft and A. J. M. D. Lambert, Physica A **127**, 72 (1984).
[44] W. D. Kraeft, D. Kremp, W. Ebeling, and G. Röpke, *Quantum Statistics of Charged Particle Systems* (Plenum Press, New York, 1986).
[45] W. Ebeling, Physica **40**, 290 (1968).
[46] W.-D. Kraeft, J Phys: Conf. Ser. **11**, 131 (2005).
[47] H. Kleinert, *Path Integrals in Quantum Mechanics, Statistics, Polymer Physics and Financial Markets* (World Scientific, Singapore, 2004).
[48] B. Jancovici, Physica A **91**, 152 (1978).
[49] W. Ebeling, W. D. Kraeft, and G. Röpke, Ann. Phys. **524**, 311 (2012).
[50] V. Ballenegger, Ann. Phys. (Berlin) **524**, 103 (2012).
[51] W. Ebeling, W. D. Kraeft, and G. Röpke, Contrib. Plasma Phys. **52**, 7 (2012).
[52] L. D. Landau and E. M. Lifchitz, *Quantum Mechanics*, 3rd ed. (Pergamon Press, Oxford, 1977), Vol. 3.
[53] A. W. Irwin, Astron. Astrophys. **182**, 348 (1987).
[54] A. Alastuey and V. Ballenegger, Contrib. Plasma Phys. **50**, 46 (2010).
[55] A. Alastuey and A. Perez, Europhys. Lett. **20**, 19 (1992).
[56] T. Kahlbaum, J. Phys. IV France **10**, 455 (2000).
[57] H-H. Lin and W. Däppen, Contrib. Plasma Phys. **52**, 126 (2012).
[58] D. Saumon and G. Chabrier, Phys. Rev. A **46**, 2084 (1992).
[59] V. K. Gryaznov, S. V. Ayukov, V. A. Baturin, I. L. Iosilevskiy, A. N. Starostin, and V. E. Fortov, J. Phys. A: Math. Gen. **39**, 4459 (2006).
[60] J. Christensen-Dalsgaard *et al.*, Science **272**, 1286 (1996).
[61] V. K. Gryaznov, S. V. Ayukov, V. A. Baturin, I. L. Iosilevskiy, A. N. Starostin, and V. E. Fortov, in Equation-of-state and phase-transition issues in models of ordinary astrophysical matter, edited by V. Celebonovic, W. Däppen, and D. Gough, AIP Conf. Proc. **731**, 147 (2004).
[62] A. N. Starostin and V. C. Roerich, Plasma Sources Sci. Technol. **15**, 410 (2006).
[63] W. Kolos and L. Wolniewicz, J. Chem. Phys. **43**, 2429 (1964).

Contrib. Plasma Phys. **52**, No. 1, 95 – 99 (2012) / **DOI** 10.1002/ctpp.201100087

Atomic Ionization and Molecular Dissociation in a Hydrogen Gas within the Physical Picture

A. Alastuey[1] and **V. Ballenegger**[2]*

[1] Laboratoire de Physique, Ecole Normale Supérieure de Lyon, CNRS, 46 allée d'Italie, 69364 Lyon Cedex 07, France
[2] Institut UTINAM, Université de Franche-Comté, CNRS, 16 route de Gray, 25030 Besançon Cedex, France

Received 14 October 2011, revised 02 December 2011, accepted 10 December 2011

Key words Hydrogen, equation of state, ionization, molecular dissociation, bound states, partition function.

We study a hydrogen gas at low densities within the physical picture. Recombination processes leading to the formation of atoms and molecules are properly taken into account *via* the well-known Ebeling function and a new four-body partition function. Our method provides a reliable equation of state which covers the plasma, atomic and molecular phases.

1 Introduction

We consider hydrogen in the low-density region of the phase diagram, where recombination processes lead to the formation of atoms or/and molecules. Contrarily to familiar chemical approaches which assume that those recombined entities are preformed objects with phenomenological internal partition functions and effective interactions, we describe hydrogen within the so-called physical picture in terms of a quantum plasma of point protons and electrons interacting *via* the Coulomb $1/r$ potential. Only the use of the physical picture can provide reliable equations of state (EOS) that reach the accuracy of experimental observations. Recent advances in helioseismology allow one for instance to probe the equation of state (EOS) of the solar matter —composed mainly of hydrogen — to an accuracy of better than 10^{-4} [1].

Atomic recombination or ionization was first considered in the framework of the physical picture by Ebeling [2]. More precisely, using Morita approach, Ebeling exactly computed the second virial coefficient (SVC) in the low-activity expansion, where atomic contributions are properly taken into account. The key quantity is some function Q, which can be expressed as an infinite sum over both bound and scattering states of the two-body Coulomb Hamiltonian. The finiteness of Q is ensured by the substraction from the two-body Gibbs factor of non-integrable powers of the Coulomb potential, while the *a priori* divergent contributions of those powers are regularized thanks to screening by the ionized charges present in the medium. We stress that the screening mechanism that regularizes both the internal atomic partition function and the ionized charge-charge interactions in the SVC is treated in an exact way. Bound and scattering contributions are intrinsically mixed in function Q. The tight interplay between recombination and screening is such that these two effects cannot be fully disentangled. [1]

If Ebeling function Q controls atomic corrections in the almost fully ionized regime at low densities and rather high temperatures, it was shown more recently that Q is still a central quantity for the ionization equilibrium $\mathrm{H} \rightleftarrows \mathrm{p} + \mathrm{e}$ in the partially ionized regime [3, 4] which can be attained by lowering the temperature along a low-density isochore. If temperature is further decreased beyond full atomic recombination, molecules form and become the most important chemical species. The main purpose of this paper is the description of that molecular recombination process within the physical picture. We show that the central quantity is now a cluster particle function $Z(2, 2)$ for two protons and two electrons, which can be viewed as a natural extension of Ebeling

---

* Corresponding author. E-mail: vincent.ballenegger@univ-fcomte.fr, Phone: +33 3 81 66 64 79, Fax: +33 3 81 66 64 75

[1] Accordingly, the extraction from Q of an internal atomic partition function remains arbitrary. For instance the well-known Planck-Larkin-Brillouin formula is a possible choice. Nevertheless, as far as thermodynamical properties are concerned, only the full contribution of Q, including scattering states, plus the contributions from screening effects makes an unambiguous sense.

function Q. Our method, briefly sketched in Sec. 2 is based on the so-called Screened Cluster Representation for the density ρ in terms of the chemical potential μ [5,7]. That formalism allows us to deal with the interplay between screening and recombination into atoms, molecules, or any other species like ions H^-, H_2^+,..., in an exact way. Cluster partition function $Z(2,2)$ involves not only contributions from bound states of the molecule, but also from its dissociation products which can be associated with the various chemical equilibria $\mathrm{H}_2 \rightleftarrows \mathrm{H}+\mathrm{H}$, $\mathrm{H}_2 \rightleftarrows \mathrm{H}_2^+ + \mathrm{e}$, $\mathrm{H}_2 \rightleftarrows \mathrm{H}^- + \mathrm{p}$, $\mathrm{H}_2 \rightleftarrows \mathrm{H} + \mathrm{p} + \mathrm{e}$, and $\mathrm{H}_2 \rightleftarrows \mathrm{p} + \mathrm{p} + \mathrm{e} + \mathrm{e}$. As exposed in Sec. 2, an adequate choice of diagrams in the Screened Cluster Representation allows one to obtain an approximate function $\rho(\beta,\mu)$ which should be reliable in the considered low-density regime. Then, $\mu(\beta,\rho)$ is determined *via* a numerical inversion, and we compute the pressure along a low-density isochore that covers the fully ionized regime at high temperatures, the atomic phase at intermediate temperatures, and the molecular phase at lower temperatures (see Sec. 3). Our predictions are compared to numerical results of quantum Monte Carlo simulations [6] and to the tabulated OPAL EOS [8] derived using the ACTEX method [9]. In Sec. 4, we make some concluding comments.

2 Method

2.1 Recombination and screening

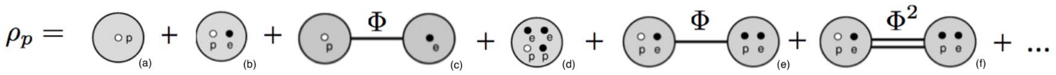


Fig. 1 A few SC diagrams in the Screened Cluster representation of the proton density. Bonds between particle clusters can be Φ, $\frac{1}{2!}\Phi^2$ or $\frac{1}{3!}\Phi^3$.

The Screened Cluster (SC) representation is a resummed activity series expressed in terms of Mayer-like diagrams built with particle clusters and a screened effective potential Φ [5]. In Fig. 1, we show a few diagrams in the SC representation of the proton density $\rho_\mathrm{p} = \rho_\mathrm{e} = \rho$. The SC diagrams account, simultaneously and consistently, for

- **Recombination**: The statistical weights of particle clusters $Z(N_p, N_e)$ incorporate the contributions of bound states, because both quantum mechanics and Coulomb interactions are treated non-perturbatively.
- **Screening**: All large-distance contributions of diffusive states are finite, because effective interactions between particle clusters are screened.

When the density tends to zero, bare Coulomb Hamiltonians $H_{N_\mathrm{p},N_\mathrm{e}} = -\sum_{i=1}^{N} \frac{\hbar^2}{2m_{\alpha_i}}\Delta_i + \frac{1}{2}\sum_{i\neq j} \frac{e_{\alpha_i}e_{\alpha_j}}{|\mathbf{r}_i-\mathbf{r}_j|}$ (where $\alpha_i = \mathrm{p}, \mathrm{e}$ is the species of the ith particle, e_{α_i} its charge and m_{α_i} its mass) emerge at leading order in the statistical weights of particle clusters. The corresponding cluster partition functions $Z(N_\mathrm{p}, N_\mathrm{e})$ reduce to finite truncated traces $\mathrm{Tr}[\exp(-\beta H_{N_\mathrm{p},N_\mathrm{e}} - \ldots)]$. The counter-terms substracted from Gibbs operator $\exp(-\beta H_{N_\mathrm{p},N_\mathrm{e}})$ cancel out long-range divergences in the trace, while their own *a priori* divergent contributions remain finite thanks to the introduction of the screened potential Φ. That exact construction, which gives rise in particular to diagrams made up of two or more clusters connected by bonds Φ, $\frac{1}{2!}\Phi^2$ or $\frac{1}{3!}\Phi^3$, is systematically performed through the whole activity series [5]. Chemical species naturally emerge in cluster partition function $Z(N_\mathrm{p}, N_\mathrm{e})$, which involves contributions from both bound and diffusive states.

The individual contributions of an atom H arise in diagram (b) shown in Fig. 1, and they are embedded in cluster partition function $Z(1,1)$. The contributions of the counter-terms associated to the truncation in $Z(1,1)$ are embodied in diagrams $\mathrm{p}-\frac{1}{n!}\Phi^n-\mathrm{e}$ with $n = 1, 2, 3$ (for instance diagram (c) when $n = 1$). Cluster partition function $Z(1,1)$ is in fact merely related to Ebeling function $Q(\beta)$. Similarly, hydrogen molecules H_2 emerge in diagram (d), and their contributions are embedded in cluster partition function $Z(2,2)$. All possible species resulting from molecular dissociation, namely p, e, H, H_2^+ and H^- arise in the truncation of $\exp(-\beta H_{2,2})$. If cluster partition function $Z(2,2)$ contains contributions from all thermally excited states, the molecular ground-state contribution prevails in the zero-temperature limit.

Diagrams with bonds between the clusters describe screened interactions between the involved entities. Graph (f) accounts for instance for long-range van der Waals interactions between two H atoms, while graph (d) contains contributions linked to their short-range interactions.

2.2 Computation of the EOS at low densities

Along a low-density isochore, $\rho < 10^{-2}$ g/cc and for not too low temperatures $T > 10^3$ K, we have estimated the SC diagrams that provide the most important contributions to the density $\rho(\beta, z)$ where $\beta = 1/(kT)$ and $z = \exp(\beta\mu)$ is the fugacity [2]. It turns out that the relevant diagrams account for ideal contributions of ionized charges, atoms H, molecules H_2, plasma polarization and interactions between ionized charges and atoms H. All other diagrams, in particular those describing ions H_2^+ and H^- or interactions between molecules can be neglected. This provides a quite reasonable approximation for $\rho(\beta, z)$, which takes a polynomial form in $\sqrt{z}$ with temperature-dependent coefficients. Most coefficients can be calculated analytically, while we used a simple approximation for $Z(2, 2)$ introduced in Ref. [13].

The pressure $P(\beta, \rho)$ is then computed as function of temperature and density by eliminating numerically the fugacity z between equations $\rho = \rho(\beta, z)$ and $\beta P = z\partial/\partial z\rho(\beta, z)$. Notice that, by construction, the present calculation reproduces the first terms of the known exact asymptotic expansions, namely, on the one hand, the virial expansion [10, 11] valid in the almost fully ionized regime $\rho \to 0$ at T fixed, and, on the other hand, the Scaled Low Temperature (SLT) expansion [4] valid in the Saha atomic regime $T \to 0$ with $\rho \sim e^{-|E_H|/(kT)}$.

3 Numerical results

3.1 At very low densities

At very low densities, molecules are very scarce along the full isochore, except at very low temperatures. Then, the pressure remains close to the predictions of Saha theory which describes an ideal mixture of ionized protons, ionized electrons and atoms in their groundstate. We plot in Fig. 2 the deviation $\beta(P - P_{\rm Saha})/\rho$, on a logarithmic scale, along a very-low density isochore $\rho = 10^{-10}$ g/cc. That deviation agrees with the first two simple terms P_1 (plasma polarization) and P_2 (molecular recombination) of the fully analytical SLT expansion [4], apart at very low temperatures where the pressure approaches $\rho kT/2$ as expected.

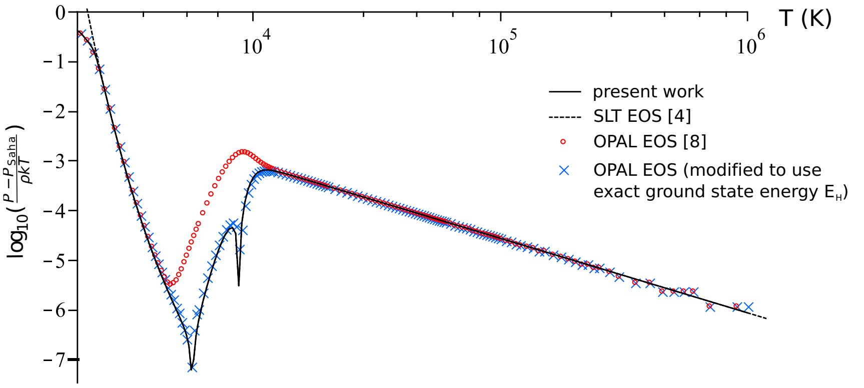


Fig. 2 Log plot of deviations to the ideal Saha pressure along isochore 10^{-10} g/cc. Our predictions (solid line) are compared to the analytical SLT expansion (dashed line) and to the tabulated OPAL equation of state (circles). The crosses correspond to the OPAL EOS modified to account for the finite mass of the proton in the binding energy of the hydrogen atom, i.e. using $E_H = -me^4/(2\hbar^2) \simeq -13.5983$ eV instead of $E_H \simeq 1$ Ry $= -m_e e^4/(2\hbar^2) \simeq -13.65057$ eV. We corrected the OPAL values only at the level of the ideal terms: $P_{\rm OPAL,\ corr} = P_{\rm OPAL} + P_{\rm Saha}[E_H] - P_{\rm Saha}[1\ \rm Ry]$.

3.2 At low and moderate densities

Now we consider an isochore at moderate densities, typically $\rho = 10^{-3}$ g/cc as shown in Fig. 3. At high temperatures, entropy wins and the gas is almost fully ionized, with pressure close to its ideal value $2\rho kT$. When temperature is lowered, energy enters into the game, so atomic recombination takes place and a first

---

[2] As rigorously proved in Ref. [12], only the mean chemical potential $\mu = (\mu_p + \mu_e)/2$ of protons and electrons is relevant in the thermodynamical limit, and it entirely determines the common particle density $\rho = \rho_{\rm p} = \rho_{\rm e}$.

plateau appears with pressure close to the ideal expression ρkT for a purely atomic gas. If we further lower the temperature, energy favors molecules since they are more stable than atoms. Then a second plateau appears with pressure close to the ideal value $\rho kT/2$ for a purely molecular gas.

The cross-over between the plasma and atomic phases is well captured by the thermal effects contained in function Q which suitably corrects the predictions of Saha theory. In that region, the agreement between our calculations, the OPAL EOS [8] and Path Integral Monte Carlo (PIMC) simulations [6] is very good.

At temperatures lower than about 10^4 K, the cross-over between the atomic and molecular phases is crucially controlled by function $Z(2,2)$. Our predictions deviate then somewhat from those of OPAL, while the numerical results of PIMC simulations have rather large uncertainties.

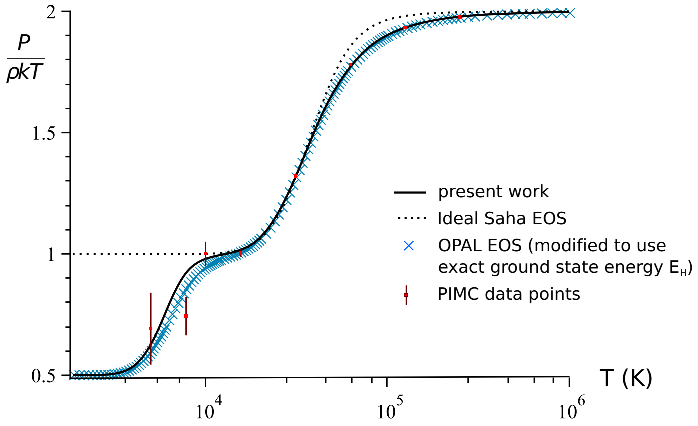


Fig. 3 Pressure along isochore 10^{-3} g/cc. Crosses correspond to the OPAL EOS, as in Fig. 1. Numerical results of PIMC simulations [6] are also shown with their error bars.

Along the two considered isochores, the ionized charges remain always weakly coupled. The order of magnitude of the coupling constant Γ can be estimated within the reasonable *ad-hoc* definition of the density of ionized charges as the sum of all SC diagrams (see Fig. 1) for which the root proton is the sole particle in the root cluster. For isochore $\rho = 10^{-3}$ g/cc, this provides $\Gamma_{\max} = 0.24$ which is reached when $T \simeq 28\,760$ K.

Notice that the order of magnitude of the cross-over temperatures between the successive ionized, atomic and molecular phases can be estimated from a simple criterion on the chemical potential, inspired by the atomic and molecular limit theorems at zero temperature [14, 15], which amounts to compare ideal densities of the corresponding chemical species in their groundstate. This provides $T \simeq 9000$ K (isochore 10^{-10} g/cc) and $T \simeq 40000$ K (isochore 10^{-3} g/cc) for atomic recombination, while $T \simeq 1700$ K (isochore 10^{-10} g/cc) and $T \simeq 3200$ K (isochore 10^{-3} g/cc) for molecular recombination. The comparison to our results displayed in Figs. 2 and 3 shows that the location of the cross-over between the plasma and the atomic gas is roughly recovered, while the discrepancy is larger for the transition to the molecular gas. This is not unexpected, since hydrogen molecules are obviously highly excited rotationally at the temperatures considered in Fig. 3.

4 Conclusions and perspectives

Within the physical picture, we have derived a reliable EOS which works rather well at low densities over a broad range of temperatures. Its key ingredients are Ebeling function Q and cluster partition function $Z(2,2)$, which only depend on temperature T and on the fundamental constants $\hbar, e, m_{\mathrm{p}}, m_{\mathrm{e}}$. Those functions are essential for describing properly the plasma-atomic and atomic-molecular cross-over regions.

Because of our poor knowledge of the exact spectrum of four-body Hamiltonian $H_{2,2}$, we had to use an approximate form for function $Z(2,2)$. An obvious improvement would be to better estimate that function, for instance through computationally exact numerical methods. Moreover, further studies of its analytical properties would be helpful for deriving simple modelizations of $Z(2,2)$. Such modelizations could then be used for constructing suitable internal molecular partition functions involved in chemical approaches applied to higher-density regimes.

References

[1] W. Däppen, J. Phys. A: Math. Gen. **39**, 4441 (2006).
[2] W. Ebeling, Ann. Phys. Leipz. **19**, 104 (1967).
[3] W. Ebeling, M. Steinberg, J. Ortner, Eur. Phys. J. D **12**, 513-520 (2000).
[4] A. Alastuey, V. Ballenegger, F. Cornu and Ph.A. Martin, J. Stat. Phys. **130**, 1119-1176 (2008).
[5] A. Alastuey, V. Ballenegger, F. Cornu and Ph.A. Martin, J. Stat. Phys. **113**, 455-503 (2003).
[6] B. Militzer and D. M. Ceperley, Phys. Rev. E **63**, 066404 (2001).
[7] A. Alastuey and V. Ballenegger, J. Phys. A: Math. Theor. **42**, 214031 (2009).
[8] F.J. Rogers and A. Nayfonov, Astrophys. J. **576**, 1064-1074 (2002).
[9] F.J. Rogers, Phys. Review A **24**, 1531 (1981).
[10] W.D. Kraeft, D. Kremp, W. Ebeling, and G. Röpke, Quantum Statistics of Charged Particle Systems (Plenum Press, New York, 1986).
[11] A. Alastuey and A. Perez, Europhys. Lett. **20**, 19-24 (1992).
[12] E.H. Lieb and J. Lebowitz, Adv. Math. **9**, 316-398 (1972).
[13] A. Alastuey and V. Ballenegger, Contrib. Plasma Phys. **50**, 46-53 (2010).
[14] D.C. Brydges and Ph.A. Martin, J. Stat. Phys. **96**, 1163-1330 (1999).
[15] I. Iosilevskiy, Contrib. Plasma Phys. **49**, 713-717 (2009).

MOLECULAR PHYSICS, 20 MARCH 2004, VOL. 102, NO. 6, 599–609

Structure and dielectric properties of polar fluids with extended dipoles: results from numerical simulations

V. BALLENEGGER* and J.-P. HANSEN

Department of Chemistry, University of Cambridge, Lensfield Road,
Cambridge CB2 1EW, UK

(*Received 8 December 2003; revised version accepted 16 February 2004*)

The strengths and shortcomings of the point dipole model for polar fluids of spherical molecules are illustrated by considering the physically more relevant case of extended dipoles formed by two opposite charges $\pm q$ separated by a distance d (dipole moment $\mu = qd$). Extensive molecular dynamics simulations on a high-density dipolar fluid are used to analyse the dependence of the pair structure, dielectric constant ϵ and dynamics as a function of the ratio d/σ (σ is the molecular diameter), for a fixed dipole moment μ. The point dipole model is found to agree well with the extended dipole model up to $d/\sigma \simeq 0.3$. Beyond that ratio, ϵ shows a non-trivial variation with d/σ. When $d/\sigma > 0.6$, a transition is observed towards a hexagonal columnar phase; the corresponding value of the dipole moment is found to be substantially lower than the value of the point dipole required to drive a similar transition.

1. Introduction

Highly polar fluids are particularly important in many areas of physical chemistry, chemical engineering and biology, because of their role as solvents leading to electrolyte and polyelectrolyte dissociation. Water is of course the most important among polar liquids, but because of its complex behaviour, primarily linked to the formation of hydrogen-bond networks, much theoretical work has focused on simpler models involving spherical molecules with point dipoles. The best known and most widely studied examples are dipolar hard spheres (DHS), dipolar soft spheres (DSS), and the Stockmayer model (dipolar + Lennard-Jones interactions). A long-standing problem, going back to the classic work of Onsager [1] and Kirkwood [2], is to relate the dielectric response of a polar fluid to molecular dipole fluctuations and correlations (for reviews, see [3]). Subtle conceptual and numerical problems arise in molecular dynamics or Monte Carlo simulations of finite samples of polar fluids, which are linked to the infinite range of the dipolar interactions, so that boundary conditions must be treated adequately. These issues were resolved in the early 1980s, for both the reaction field and the Ewald summation implementations of boundary-conditions [4–6]. Despite this theoretical progress, accurate estimates of the dielectric permittivity of simple polar fluids by numerical simulation remain a very challenging task, because large fluctuations of the total dipole moment of the sample occur on a relatively long time scale (of the order of 10 ps), leading to a very slow convergence rate for the dielectric constant [7, 8] (see also § 3).

More recently, it was realized that simple dipolar liquids can form a ferroelectric nematic phase for sufficiently large dipole moments [9–11]. This transition is intimately related to the formation of chains of dipoles aligned head-to-tail, which prevent the formation of a proper liquid phase in the Stockmayer model if the dispersive energy is below a certain threshold [12].

However, point dipoles represent a limiting situation, never achieved in real polar molecules, which are characterized by extended charge distributions linked to electronic charge transfer from electron donors to electron acceptor atoms. In simple heteronuclear diatomic molecules such as CO or HF, this situation can be modelled by assigning fractional charges of opposite sign to sites that are separated by a distance d, typically of the order of 0.1 nm [13]. Such situations, or more complicated ones involving more than two atoms, can be mimicked by adding higher-order point multipoles to a point dipole [14], but such an expansion will require more and more high-order multipoles as two molecules approach each other.

In this paper, we present a systematic investigation of the structure, dielectric response and phase behaviour of a simple model involving spherical molecules carrying extended (rather than point) dipoles resulting from opposite charges $\pm q$, each displaced symmetrically by

---

*Author for correspondence. e-mail: vcb25@cam.ac.uk

Molecular Physics ISSN 0026–8976 print/ISSN 1362–3028 online © 2004 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/00268970410001675554

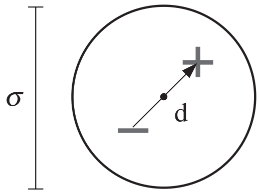


Figure 1. A polar molecule with an extended dipole moment.

a distance $d/2$ from the centre of the molecule. We study how the properties of the polar liquid change when d is increased from zero, varying q simultaneously so that the dipole moment $|\boldsymbol{\mu}| = qd$ remains constant. Although polar molecules are never spherical, the model investigated in this paper, which focuses on the electrostatic rather than steric interactions, is the simplest 'natural' extension of the dipolar sphere model towards a more realistic representation of highly polar fluids. Some studies on the structure of similar models with extended dipoles have been published previously, but without an investigation of their bulk dielectric properties [15, 16].

2. The model and simulation details

We consider a polar fluid made up of spherical molecules with two embedded point charges $\pm q$ located at $\pm \mathbf{d}/2$ from the centre of the sphere (see figure 1). The distance $|\mathbf{d}|$ is assumed fixed, so the molecule is not polarizable and carries a permanent dipole moment $\boldsymbol{\mu} = q\mathbf{d}$.

Placing the origin at the centre of the sphere, the multipole moments $q_{lm} = \int Y^*_{lm}(\theta, \phi) r^l \rho(\mathbf{r})\, \mathrm{d}^3\mathbf{r}$, where $\rho(\mathbf{r}) = q\delta(\mathbf{d}/2) - q\delta(-\mathbf{d}/2)$ is the molecular charge distribution [17], are

$$q_{lm} = \begin{cases} 2q\left(\dfrac{d}{2}\right)^l \sqrt{\dfrac{2l+1}{4\pi}} & \text{if l odd and m} = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The next non-vanishing moment after the dipole is thus the octopole, since the quadrupole moment vanishes by symmetry for this choice of origin.

The interaction energy between two molecules is given by the sum of a Lennard-Jones interaction

$$V_{\mathrm{LJ}}(r) = 4u\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{2}$$

and the four Coulombic energies due to the point charges. On figure 2, the electrostatic energy at contact is compared to a truncated multipolar expansion containing the dipole–dipole and dipole–octopole interactions. The configuration of lowest energy occurs when the molecular dipoles are aligned head-to-tail ($\theta = \theta_2 = 0$). This minimum energy is lower for extended than for point dipoles.

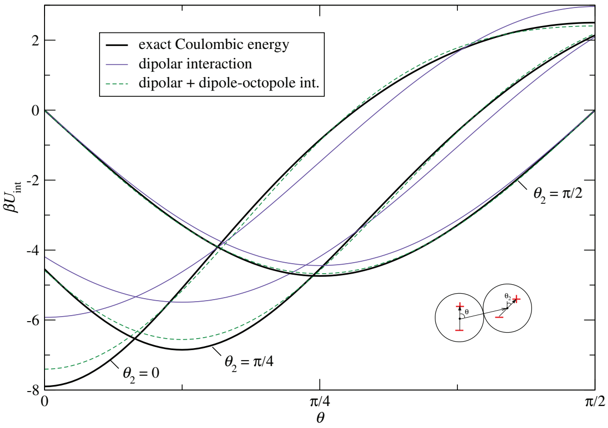


Figure 2. Electrostatic interaction energy of two molecules at contact ($|\mathbf{r}| = \sigma$) for $d = \sigma/2$.

A thermodynamic state of the fluid is specified by the values of the dimensionless parameters:

- reduced density: $\rho^* = \rho\sigma^3$,
- reduced temperature: $T^* = kT/u$,
- reduced dipole moment: $\mu^* = \sqrt{\mu^2/\sigma^3 u}$,
- reduced dipole elongation: $d^* = d/\sigma$.

We studied the influence of dipole elongation on properties of a dense highly polar fluid phase characterized by $\rho^* = 0.82$, $T^* = 1.15$, $\mu^* = 1.82$.† The reduced moment of inertia of our molecules was $I^* = I/m\sigma^2 = 0.117$, but equilibrium quantities, such as the dielectric constant and distribution functions, are independent of I^*. We also performed simulations of a dipolar *soft* sphere fluid at $\rho^* = 0.8$, $T^* = 1.35$ and $\mu^* = 2$. This thermodynamic state point of the DSS fluid has been extensively studied by Kusalik in the case of point dipoles [7, 18].

In all calculations, we employed periodic boundary conditions. We choose the spherical geometry, that is the periodic replications of the basic cubic simulation cell form an infinite sphere, which is itself embedded in an infinite region of dielectric constant ϵ'. In this case, the Hamiltonian of the system is

$$H = \sum_{i<j=1}^{N} \left(V_{\mathrm{LJ}}(\mathbf{r}_{ij}) + q_i q_j \Psi(\mathbf{r}_{ij})\right) - \frac{\kappa}{\pi^{1/2}} \sum_{i=1}^{N} q_i^2 + \frac{2\pi \mathbf{M}^2}{(2\epsilon' + 1)L^3}, \tag{3}$$

---

† Our parameters in dimensioned units were $T = 300\,\mathrm{K}$, $\mu = 2.45\,\mathrm{D}$, $\sigma = 0.3668\,\mathrm{nm}$, $m = 10\,\mathrm{u}$, $I = 0.156\,\mathrm{u\,nm}^2$, $u = 2.1747\,\mathrm{kJ\,mol}^{-1}$.

Table 1. Influence of dipole elongation on some properties of a Stockmayer fluid at $\rho^* = 0.82$, $T^* = 1.15$ and $\mu^* = 1.82$.

d/σ	ϵ	τ_M (ps)	τ_μ (ps)	$D\,(10^{-5}\,\mathrm{cm}^2/\mathrm{s})$	p^*	U^*
0.02	99.6(± 1.4)	2.21	0.50	11.7	0.42	−10.1
0.3	98.4(± 1.5)	2.64	0.54	11.6	0.45	−10.2
0.4	94.0(± 1.5)	2.88	0.63	11.5	0.45	−10.3
0.5	92.4(± 1.7)	4.14	0.88	10.6	0.42	−10.6
0.6	102.3(± 3.2)	11.44	2.01	8.7	0.26	−11.7
0.61	104.7(± 3.6)	13.97	2.36	8.5	0.23	−11.9
0.62	100.8(± 3.5)	14.81	2.79	7.9	0.19	−12.1

where L is the side of the box, $\mathbf{M} = \sum_i q_i \mathbf{r}_i$ is the total dipole moment, and

$$\Psi(\mathbf{r}) = \sum_{\mathbf{n}\in\mathbb{Z}^3} \frac{\operatorname{erfc}(\kappa|\mathbf{r}+\mathbf{n}L|)}{|\mathbf{r}+\mathbf{n}L|} + \frac{1}{\pi L}\sum_{\mathbf{n}\neq 0} \frac{1}{|\mathbf{n}|^2}\exp\left(\frac{-\pi^2|\mathbf{n}|^2}{\kappa^2 L^2} + \frac{2\pi i}{L}\mathbf{n}\cdot\mathbf{r}\right). \quad (4)$$

The last term in (3) accounts for the work done against the depolarizing field created by surface charges induced on the spherical boundary. This term vanishes only for metallic boundary conditions ($\epsilon' = \infty$). The Ewald sums in $\Psi(\mathbf{r})$ were evaluated using the smooth particle mesh Ewald method [19] (Ewald coefficient $\kappa = 3.4705\,\mathrm{nm}^{-1}$, grid size $32 \times 32 \times 32$, interpolation order 6). The interactions were truncated beyond 0.9 nm, both for the real space Ewald sum and for the Lennard-Jones interactions.

Molecular dynamics simulations were carried out using the simulation package *gromacs* [20]. The equations of motion were integrated using the so-called leap-frog algorithm with a reduced time step of $\mathrm{d}t^* = \mathrm{d}t/(m\sigma^2/u)^{1/2} = 0.0025$. The temperature was kept constant using a Berendsen thermostat. Equilibration periods lasted at least 100 ps (50 000 time steps), and were followed by data-producing runs of 8 ns or more. The number of molecules was 512 in calculations of the dielectric constant (§ 3), and 5555 in calculations of correlation functions (§ 4).

3. Dipole fluctuations and dielectric constant

The dielectric constant of a homogeneous and isotropic fluid can be calculated from the fluctuation formula (see, for example, [21])

$$\frac{(\epsilon - 1)(2\epsilon' + 1)}{2\epsilon' + \epsilon} = \frac{4\pi}{3V}\frac{\langle \mathbf{M}^2 \rangle}{kT}, \quad (5)$$

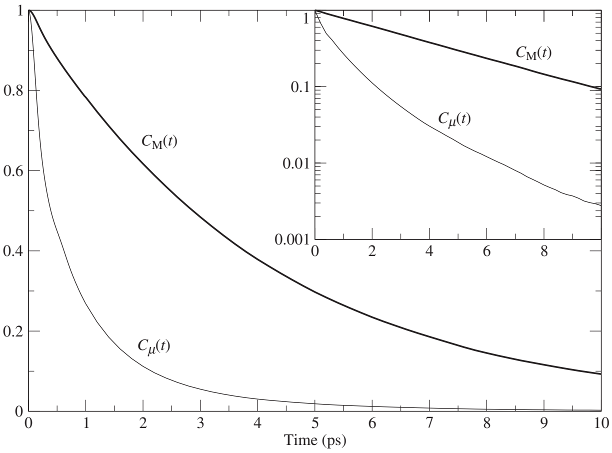


Figure 3. Autocorrelation functions $C_\mathbf{M}(t)$ and $C_\boldsymbol{\mu}(t)$, for the value $d = \sigma/2$. The inset shows a logarithmic plot, confirming the exponential behaviour of $C_\mathbf{M}(t)$.

which holds for a macroscopic spherical sample of volume V surrounded by a medium of dielectric constant ϵ'. The results obtained for the dielectric constant are independent of the choice of ϵ', provided the boundary term in equation (3) is properly taken into account. We employed metallic boundary conditions, because they are known to produce smaller uncertainties in estimates of ϵ than finite values of ϵ' [5, 22] (see also below). The fluctuation formula reduces in this case to

$$\epsilon = 1 + 3y\,\langle g \rangle, \quad g = \frac{M^2}{N\mu^2}, \quad (6)$$

where the dimensionless parameter $y = 4\pi\beta\rho\mu^2/9 \simeq 3.31$ at the state point under consideration.

Table 1 shows the influence of the dipole elongation on some properties of the Stockmayer fluid, namely on the dielectric constant ϵ, the diffusion constant D, the dielectric relaxation times τ_M and τ_μ, the reduced configurational energy $U^* = U/(Nu)$, and the reduced pressure $p^* = p\sigma^3/u$. The diffusion constant was calculated from Einstein's relation

$$\langle |\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2 \rangle = 6Dt, \qquad t \to \infty. \quad (7)$$

The relaxation times τ_M and τ_μ were determined from the autocorrelation functions $C_\mathbf{M}(t) = \langle \mathbf{M}(t) \cdot \mathbf{M}(0) \rangle /$

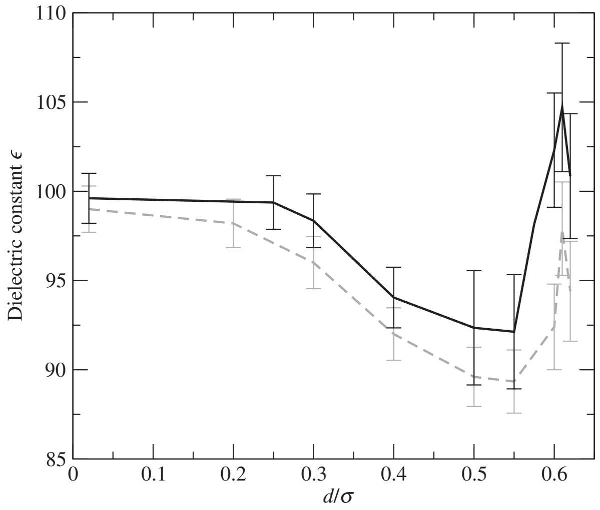


Figure 4. Dielectric constant of a Stockmayer fluid ($\rho^* = 0.82$, $T^* = 1.15$, $\mu^* = 1.82$, continuous line) and of a dipolar soft sphere fluid $\rho^* = 0.8$, $T^* = 1.35$, $\mu^* = 2$, dashed line) as a function of dipole extension.

$\langle M^2 \rangle$ and $C_{\boldsymbol{\mu}}(t)$ (see figure 3). For $t > 0.3\,\mathrm{ps}$, $C_{\mathbf{M}}(t)$ exhibits an exponential decay $\exp(-t/\tau_M)$ typical of a Debye dielectric. The relaxation of $C_{\boldsymbol{\mu}}(t)$ is not well fitted by a single exponential, and the corresponding relaxation time was estimated from the integral of $C_{\boldsymbol{\mu}}(t)$.

We show in figure 4 the variation of the dielectric constant with dipole elongation for a Stockmayer fluid and for a dipolar soft sphere fluid. For almost point dipoles ($d^* = 0.02$), our result for the dielectric constant of the DSS fluid is in good agreement with the value 98 ± 2 reported by Kusalik *et al.* [23]. At the state point under consideration, the Stockmayer fluid has almost the same dielectric constant: $\epsilon_{\text{point}} = 99.6 \pm 1.4$. Our data show that when d^* increases, the dielectric constant decreases and reaches a minimum about 6% lower than ϵ_{point} at $d^* \simeq 0.55$. When d^* is further increased, the dielectric constant increases rapidly above ϵ_{point}, up to the critical distance $d^*_{\text{c}} \simeq 0.63$. At this critical distance, a phase transition occurs from an isotropic fluid to an orientationally ordered 'liquid crystal' phase (see § 5).

The simulations show that the point dipole model gives a reliable estimate of the dielectric constant over a very wide range of extensions d, namely up to the point where the system undergoes a phase transition. The weak sensitivity of the dielectric constant on the extension of the dipole, which contrasts with the large sensitivity observed in water models [24], may be due to the absence of a quadrupole moment in our molecules.

It is clear from table 1 that the dynamics of the fluid slows down when d is increased: the diffusion coefficient D drops and the relaxation times increase. This slowdown is due to the formation of head-to-tail dipolar chains in the system. Their entanglement makes these chains less mobile than individual molecules in the present high-density regime.

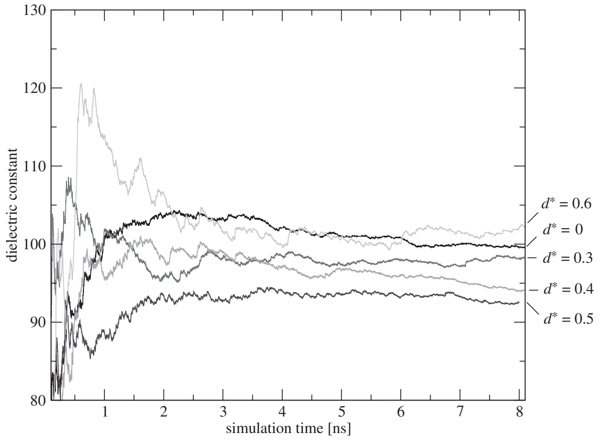


Figure 5. Convergence of ϵ with simulation time, for dipole elongations $d^* = d/\sigma = 0$, 0.3, 0.4, 0.5 and 0.6.

Long runs were needed to obtain even moderate accuracy (about 2%) in the estimated dielectric constants. Figure 5 shows the running estimate of ϵ as a function of simulation time. The slow convergence, especially for large elongations of the dipole, can be traced back to the long relaxation times τ_M, as shown by the following error analysis.

By definition, the probability distribution of the sample having a total dipole moment of magnitude M and arbitrary orientation is given by

$$P(M) \propto 4\pi M^2 \mathrm{e}^{-\beta F(M)}, \tag{8}$$

where $F(M)$ is the free energy of the system. From macroscopic electrostatics, the energy of a spherical dielectric sample, of dielectric constant ϵ and carrying a uniform polarization M/V, is

$$U(M) = \frac{2\pi M^2}{V} \frac{2\epsilon' + \epsilon}{(\epsilon - 1)(2\epsilon' + 1)}, \tag{9}$$

where ϵ' is the dielectric constant of the surrounding medium. Following Kusalik [25], we combine equations (8) and (9) with the approximation $F(M) \simeq F(0) + U(M)$. This leads to the following expression for the probability distribution of fluctuations $g = M^2/(N\mu^2)$:

$$P(g) = Ag^{1/2}\,\mathrm{e}^{-\kappa g}, \quad \kappa \equiv \frac{9y}{2} \frac{2\epsilon' + \epsilon}{(\epsilon - 1)(2\epsilon' + 1)}, \tag{10}$$

where the normalization constant is $A = 2(\kappa^3/\pi)^{1/2}$. The mean of this distribution is $\langle g \rangle = 3/(2\kappa)$, in agreement with the fluctuation formula (5). Though the distribution (10) neglects changes in entropy and is valid only in the linear regime, it gives a good description of fluctuations of the total dipole moment observed in simulations of highly polar fluids [23].

The dielectric relaxation time τ_M gives a time scale for two measurements of M^2 to be independent. In a simulation of total duration t, the distribution (10) is thus sampled $n \simeq t/\tau_M$ times. After n such independent measurements, the standard deviation in the average $\sum_{i=1}^{n} g_i/n$ of the g factor is $\sigma_{g,n} = \sigma_g/n^{1/2}$ where $\sigma_g^2 = \langle (g - \langle g \rangle)^2 \rangle = 3/(2\kappa^2)$ is the variance of the distribution (10). The expected relative uncertainty in the g factor,

$$I_{\langle g \rangle} \equiv \frac{\sigma_{g,n}}{\langle g \rangle} = \sqrt{\frac{2}{3n}} = \sqrt{\frac{2}{3}\frac{\tau_M}{t}}, \tag{11}$$

depends therefore on the boundary condition ϵ' only via the relaxation time τ_M. Solving (5) for ϵ, one has

$$\epsilon - 1 = \frac{3y\langle g \rangle (2\epsilon' + 1)}{2\epsilon' + 1 - 3y\langle g \rangle}. \tag{12}$$

By the rules of propagation of errors, the relative uncertainty in the dielectric constant minus one is thus

$$I_{\epsilon-1} = \frac{2\epsilon' + \epsilon}{2\epsilon' + 1}\sqrt{\frac{2}{3}\frac{\tau_M}{t}}. \tag{13}$$

The error bars in figure 4 were determined from this formula, and are in agreement with the fluctuations observed in figure 5. In a Debye dielectric, the relaxation time τ_M is related to the Debye relaxation time τ_D (which is independent of boundary conditions) by [26]

$$\tau_M = \frac{2\epsilon' + 1}{2\epsilon' + \epsilon}\tau_D. \tag{14}$$

Inserting (14) into (13), we see that larger values of ϵ' will lead to smaller uncertainties in the dielectric constant. This explains the faster convergence of ϵ observed when using metallic boundary conditions [5, 22].

According to the present analysis, the slow convergence of ϵ, as determined from the fluctuation formula, is due to the large value of the Debye dielectric relaxation time [8] and the rather broad distribution $P(g)$. Moreover, the uncertainties in ϵ are independent of system size, as long as it is macroscopic. In large systems, it may therefore be favourable to determine ϵ from correlation functions rather than from the fluctuation formula.

4. Structure

4.1. *The pair distribution function*

The pair distribution function $h(1,2) = h(\mathbf{r}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ of the infinite system can be expanded in rotational invariants [27]:

$$h(1,2) = h^{000}(r) + h^{110}(r)\Phi^{110}(1,2) + h^{112}(r)\Phi^{112}(1,2) + \cdots,$$

where

$$\Phi^{110}(1,2) = \hat{\boldsymbol{\mu}}_1 \cdot \hat{\boldsymbol{\mu}}_2, \tag{15}$$

$$\Phi^{112}(1,2) = 3(\hat{\boldsymbol{\mu}}_1 \cdot \hat{\mathbf{r}})(\hat{\boldsymbol{\mu}}_2 \cdot \hat{\mathbf{r}}) - \hat{\boldsymbol{\mu}}_1 \cdot \hat{\boldsymbol{\mu}}_2. \tag{16}$$

The functions $\Phi^{l_1 l_2 l}$ form an orthogonal basis for the angular dependence of $h(1,2)$. The first projections are

$$h^{000}(r) = \langle h(1,2) \rangle_{\mu_1,\mu_2} = g(r) - 1, \tag{17}$$

$$h^{110}(r) = 3\langle h(1,2)\Phi^{110}(1,2) \rangle_{\mu_1,\mu_2}, \tag{18}$$

$$h^{112}(r) = \frac{3}{2}\langle h(1,2)\Phi^{112}(1,2) \rangle_{\mu_1,\mu_2}, \tag{19}$$

where $\langle \cdots \rangle_\mu = \int \cdots \mathrm{d}\Omega_\mu / 4\pi$ denotes an unweighted angular average over the orientations of $\boldsymbol{\mu}$.

Plots of $h^{000}(r)$ and $h^{112}(r)$ are shown in figure 6 for three elongations d of the dipole. As d is increased, the stronger multipolar moments carried by the molecules lead to a slight reduction of the fluid structure as measured by the centre-to-centre distribution $g(r)$, but more orientational order, as measured by the projections $h^{112}(r)$ and $h^{110}(r)$ (the latter projection, not shown in the figure, closely resembles $h^{112}(r)$).

The projection $h^{112}(r)$ is related to the dielectric constant of the fluid by the formula

$$\lim_{r\to\infty} r^3 h^{112}(r) = \frac{(\epsilon - 1)^2}{\epsilon}\frac{1}{4\pi\rho y}, \tag{20}$$

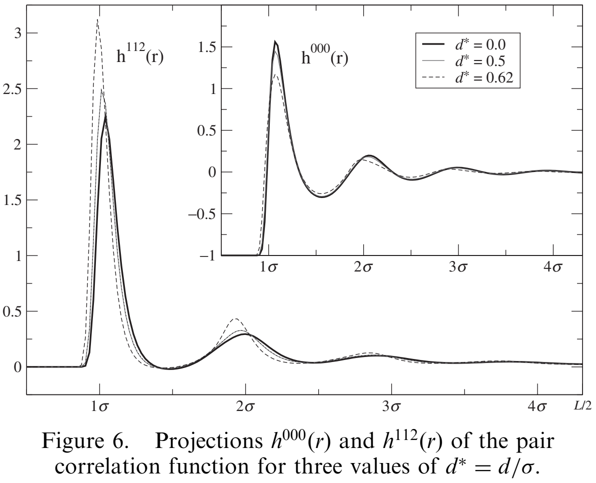


Figure 6. Projections $h^{000}(r)$ and $h^{112}(r)$ of the pair correlation function for three values of $d^* = d/\sigma$.

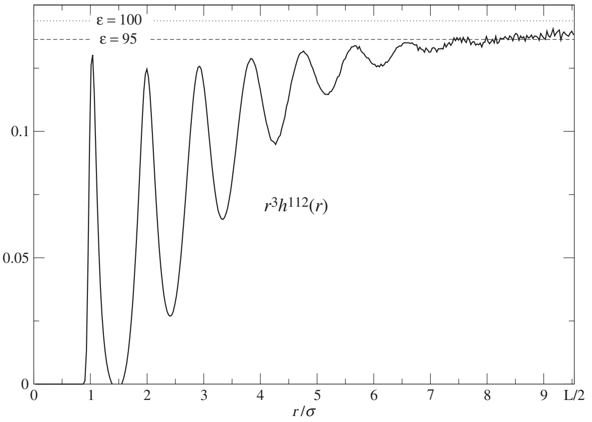


Figure 7. Convergence of $r^3h^{112}(r)$ at large distances towards the limit (20). Data from a 6 ns long simulation of a system of 5555 molecules ($\rho^* = 0.82$, $T^* = 1.15, \mu^* = 1.82$, $d^* = 0.5$).

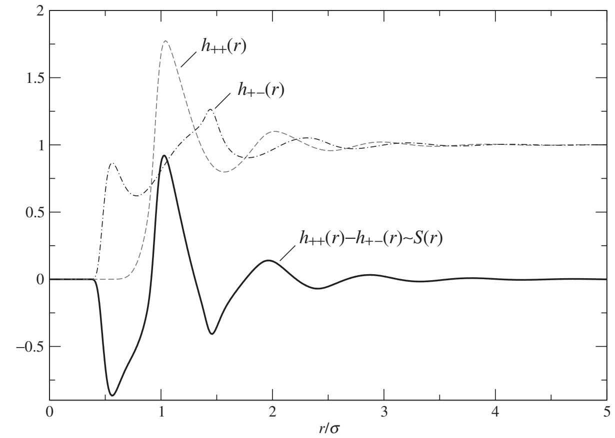


Figure 9. Site–site distribution functions. Same system as in figure 7.

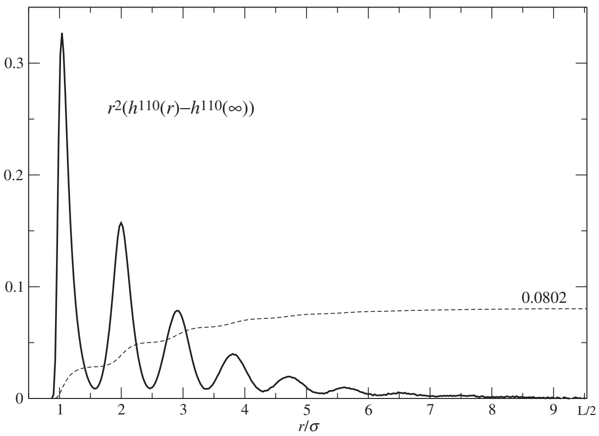


Figure 8. The function $r^2(h^{110}(r) - h^{110}(\infty))$ and its integral (dashed line). Same system as in figure 7.

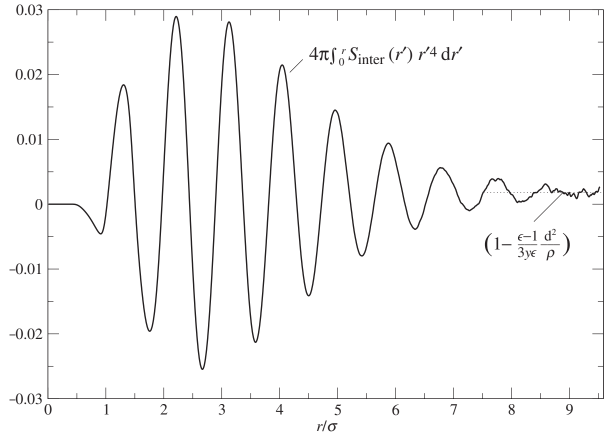


Figure 10. Integration of the second moment of $S_{\text{inter}}(r)$.

first derived by Nienhuis and Deutch [28]. A 512-molecule system with a half box size of $L/2\sigma = 4.3$ is too small to reach the asymptotic limit (20). The results for the correlation function shown in figures 7–10 were hence obtained using a larger system (simulation of 5555 molecules during 6 ns) under the same conditions ($\rho^* = 0.82$, $T^* = 1.15$, $\mu^* = 1.82$, $d = \sigma/2$, $\epsilon' = \infty$). Now $L/2\sigma = 9.55$, and figure 7 shows that $r^3h^{112}(r)$ does reach the asymptotic value (20) at a distance $r \simeq 7\sigma$, as in the case of point dipoles [18]. In [18], it was observed that $r^3h^{112}(r)$ drops sharply for r greater than $L/2$, even when the reduced size of the volume element is properly taken into account in the normalization. In a system with long range forces, care must be exercised in the interpretation of correlation functions at distances larger than half the box length (since the Ewald potential differs strongly from the Coulomb potential at these distances). Caillol analysed this problem carefully, and derived a formula for the asymptotic behaviour of $h^{112}(r)$ valid for distances up to $\sqrt{2}L/2$ [29]. Estimations of the dielectric constant from equation (20) become more accurate when the size of the system is increased, contrary to estimations based on the fluctuation formula.

The projection $h^{110}(r)$ is also related to the dielectric constant, since the fluctuation formula (5) can be written in the Kirkwood form†

$$\frac{(\epsilon-1)(2\epsilon'+1)}{2\epsilon'+\epsilon} = 3y\left(1+\frac{4\pi\rho}{3}\int_0^\infty h_{\epsilon'}^{110}(r)r^2\,dr\right). \quad (21)$$

---

† For a cubic simulation cell, the upper limit in the integral becomes $3^{1/2}L/2$.

Since the LHS of equation (21) depends on ϵ', the projection $h^{110}(r)$ must also be sensitive to this boundary condition, whence the introduction of a subscript ϵ'. Figure 8 shows a plot of $r^2 h^{110}_{\infty}(r)$, and the integral of this function, for the same system as in figure 7. The correlations extend up to $r \simeq 7\sigma$, just as in the case of $h^{112}(r)$. This distance corresponds to the scale beyond which the fluid behaves as a continuum dielectric and obeys the equations of macroscopic electrostatics.

Though formula (21) is equivalent to (5), it is worthwhile to discuss how the pair correlation function depends on the dielectric constant of the external medium. This problem has been addressed in [28] (see also the perturbation theory presented in [5]); here, we hope to give a clear and concise answer to the above question, using simple physical arguments to justify the formulas.

In a spherical sample, $h^{110}_{\epsilon'}(r)$ is in fact the only projection, among all the $h^{l_1 l_2 l}$, to be strongly affected by the boundary condition ϵ'. This is due to the surface term in the Hamiltonian (3), which corresponds to an interaction energy between two molecules

$$\frac{4\pi}{2\epsilon'+1}\frac{\boldsymbol{\mu}_1\cdot\boldsymbol{\mu}_2}{V}, \tag{22}$$

which has the angular dependence of $\Phi^{110}(1,2)$.

As the interaction (22) is independent of the distance between the molecules, $h^{110}_{\epsilon'}(r)$ does not decay in general to zero at infinity, but rather to an ϵ'-dependent constant. We will prove below that this constant is

$$\lim_{r\to\infty} h^{110}_{\epsilon'}(r) = \frac{2}{V}\frac{(\epsilon-1)^2}{3\rho y\epsilon}\frac{\epsilon'-\epsilon}{2\epsilon'+\epsilon} \tag{23}$$

in the spherical geometry [5, 28, 29]. The constant (23) vanishes only when using the boundary condition $\epsilon' = \epsilon$, which mimics an infinite sample, or in the thermodynamic limit $V \to \infty$ (which is never reached in simulations). The limit (23) is achieved in practice at distances large compared to the decay length of $h^{110}(r)$, but small compared to the size of the system.

The fact that $h^{110}_{\epsilon'}(r)$ contains the $\mathcal{O}(1/V)$ constant contribution (23) at large distances ensures that Kirkwood formula (21) gives consistent results for different boundary conditions. Indeed, when $h^{110}_{\epsilon'}(r)$ is integrated over the volume $V = 4\pi R^3/3$ of a large sample, as in the RHS of equation (21), the constant (23) gives a finite contribution to the integral that is precisely what is required to match the ϵ'-dependence of the LHS of the equation. In other words, equations (22) and (23) imply that

$$4\pi\int_0^R h^{110}_{\epsilon'}(r) r^2\,\mathrm{d}r = 4\pi\int_0^R h^{110}_{\epsilon}(r) r^2\,\mathrm{d}r + V h^{110}_{\epsilon'}(\infty),$$

when the samples are large enough (i.e. in the limit $R \to \infty$). When this identity is inserted into Kirkwood formula (21), it is immediately clear that the predicted dielectric constant is independent of ϵ'.

In order to prove equation (23) with simple physical arguments, we need to recall two basic results from the statistical mechanics of polar liquids. The first result is the expression of the density $\rho(\mathbf{r},\boldsymbol{\mu})$ of molecules at $\mathbf{r}$ with orientation $\boldsymbol{\mu}$ in a polarized sample permeated by a macroscopic field $\mathbf{E}(\mathbf{r})$: [30]

$$\rho(\mathbf{r},\boldsymbol{\mu}) = \frac{\rho}{4\pi}\left(1 + \frac{\epsilon-1}{3y}\beta\boldsymbol{\mu}\cdot\mathbf{E}(\mathbf{r})\right) + \mathcal{O}(E^2) \tag{24}$$

(y is defined after equation (6)). This formula is consistent with the constitutive relation $\mathbf{P}(\mathbf{r}) = (\epsilon-1)/(4\pi)\mathbf{E}(\mathbf{r})$ of macroscopic electrostatics, since the average polarization density in the fluid is by definition $\mathbf{P}(\mathbf{r}) = \int \rho(\mathbf{r},\boldsymbol{\mu})\boldsymbol{\mu}\,\mathrm{d}\Omega_{\mu}$.

The second result we need is the expression of the effective dipole moment $\boldsymbol{\mu}^{\mathrm{eff}}$ of a polar molecule held fixed in a polar liquid ($\boldsymbol{\mu}^{\mathrm{eff}}$ is defined as $\boldsymbol{\mu}$, the dipole moment of the fixed molecule, plus the total dipole moment of the screening cloud around $\boldsymbol{\mu}$). One may first think naively that $\boldsymbol{\mu}^{\mathrm{eff}} = \boldsymbol{\mu}/\epsilon$: the fluid would screen the dipole according to its dielectric constant. But this would be treating the polar fluid as a dielectric continuum everywhere, including in the interior of the fixed molecule, which is obviously wrong. An exact statistical mechanical calculation shows that the right answer is [31]

$$\boldsymbol{\mu}^{\mathrm{eff}} = \frac{\epsilon-1}{3y\epsilon}\boldsymbol{\mu}. \tag{25}$$

(The expression $(\epsilon-1)/3y\epsilon$ can itself be interpreted as being composed of a factor $(\epsilon-1)/3y$ arising from local correlations around $\boldsymbol{\mu}$, times the expected factor $1/\epsilon$ due to screening by distant molecules). With these two results in mind, we can now understand easily formulas (20), (22) and (23).

The result (20) for the large-distance behaviour of the pair correlation function $h(1,2)$ can be seen as a straightforward consequence of the screening effect (25). By definition of the distribution functions, the density of molecules at $\mathbf{r}_2$ with orientation $\boldsymbol{\mu}_2$, when a molecule is known to be located at $1 = (\mathbf{r}_1,\boldsymbol{\mu}_1)$ is

$$\rho(2|1) = \frac{\rho(2,1)}{\rho(1)} = \frac{\rho}{4\pi}(1 + h(1,2)). \tag{26}$$

From (25), the electric field due to the fixed molecule $\boldsymbol{\mu}_1$ is equivalent, at large distances $\mathbf{r}_{12} = \mathbf{r}_2 - \mathbf{r}_1$, to that of a renormalized dipole moment $\boldsymbol{\mu}_1^{\mathrm{eff}}$. This dipolar field

$-\boldsymbol{\nabla}_2(\boldsymbol{\mu}_1^{\rm eff}\cdot\boldsymbol{\nabla}_1)|\mathbf{r}_{12}|^{-1}$ is locally uniform and weak, so we can apply the linear response result (24). Using (24) and (25), we find that

$$\rho(2|1) \underset{|\mathbf{r}_{12}|\to\infty}{\sim} \frac{\rho}{4\pi}\left(1-\frac{(\epsilon-1)^2}{9y^2\epsilon}\beta v_{\rm dip}(1,2)\right), \tag{27}$$

where $v_{\rm dip}(1,2)=(\boldsymbol{\mu}_1\cdot\boldsymbol{\nabla}_1)(\boldsymbol{\mu}_2\cdot\boldsymbol{\nabla}_2)|\mathbf{r}_{12}|^{-1}$ is the dipolar potential, and we assume an infinitely extended sample. Comparing (26) and (27), we obtain the asymptotic behaviour of the pair correlation function:

$$h(1,2) \underset{|\mathbf{r}_{12}|\to\infty}{\sim} \frac{(\epsilon-1)^2}{9y^2\epsilon}(-\beta v_{\rm dip}(1,2)). \tag{28}$$

The result (20) follows then upon inserting (28) into (19).

Formula (23) can be interpreted in a similar way, namely as arising from the interaction of a molecule with the reaction field produced by the screened dipole moment of another molecule. We recall from macroscopic electrostatics that a dipole $\boldsymbol{\mu}_1$ at the centre of a spherical sample of radius R and dielectric constant ϵ, surrounded by a dielectric medium ϵ', produces a uniform reaction field

$$\mathbf{E}_R^{[\epsilon,\epsilon']}(\boldsymbol{\mu}_1)=\frac{2}{\epsilon}\frac{\epsilon'-\epsilon}{2\epsilon'+\epsilon}\frac{\boldsymbol{\mu}_1}{R^3} \tag{29}$$

inside the sample, because of the surface charge density induced at the dielectric discontinuity [32]. In a finite spherical sample, a molecule at a position $\mathbf{r}_2$ far enough from $\boldsymbol{\mu}_1$—so that it does not disturb the screening cloud around it—will interact therefore not only with the dipolar field of $\boldsymbol{\mu}_1^{\rm eff}$, as in (27), but also with the reaction field $\mathbf{E}_R(\boldsymbol{\mu}_1^{\rm eff})$ produced by the screened dipole moment of this molecule. From (24), a term

$$\frac{\rho}{4\pi}\cdot\frac{\epsilon-1}{3y}\beta\boldsymbol{\mu}_2\cdot\mathbf{E}_R^{[\epsilon,\epsilon']}(\boldsymbol{\mu}_1^{\rm eff}), \tag{30}$$

must hence be added to (27) in a finite spherical sample. The pair correlation at large distances, equation (28), includes then the additional contribution

$$\frac{4\pi}{3V}\frac{2(\epsilon-1)^2}{9y^2\epsilon}\frac{\epsilon'-\epsilon}{2\epsilon'+\epsilon}\beta\boldsymbol{\mu}_2\cdot\boldsymbol{\mu}_1, \tag{31}$$

where we used (29), (25) and $V=4\pi R^3/3$. Formula (23) follows now from projecting (31) onto $\Phi^{110}(1,2)=\hat{\boldsymbol{\mu}}_1\cdot\hat{\boldsymbol{\mu}}_2$ according to (18).

We conclude this discussion by noting that the interaction energy (22) used in the simulations, or equivalently the surface term in the Hamiltonian (3), can also be interpreted in terms of a reaction field effect. Indeed, each polar molecule in the sample will create a reaction field, acting on itself and on all other molecules, that are given by equation (29) with $\epsilon=1$. Since the Ewald sums (4) give the electrostatic energy between the molecules in the case of a sample surrounded by a metal ($\epsilon'=\infty$), the correction to this energy to be used in the Hamiltonian of a spherical system with boundary condition ϵ' is

$$\frac{1}{2}\sum_{i,j=1}^{N}(-\boldsymbol{\mu}_i)\cdot\left[\mathbf{E}_R^{[1,\epsilon']}(\boldsymbol{\mu}_j)-\mathbf{E}_R^{[1,\infty]}(\boldsymbol{\mu}_j)\right]=\frac{2\pi\mathbf{M}^2}{(2\epsilon'+1)v}, \tag{32}$$

in agreement with (3).

4.2 *Site–site correlations*

Contrary to the point dipolar fluid model, the present model with extended dipoles has well-defined site–site distribution functions $h_{++}(r)=h_{--}(r)$ and $h_{+-}(r)$ [33]. From these we get the charge–charge correlation function $S(r)=S_{\rm intra}(r)+S_{\rm inter}(r)$, where

$$S_{\rm inter}(r)=2q^2\rho^2(h_{++}(r)-h_{+-}(r)) \tag{33}$$

describes the intermolecular correlations, while

$$S_{\rm intra}(r)=2q^2\rho\delta(\mathbf{r})-2q^2\rho\frac{\delta(|\mathbf{r}|-d)}{4\pi d^2} \tag{34}$$

is the intramolecular correlation function for a molecule with a rigid dipole of extension d. The charge–charge correlation is of special interest, because it satisfies the two sum rules [34]:

$$\text{neutrality}: \int S(r)\,{\rm d}^3\mathbf{r}=0, \tag{35}$$

$$\text{Stillinger–Lovett}: \frac{1}{\epsilon}=1+\frac{2\pi\beta}{3}\int {\rm d}^3\mathbf{r}\,\mathbf{r}^2\,S(r). \tag{36}$$

The site–site correlation functions and $S(r)$ are shown on figure 9 for $d=\sigma/2$. The charge neutrality sum rule is found to be accurately satisfied:

$$\rho\int_0^\infty(h_{++}(r)-h_{+-}(r))r^2\,{\rm d}r\simeq 8.8\times10^{-5}. \tag{37}$$

The Stillinger–Lovett sum rule allows in principle the determination of the dielectric constant from $S(r)$, but this route is not practicable in a computer simulation, because of the unfavourable ratio $(1-\epsilon)/\epsilon$ which saturates for large ϵ, and also because it is difficult to determine the second moment of $S_{\rm inter}(r)$ accurately. Figure 10 shows, however, that equation (36) is satisfied within the uncertainties of our data.

5. Orientational order

When the molecular dipole has an extension greater than $d \simeq 0.64\sigma$, the simulations show spontaneous formation of orientationally ordered phases, starting from random initial configurations. At the state point under consideration ($\rho^* = 0.82$, $T^* = 1.15$), we observed phase coexistence between a dense liquid crystal and a very dilute gas. In order to deal with pure phases, we performed simulations at constant pressure ($p^* = p\sigma^3/u = 0.22$), rather than constant volume, for $d \geq 0.62\sigma$.

The occurrence of orientational order was monitored by computing two order parameters. The rank-one order parameter P_1 is defined as

$$P_1 = \frac{\langle M_\parallel \rangle}{N\mu}, \qquad (38)$$

where $M_\parallel \equiv \mathbf{M} \cdot \hat{\mathbf{n}}$ is the projection of the total dipole moment along the director $\mathbf{n}$ ($P_1 = 1$ for a completely polarized system). The second-rank order parameter P_2 is the largest eigenvalue of the matrix

$$Q_{\alpha\beta} = \frac{1}{N\mu^2}\left\langle \sum_{i=1}^{N} \frac{1}{2}\left(3\mu_i^\alpha \mu_i^\beta - \mu^2 \delta_{\alpha\beta}\right)\right\rangle, \qquad (39)$$

where μ_i^α is the α component of the vector $\boldsymbol{\mu}_i$. The corresponding eigenvector $\mathbf{n}$ is the director. $P_2 = 1$ when all dipoles are oriented parallel to $\mathbf{n}$ or $-\mathbf{n}$.

Table 2 lists our results for these order parameters, as well as for the dielectric tensor $\boldsymbol{\epsilon} = (\epsilon_\parallel, \epsilon_\perp)$. In a liquid crystal with director $\mathbf{n}$, the latter is determined by the following generalization of equation (6):

$$\epsilon_\parallel = 1 + y\,\frac{\langle M_\parallel^2 \rangle - \langle M_\parallel \rangle^2}{N\mu^2}, \qquad (40)$$

and a similar equation for $\epsilon_\perp$ in terms of the perpendicular fluctuations $\left(\langle \mathbf{M}_\perp^2 \rangle - \langle \mathbf{M}_\perp \rangle^2\right)/N\mu^2$.

When d is increased above the critical distance $d_c \simeq 0.63\sigma$, the order parameter P_2 jumps from essentially zero to about almost one, indicating a first-order transition to a highly orientationally ordered phase. Figure 11 provides snapshots of the simulation cell for $d = 0.64\sigma$. It is clear from the snapshots that the molecules are associated into columns, composed of chains of dipoles oriented head-to-tail. These columns are all parallel to the director, and are arranged in a hexagonal lattice in the perpendicular plane. The simulations for $d > 0.64\sigma$ yielded similar liquid crystal phases with columnar order, each with a different

Table 2. Constant pressure simulations of a Stockmayer fluid at $\rho^* = 0.22$, $T^* = 1.15$ and $\mu^* = 1.82$. Data from 8 ns long simulations of 512 molecules, collected after an equilibration period that lasted up to 10 ns. For $d \geqslant 0.64\sigma$, the system is a ferro-electric liquid crystal with columnar order. Uncertainties in $\epsilon_\parallel$ and $\epsilon_\perp$ are about ± 0.01 and ± 0.02 respectively.

d/σ	$\langle \rho^* \rangle$	P_1	P_2	$\epsilon_\parallel$	$\epsilon_\perp$
0.62	0.80	0.08	0.07	$\epsilon = 103.8(\pm 4)$	
0.63	0.80	0.09	0.08	$\epsilon = 112.6(\pm 5)$	
0.64	0.95	0.97	0.91	1.16	1.46
0.65	0.98	0.66	0.91	1.28	1.56
0.66	1.03	0.98	0.94	1.02	1.43

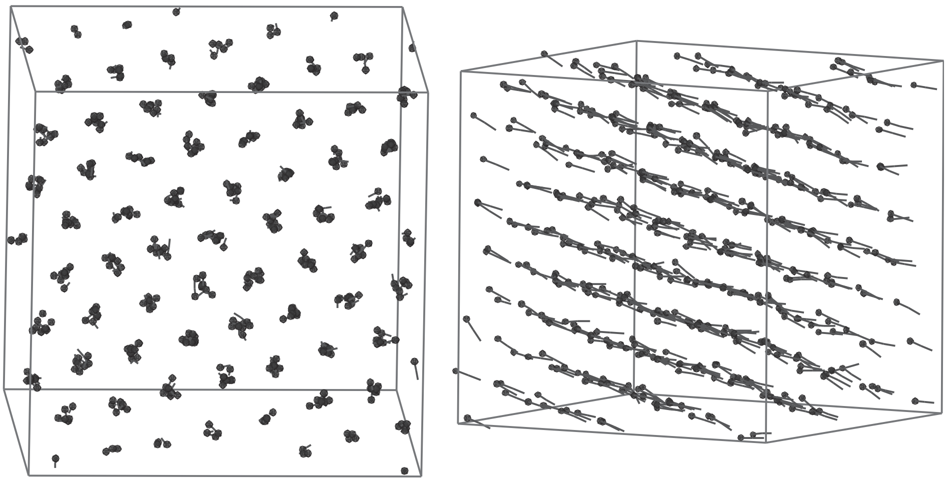


Figure 11. Snapshots of the simulation cell of the Stockmayer fluid in the columnar phase. The hexagonal lattice in the plane orthogonal to the director is apparent in the first snapshot. The dipoles are represented by a line joining the minus charge (shown as a small bead) to the plus charge.

orientation of the director. The system shows strong spatial correlations in the direction of the director, but it is still fluid in that direction, as indicated by the mean-square displacement of the molecules. The latter increases indeed linearly with time, with a diffusion constant of about $D_\parallel \simeq 0.09 \times 10^{-5}\,\mathrm{cm}^2\,\mathrm{s}^{-1}$.

In some runs (not listed in table 2), the system formed two liquid crystal domains characterized by different orientations of the director. As the transition to a single domain is expected to occur on a time scale much larger than our simulation time (8 ns), since it requires the collective motion of many molecules, we included in table 2 only results from runs where a single domain formed spontaneously in less than 10 ns. Most runs yielded fully polarized liquid crystals (P_1 close to 1). It is likely that the lower value of P_1 measured in the case $d = 0.65\sigma$ is due to insufficient sampling of phase space: the probability of a column inverting its orientation during our simulation time is indeed very small.

As the column configuration is energetically more favourable for extended than for point dipoles (see figure 2), it is not surprising that liquid crystal columnar phases form at a much lower dipolar coupling constant $\lambda = \mu^{*2}/T^*$ than previously reported for point dipolar fluid models; here $\lambda \sim 2.9$, while columnar phases were observed in the dipolar soft sphere fluid at $\lambda = 9$, and in the dipolar hard sphere fluid at $\lambda = 6.25$ [9, 35]. A nematic ferroelectric liquid phase has been identified in the Stockmayer model at $\lambda \simeq 4$ (a columnar phase has not been seen previously in this model to our knowledge) [36, 37]. The hexagonal lattice arrangement found here is to be contrasted with the square lattice reported in [35].

6. Conclusion

We have extended the considerable body of earlier work on dipolar fluids by replacing the usual point dipole on spherical molecules by physically more relevant extended dipoles obtained by placing two opposite charges symmetrically with respect to the centre. The structural, dielectric and dynamical behaviour was monitored as the spacing d of the charges was increased, keeping the total dipole moment $\mu = qd$ fixed. Periodic boundary conditions were used with proper Ewald summations of the Coulombic interactions within an infinitely large sphere bounded by a dielectric medium of permittivity ϵ'. The key findings are the following:

a) Runs of several nanoseconds, longer than in most previous studies, were required to obtain estimates of the dielectric constant ϵ within about 2% using the standard fluctuation formula (5). A careful error analysis shows that 'metallic' boundary conditions (where $\epsilon' = \infty$ at infinity) yield optimal estimates of ϵ.

b) The values of ϵ deduced from the h^{112} and h^{110} correlation functions agree with the fluctuation results within statistical errors, provided a sufficiently large simulation cell is used to obtain proper estimates of the asymptotic behaviour of these correlation functions. The strong influence of the boundary condition ϵ' on the projection $h^{110}(r)$ arises from the interaction of the polar molecules with the reaction field to the dielectric discontinuity between the fluid and the external medium ϵ'. When $\epsilon' \neq \epsilon$, $h^{110}(r)$ does not decay to zero at large distances, but rather to a finite constant of order $1/V$ [5, 28, 29]. The value of this constant (equation (23)) was derived using simple physical arguments based on macroscopic electrostatics and linear response theory.

c) ϵ has a non-trivial dependence on the ratio $d^* = d/\sigma$. Up to $d^* \simeq 0.25$, ϵ agrees with the point dipole result within statistical uncertainties, thus illustrating the practical usefulness of the point dipole limit. For $d^* \gtrsim 0.3$, ϵ drops to a minimum value roughly 6% below the point dipole result when $d^* \simeq 0.55$. When d^* is further increased, ϵ increases sharply and reaches a maximum at $d^* \simeq 0.6$.

d) For still larger extensions d^*, the system is seen to undergo a transition, at constant pressure, to an orientationally ordered state similar to a columnar phase with a hexagonal ordering in the plane orthogonal to the director. This phase is characterized by large values of the usual orientational order parameters P_1 and P_2. At the same time the dielectric tensor becomes anisotropic, and the mean dielectric constant is very low ($\epsilon \simeq 1.4$), signalling the strong suppression of dipole moment fluctuations. The transition to the columnar phase occurs at a value of the dipole moment well below that required to observe the transition with point dipoles [9, 35].

e) The dynamics, characterized by the relaxation times τ_M and τ_μ of the total and individual dipole moments, as well as by the self-diffusion constant D, slows down very significantly as d^* increases, due to the enhanced tendency of the system to form parallel strings, which eventually lead to the columnar phase. In the latter, the diffusion coefficient $D_\parallel$ parallel to the director is about two orders of magnitude smaller than D in the isotropic phase, but still substantial, showing that the system behaves like a one-dimensional fluid.

The present results are for a single high-density $\rho^* = 0.82$, and a single pressure in the anisotropic phase ($p^* = 0.22$, corresponding to $\langle \rho^* \rangle \simeq 1$). Clearly more work is needed to be able to map out a complete phase diagram of the Stockmayer fluid, in view of the additional variable d^*. The present work illustrates the strengths and deficiencies of the point–dipole model. Many simple molecular systems fall in the region $d^* \simeq 0.5$, where the deviations from point dipole behaviour begin to be substantial.

The authors thank Joachim Dzubiella, Reimar Finken, Michael Klein and Mark Miller for fruitful discussions. The financial support of the Swiss National Science Foundation is also gratefully acknowledged.

References

[1] Onsager, L., 1936, *J. Am. chem. Soc.*, **58,** 1486.
[2] Kirkwood, J., 1939, *J. chem. Phys.*, **7,** 911.
[3] Stell, G., Patey, G. N., and Høye, J. S., 1981, *Adv. Chem. Phys.*, **48,** 183; Madden, P. A., and Kivelson, D., 1984, *Adv. Chem. Phys.*, **56,** 467.
[4] Patey, G. N., Levesque, D., and Weis, J. J., 1982, *Molec. Phys.*, **45,** 733.
[5] de Leeuw, S. W., Perram, J. W., and Smith, E. R., 1980, *Proc. R. Soc. Lond.* A, **373,** 27; *ibid.*, **373,** 57; *ibid.*, **388,** 177.
[6] Neumann, M., 1983, *Molec. Phys.*, **50,** 841.
[7] Kusalik, P. G., 1990, *J. chem. Phys.*, **93,** 3520.
[8] Kurtović, Z., Marchi, M., and Chandler, D., 1993, *Molec. Phys.*, **78,** 1155.
[9] Wei, D., and Patey, G. N., 1992, *Phys. Rev. Lett.*, **68,** 2043.
[10] Weis, J. J., and Levesque, D., 1993, *Phys. Rev.* E, **48,** 3728.
[11] Groh, B., and Dietrich, S., 1994, *Phys. Rev. Lett.*, **72,** 2422.
[12] van Leeuwen, M. E., and Smit, B., 1993, *Phys. Rev. Lett.*, **71,** 3991.
[13] Klein, M. L., and McDonald, I. R., 1979, *J. chem. Phys.*, **71,** 298.
[14] Wei, D., 1995, *Molec. Crystal Liq. Crystal*, **269,** 89.
[15] Ciccotti, G., Ferrario, M., Hynes, J. T., and Kapral, R., 1989, *Chem. Phys.*, **129,** 241.
[16] Hertzner, A. W., Schoen, M., and Morgner, H., 1991, *Molec. Phys.*, **73,** 1011.
[17] Jackson, J. D., 1998, *Classical Electrodynamics*, 3rd edn (New York: John Wiley and Sons).
[18] Kusalik, P. G., 1991, *Molec. Phys.*, **73,** 1349.
[19] Essman, U., Perela, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G., 1999, *J. chem. Phys.*, **103,** 8577.
[20] Lindahl, E., Hess, B., and van der Spoel, D., 2001, *J. molec. Mod.*, **7,** 306 (http://www.gromacs.org).
[21] Frenkel, D., and Smit, B., 2002, *Understanding Molecular Simulation*, 2nd edn (London: Academic Press).
[22] Wang, Z., Holm, C., and Müller, H. W., 2003, *J. chem. Phys.*, **119,** 379.
[23] Kusalik, P. G., Mandy, M. E., and Svishchev, I. M., 1994, *J. chem. Phys.*, **100,** 7654.
[24] Höchtl, P., Boresch, S., Bitomsky, W., and Steinhauser, O., 1998, *J. chem. Phys.*, **109,** 4927.
[25] Kusalik, P. G., 1993, *Molec. Phys.*, **80,** 225.
[26] Neumann, M., and Steinhauser, O., 1983, *Chem. Phys. Lett.*, **102,** 508.
[27] Gray, C. G., and Gubbins, K. E., 1984, *Theory of Molecular Fluids*, Vol. 1 (Oxford: Clarendon Press).
[28] Nienhuis, G., and Deutch, J. M., 1971, *J. chem. Phys.*, **55,** 4213.
[29] Caillol, J. M., 1992, *J. chem. Phys.*, **96,** 7039.
[30] Alastuey, A., and Ballenegger, V., 2000, *Physica* A, **279,** 268.
[31] Finken, R., Ballenegger, V., and Hansen, J.-P., 2003, *Molec. Phys.*, **101,** 2559.
[32] Fröhlich, H., 1949, *Theory of Dielectrics* (Oxford: Oxford University Press).
[33] Hansen, J.-P., and McDonald, I. R., 1986, *Theory of Simple Liquids*, 2nd edn (London: Academic Press).
[34] Martin, Ph. A., 1988, *Rev. Mod. Phys.*, **60,** 1075.
[35] Weis, J. J., Levesque, D., and Zarragoicoechea, G. J., 1992, *Phys. Rev. Lett.*, **69,** 913.
[36] Stevens, M. J., and Grest, G. S., 1995, *Phys. Rev.* E, **51,** 5976.
[37] Gao, G. T., and Zeng, X. C., 2000, *Phys. Rev.* E, **61,** R2188.

THE JOURNAL OF CHEMICAL PHYSICS **122**, 114711 (2005)

Dielectric permittivity profiles of confined polar fluids

V. Ballenegger and J.-P. Hansen
Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom

(Received 22 September 2004; accepted 10 November 2004; published online 22 March 2005)

The dielectric response of a simple model of a polar fluid near neutral interfaces is examined by a combination of linear response theory and extensive molecular dynamics simulations. Fluctuation expressions for a local permittivity tensor $\epsilon(\mathbf{r})$ are derived for planar and spherical geometries, based on the assumption of a purely local relationship between polarization and electric field. While the longitudinal component of ϵ exhibits strong oscillations on the molecular scale near interfaces, the transverse component becomes ill defined and unphysical, indicating nonlocality in the dielectric response. Both components go over to the correct bulk permittivity beyond a few molecular diameters. Upon approaching interfaces from the bulk, the permittivity tends to increase, rather than decrease as commonly assumed, and this behavior is confirmed for a simple model of water near a hydrophobic surface. An unexpected finding of the present analysis is the formation of "electrostatic double layers" signaled by a dramatic overscreening of an externally applied field inside the polar fluid close to an interface. The local electric field is of opposite sign to the external field and of significantly larger amplitude within the first layer of polar molecules. © *2005 American Institute of Physics*. [DOI: 10.1063/1.1845431]

I. INTRODUCTION

The dielectric permittivity of a medium is a macroscopic concept which is defined by the relationship between the polarization **P** and the electric field **E** inside the medium.[1] When the dielectric medium is inhomogeneous over distances much larger than molecular scales, a space-dependent (local) permittivity $\epsilon(\mathbf{r})$ may be defined when dealing with mesoscopic electrostatic problems. The question of how far towards molecular scales a local permittivity remains a meaningful concept, and how $\epsilon(\mathbf{r})$ is related to dipolar fluctuations is a long-standing problem[2] which we have recently addressed in the case of a polar fluid near a sharp interface.[3] We showed that a necessary condition for the existence of a meaningful, statistical definition of a local permittivity is that the local electric field inside the medium does not vary appreciably on the scale of the molecular correlation length, as already noted by Nienhuis and Deutch.[2]

The ability to give a clear-cut statistical definition of a local permittivity is crucial for any coarse-graining strategy, whereby large parts of a complex multicomponent system are treated as continuous dielectric media, while the remaining parts are described in molecular detail. An important example is provided by implicit solvent models of biomolecular assemblies, where water is considered as a continuous dielectric medium, characterized by a local permittivity in the immediate vicinity of biomolecules or membranes. A spatially varying permittivity then determines the electrostatic interactions between charged residues and ions.[4] Conversely one may wish to describe a polar solvent trapped within a dielectric matrix, as in the case of water confined between membranes or clay platelets, or within narrow pores. In these circumstances it may be advantageous to describe the confining matrix as a dielectric continuum, while the confined polar liquid is modeled with molecular resolution.[5]

In this paper we consider the case of polar fluids confined by continuous dielectric media characterized by a permittivity ϵ'. We relate the local dielectric permittivity $\epsilon(\mathbf{r})$ to the dipolar fluctuations within the inhomogeneous fluid, along the lines of the classic Kirkwood–Fröhlich (KF) linear response treatment of the bulk permittivity.[6,7] More specifically, we shall consider the cases of a simple polar fluid in an infinite slab confined by two semi-infinite dielectric media and of a polar fluid confined to a spherical cavity inside a uniform, macroscopic dielectric continuum. Numerical results based on long Molecular Dynamics (MD) simulations will illustrate the limitations of the concept of a local permittivity in the two geometries.

All considerations in this paper will be restricted to sharp interfaces. Like most previous theoretical and numerical work in the field, the present coarse-grained treatment suffers from the inconsistency of ignoring the molecular graininess of the confining media, while using a fully molecular description of the polar fluid.

II. POLARIZATION IN LINEAR RESPONSE

Consider a classical fluid at temperature $T=1/(k_B\beta)$, made up of N polar molecules carrying dipole moments $\boldsymbol{\mu}_i$, confined to a cavity of arbitrary shape and volume V, carved out of a macroscopic dielectric medium of uniform permittivity ϵ'. The molecules may be polarizable; their interactions are arbitrary at short distances, but tend towards the dipolar interaction at larger distances. The microscopic polarization density is

$$\mathbf{m}(\mathbf{r}) = \sum_{i=1}^{N} \boldsymbol{\mu}_i \delta(\mathbf{r}-\mathbf{r}_i), \tag{1}$$

where $\mathbf{r}_i$ is the position of the ith molecule inside the cavity. The corresponding total dipole moment is

$$\mathbf{M} = \int_{\mathcal{D}_{\text{cavity}}} \mathbf{m}(\mathbf{r})\mathrm{d}\mathbf{r} = \sum_i \boldsymbol{\mu}_i. \tag{2}$$

Let $\mathbf{P}_0(\mathbf{r}) = \langle \mathbf{m}(\mathbf{r}) \rangle$ be the average local polarization of the fluid in the absence of an externally applied electric field (by definition, $\mathbf{E}'$ is the field far away from the cavity, created by charges at $|\mathbf{r}| \to \infty$) $\mathbf{E}'$. In an isotropic phase, $\mathbf{P}_0(\mathbf{r}) = 0$ for points in the bulk of the fluid. Close to the confining boundaries, $\mathbf{P}_0(\mathbf{r})$ is nonzero in general, but may vanish for symmetry reasons, as in the case of linear polar molecules confined in a slab or a spherical cavity (see Sec. III). When a uniform external field is applied to the system, it induces a polarization density defined by

$$\Delta\mathbf{P}(\mathbf{r}) = \mathbf{P}(\mathbf{r}) - \mathbf{P}_0(\mathbf{r}) = \langle \mathbf{m}(\mathbf{r}) \rangle_{\mathbf{E}'} - \langle \mathbf{m}(\mathbf{r}) \rangle = \frac{\int [\mathbf{m}(\mathbf{r}) - \langle \mathbf{m}(\mathbf{r}) \rangle] \exp[-\beta(U_{\epsilon'}(1,\ldots,N) - \mathbf{M}\cdot\mathbf{E}_c)]\mathrm{d}1\cdots\mathrm{d}N}{\int \exp[-\beta(U_{\epsilon'}(1,\ldots,N) - \mathbf{M}\cdot\mathbf{E}_c)]\mathrm{d}1\cdots\mathrm{d}N}, \tag{3}$$

where we have used the short-hand notation i for the degrees of freedom of the ith molecule. For linear nonpolarizable molecules, $i = (\mathbf{r}_i, \boldsymbol{\mu}_i)$ reduces to the position and orientation of the permanent dipole moment, and integration with phase space element $\mathrm{d}i = \mathrm{d}^3\mathbf{r}_i \mathrm{d}\Omega_{\boldsymbol{\mu}_i}$ is performed over all possible positions and orientations of the molecule inside the cavity. $U_{\epsilon'}$ is the total interaction energy of the N polar molecules of the fluid within the cavity in the absence of $\mathbf{E}'$; it depends obviously on the permittivity ϵ' of the surrounding dielectric. The instantaneous total dipole moment $\mathbf{M}$ couples to the cavity field $\mathbf{E}_c$, i.e., the electric field inside the cavity in the absence of polar fluid, when the external (applied) field in the embedding dielectric is $\mathbf{E}'$. The two fields are related by the usual boundary conditions of macroscopic electrostatics.

Let $\Delta\mathbf{E}(\mathbf{r}) = \mathbf{E}(\mathbf{r}) - \mathbf{E}_0(\mathbf{r})$ be the difference between the mean local electric field inside the cavity, due to the external field and all the dipoles, and the mean electric field when no external field is applied [note that $\mathbf{E}_0(\mathbf{r}) = 0$ if $\mathbf{P}_0(\mathbf{r}) = 0$ everywhere]. Then, within the linear regime [i.e., for not too strong $\Delta\mathbf{E}(\mathbf{r})$], the induced polarization density is related to $\Delta\mathbf{E}(\mathbf{r})$ via

$$\Delta\mathbf{P}(\mathbf{r}) = \frac{1}{4\pi} \int_{\mathcal{D}_{\text{cavity}}} \underline{\chi}(\mathbf{r}, \mathbf{r}') \cdot \Delta\mathbf{E}(\mathbf{r}')\mathrm{d}\mathbf{r}', \tag{4}$$

where $\underline{\chi}$ is the dielectric susceptibility tensor. In the slow modulation limit, i.e., for slowly varying $\Delta\mathbf{E}(\mathbf{r})$, the integral factorizes approximately, and Eq. (4) reduces to the local form

$$\Delta\mathbf{P}(\mathbf{r}) = \frac{1}{4\pi} \underline{\chi}(\mathbf{r}) \cdot \Delta\mathbf{E}(\mathbf{r}), \tag{5}$$

where, formally, $\underline{\chi}(\mathbf{r}, \mathbf{r}') = \underline{\chi}(\mathbf{r})\delta(\mathbf{r} - \mathbf{r}')$. The local permittivity tensor is defined by

$$\underline{\chi}(\mathbf{r}) = \underline{\epsilon}(\mathbf{r}) - \underline{I}. \tag{6}$$

Linearization of Eq. (3) with respect of $\mathbf{E}_c$ leads to the following relation between the components of $\Delta\mathbf{P}(\mathbf{r})$ and $\mathbf{E}_c(\mathbf{r})$:

$$\Delta P_\alpha(\mathbf{r}) = \beta \sum_{\gamma = x,y,z} [\langle m_\alpha(\mathbf{r}) M_\gamma \rangle - \langle m_\alpha(\mathbf{r}) \rangle \langle M_\gamma \rangle] E^c_\gamma, \tag{7}$$

where $\alpha, \gamma = x$, y, or z and the statistical averages are understood to be taken at zero external (and hence cavity) field, i.e., with a Bolzmann weight $\exp(-\beta U_{\epsilon'})$. As expected for the linear response to a uniform external field, Eq. (7) involves the average correlation between a fluctuation in the local polarization density $\mathbf{m}(\mathbf{r})$ and a fluctuation in the global dipole moment $\mathbf{M}$ of the system, as has been recognized recently by Stern and Feller.[8] Note that $\langle \mathbf{M} \rangle$ will be zero by symmetry in all systems we shall consider.

Comparison between Eqs. (5) and (7) does not provide a fluctuation formula for $\chi(\mathbf{r})$ or $\epsilon(\mathbf{r})$, since they involve the total and cavity fields, respectively. The relation between these two fields depends on the geometry of the cavity, and can be established within macroscopic electrostatics. We consider successively the case of slab and spherical geometries.

A. Slab geometry

We consider a cavity in the form of an infinite slab where a fluid of $\rho = N/V$ polar molecules per unit volume is confined in the z direction by two infinite dielectric walls of permittivity ϵ'. The distance between the dielectric walls is L. The confined fluid is inhomogeneous in the z direction (orthogonal to the walls) only. By symmetry, the permittivity tensor reduces to the diagonal form

$$\epsilon(z) = \begin{pmatrix} \epsilon_\parallel(z) & 0 & 0 \\ 0 & \epsilon_\parallel(z) & 0 \\ 0 & 0 & \epsilon_\perp(z) \end{pmatrix}, \tag{8}$$

where $\epsilon_\parallel$ and $\epsilon_\perp$ denote the components parallel and orthogonal to the walls. Equations (5) and (6) then combine into two independent relations,

$$\mathbf{P}_\parallel(z) = \frac{\epsilon_\parallel(z) - 1}{4\pi} \mathbf{E}_\parallel(z), \tag{9a}$$

$$\Delta\mathbf{P}_\perp(z) = \frac{\epsilon_\perp(z) - 1}{4\pi}\Delta\mathbf{E}_\perp(z). \tag{9b}$$

We dropped the symbol Δ in Eq. (9a) because isotropy in the (x, y)-plane implies that $\mathbf{P}_0(\mathbf{r})$ (the average polarization in the absence of external field) has no parallel components. Using the standard boundary conditions on the normal and tangential components of the electric field, one finds the following relations between the components of the uniform external field $\mathbf{E}'$ and the cavity field $\mathbf{E}_c$,

$$\mathbf{E}_\parallel^c = \mathbf{E}'_\parallel, \quad \mathbf{E}_\perp^c = \epsilon' E'_\perp, \tag{10}$$

where $E_\parallel^c$ and $E'_\parallel$ are two-dimensional vectors in the (x,y) plane; the orthogonal components are along the z direction. Maxwell's equation $\nabla \times \mathbf{E}(z) = 0$ implies

$$\frac{\partial E_x(z)}{\partial z} = \frac{\partial E_y(z)}{\partial z} = 0, \tag{11}$$

so that $\mathbf{E}_\parallel = (E_x, E_y)$ is independent of z, i.e., $\mathbf{E}_\parallel(z) = E'_\parallel$ everywhere in space. In other words, Eq. (9a) leads to

$$\mathbf{P}_\parallel(z) = \frac{\epsilon_\parallel(z) - 1}{4\pi}\mathbf{E}'_\parallel = \frac{\epsilon_\parallel(z) - 1}{4\pi}\mathbf{E}^c_\parallel. \tag{12}$$

Comparison of Eqs. (7) and (12), together with isotropy in the (x,y) plane then leads to the desired fluctuation formula for $\epsilon_\parallel(z)$:

$$\epsilon_\parallel(z) = 1 + 2\pi\beta[\langle \mathbf{m}_\parallel(z) \cdot \mathbf{M}_\parallel \rangle - \langle \mathbf{m}_\parallel(z) \rangle \cdot \langle \mathbf{M}_\parallel \rangle]. \tag{13}$$

The orthogonal component may be determined from Maxwell's equation $\nabla \cdot \mathbf{D}(z) = 0$, where $\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P}$ is the displacement vector, leading to

$$\frac{\mathrm{d}}{\mathrm{d}z}[E_\perp(z) + 4\pi P_\perp(z)] = 0. \tag{14}$$

Integration of Eq. (14) from $-\infty$ to z yield

$$\begin{aligned} E_\perp(z) - E'_\perp &= -4\pi P_\perp(z) + 4\pi P_\perp(z = -\infty) \\ &= -4\pi P_\perp(z) + (\epsilon' - 1)E'_\perp. \end{aligned} \tag{15}$$

Substracting from Eq. (15) the same equation without external field and using Eq. (9b) gives $\epsilon_\perp(z)\Delta E_\perp(z) = \epsilon' E'_\perp = E^c_\perp$, where the second equality follows from Eq. (10). Equation (9b) may hence be rewritten as

$$P_\perp(z) = \frac{1}{4\pi}\frac{\epsilon_\perp(z) - 1}{\epsilon_\perp(z)}E^c_\perp. \tag{16}$$

Comparison of Eq. (16) with the transverse ($\alpha = \perp$) version of Eq. (7) leads to the desired fluctuation formula for $\epsilon_\perp(z)$,

$$\frac{\epsilon_\perp(z) - 1}{\epsilon_\perp(z)} = 4\pi\beta[\langle m_\perp(z) M_\perp \rangle - \langle m_\perp(z) \rangle \langle M_\perp \rangle]. \tag{17}$$

Equations (13) and (17) are the appropriate fluctuation formulas to compute the permittivity tensor $\epsilon(z)$ of a system inhomogeneous in one direction. Note that these formulas depend only implicitly, via the statistical averages weighted by the Bolzmann factor $\exp(-\beta U_{\epsilon'})$, on the permittivity ϵ' of the confining medium. This is to be contrasted with the results for spherical samples to be discussed below.

The key finding is that a local expression of the permittivity involves correlations of the local and total polarization of the form $\langle m_\alpha(z) M_\gamma \rangle$, and not of the local polarization alone, as has sometimes been wrongly assumed in the literature. This was already recognized by Stern and Feller,[8] but their expression for the permittivity tensor $\epsilon(z)$ differs from the one derived here, because they did not consider a single slab, but a system which is periodically replicated in space to form an infinite spherical array of the original slab.

We stress that formulas (13) and (17) were derived for a uniform external field under the local assumption (5). If the local assumption is not valid, definitions (9a) and (9b) become purely formal, and the permittivity $\epsilon(z)$ may take values that are unphysical and specific to the case of a uniform external field permeating a planar interface.

B. Spherical geometry

We now consider a system of N polar molecules confined to a spherical cavity of radius R, surrounded by a dielectric medium of permittivity ϵ'. We first recall the fluctuation formula for the *bulk* dielectric constant of this system. The cavity field is now $\mathbf{E}^c = 3\epsilon'\mathbf{E}'/(2\epsilon' + 1)$. For a macroscopic spherical sample of uniform permittivity ϵ, the local field, far from the boundaries, is uniform and equal to $\mathbf{E} = 3\epsilon'\mathbf{E}'/(2\epsilon' + \epsilon)$. Substituting $\mathbf{E}^c$ and $\mathbf{E}$ into Eq. (7) and into the definition (5) of the polarization, one arrives, upon identification and use of the isotropy of the system, at

$$\frac{(\epsilon - 1)(2\epsilon' + 1)}{(2\epsilon' + \epsilon)} = \frac{4\pi\epsilon}{3}[\langle \mathbf{m}(\mathbf{r}) \cdot \mathbf{M} \rangle - \langle \mathbf{m}(\mathbf{r}) \rangle \cdot \langle \mathbf{M} \rangle]. \tag{18}$$

In this formula $\mathbf{r}$ can be any point in the bulk of the sample, so that $\mathbf{m}(\mathbf{r})$ may be replaced by $\mathbf{M}/V$ if boundary effects are negligible. This leads back to the well known KF formula for the bulk dielectric constant in terms of fluctuations $(\langle \mathbf{M}^2 \rangle - \langle \mathbf{M} \rangle^2)/V$ of the total dipole moment of the system.[7] Since boundary conditions in computer simulations are designed to minimize finite size effects, the KF formula can be used, as expected, to compute the dielectric constant in a simulation performed with a reaction field or periodic boundary conditions (if the Ewald sums are performed using the spherical convention for the order of summation).[9] For a confined spherical system which is *not* periodically repeated, the dielectric constant should be computed from Eq. (18) (see Sec. III B).

Attempts have been made to generalize the KF relation, valid for a macroscopic spherical system, to mesoscopic samples, where $\mathbf{P}(\mathbf{r})$, $\mathbf{E}(\mathbf{r})$, and the resulting $\epsilon(\mathbf{r})$ are nonuniform near the confining surface.[10,11] Strictly speaking, $\epsilon(\mathbf{r})$ is, however, no longer a scalar near the sample boundary, but a tensor with radial and tangential components.

We consider here the case where the external field is radial, preserving thus the spherical symmetry of the problem. Such a radial field can arise from an external charge q placed at the center of the spherical cavity filled with polar molecules, or polar residues of a globular macromolecule (e.g., a protein); in that case the nonuniform external field is $\mathbf{E}'(\mathbf{r}) = (q/r^2)\hat{\mathbf{r}}$ where $\hat{\mathbf{r}} = \mathbf{r}/r$. The dipoles of the confined fluid or macromolecule couple to this field with energy,

$$U_{\text{ext}} = -\sum_i \boldsymbol{\mu}_i \cdot \mathbf{E}'(\mathbf{r}) = -\int_{\mathcal{D}_{\text{cavity}}} m(\mathbf{r}) E'(\mathbf{r}) \mathrm{d}\mathbf{r}, \tag{19}$$

where $m(\mathbf{r}) = \mathbf{m}(\mathbf{r}) \cdot \hat{\mathbf{r}}$ is the radial component of the microscopic polarization density (1), and $E'(r) = q/r^2$. Substituting (19) into the Boltzmann factor in the definition (3) of the polarization, and linearizing (which is valid in the limit of small q), one arrives at the following relation between the radial component of $\mathbf{P}(\mathbf{r})$ and the radial component of the external field:

$$P(r) = \mathbf{P}(\mathbf{r}) \cdot \hat{\mathbf{r}} = \langle m(\mathbf{r}) \rangle + \beta \int_{\mathcal{D}_{\text{cavity}}} \mathrm{d}\mathbf{r}' [\langle m(\mathbf{r}) m(\mathbf{r}') \rangle - \langle m(\mathbf{r}) \rangle \langle m(\mathbf{r}') \rangle] E'(\mathbf{r}'), \tag{20}$$

where the statistical averages are once more taken at zero external field. We consider the case of molecules carrying linear dipoles, so that $\langle m(\mathbf{r}) \rangle = 0$ by symmetry. We now assume a local relationship between $P(\mathbf{r})$ and the radial local field $E(\mathbf{r})$ in the general form defined by Eqs. (5) and (6), and involving a local dielectric permittivity $\epsilon(r)$,

$$P(r) = \frac{\epsilon(r) - 1}{4\pi} E(r). \tag{21}$$

This relationship, together with $\nabla \cdot \mathbf{D}(\mathbf{r}) = 4\pi q \delta(\mathbf{r})$, implies that the fields $E(r)$ and $E'(r)$ are related by

$$E(r) = \frac{E'(r)}{\epsilon(r)}. \tag{22}$$

Combination of Eqs. (20), (21), and (22) then leads to the following relation for $\epsilon(r)$:

$$\frac{1}{4\pi} \frac{\epsilon(r) - 1}{\epsilon(r)} E'(r) = \beta \int_{\mathcal{D}_{\text{cavity}}} \mathrm{d}\mathbf{r}' \langle m(\mathbf{r}) m(\mathbf{r}') \rangle E'(r') \tag{23}$$

or, substituting $E'(r) = q/r^2$,

$$\frac{\epsilon(r) - 1}{\epsilon(r)} = 4\pi\beta \int_{\mathcal{D}_{\text{cavity}}} \mathrm{d}\mathbf{r}' \langle m(\mathbf{r}) m(\mathbf{r}') \rangle \left(\frac{r}{r'} \right)^2. \tag{24}$$

Note that, contrary to Eq. (18), this relation does not depend explicitly on the permittivity ϵ' of the confining medium. The present space-dependent dielectric constant $\epsilon(r)$ describes the screening by the polar fluid of the external field created by the point charge q, as is obvious from Eq. (22). It reduces to the bulk dielectric constant when r is sufficiently large, but still small compared to the radius R of the spherical cavity.

III. MOLECULAR DYNAMICS RESULTS

A. Slab geometry

We have carried out a number of long MD simulations (spanning tens of nanoseconds) to obtain estimates of $\epsilon_{\parallel}(z)$ and $\epsilon_{\perp}(z)$ from the fluctuation formulas (13) and (17). In a slab of width L, 3500 dipolar soft spheres (DSS) were placed. The confining walls at $z=0$ and $z=L$ are assumed to be nonpolarizable ($\epsilon'=1$). The simulation cell is a cube with edges of length L, and periodic boundary conditions are imposed in the (x,y) directions. Each molecule carries an extended dipole $\boldsymbol{\mu}$ made up of two opposite charges $\pm q$ placed at $\pm \mathbf{d}/2$ from the center, such that $\mu = qd$; the elongation was chosen to be $d/\sigma = 1/3$, where σ is the molecular diameter. The bulk dielectric behavior of fluids with such extended dipoles has been shown to be very similar to that of fluids with point dipoles, as long as $d/\sigma < 1/2$.[12] The short-range repulsion between the spherical molecules is chosen to be of the "soft sphere" form

$$u_{\text{Sr}}(r) = 4u \left(\frac{\sigma}{r} \right)^{12} \tag{25}$$

with $\sigma = 0.366$ nm and $u = 1.85$ kJ/mole [the charges $\pm q$ carry a mass $m = 5$ amu; a molecule has hence a total mass of $2m$ and a reduced moment of inertia $I^* = I/2m\sigma^2 = 1/9$. The simulations were performed with the simulation package GROMACS.[13] The equations of motion were integrated with a time step $dt = 2$ fs (reduced time step $dt^* = dt/(m\sigma^2/u)^{1/2} = 0.0024$)]. The walls exert a force on the center of mass of the molecules that derives from the potential,

$$U_{\text{walls}}(z) = \frac{4\pi u}{45} \left[\frac{\sigma^9}{z^9} + \frac{\sigma^9}{(L-z)^9} \right]. \tag{26}$$

This potential follows from integrating the soft-sphere repulsion potential (25) over the regions $z<0$ and $z>L$, for a wall of density $\rho_{\text{wall}}\sigma^3 = 1$. The long range Coulomb interactions between the charges $\pm q$ and the infinite array of periodic images are computed by a slab-adapted version of the usual 3D Ewald summation, as explained in the Appendix.

The structure of the fluid inside the slab is best characterized by the density-orientation profile $\rho(z,\theta)$, where θ is the angle between a molecular dipole and the z axis. This may be expanded in Legendre polynomials according to

$$\rho(z,\theta) = \sum_{\ell=0}^{\infty} \rho_\ell(z) P_\ell(\cos\theta), \tag{27}$$

where $\rho_\ell(z) = 1/2(2\ell+1)\int_{-1}^{+1} \rho(z,\theta) P_\ell(\cos\theta) \mathrm{d}(\cos\theta)$. In the case of uncharged walls, only even coefficients appear in the series (27) because of the symmetry $\rho(z,\theta) = \rho(z,\pi-\theta)$. The $\ell=0$ coefficient $\rho_0(z) = \rho(z)/2$ is one half of the density profile $\rho(z) = \int_0^\pi \rho(z,\theta) \sin\theta d\theta$. The ratio

$$\alpha(z) = \frac{\rho_{\ell=2}(z)}{\rho(z)} \tag{28}$$

provides a measure of the mean alignment of the dipoles. Since $P_2(x) = (3x^2-1)/2$, $\alpha(z)$ is negative if the dipoles are predominantly aligned parallel to the interface [$\alpha(z) = -5/4$ for complete alignment], while $\alpha(z)$ is positive for predominantly orthogonal alignment [$\alpha(z) = 5/2$ for full alignment orthogonal to the interface].

We performed the simulations at a constant temperature $T = 300$ K (reduced temperature $T^* = kT/u = 1.35$), and for two values of the dipole moment: $\mu = 1.47$ and 2.45 D, corresponding to a reduced dipole $\mu^* = \sqrt{\mu^2/\sigma^3 u} = 1.2$ and 2, respectively. The width of the slab was adjusted to $L = 16.62\sigma$, so that the reduced density of the fluid far from the walls is $\rho^*_{\text{bulk}} = \rho_{\text{bulk}}\sigma^3 = 0.8$. The bulk dielectric constant of this polar fluid is 98±2 at $\mu^* = 2$,[12,14] and about 10 at $\mu^* = 1.2$.

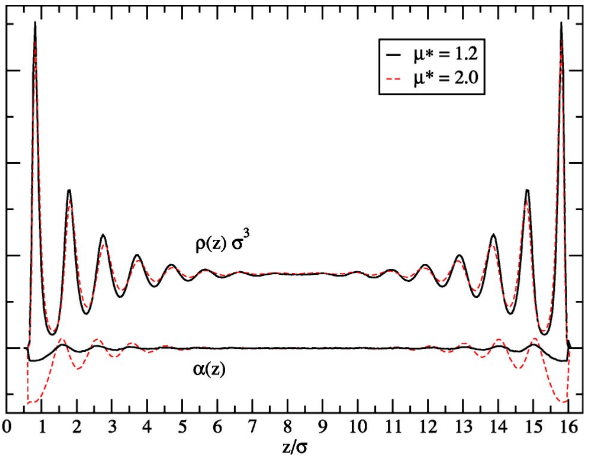


FIG. 1. Density and orientational profile of a DSS fluid in a slab ($T^*=1.35$, $\rho^*_{\text{bulk}}=0.8$) for two values of the reduced dipole moment.

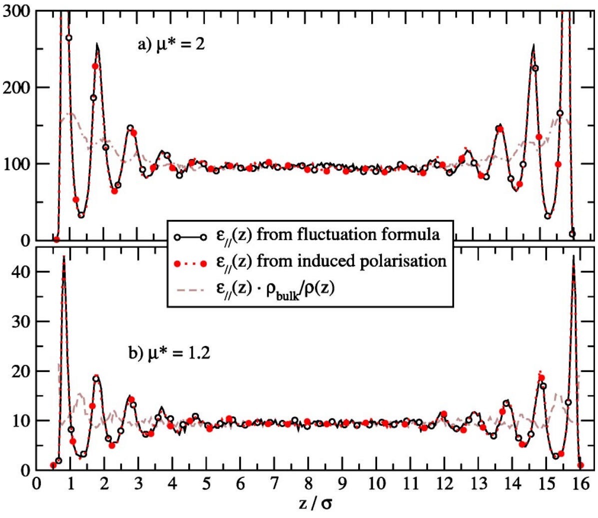


FIG. 2. Parallel component of the permittivity tensor (same system as in Fig. 1), from fluctuation formula (13) and from the response to an external field $E'=0.1$ V/nm along x axis.

The resulting density and orientation profiles $\rho(z)$ and $\alpha(z)$ in zero applied field are plotted in Fig. 1. Layering along the walls occurs in an interfacial region of five to six molecular diameters. The first layer of molecules is seen to align its dipoles parallel to the interface, but orientational ordering is rapidly lost further away from the dielectric walls. The ordering of molecular dipoles parallel to the wall in the first layer may be understood qualitatively in terms of electrostatic interactions of these dipoles with their images.[15] Strictly speaking, there are no images on the molecular scale since the walls are not polarizable, but image charge interactions arise on the mesoscopic scale because of the dielectric discontinuity between the polar fluid ($\epsilon>1$) and the walls ($\epsilon'=1$). The behavior of the density-orientation profile far from a single dielectric wall has been studied by Badiali, who showed that $\rho(z)$ is given asymptotically by its bulk value plus an A/z^3 tail arising from the dielectric discontinuity between ϵ and ϵ'.[16]

Parallel and perpendicular permittivity profiles $\epsilon_{\parallel}(z)$ and $\epsilon_{\perp}(z)$ were estimated from the simulations in zero external field using the fluctuation formulas (13) and (17), as well as from simulations in the presence of an external field, by evaluation of the ratio $P(z)/E(z)$ of the induced local polarization and electric field [cf. Eqs. (9)].

Results for the parallel permittivity $\epsilon_{\parallel}(z)$ obtained by both routes are shown in Fig. 2 (the simulation with $E'=0$ lasted 28 ns, while that with $E'=0.1$ V/nm along the x axis was 3.5 ns long). The agreement between the two independent estimates is seen to be perfect. The pronounced oscillations of $\epsilon_{\parallel}(z)$ near the walls closely mirror the oscillations in the density profile apparent in Fig. 1. In fact the ratio $\epsilon_{\parallel}(z)\rho_{\text{bulk}}/\rho(z)$, also shown in Fig. 2, shows much less structure. Towards the middle of the slab, $\epsilon(z)$ is seen to be constant and to take a value $\epsilon\simeq 10$ (for $\mu^*=1.2$) and $\epsilon\simeq 96$ (for $\mu^*=2$), in agreement with the bulk value derived from MD simulations of a periodic nonconfined fluid at the same state point. Note that on average, $\epsilon_{\parallel}(z)$ tends to increase above its bulk value close to the confining walls; in other words, parallel dipolar fluctuations tend to be enhanced near a dielectric wall with $\epsilon'=1$. This is contrary to the prediction of a generalization of the familiar Onsager cavity model to the case of a dipolar fluid near a dielectric wall.[17]

Turning to the perpendicular permittivity $\epsilon_{\perp}(z)$, we consider first the MD results obtained in the absence of external field. The structure of the fluctuation formula (17) is very unfavorable for the estimation of large values of the permittivity since the ratio of the left-hand side is then close to 1, so that statistical uncertainties on the fluctuation expression on the right-hand side, which we shall henceforth denote by $f(z)$, will be dramatically amplified on the quantity of interest $\epsilon_{\perp}(z)=1/[1-f(z)]$. This shortcoming is illustrated in Fig. 3 for the case $\mu^*=1.2$ (we did not succeed in extracting $\epsilon_{\perp}(z)$ from dipolar fluctuations in the case $\mu^*=2$). The signal $f(z)$ calculated by dividing the slab width L into 300 "bins" is seen to be rather noisy, despite the length of the simulation (28 ns). The fluctuations are strongly enhanced near the

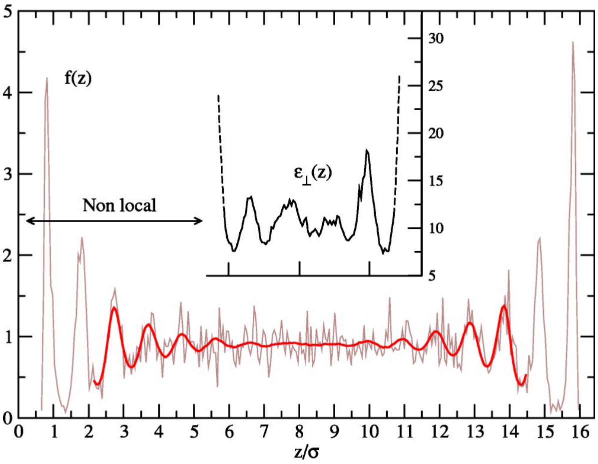


FIG. 3. Dipolar fluctuations $f(z)$ (thin line) and smoothed curve $\bar{f}(z)$ (thick line) [see text]. Inset: orthogonal component of the permittivity tensor ($\mu^*=1.2$, $T^*=1.35$, $\rho^*_{\text{bulk}}=0.8$).

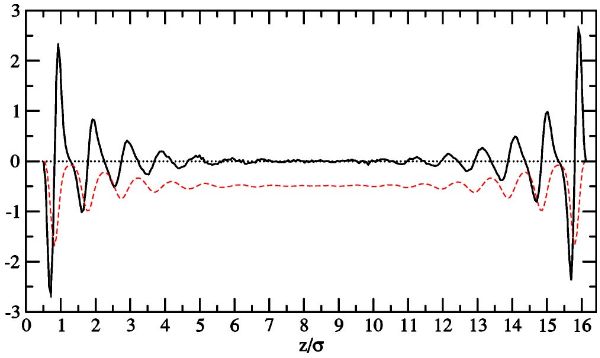


FIG. 4. Continuous line: polarization charge density $c(z)$ [e/nm^3] induced inside a slab of polar fluid ($\mu^*=1.2$, $\rho^*_{\mathrm{bulk}}=0.8$, $T^*=1.35$) by an external electric field $E'=1$ V/nm along the z direction. Dashed line: ten times the integral $\int_0^z c(z')\mathrm{d}z'$.

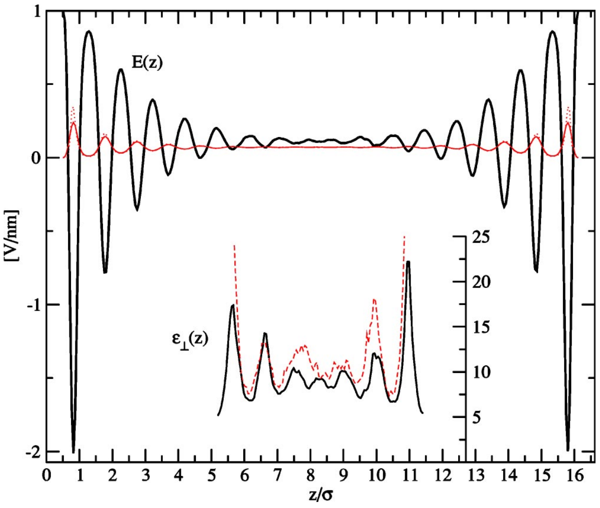


FIG. 5. Electric field $E(z)$ (thick line) and polarization density $P(z)$ computed using Eq. (30) (thin line) and from the statistical average $\langle m(z)\rangle$ (dotted line), for the same system as in Fig. 4. Inset: $\epsilon_\perp(z)$ from Eq. (31) and from fluctuation formula (17) (dashed line).

walls, but these values lead to nonphysical, negative values of $\epsilon_\perp(z)$, pointing to the inadequacy of the local assumption (9). A smoothed curve $\bar{f}(z)$ is obtained by averaging the signal $f(z)/\rho(z)$ over intervals of width 3σ, and multiplying the result by the local density $\rho(z)$. The resulting estimate of $\epsilon_\perp(z)$ in the central region of the slab is shown in the inset of Fig. 3.

The statistical uncertainties are still large, but the average value is compatible with a bulk permittivity $\epsilon_\perp=\epsilon_\parallel \simeq 10$. Outside this central "bulk" interval, $\epsilon_\perp(z)$ becomes unphysical whenever $f(z)$ exceeds 1.

We have investigated the transverse dielectric response by applying an external field $E'=1$ V/nm along the z axis of the moderately polar fluid ($\mu^*=1.2$, simulation time: 12 ns). This large value was chosen such that the local (screened) field near the center of the slab, $E(z)=E'/\epsilon_\perp(z)$, is still sufficiently strong to induce a sizeable polarization. The local charge density $c(z)$ induced by the applied field is plotted in Fig. 4. As expected it is antisymmetric with respect to midplane ($z=L/2$), and integration of $c(z)$ over the left-hand and right-hand halves of the slab leads to induced surface charge densities $c_S \simeq \pm 0.05$ e/nm^2. The local electric field $E(z)$ is calculated from

$$E(z) = E' + 4\pi\int_0^z c(z')\mathrm{d}z' \tag{29}$$

and the polarization density may then be deduced from

$$P(z) = \frac{E' - E(z)}{4\pi} = -\int_0^z c(z')dz' \tag{30}$$

while the permittivity profile follows from

$$\epsilon_\perp(z) = \frac{E'}{E(z)}. \tag{31}$$

Results obtained in this way for $E(z)$, $P(z)$, and $\epsilon_\perp(z)$ are plotted in Fig. 5. The polarization profile $P(z)$ can also be determined from the statistical average $\langle m(z)\rangle$ of the microscopic polarization density alone (dotted line in Fig. 5). The local electric field $E(z)$ exhibits large oscillations close to the walls. Inside the first layer, $E(z)$ is strongly negative ($\simeq -2$ V/nm), i.e., the external field E' is *overscreened* by a factor of 2! We refer to this remarkable effect as the formation of an "electrostatic double layer" (EDL). Beyond the first layer, the oscillations in $E(z)$ are gradually damped, but are still visible in the central part of the slab, where the oscillations are around a mean value of about 0.1 V/nm. The polarization $P(z)$ oscillates out of phase with $E(z)$ as expected, and the two estimates, based on Eq. (30) and on $\langle m(z)\rangle_{E'}$, are in excellent agreement. So are the estimates of $\epsilon_\perp(z)$ based on Eqs. (17) and (31), despite the large statistical uncertainties. As pointed out above, the relation between $P_\perp(z)$ and $E_\perp(z)$ is nonlocal [i.e., of the more general form (4)] near the walls, where $\epsilon(z)=1+4\pi P(z)/E(z)$ can take negative values, and diverges whenever $E(z)=0$. Hence $\epsilon_\perp(z)$ is not a useful quantity near the walls. A more relevant quantity is the normalized EDL profile $E(z)/E'$. For weak fields, it is given by linear response theory, viz.,

$$\begin{aligned}\frac{E_\perp(z)}{E'} &= 1 - 4\pi\beta[\langle m_\perp(z)M_\perp\rangle - \langle m_\perp(z)\rangle\langle M_\perp\rangle]\\ &= 1 - f(z).\end{aligned} \tag{32}$$

Figure 6 illustrates the good agreement between the EDL profiles computed from the fluctuations $f(z)$ (properly smoothed as in Fig. 3) and directly from the ratio $E_\perp(z)/E'$. The small discrepancies close to the walls are probably due to the fact that the extended dipoles are treated as point dipoles in the evaluation of $f(z)$. In order to investigate the influence of the dipole extension d/σ on the striking overscreening effect observed in the EDL profile near the walls, we have also carried out simulations for an extension $d/\sigma = 0.1$ (compared to 0.3 in the previous simulations), keeping the dipole moment fixed at $\mu^*=1.2$. An oscillatory profile of $E(z)$ similar to that in Fig. 5 is once more observed, but the amplitude of the oscillation closest to the two walls increases by ≈50% (from −2 to about −3 V/nm), thus pointing to an

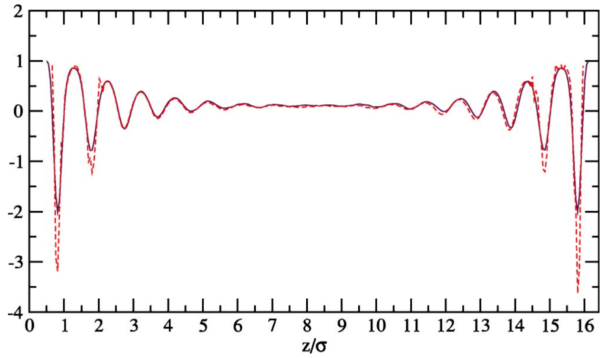


FIG. 6. EDL ratio $E_{\perp}(z)/E'$ from Eq. (29) (continuous line) and from the fluctuation formula (32) (dashed line) [same system as in Fig. 3].

enhancement of the overscreening effect as the point dipole limit is approached.

We also carried out one long (30 ns) MD run of 2076 SPC water molecules,[18] confined between the same hydrophobic walls (26). Figure 7 shows the resulting profiles for the Oxygen density, molecular orientation, and parallel permittivity (the orthogonal permittivity could not be obtained from the fluctuations). The density profiles are not unlike those observed in the simulations of Lee, McCammon, and Rossky,[19] who used a different water model, different wall-water interaction, and a narrower slit. The symmetry $\rho(z,\theta)=\rho(z,\pi-\theta)$ no longer holds, and there is indeed a nonzero average polarization $P_z(z)=\rho(z)\mu\langle\cos\theta\rangle$ close to the walls, even with no applied external field. The permittivity profile $\epsilon_{\perp}(z)$ shows that the bulk permittivity of SPC water ($\epsilon \simeq 65$),[20] is approximately reached after just one molecular layer inside the fluid. It is a striking result that, despite the large dielectric discontinuity between the fluid and the surrounding medium of dielectric constant unity, the local dipolar fluctuations in the vicinity of the interface remain almost bulklike, except in the very first layer. This behavior is in marked contrast to that observed for linear extended dipoles, illustrating the dominant influence of hydrogen bonding in the case of water.

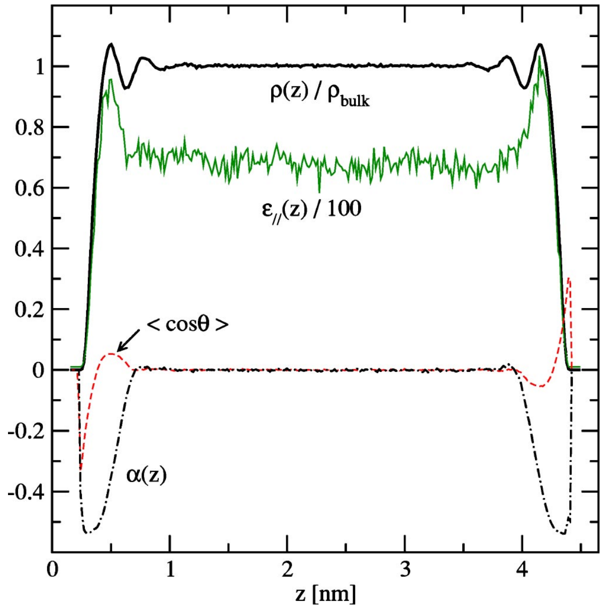


FIG. 7. Density, orientation, and permittivity profiles of 2076 SPC water molecules confined in a slab of width 4.65 nm by hydrophobic walls ($\rho_{\mathrm{bulk}}=33.6\ \mathrm{nm}^{-3}$, $T=300$ K).

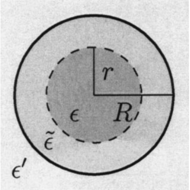


FIG. 8. Geometry for the Berendsen formula: a droplet of radius R is surrounded by a continuous medium of dielectric constant ϵ'. Dipolar fluctuations in the fluid are measured inside a concentric subsphere of radius r, while the remaining fluid in the outer shell is assumed to behave as a dielectric continuum of permittivity $\tilde{\epsilon}$.

B. Spherical geometry

The bulk dielectric constant of a spherical droplet of a polar fluid (possibly surrounded by a continuous medium of permittivity ϵ') is given in terms of dipolar fluctuations by Eq. (18). Previous workers have used another approach, which is approximate: they get the dielectric constant of a droplet from the mean square dipole $\langle m_{\mathrm{B}}^2(r)\rangle$ of an inner spherical region of radius r, assuming that the remaining surrounding fluid (a shell of thickness $R-r$) can be treated as a dielectric continuum of permittivity $\tilde{\epsilon}$ (see Fig. 8). The latter approach yields the Berendsen formula,[21–23]

$$(\epsilon-1)\frac{(2\epsilon'+\tilde{\epsilon})(2\tilde{\epsilon}+1)-2(r/R)^3(\epsilon'-\tilde{\epsilon})(1-\tilde{\epsilon})}{(2\epsilon'+\tilde{\epsilon})(2\tilde{\epsilon}+\epsilon)-2(r/R)^3(\epsilon'-\tilde{\epsilon})(\epsilon-\tilde{\epsilon})}$$

$$=\frac{4\pi\beta\langle m_{\mathrm{B}}^2(r)\rangle}{3V_r}, \qquad (33)$$

when $V_r=4\pi r^3/3$. Equation (33) reduces to the KF formula if $\tilde{\epsilon}=\epsilon'$ or if $r=R$. When $\tilde{\epsilon}=\epsilon$, it interpolates between the KF formula for a sphere surrounded by a medium of permittivity ϵ (for $r \ll R$), and the KF formula for a sphere surrounded by a medium ϵ' (for $r=R$).

We compared the predictions of Eq. (18) with those of Eq. (33) [with $\tilde{\epsilon}=\epsilon$], for a dipolar soft sphere fluid with parameters $\mu^*=2$, $T^*=1.35$, and $\rho^*_{\mathrm{bulk}}=0.2$. The fluid was confined to a spherical region by external forces deriving from the potential,

$$V_R(r)=4u\sigma^9\pi\left(\frac{1}{360}\frac{1}{r}\left[\frac{9R-r}{(R-r)^9}-\frac{9R+r}{(R+r)^9}\right]-\frac{4}{9R^9}\right), \qquad (34)$$

which arise from integrating the soft-phere repulsion potential (25) over the region $r>R$, assuming a confining medium of density $\rho_{\mathrm{wall}}\sigma^3=1$. The results of a 30 ns long MD simulation of a droplet of $N=1000$ molecules confined in a sphere of radius $R=11.4\sigma$ surrounded by vacuum ($\epsilon'=1$) are shown in Fig. 9. Interactions were computed without the introduction of any cutoff. The estimate for ϵ based on Eq. (18)

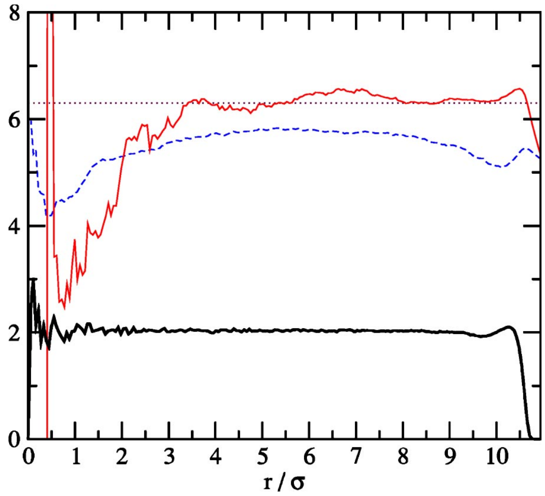


FIG. 9. Density profile [thick curve: ten times $\rho(r)\sigma^3$] and estimate of the *bulk* dielectric constant of a droplet of a polar fluid from Eq. (18) with $\mathbf{m}(\mathbf{r})$ the total dipole moment of a concentric subsphere of radius r (thin curve), and from the Berendsen formula (33) (dashed line). The expected bulk value is indicated by the dotted line.

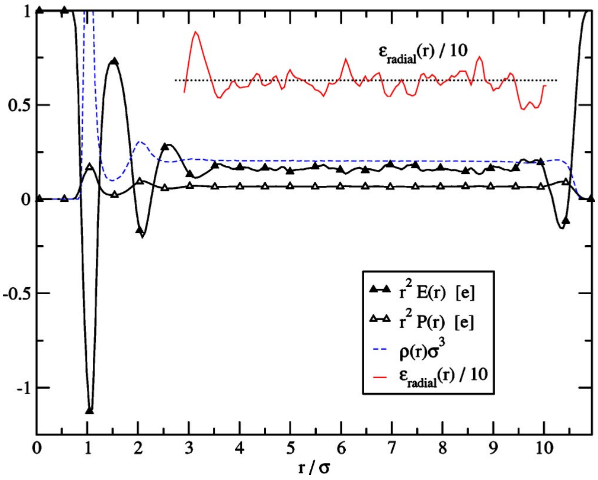


FIG. 10. Radial electric field, polarization density, molecular density, and permittivity profiles for a spherical droplet of polar fluid ($\mu^*=2$, $T^*=1.35$, $\rho^*_{\text{bulk}}=0.2$) when an ion of unit electronic charge (reduced charge $q^*=28.7$) is present at the origin. The dotted line indicates the bulk dielectric constant $\epsilon=6.3$ (divided by 10).

agrees very well with the bulk dielectric constant $\epsilon=6.3\pm0.2$ obtained from a simulation performed under periodic boundary conditions using Ewald sums. The Berendsen approximation is seen to underestimate the dielectric constant by about 8%. It yields the most accurate estimate when r is large enough to encompass the structural angular correlations, but small compared to R so that the outer shell of the fluid can screen effectively the dipolar fluctuations. This is in contrast with Eq. (18), which provides in principle an estimate for ϵ that is independent of the chosen radius r of the inner subsphere, as long as it is small enough for boundary effects to be negligible. The variations of the estimate observed in Fig. 9, especially at small r, are to be attributed to statistical uncertainties. Notice that when $\epsilon'=1$, Eq. (18) involves the unfavorable ratio $(\epsilon-1)/(\epsilon+2)$, so that uncertainties are greatly enhanced in this case. Better estimates for the dielectric constant would be obtained by surrounding the droplet with a medium of permittivity $\epsilon'=\epsilon$, or even $\epsilon'=\infty$ (metallic boundary conditions), similarly to the case of periodically repeated systems (see discussion in Ref. 12).

We turn now to the radial dielectric permittivity profile $\epsilon(r)$ of the spherical droplet, which is defined formally by Eq. (22). The profiles for the average electric field, polarization density, molecular density, and radial permittivity are shown in Fig. 10. These results were obtained from a 35 ns long MD simulation of the previous droplet, when an ion of reduced charge $q^*=q\mu/\sigma^2u=28.7$ is present at the origin. The total charge enclosed in a sphere a radius r around the central ion is $q(r)=q-4\pi r^2P(r)$, since the induced charge density is $-\boldsymbol{\nabla}\cdot\mathbf{P}(\mathbf{r})$. The polarization, and hence also the electric field profile $E(r)=q(r)/r^2$, can be measured directly in the simulation from the average radial polarization density $P(r)=\langle\mathbf{m}(\mathbf{r})\cdot\hat{\mathbf{r}}\rangle$. A better method, however, is to compute the profiles from the measured charge $q(r)$ enclosed in a sphere of radius r, as this yields smoother profiles thanks to the integration over the sphere.

The electric field oscillates strongly close to the central ion. Overscreening of the ion's charge occurs in the first molecular layers, similarly to the behavior observed near a planar interface (Fig. 5). At distances $3\sigma<r<R$, the electric field and polarization density decay as $1/r^2$, as required by macroscopic electrostatics. The radial permittivity profile, obtained from the ratio $P(r)/E(r)$, shows qualitatively the same behavior than the orthogonal permittivity close to a planar interface: it reduces to the bulk dielectric constant far from the interfaces (here $\epsilon_{\text{bulk}}\simeq6.3$), while it is ill defined close to the central ion and close to the surface of the droplet since it diverges whenever $E(r)$ changes sign.

IV. CONCLUSION

We have combined linear response theory with extensive molecular dynamics simulations of a simple model for dipolar molecules, in an attempt to validate the concept of a local dielectric permittivity of a confined (inhomogeneous) polar fluid. The geometries which were specifically investigated correspond to a polar fluid confined between two parallel walls (slit geometry), and in a spherical cavity surrounded by a dielectric continuum. Rather large samples (thousands of polar molecules) were considered, with confinement lengths on the scale of a few nanometers, while statistics were gathered over tens of nanoseconds. The key findings of our work are the following.

(a) There is excellent agreement between the results for the permittivity tensor $\epsilon(\mathbf{r})$ obtained from the zero field fluctuation formulas derived in Sec. II, and from the explicit response of the system to an external field. Simulation results based on the latter method converge generally faster if the amplitude of the applied field is large enough. The efficiency of both methods is comparable when we take into account the computational cost of ensuring that the applied field results are not affected by nonlinearity.[24]

(b) The purely local assumption is found, perhaps not surprisingly, to break down in the vicinity of an interface. In the case of the slit geometry, the parallel component $\epsilon_{\|}(z)$ of the dielectric tensor is found to oscillate strongly, mirroring the oscillatory density profile. Contrary to a wide-spread belief, the coarse-grained envelope of $\epsilon_{\|}(z)$ tends, if anything, to increase rather than to decrease near the dielectric interfaces. This trend is observed both for dipolar soft spheres and for the SPC model of water. The tendency is opposite of that recently observed in water close to mica surfaces,[25] but in those experiments, the mica surface is highly charged, leading to local electric fields on the order of 0.2 V/nm, a strength at which nonlinear effects cannot be neglected. The orthogonal component $\epsilon_{\perp}(z)$ is subject to very large statistical fluctuations. It tends to take unphysical (negative) values close to the interfaces, illustrating the necessity of considering a nonlocal relation between the local polarization and the local electric field. However both $\epsilon_{\perp}(z)$ and $\epsilon_{\|}(z)$ go over to the bulk values beyond a few molecular diameters from the surfaces. Similar conclusions hold in spherical geometry.

(c) The fluctuation formula (18), which is closely related to the classic Kirkwood–Fröhlich formula, is an exact result that can be used to compute the bulk dielectric constant of a homogeneous spherical drop of a polar fluid. It yields an estimate of ϵ that is in better agreement with the results from simulations performed under periodic boundary conditions, than the traditional route based on the approximate Berendsen formula. For highly polar fluids, metallic boundary conditions should be used to minimize the propagation of statistical errors when solving Eq. (18) for the dielectric constant.

(d) If the local electric field $\mathbf{E}(\mathbf{r})$ varies on molecular scales, a local $\epsilon(\mathbf{r})$ may always be formally defined from the ratio of $\mathbf{P}(\mathbf{r})$ over $\mathbf{E}(\mathbf{r})$, but the resulting $\epsilon(\mathbf{r})$ then generally depends on the particular interface under consideration.

(e) In the case of a uniform external field $\mathbf{E}'$, the ratio $\mathbf{E}(\mathbf{r})/\mathbf{E}'$ is a more informative quantity, which can be computed from dipolar fluctuations [see Eq. (32)]. Our simulations point to a dramatic and unexpected "overscreening" of the applied field in the vicinity of the interfaces, which is reminiscent of overscreening of surface charges by electric double layers in highly correlated ionic systems.[26]

The main message is that the use of local permittivities in "implicit solvent" models of biomolecular aggregates is highly questionable on the nanometer scale. Future work will consider the dielectric response of polar fluids near charged (rather than hydrophobic) surfaces, and will examine the influence of the molecular graniness and the polarizability of the interface on dielectric properties of the adjacent polar fluid.

APPENDIX: EWALD SUMS IN THE SLAB GEOMETRY

There exist several exact methods for computing electrostatic interactions in systems periodic in two dimensions (2D), but most are only slowly convergent: 2D Ewald sums and the Lekner method both have an $\mathcal{O}(N^2)$ scaling, while the MMM2D method scales as $\mathcal{O}(N^{5/3})$.[27,28] A faster method is to use an efficient implementation of the full 3D Ewald summations, leaving empty space in the simulation box outside the slab, in an attempt to decouple the interactions between the original slab and its periodic images in the z direction.[29] Adopting this approach, Yeh and Berkowitz showed that the interactions with the periodic images (in all three dimensions) must not be summed with the usual spherical convention, but rather in a slabwise manner.[30] The effect of this change in summation order is simply to replace the boundary term $U_{\text{sphere}}(\mathbf{M})=2\pi\mathbf{M}^2/(1+2\tilde{\epsilon})V$ by $U_{\text{slab}}(\mathbf{M})=2\pi\mathbf{M}_z^2$ in the Coulomb energy, where $\tilde{\epsilon}$ is the dielectric constant at infinity (not be be confused with the $\tilde{\epsilon}$ introduced in Sec. III B). This simple result can be seen as a particular case of the general formula

$$U_{\text{ellipsoid}}(\mathbf{M}) = \sum_{\alpha=x,y,z} \frac{B_\alpha}{\tilde{\epsilon}+(1-\tilde{\epsilon})B_\alpha}\frac{2\pi M_\alpha^2}{V} \tag{A1}$$

for the boundary term when the Coulombic interactions are summed with ellipsoidal summation order. In Eq. (A1), which is a generalization to arbitrary values of the dielectric constant $\tilde{\epsilon}$ at infinity of a result due to Smith,[31] the dimensionless coefficients B_α are related to the semiaxes a_x, a_y, a_z of the ellipsoid by

$$B_\alpha = \frac{a_x a_y a_z}{2}\int_0^\infty \frac{1}{x+a_\alpha^2}\frac{1}{[(x+a_x^2)(x+a_y^2)(x+a_z^2)]^{1/2}}\mathrm{d}x. \tag{A2}$$

These numbers satisfy $B_x+B_y+B_z=1$. By symmetry, $B_x=B_y=B_z=1/3$ in the case of a sphere; $B_x=B_y=1/2$, $B_z=0$ in the case of a cylinder ($a_z\to\infty$); and $B_x=B_y=0$, $B_z=1$ in the case of a slab (a_x, $a_y\to\infty$). Expression (A1) is simply the interaction energy of the charges with the depolarizing field produced by the uniform polarization density $\mathbf{M}/V$ of the infinite ellipsoidal collection of cells, when the latter is surrounded by a medium of dielectric constant $\tilde{\epsilon}$ (see Ref. 32 for the formula of the depolarizing field in a uniformly polarized ellipsoid). In the particular case of a slabwise summation order ($B_x=B_y=0$, $B_z=1$), expression (A1) becomes independent of the permittivity $\tilde{\epsilon}$ at infinity, as had been suspected by J de Joannis, Arnold, and Holm.[33] In the latter reference, explicit formulas to correct for the unwanted interlayer interactions in the 3D Ewald sums have been obtained, making the present approach exact in principle, while preserving its $\mathcal{O}(N\ln N)$ complexity.

Our simulation cell contained an interlayer gap of width $2L$, so that it was a parallelipiped of sides $L\times L\times 3L$. No correction term for substracting interlayer interactions was used, apart from the slabwise summation order. The Ewald sums were computed using the smooth particle mesh Ewald method (Ewald coefficient $\alpha=3.4705\ \text{nm}^{-1}$, grid size $13\times 13\times 40$ cells, interpolation order 6).[34] The interactions were truncatd beyond 0.9 mm, both in the real space Ewald sum and in the soft-sphere interactions. Assuming that the forces computed using a gap of width $10L$ are the exact answers, a gap a width $2L$ leads to a relative rms error in the forces of the order of 10^{-6} for the system simulated in Sec. III A (polar fluid with $\mu^*=2$, $\rho^*_{\text{bulk}}=0.8$, confined to a slit of width 16.6σ).

[1] J. D. Jackson, *Classical Electrodynamics*, 3rd ed. (Wiley, New York, 1998)
[2] G. Nienhuis and J. M. Deutch, J. Chem. Phys. **55**, 4213 (1971).
[3] V. Ballenegger and J. P. Hansen, Europhys. Lett. **63**, 381 (2003).
[4] For a recent illustration, see, e.g., B. Mallik, A. Masunov, and T. Lazaridis, J. Comput. Chem. **23**, 1090 (2002).
[5] For a recent example, see, e.g., R. J. Allen, J. P. Hansen, and S. Melchionna, J. Chem. Phys. **119**, 3905 (2003).
[6] J. Kirkwood, J. Chem. Phys. **7**, 911 (1939).
[7] H. Fröhlich, *Theory of Dielectrics* (Clarendon, Oxford, 1949).
[8] H. A. Stern and S. E. Feller, J. Chem. Phys. **118**, 3401 (2003).
[9] M. Neumann, Mol. Phys. **50**, 841 (1983).
[10] J. G. Powles, M. L. Williams, and W. A. B. Evans, J. Phys. C **21**, 1639 (1988).
[11] T. Simonson and C. L. Brooks III, J. Am. Chem. Soc. **118**, 8452 (1996).
[12] V. Ballenegger and J. P. Hansen, Mol. Phys. **102**, 599 (2004).
[13] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. [Electronic Publication] **7**, 306 (2001), http://www.gromacs.org
[14] P. G. Kusalik, M. E. Mandy, and I. M. Svishchev, J. Chem. Phys. **100**, 7654 (1994).
[15] R. J. Allen and J. P. Hansen, Mol. Phys. **101**, 1575 (2003).
[16] J. P. Badiali, J. Chem. Phys. **90**, 4401 (1989).
[17] R. Finken, V. Ballenegger, and J. P. Hansen, Mol. Phys. **101**, 2559 (2003).
[18] G. W. Robinson, S. B. Zhu, S. Singh, and M. W. Evans, *Water in Biology, Chemistry and Physics: Experimental Overviews and Computational Methodologies* (World Scientific, Singapore, 1996).
[19] C. Y. Lee, J. A. McCammon, and P. J. Rossky, J. Chem. Phys. **80**, 4448 (1984).
[20] P. Höchtl, S. Boresch, W. Bitomsky, and O. Steinhauser, J. Chem. Phys. **109**, 4927 (1998).
[21] H. J. C. Berendsen, CECAM Report No. 29, 1972 (unpublished).
[22] D. J. Adams and I. R. McDonald, Mol. Phys. **32**, 931 (1976).
[23] J. G. Powles, R. F. Fowler, and W. A. B. Evans, Chem. Phys. Lett. **107**, 280 (1984).
[24] Z. Wang, C. Holm, and H. W. Müller, J. Chem. Phys. **119**, 379 (2003).
[25] O. Teschke, G. Ceotto, and E. F. de Souza, Phys. Chem. Chem. Phys. **3**, 3761 (2001).
[26] J. P. Hansen and H. Löwen, Annu. Rev. Phys. Chem. **51**, 209 (2000).
[27] J. J. Weis and D. Levesque, J. Phys.: Condens. Matter (submitted).
[28] A. Arnold and Ch. Holm, Comput. Phys. Commun. **148**, 327 (2002).
[29] E. Spohr, J. Chem. Phys. **107**, 6342 (1997).
[30] L. C. Yeh and M. L. Berkowitz, J. Chem. Phys. **111**, 3155 (1999).
[31] E. R. Smith, Proc. R. Soc. London, Ser. A London, Ser. A **375**, 475 (1981).
[32] C. J. F. Böttcher, *Theory of Electric Polarization*, 2nd ed. (Elsevier, New York, 1973).
[33] J. de Joannis, A. Arnold, and C. Holm, J. Chem. Phys. **117**, 2496 (2002).
[34] U. Essman, L. Perela, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).

THE JOURNAL OF CHEMICAL PHYSICS **131**, 094107 (2009)

Simulations of non-neutral slab systems with long-range electrostatic interactions in two-dimensional periodic boundary conditions

V. Ballenegger,[1,a)] A. Arnold,[2] and J. J. Cerdà[3]
[1]*Institut UTINAM, Université de Franche-Comté, UMR 6213, 16 route de Gray, 25030 Besançon Cedex, France*
[2]*Fraunhofer Institute SCAI, Schloss Birlinghoven, 53754 Sankt Augustin, Germany*
[3]*Institute for Computational Physics, Universität Stuttgart, 70569 Stuttgart, Germany*

(Received 11 June 2009; accepted 12 August 2009; published online 3 September 2009)

We introduce a regularization procedure to define electrostatic energies and forces in a slab system of thickness h that is periodic in two dimensions and carries a net charge. The regularization corresponds to a neutralization of the system by two charged walls and can be viewed as the extension to the two-dimensional (2D)+h geometry of the neutralization by a homogeneous background in the standard three-dimensional Ewald method. The energies and forces can be computed efficiently by using advanced methods for systems with 2D periodicity, such as MMM2D or P3M/ELC, or by introducing a simple background-charge correction to the Yeh–Berkowitz approach of slab systems. The results are checked against direct lattice sum calculations on simple systems. We show, in particular, that the Madelung energy of a 2D square charge lattice in a uniform compensating background is correctly reproduced to high accuracy. A molecular dynamics simulation of a sodium ion close to an air/water interface is performed to demonstrate that the method does indeed provide consistent long-range electrostatics. The mean force on the ion reduces at large distances to the image-charge interaction predicted by macroscopic electrostatics. This result is used to determine precisely the position of the macroscopic dielectric interface with respect to the true molecular surface. © *2009 American Institute of Physics*. [doi:10.1063/1.3216473]

I. INTRODUCTION

The high interest in simulating surface and interfacial problems has led to an active search for efficient methods to compute long-range interactions in two-dimensional (2D) periodic systems that have a finite thickness (2D+h geometry). 2D Ewald-type formulas were derived by a number of workers.[1–6] As most of these methods scale quadratically with the number N of particles, other faster methods were proposed. The MMM2D method has an adjustable preset accuracy and scales as $N^{5/3}$,[7] see also Ref. 8. An approach with an almost linear scaling was proposed by Yeh and Berkowitz: It consists in introducing a gap in the simulation box along the nonperiodic dimension and to use a standard Ewald code for three-dimensional (3D) periodic systems with the summation order changed to slabwise (EW3DC method).[9] That method is accurate only when the gap in the simulation box is large enough. The errors introduced by the unwanted interactions with the layers artificially replicated in the nonperiodic direction can be exactly compensated for by adding an electrostatic layer correction (ELC) term, whose computational cost is linear in the number of particles. When that term is used in conjunction with a fast particle-mesh implementation of the Ewald sum, such as the P3M (Ref. 10) or SPME (Ref. 11) algorithm, the resulting methods P3M/ELC and SPME/ELC scale as $N \log(N)$. In the case of the P3M/ELC method, an error estimate is moreover available;[12] it allows the method to be tuned automatically to its optimal operation point, which minimizes the computational time at the desired accuracy. Another slab method with a $N \log(N)$ scaling was proposed recently by Ghasemi *et al.*[13] The distribution of errors is uniform across the slab in that latter method, contrary to P3M/ELC, but it has no *a priori* error estimate.

In all the aforementioned works, the system is assumed to be overall charge neutral. In some situations, for example, when studying surfaces with charged defects or when computing ionic solvation free energies near an interface, one needs a method applicable to a slab system carrying a net charge. The purpose of this paper is to define a regularization of the (divergent) lattice sum for the energy and to have an efficient way to compute the regularized energies and forces. This allows, for example, Monte Carlo and molecular dynamics simulations of non-neutral slab systems to be performed with long-range electrostatic interactions that are fully consistent with the 2D periodic boundary conditions (2D-PBC).

The regularization of the lattice sum via a suitable neutralization of the simulation cell is presented in Sec. II. The practical computation of these regularized sums is discussed in Sec. III for three recent efficient algorithms: EW3DC, P3M/ELC, and MMM2D. Numerical tests performed in Sec. IV show that the three methods give consistent results for energies and forces that agree as well with direct lattice sum calculations. In Sec. V, a molecular dynamics simulation of a sodium ion close to an air/water interface is performed to demonstrate that the method does indeed provide consistent long-range electrostatics.

[a)] Electronic mail: vincent.ballenegger@univ-fcomte.fr.

II. CHARGED SLAB IN 2D PERIODIC BOUNDARY CONDITIONS

A. Definition of energy and forces

We consider a slab system composed of N particles with charges q_i and positions $\boldsymbol{r}_i=(x_i,y_i,z_i)$ $(i=1,\ldots,N)$ in a simulation cell which is periodically replicated in the x and y directions with period L_x and L_y, respectively. In the nonperiodic z-direction, the system has a finite extent h, which can be arbitrarily large. Particle positions lie in the region delimited by $-L_x/2<x\leq L_x/2$, $-L_y/2<y\leq L_y/2$, and $-h/2\leq z\leq h/2$. The regions outside the slab, $z>h/2$ and $z<-h/2$, are assumed to be void; their dielectric permittivity is that of vacuum. The Coulomb energy in Gaussian units is given by the sum over periodic images,

$$E=\frac{1}{2}\sum_{n_x,n_y\in\mathbb{Z}}\sum_{i,j=1}^{N}{}'\frac{q_iq_j}{|\boldsymbol{r}_i-\boldsymbol{r}_j+n_x\boldsymbol{L}_x+n_y\boldsymbol{L}_y|},\tag{1}$$

where $\boldsymbol{L}_x=L_x\hat{\boldsymbol{e}}_x$, $\boldsymbol{L}_y=L_y\hat{\boldsymbol{e}}_y$, and the prime denotes the omission of the $i=j$ term in the primary cell $n_x=n_y=0$. Notice that interactions between charges in the primary cell are accounted for with a factor 1, since they appear twice in the sum (1), while interactions of charges in the primary cell with charges in image cells $(n_x,n_y)\neq(0,0)$ are accounted for with a factor 1/2 since they appear only once in the sum. That factor 1/2 is necessary to avoid double counting in the energy of the simulated macroscopic sample, as explained in Appendix A.

The sum (1) is only defined for a charge neutral system $(\Sigma_iq_i=0)$, and even then, the sum is only conditionally convergent. This means that we need to specify the order in which we perform the sum. We adopt here the cylindrical limit, i.e.,

$$E=\lim_{R\to\infty}E(R),\tag{2}$$

with

$$E(R)=\frac{1}{2}\sum_{\substack{n_x,n_y\in\mathbb{Z}\\(n_xL_x)^2+(n_yL_y)^2\leq R^2}}\sum_{i,j=1}^{N}{}'\frac{q_iq_j}{|\boldsymbol{r}_i-\boldsymbol{r}_j+n_x\boldsymbol{L}_x+n_y\boldsymbol{L}_y|}.\tag{3}$$

Notice that the energy does not depend in the limit $R\to\infty$ on the dielectric constant ϵ' of the external medium in region $x^2+y^2>R^2$, contrary to the usual Ewald method for 3D-periodic systems, where the dielectric constant of the outer spherical medium at $x^2+y^2+z^2>R^2$ does intervene. The reason for this can easily be understood on physical grounds. Indeed, when the simulation cell is polarized, macroscopic electrostatics tells us that surface charges appear at the dielectric discontinuity between the external medium and the macroscopic sample made up of copies of the central simulation cell. The magnitude of these polarization surface charges depend on ϵ' and may produce a depolarizing field in the sample that contributes to the energy. In the case of a slab system, the charges induced on the side surfaces at $x^2+y^2=R^2$ are proportional to the area $2\pi Rh$, while the electric field created by them in the sample decays as R^{-2}. In the limit $R\to\infty$, that field vanishes, so that the energy is indeed independent of ϵ'. The energy of a slab system depends, however, on the dielectric permittivities ϵ_1 and ϵ_2 of the two regions above $(z>h/2)$ and below $(z<-h/2)$ the slab, which are here assumed to be empty $(\epsilon_1=\epsilon_2=1)$.

If some charges (electrons, counterions, etc.) are treated implicitly in the system, the total charge

$$Q_{\text{tot}}=\sum_{i=1}^{N}q_i\tag{4}$$

of the simulation box can differ from zero. The energy (1) remains finite only if the background charge provided by the implicit particles is properly accounted for.

In systems with 3D periodicity where the implicit charges are assumed to provide a homogeneous and isotropic charge distribution, the background charge reduces to a uniform charge density $\rho_b=-Q_{\text{tot}}/V$, where V is the volume of the simulation box. The standard Ewald method can be used to sum the Coulomb interactions in such systems, including the interaction with the neutralizing background, see e.g., Refs. 14 and 15. The Ewald formula for the energy of the 3D-periodic system takes then the form $E_{\text{3D}}=E^{(r)}+E^{(k)}+E^{(d)}+E^{(n)}$, where the three first terms are the usual real-space, reciprocal-space, and surface contributions, while the last term is an electroneutrality contribution

$$E^{(n)}=-\frac{\pi Q_{\text{tot}}^2}{2\alpha^2V}\tag{5}$$

that depends on the Ewald splitting parameter α. The energy E_{3D} is independent of the free Ewald parameter α only if the contribution $E^{(n)}$ is included. As the neutralizing background is homogeneous, the energy $E^{(n)}$ is independent of the particle positions and does thus not lead to any force.

For future reference, it is useful to recall the origin and physical content of term $E^{(n)}$. It corresponds merely to the sum of the direct-space interaction $E_{c\text{-}b}^{(r)}$ of the charged particles with the neutralizing background, and to the direct-space background-background interactions $E_{b\text{-}b}^{(r)}$ on the other hand. Thus $E^{(n)}=E_{c\text{-}b}^{(r)}+E_{b\text{-}b}^{(r)}$, where

$$E_{c\text{-}b}^{(r)}=\sum_i q_i\int_{\mathbb{R}^3}\psi(r)\rho_bd^3\boldsymbol{r}=-\frac{\pi Q_{\text{tot}}^2}{\alpha^2V}\tag{6}$$

[see, e.g., Eq. (3.5) of Ref. 16] and

$$\begin{aligned}E_{b\text{-}b}^{(r)}&=\frac{1}{2}\int_V d\boldsymbol{r}\int_{\mathbb{R}^3}d\boldsymbol{r}'\rho_b\psi(|\boldsymbol{r}-\boldsymbol{r}'|)\rho_b\\&=\frac{\rho_b^2}{2}\int_V d\boldsymbol{r}\frac{\pi}{\alpha^2}\\&=\frac{\pi Q_{\text{tot}}^2}{2\alpha^2V},\end{aligned}\tag{7}$$

where $\psi(r)=\text{erfc}(\alpha r)/r$ is the direct-space interaction. Notice the factor 1/2 in the background-background energy that is needed to avoid double counting (see Appendix A). The reciprocal-space interaction of the particles with the background and the reciprocal-space background-background energy exactly cancel the corresponding singularity in the charge-charge interactions. In the Ewald formula for a nonneutral system, they are included in the canceling of the (otherwise divergent) $\boldsymbol{k}=\boldsymbol{0}$ term in the reciprocal sum.

In a system with 2D periodicity, the background charge associated with the implicit particles can take different forms. For example, if these particles are homogeneously distributed in some confined region $[z_1,z_2]$ along the nonperiodic direction, they may be replaced by a uniform neutralizing slab in that region. That charged background slab will clearly give rise an inhomogeneous electric potential $V_0(z)$ and to forces on charged particles. If the implicit particles are not confined to some particular region of space, it is appropriate to introduce a neutralization of the simulation cell that does not introduce any background force on the particles. Having a neutralization with this property is especially useful because it is generic; it can be used to simulate systems with any other neutralizing background by simply adding the corresponding background electric potential $V_0(z)$ as an external field in the simulation.

Unlike the 3D case, a neutralization that does not exert forces cannot be realized in 2D-PBC by adding a uniform charge distribution over all space; therefore a different route is needed to deal with 2D systems with a net charge. We propose to simply substract the singularity that arises due to the excess charge, which can be calculated analytically as follows. The asymptotically diverging behavior of $E(R)$ as $R\to\infty$ can be found by approximating the slab as a charged sheet carrying a surface charge density $\sigma_{\text{sheet}}=Q_{\text{tot}}/(L_xL_y)$ and by noting that the divergence arises from the large distance contributions to the interaction energy of the simulation cell with this sheet. This divergence is

$$\frac{1}{2}Q_{\text{tot}}\int_0^R \frac{\sigma_{\text{sheet}}}{r}2\pi r dr=\frac{\pi Q_{\text{tot}}^2}{L_xL_y}R, \tag{8}$$

where the factor 1/2 ensures, as in Eq. (1), that one measures only half of the interaction energy of the simulation cell with the periodic copies of the cell (see Appendix A). Subtracting Eq. (8) from Eq. (3), we define the energy of a charged system in 2D-PBC as

$$E=\lim_{R\to\infty}\left(E(R)-\frac{\pi Q_{\text{tot}}^2}{L_xL_y}R\right). \tag{9}$$

Notice that Eq. (9) is the result of a formal regularization of the divergent lattice sum (1). We will show in Sec. II B that this formal regularization is equivalent to a neutralization of the system by two charged walls.

Forces follow by differentiation of this potential energy function with respect to particle positions. As the regularizing term in Eq. (9) is independent of particle positions, it does not give rise to any contribution to the forces, as required. Notice that the lattice sum for the force, given by Eq. (1) with an additional gradient operator $-\nabla_i$, is conditionally convergent from the start in systems with a net charge, and hence well defined once the summation order is specified.

B. Interpretation of energies

The regularization (9) of the energy can be interpreted as a neutralization of the system by two charged walls located on each side of the slab, at $z=\pm h/2$, that carry a surface charge density

$$\sigma=-\frac{Q_{\text{tot}}}{2L_xL_y}. \tag{10}$$

These walls can be thought of as arising from a uniform neutralizing charge density created by the particles treated implicitly in the regions above $(z>h/2)$ and below $(z<-h/2)$ the slab. Since these walls exert (constant) equal and opposite forces on charged particles located in between them, they have no net physical effect, apart for a shift in the energy. Let us show that the presence of the two walls regularizes the energy (3) in accordance with the prescription (9) and determine the shift in the energy that results from the particle-wall and wall-wall interactions.

As interactions are summed in a cylindrical limit, we treat the walls as large circular plates of radius $R\to\infty$. The electrostatic potential created by a plate of radius R with surface charge σ reads

$$\Phi_R(z)=\int_0^R\int_0^{2\pi}\frac{\sigma}{\sqrt{r^2+z^2}}rdrd\phi=2\pi\sigma(\sqrt{R^2+z^2}-|z|) \tag{11}$$

at a distance z from the plate on its symmetry axis. When R is large, $\Phi_R(z)$ behaves as

$$\Phi_R(z)=2\pi\sigma(R-|z|)+\mathcal{O}\left(\frac{1}{R}\right). \tag{12}$$

Two large circular plates at $\pm h/2$ create therefore a constant electrostatic potential in between the plates given by

$$\Phi_{\text{slab}}=2\pi\sigma(2R-h),\quad R\to\infty. \tag{13}$$

The interaction energy of the charges in the simulation box with the two neutralizing walls is therefore

$$E_{c\text{-}w}(R)=Q_{\text{tot}}\Phi_{\text{slab}}=-\frac{\pi Q_{\text{tot}}^2}{L_xL_y}(2R-h). \tag{14}$$

The mutual interaction energy of the walls is (per cell)

$$E_{w_1\text{-}w_2}(R)=\Phi_R(h)\sigma L_xL_y=\frac{\pi Q_{\text{tot}}^2}{2L_xL_y}(R-h), \tag{15}$$

while the self-energies of the two walls read (per cell)

$$E_{w_1\text{-}w_1}(R)+E_{w_2\text{-}w_2}(R)=2\frac{1}{2}\sigma L_xL_y\Phi_R(0)=\frac{\pi Q_{\text{tot}}^2}{2L_xL_y}R. \tag{16}$$

Combining Eqs. (14)–(16) with Eq. (3), the electrostatic energy of the system complemented with the two neutralizing walls is

$$\tilde{E}=\lim_{R\to\infty}\left(E(R)-\frac{\pi Q_{\text{tot}}^2}{L_xL_y}R\right)+\frac{\pi Q_{\text{tot}}^2h}{2L_xL_y}. \tag{17}$$

The neutralization by two walls gives therefore energies that agree with the regularized energies (9), apart for an

h-dependent term that does not depend on the particle positions. That term is a constant shift in the energy that can be dropped because it has no physical significance.

In summary, the energy of a system carrying a net charge in 2D-PBC is defined by Eq. (9), which can be interpreted as the energy of the system neutralized by two walls, with the prescription that the constant h-dependent shift in the energy is omitted. The present regularization procedure to simulate slab systems with a net charge constitutes the first significant result of this paper.

C. Systems with a changing net charge

The computation of the solvation free energy of ions at infinite dilution is an example where a simulation box with a net charge is used.[17] When computing this free energy via the thermodynamic integration method, the ion is progressively charged, and the charging free energy can be determined from the measured fluctuations of the solute electrostatic energy.[14] Let us show that charging free energies can be correctly computed in 2D periodic systems using definition (9) of the energy, despite the fact that the net charge Q_{tot} varies during the simulation.

As shown in Appendix B, the energy (9) can be written using an effective kernel $\phi_{\text{PBC}}(\boldsymbol{r})$ which takes into account the 2D-PBC:

$$E = \frac{1}{2}\sum_{i,j=1}^{N} q_i q_j \phi_{\text{PBC}}(\boldsymbol{r}_i - \boldsymbol{r}_j). \tag{18}$$

We assume now that a charge q is added at position $\boldsymbol{r}$. The energy of this new system is

$$E_q = E + \frac{1}{2}\left(2\sum_{i=1}^{N} q_i q \phi_{\text{PBC}}(\boldsymbol{r}_i - \boldsymbol{r}) + q^2 \phi_{\text{PBC}}(\boldsymbol{0})\right), \tag{19}$$

where $\phi_{\text{PBC}}(\boldsymbol{0})$ is the regularized interaction energy of a single unit charge with its own periodic images [the value of $\phi_{\text{PBC}}(\boldsymbol{0})$ is computed in Sec. IV A]. One can rewrite Eq. (19) as

$$E_q = E + \int_0^q \Phi_{q'}(\boldsymbol{r}) dq', \tag{20}$$

where

$$\Phi_q(\boldsymbol{r}) = \frac{dE_q}{dq} = \sum_i q_i \phi_{\text{PBC}}(\boldsymbol{r}_i - \boldsymbol{r}) + q\phi_{\text{PBC}}(\boldsymbol{0}). \tag{21}$$

$\Phi_{q'}(\boldsymbol{r})$ is nothing but the electrostatic potential at $\boldsymbol{r}$ created by all other charges in the system and by the periodic images of the charge q' already present at $\boldsymbol{r}$. Equation (20) gives the change in energy when the charge q is added progressively in the system. From Eq. (21), we see that this change in energy is given by the correct expression associated with the 2D-PBC. In particular, the last term in $\Phi_q(\boldsymbol{r})$ gives rise to the charge self-energy $\frac{1}{2}q^2\phi_{\text{PBC}}(\boldsymbol{0})$ which is a known contribution to the energy in PBC.[14,16] The fact that the regularized energy (9) can be written in the form (18) is therefore sufficient to conclude that the energy difference between two systems characterized by a different net charge takes a meaningful value (which includes variations in the self-energies of the charges). Equations (18)–(21) are the counterparts of formally identical formulas that hold in the standard Ewald method for neutral and non-neutral systems in 3D-PBC. As shown in Ref. 14, the charging free energies computed using the Ewald method converge very quickly to their thermodynamical values once a simple correction for finite-size effects is included in the calculations (see also the detailed analysis of finite-size effects of Ref. 18).

The results of this section, together with Eq. (9), show that the thermodynamic integration method can be used to compute charging free energies in inhomogeneous systems with slab geometry. In the present approach, the long-range electrostatic interactions are treated in a manner fully consistent with the 2D periodicity.

III. FAST COMPUTATIONS OF ENERGIES AND FORCES

The regularized energy (9), and the forces on particles derived from it, can be computed quickly in simulations, thanks to advanced algorithms, as detailed in the next sections.

A. MMM2D method

In the MMM2D method,[7] a convergence factor $\exp(-\beta|\boldsymbol{r}_i - \boldsymbol{r}_j + n_x \boldsymbol{L}_x + n_y \boldsymbol{L}_y|)$ is introduced in the energy (1), and the sum is computed in the limit $\beta \to 0$ using an $\mathcal{O}(N^{5/3})$ algorithm.

When the system is non-neutral, the large distance charge-charge interactions generate a divergent contribution as $\beta \to 0$. By the same reasoning as that which leads to Eq. (8), the diverging behavior is given by

$$\frac{1}{2}Q_{\text{tot}}\int_0^\infty \sigma_{\text{sheet}} \frac{\exp(-\beta r)}{r} 2\pi r dr = \frac{\pi Q_{\text{tot}}^2}{L_x L_y}\frac{1}{\beta}. \tag{22}$$

The convergence factor approach is thus equivalent to the cylindrical limit approach with the role of variable R played by $1/\beta$. The MMM2D method can therefore be used to compute the regularized energy (9) and forces by simply dropping the divergent contribution (22) in the MMM2D formulae for the energy. As contribution (22) vanishes in a charge neutral system, it is already omitted in the MMM2D method, which can hence be used without any modification to simulate non-neutral slab systems, with energies and forces defined according to Sec. II.

B. The Yeh–Berkowitz approach: EW3DC method

In the EW3DC method of Yeh and Berkowitz,[9] a gap is introduced in the simulation box along the nonperiodic z-direction and a standard Ewald code for 3D-periodic systems is used, typically in a fast $\mathcal{O}(N \log N)$ particle-mesh implementation such as P3M[10] or SPME.[11] The interactions must be summed in a slabwise order. The surface term (also known as the dipole term) in the Ewald formula for the energy takes then the form

$$E^{(d)} = \frac{2\pi}{V} M_z^2, \tag{23}$$

where $M_z = \Sigma_i q_i z_i$ is the total dipole moment of the simulation box along the direction normal to the slab and $V = L_x L_y L_z$ is the volume of the box. The dipole term contributes not only to the energy but also to forces via the expression $\boldsymbol{F}_i^{(d)} = -\boldsymbol{\nabla}_i E^{(d)}$, i.e., it leads to a normal force

$$\boldsymbol{F}_i^{(d)} = -\frac{4\pi M_z}{V} q_i \hat{\boldsymbol{e}}_z \tag{24}$$

on particle i. Since M_z depends on the choice of the origin in systems with a net charge, the EW3DC method obviously cannot be applied to such systems without modification.

When using the Ewald method for a system with a net charge in 3D-PBC, it is assumed implicitly that there are particles distributed uniformly in the simulation box so that they create a constant background charge density $\rho_b = -Q_{\mathrm{tot}}/V$, which neutralizes the system. This neutralization obviously differs from the two walls neutralization defined in Sec. II. Results of the EW3DC approach applied to a system with a net charge will hence need to be corrected [see Eq. (28)] to avoid spurious effects arising from unwanted interactions of the particles with the background charge ρ_b (in particular, the background charge located in region $-L_z/2 \leq z \leq L_z/2$).

The background charge ρ_b contributes to the Ewald energy not only via Eq. (5) (together with the cancellation of the $\boldsymbol{k} = \boldsymbol{0}$ term in the reciprocal Ewald sum), but also via the dipole term (23), because the background charge has to be accounted for when computing the total dipole moment M_z of the simulation box. Let us fix the origin of our Cartesian coordinate system at the center of the box. The contribution of the background charge to the total dipole moment vanishes then by symmetry, allowing expression (23) to be used without modification.

We assume that the gap in the simulation box is large enough so that interactions between the slab and the unwanted replicas of the slab in the z-direction are entirely negligible (see Sec. IV). The simulated system can then be viewed as a charged slab, of width h, embedded into another slab of width L_z carrying a neutralizing charge density ρ_b. This charge density does not contribute to the dipole term, as with our choice of the coordinate system, its dipole moment is zero. Moreover, the Yeh–Berkowitz approach relies on the fact that for a sufficiently large L_z, the interaction of the charges in the primary slab with the image charges in the z-direction can be approximated by a homogeneous charge distributions. Under this assumption, the interactions with the charges and the background ρ_b in the image slabs cancel, and we are left with the contribution of the background only in the primary slab itself. The electrostatic energy of the system clearly depends on the position of the particles within the larger neutralizing slab, because the latter slab produces a parabolic electrostatic potential

$$V_0(z) = -2\pi\rho_b \int_{-L_z/2}^{L_z/2} |z - z'| dz' = -2\pi\rho_b\left(z^2 + \left(\frac{L_z}{2}\right)^2\right). \tag{25}$$

The fact that the energy depends on the position of the charged slab system along the z-axis can also be seen on the level of the Ewald formula: The total dipole moment M_z in the surface term (23) clearly depends on the position of the slab in the simulation box. Note that we use here again Eq. (12) but leaving out the R-dependent term, which leads to the divergent contribution (8) already accounted for in the 3D Ewald method with slabwise summation order. To recover the energy of the system defined by the regularization procedure of Sec. II, we only have to subtract the interaction energy of the primary charges with the neutralizing slab, namely,

$$E_{c\text{-}b} = \sum_i q_i V_0(z_i) = -2\pi\rho_b\left(\sum_i q_i z_i^2 + Q_{\mathrm{tot}}\frac{L_z^2}{4}\right), \tag{26}$$

and the self-interaction of the background slab, i.e.,

$$E_{b\text{-}b} = \frac{1}{2} L_x L_y \int_{-L_z/2}^{L_z/2} \rho_b V_0(z) dz = 2\pi\rho_b Q_{\mathrm{tot}}\frac{L_z^2}{6}. \tag{27}$$

We find thus that the contribution

$$E_{\mathrm{bcc}} := -(E_{c\text{-}b} + E_{b\text{-}b}) = 2\pi\rho_b\left(\sum_i q_i z_i^2 + Q_{\mathrm{tot}}\frac{L_z^2}{12}\right), \tag{28}$$

which we call the background charge correction (bcc) term, must be added to the Ewald energies when applying the EW3DC approach to a system with a net charge.

In summary, to simulate a charged slab system using the EW3DC approach, one uses a standard 3D-Ewald code with slabwise summation order, i.e., with surface term (23), in a simulation box with a large enough gap. The interaction energy (5) with the neutralizing background must be included (since the EW3DC approach computes the 3D Ewald energy $E_{\mathrm{3D}} = E^{(r)} + E^{(k)} + E^{(d)} + E^{(n)}$), and the correction term (28) must be added to remove the effect of the parabolic potential created by the neutralizing background slab. Combining Eq. (28) with Eq. (23) yields a corrected surface term given by expression

$$E_{\mathrm{bcc}}^{(d)} := E^{(d)} + E_{\mathrm{bcc}} = \frac{2\pi}{V}\left(M_z^2 - Q_{\mathrm{tot}}\sum_i q_i z_i^2 - Q_{\mathrm{tot}}^2\frac{L_z^2}{12}\right). \tag{29}$$

Notice that $E_{\mathrm{bcc}}^{(d)}$ is independent of the z-position of the slab in the simulation box, i.e., it is invariant under translations $z \to z + a$ for any value of a, as it should for a system in the 2D+h geometry. There is thus no specific condition on the origin of the coordinate system when using Eq. (29). Forces follow by differentiation of the energy. Taking the negative gradient of Eq. (29) with respect to z_i, one obtains the expression

$$-\nabla_i(E^{(d)}_{\text{bcc}}) = -\frac{4\pi}{V}(M_z - Q_{\text{tot}}z_i)q_i\hat{\boldsymbol{e}}_z \tag{30}$$

for the contribution to the force on the ith particle arising from the surface term and from the bcc. The corrected surface terms (29) and (30) are needed to extend the widely used EW3DC method to non-neutral systems, in accordance with the regularization scheme introduced in Sec. II. This constitutes the second important point of this paper. It is interesting to note that the final result (30) for the force agrees with a new interpretation of the surface term of the Ewald method proposed in Ref. 19. In that reference, the authors derive Eq. (30) from an analysis of the surface term in the 3D-Ewald method for different summation orders (in particular, slabwise), considering also non-neutral systems but without referring to any neutralizing background. We stress, however, that, whatever the summation order, the real-space interaction energy (5) with the neutralizing background charge must be included when computing Ewald energies to avoid results that depend on the free Ewald parameter α.

C. The electrostatic layer correction: P3M/ELC method

The P3M/ELC method[12] uses the same Yeh–Berkowitz approach based on a gap, but it is much more efficient than the EW3DC method because it includes an ELC term that subtracts the interactions of the slab with the unwanted replicas of the slab in the nonperiodic direction. The ELC term allows to reduce significantly the size of the gap needed in the simulation box, and hence the computation effort. Error estimates exist moreover for the ELC term.[12] Since the EW3DC method is a particular case of the P3M/ELC method where the ELC is entirely neglected, the error estimate of the latter method can actually be used to control the accuracy of the EW3DC method as well. As shown by Eq. (18) of Ref. 12, the accuracy of these gap methods is essentially proportional to $\exp(-2\pi Kg)$, where K is the ELC cutoff and g the width of the gap. The EW3DC method does not involve any ELC terms; therefore the cutoff is essentially $K = \min(L_x^{-1}, L_y^{-1})$, and consequently the error is controlled by the ratio $g/\max(L_x, L_y)$ and not by the ratio h/L_z as claimed in the original EW3DC paper[9] and still commonly believed.

Not surprisingly, for non-neutral systems, one has to use in P3M/ELC exactly the same bcc (29) as in the EW3DC method. This correction is included in the formula for the ELC term derived recently in the most general case of a (possibly non-neutral) slab system in-between dielectric walls.[20] The algorithm described in Ref. 20 can thus be used without modification[21] to compute energies and forces in charged slab systems, in agreement with the regularization procedure of Sec. II.

IV. NUMERICAL TESTS

In this section the correctness and accuracy of the results stated in Sec. II are tested numerically by computing energies and forces in simple systems, using Eq. (9) with direct summation on one hand and the methods MMM2D, EW3DC, and P3M/ELC on the other hand. To specify what electrostatic method is used in the EW3DC approach, we add a suffix "slab" to the method, which yields denominations such as Ewald/slab, P3M/slab, or SPME/slab for the various incarnations. When the ELC term is employed in these methods to substract the unwanted interlayer interactions, we denote the methods by Ewald/ELC, P3M/ELC, and SPME/ELC.

A. Madelung energy of a 2D square charge lattice

We consider a square simulation box ($L_x = L_y = L$) containing a single charge q under 2D-PBC. The regularized electrostatic energy (9) of this system is equal to the sum of the Coulomb interactions of the charge with all its periodic images and with a neutralizing sheet in the plane of the 2D charge lattice (because the two neutralizing walls coalesce into a single sheet since $h=0$). The energy of this system takes the form

$$E = -\frac{q^2}{2}\phi_{\text{PBC}}(\mathbf{0}) = -\frac{q^2}{2}\frac{\mathcal{F}}{L}, \tag{31}$$

where $\mathcal{F}$ is a dimensionless constant. Using the regularized direct-space sum (9) to compute this Madelung energy, we obtain $\mathcal{F} \simeq 3.9$. A high accuracy is difficult to attain with formula (9) because it expresses the energy as a difference between two large numbers and also because the convergence with R is slow. Nijboer and Ruijgrok found $\mathcal{F} \simeq 3.900\,26$ by using another formula better suited for this purpose.[22]

Using the MMM2D method implemented in the simulation package ESPRESSO (Ref. 23) and tuning the algorithm to accuracy 10^{-10}, we obtain

$$\mathcal{F} \simeq 3.900\,264\,920\,0, \tag{32}$$

where the precision is mainly limited by the employed implementation of the Bessel functions. The MMM2D method clearly works for non-neutral systems and is able to compute quickly 2D Madelung energies to a very high accuracy.

We computed also the energy of this Madelung system by using the P3M/ELC method also implemented in ESPRESSO. We used an accuracy goal of 10^{-4}, which resulted in the following values for the various parameters: Ewald splitting coefficient $\alpha = 7.251\,84L^{-1}$, real-space cutoff $r_{\text{cut}} = 0.45L$, mesh with 32 points in each direction, charge assignment order=6, gap size=$0.1L$, and ELC far cutoff $=26/L$. This calculation reproduced the first five digits of $\mathcal{F}$, thus validating the bcc (29) for the EW3DC and P3M/ELC methods.

B. Forces between two particles

In molecular dynamics simulations, one is mainly interested in the accuracy of the force computation, since the forces govern the dynamics of the system. Figure 1 shows the pair force between two like charged unit point charges at (0,0,0) and $(1,1,z)$ in a unit cell of dimensions $L_x = L_y = L_z = L = 10$ under 2D-PBC; z changes from 0 to L. This system is characterized by a net charge $Q_{\text{tot}} = 2$ and a varying dipole

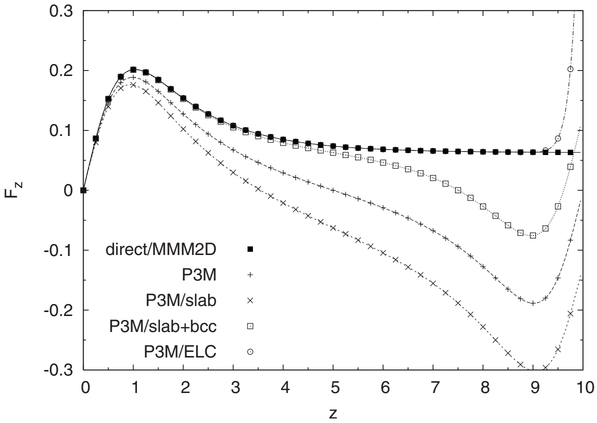


FIG. 1. The force F_z between two unit charges at a relative position of $(1,1,z)$, as a function of z for various algorithms. P3M/slab stands for the EW3DC approach using P3M as the underlying electrostatic method, while P3M/slab+bcc is the same method with the addition of the bcc (30).

moment $M_z=qz$; therefore it is well suited to demonstrate the effect of the net charge in the various electrostatic methods that we tested:

- Well-converged direct summation taking into account 100^2 images,
- MMM2D method of Ref. 7,
- P3M with metallic boundary conditions, i.e., without surface term,
- P3M with the slab-corrected surface term following Ref. 9,
- P3M with the slab-corrected surface term and bcc (30),
- P3M with the ELC extension following Ref. 20.

MMM2D and the direct summation agree in at least two digits; therefore only the results for MMM2D are shown. MMM2D was again tuned for a precision of at least 10^{-10}, and P3M and ELC for a precision better than 10^{-4}. When $z > L/2$, the force computed with direct summation and MMM2D takes its asymptotic value $2\pi/(L_xL_y)$, which corresponds to the force exerted on a unit charge by a charged sheet with surface charge density $1/(L_xL_y)$.

P3M with slabwise summation order and bcc (30) agrees well with the direct summation in the expected range of validity, namely, for $z < L/3$. Correspondingly, P3M with the ELC extension, which was tuned to agree with high precision for $z < L_z/2$, agrees well up to $z=9$. In contrast to this, applying only the slab correction of Ref. 9 actually *increases* the error of the plain P3M method with metallic boundary conditions.

Would the two unit charges be oppositely charged and hence the system neutral, the force between the charges would simply be the opposite of those presented in Fig. 1. This is indeed the case for all methods, again with the exception of P3M/slab, which gives correct forces only in systems with no net charge.

V. EXAMPLE: MEAN FORCE ON AN ION NEAR AN INTERFACE

A precise understanding of the adsorption, or depletion, of ions near an interface is important in several fields, for example, in biophysics, where the hydrophobic solid/water interface governs the aggregation and folding of apolar molecules, and in atmospheric chemistry, where physicochemical processes occur at the air/water and ice/water interfaces. Potentials of mean force for ions at infinite dilution have been obtained recently at such interfaces.[24–26] In these works, the simulated systems had a net charge and the long-range nature of the Coulomb force was taken into account by using 3D Ewald summations with a gap in the simulation box (EW3DC method if the summation order was changed to slabwise). As shown in Sec. IV, it is essential to include the newly derived bcc (30) for the electrostatic forces to be correct in such non-neutral simulations.

According to macroscopic electrostatics, an ion close to a dielectric interface sees an electrostatic potential that corresponds to a (fictitious) image charge located symmetrically on the opposite side of the interface.[27] In the case of an ion close to an air/water interface, the dielectric contrast is about 80, and this image-charge effect is expected to contribute to the mean force on the ion as it approaches the surface. We performed a molecular dynamics simulation of this non-neutral system to show that the EW3DC method, when complemented with the bcc (30), does lead to results that agree with the prediction of macroscopic electrostatics.

The simulation box was setup similarly as in Refs. 24 and 26: A sodium ion carrying charge $q=+e$ is at a distance z from the surface of a water slab of width $h=2.483$ nm made up of 2048 water molecules. The dimensions of the simulation box are $L_x=L_y=2h$ and $L_z=5h$, i.e., a gap of width $4h$ is introduced in the nonperiodic direction normal to the slab. Water molecules are described according to the SPC/E model[28] and the Lennard-Jones parameters for the sodium-water interaction are taken from Ref. 29 in their nonpolarizable form. The simulations were performed using a modified version of the GROMACS simulation package,[30] in which we implemented the bcc. The equations of motion were integrated using the leap-frog algorithm with a time step of 2 ps, and the temperature was kept constant at 300 K using the Berendsen method. The Coulomb interactions were computed using the SPME method[11] with a slabwise summation order and bcc (SPME/slab+bcc). The cutoff for the van der Waals and real-space Coulomb forces are both set to 1 nm.

The sodium ion is free to move in the x-y directions, but its distance to the water surface is kept fixed during each simulation (a rigid constraint prevents the z coordinate of the ion to change). The position of the surface is defined by the criterion that the density of water at the surface is half that of the bulk. For each considered distance of the ion, we performed a 5 ns long simulation to determine the mean force F_z acting on the ion. We simulated also the case of a water surface confined by a (hydrophobic) hard wall located 0.1 nm above the water surface. In the latter case, the water

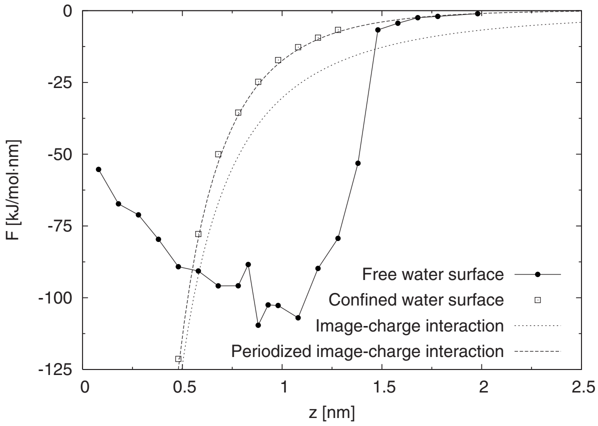


FIG. 2. Mean force on a sodium ion at a distance z from a water surface, which is either free (black circles) or confined by a hydrophobic hard wall located 0.1 nm above the surface (white squares). The image-charge force predicted by macroscopic electrostatics for a dielectric interface located at $z=-0.22$ Å is shown for two different choices of boundary conditions: zero field at infinity (dotted line) and 2D-PBC (dashed line).

molecules whose oxygen atom hit the hard wall were reflected back into the liquid by inverting the z-component of their velocity.

When the water molecules are prevented by the hard wall to partially solvate the ion, the measured force on the ion is well predicted by the image-charge interaction of macroscopic electrostatics (see Fig. 2). According to this theory, an ion of charge q in a medium of permittivity $\epsilon_1 \simeq 1$ (air) and at a distance z from a slab of width h and permittivity $\epsilon_2 \simeq 80$ (water) sees an electrostatic potential[31]

$$\Phi(z) = \frac{q}{\epsilon_1}\left\{ \frac{\epsilon_1-\epsilon_2}{\epsilon_1+\epsilon_2}\frac{1}{2z} - \frac{4\epsilon_1\epsilon_2}{(\epsilon_1+\epsilon_2)^2}\sum_{n=1}^{\infty}\left(\frac{\epsilon_1-\epsilon_2}{\epsilon_1+\epsilon_2}\right)^{2n-1}\frac{1}{2z+2nh}\right\}. \qquad (33)$$

The first term in (33) is an image-charge due to the first interface, while the infinite sum corresponds to additional image charges that are due to the presence of the second interface at the other side of the water film. The force on the ion derives from its electrostatic potential energy $\int_0^q \Phi(z)dq = \frac{1}{2}q\Phi(z)$ and is plotted as a dotted line in Fig. 2. The potential (33) corresponds to the interaction of an ion with a dielectric slab under the boundary condition of zero field at infinity, while the electrostatic potential is computed in the simulations under 2D-PBC. To account for this periodicity in the x-y directions, one has to sum up over the periodic images of the image charges. This was done by setting up a cell containing the ion and the infinite set of image charges[32] that appear in Eq. (33) and by computing the force on the ion using the MMM2D method. The result, assuming the continuous dielectric interface to be placed at a distance $d' = 0.22$ Å below the molecular water surface, is shown as a dashed line in Fig. 2. That prediction of macroscopic electrostatics agrees very well with the simulation results. The agreement holds at all distances in the case of the confined water surface, while it holds only at large distances in the case of the free water surface. The simulations show that the ion is indeed partially solvated even when quite far away from the water surface (up to $z \simeq 1.45$ nm).

The fact that the dielectric interface is located very close to the molecular water surface agrees with the very sharp dielectric permittivity profile of a water slab determined in Ref. 33. A more detailed analysis of the relative position between the molecular water surface and the position of the dielectric interface (in the case of a confinement more realistic than a hard wall) will be performed in a future work.

We note that finite-size effects are quite pronounced with the size of the simulation box used in this example. The prediction of macroscopic electrostatics changes indeed substantially once the periodicity in the x-y directions is introduced (compare dotted and dashed lines in Fig. 2). This is not unexpected since Coulomb interactions with image charges in the neighboring cells are negligible only if L_x and L_y are sufficiently large. To determine a force profile that is free from finite-size effects, a larger simulation box should clearly be used.

Because of the use of the slab geometry, the ion sees two dielectric interfaces instead of a single one. This unwanted feature can be avoided by introducing a hard wall with the same dielectric permittivity as water at the other side of the slab. This will remove almost entirely the unwanted effects of the second dielectric interface. The latest versions of the MMM2D and P3M/ELC methods that include image charges can be used to simulate such a slab system with different continuous dielectric media of permittivities ϵ_1 and ϵ_2 at the two sides of the slab.[20,34]

VI. CONCLUSIONS

A regularization of the electrostatic energy of 2D periodic slab systems bearing a net charge has been proposed. That regularization is the analog in 2D-PBC of the usual regularization of Ewald sums in three dimensions by a homogeneous neutralizing background charge density. The regularized energies, given by Eq. (9), can be interpreted as resulting from a neutralization of the system by two charged walls on each side of the slab, with the prescription that the constant shift in the energy proportional to the wall-wall separation is omitted. We proved that energy differences between two systems characterized by a different net charge take meaningful values, which include variations in the self-energies of the charges in 2D-PBC. The method can thus be used to simulate systems with a varying net charge, for example, to determine charging free energies in inhomogeneous systems with slab geometry.

The regularized energies, and the forces that derive from them, can be computed efficiently by using advanced electrostatic algorithms for 2D periodic systems, such as MMM2D, EW3DC, and P3M/ELC. We proved that the MMM2D method can be applied without any modification to non-neutral systems, while methods based on the EW3DC approach necessitate the use of a bcc, given by Eq. (28), which affects both energies and forces. The bcc removes the effect of the parabolic electrostatic potential inherited from the neutralization of the cell implicit in the 3D-Ewald method applied to a charged slab system.

Numerical tests have been done to demonstrate the correctness and the accuracy of the results. When the bcc is added to the EW3DC results, the three methods (MMM2D, EW3DC+bcc, and P3M/ELC) give consistent results for the energies and forces, that agree also with well converged direct-sum calculations. As shown in Fig. 1, the application to non-neutral systems of the uncorrected EW3DC approach leads to incorrect forces which are even further away from the reference direct-sum result than the plain 3D-Ewald method with metallic boundary conditions. We stress that the surface terms with bccs (29) and (30) are easy to evaluate and have no computational cost in a simulation.

The three methods reproduce the known value of the Madelung self-energy on an ion in 2D-PBC, which can be computed quickly to very high accuracy using the MMM2D method. That energy, given by $E_{\mathrm{self}}=-q^2\mathcal{F}/(2L)$ with $\mathcal{F}\simeq 3.900\,264\,920\,0$ in a square simulation box ($L_x=L_y=L$), plays an important role in the stability of 2D Wigner crystals, in charging free energy calculations under PBC and in the study of finite-size effects.[18,22,35]

A further demonstration that our method provides correct long-range electrostatic forces is provided in the form of a calculation via molecular dynamics simulations of the mean force on a sodium ion close to an air/water interface. A nice agreement with the image-charge prediction of macroscopic electrostatics is observed within the limits of validity of the theory. The dielectric discontinuity at the air/water interface in the macroscopic model is found to be situated at the very surface of the water slab (at a distance of 0.22 Å under the molecular water surface defined by the criterion used in Sec. V).

We expect the results of the present work to be of wide applicability to simulations of slab systems whenever an implicit uniform background of charges can be assumed. Research works related to the calculation of ionic solvation free energies, study of surfaces with charged defects, bilayers of charged particles, etc., can substantially benefit from the derivations made in the present study.

ACKNOWLEDGMENTS

J. J. Cerdà wants to thank the financial support of the Spanish Ministerio de Educación *y* Ciencia, postdoctoral Grant No. EXP2006-0931. All authors are grateful to the DAAD organization and the French *Ministère des affaires étrangères et européennes* for providing financial support.

APPENDIX A: ENERGY DENSITY AND FACTORS 1/2

This appendix explains the origin of factor 1/2 in some formulas, for example Eq. (1) when $(n_x,n_y)\neq(0,0)$ and Eqs. (7) and (8), but not in others such as Eqs. (6) and (14). We first note that, as a rule of thumb, interaction energy $E_{a\text{-}b}$ between two identical objects ($a=b$) gets a factor 1/2, while no such factor must be included when $a\neq b$.

The factor 1/2 is a consequence of the fact that, when using PBC, the energy that one computes is not the total energy of the simulation box itself, but the energy *per box* of the macroscopic sample made up of many copies of the primary cell. Indeed, dividing the macroscopic sample in B boxes, its energy can be written as

$$E_{\mathrm{sample}}=\sum_{b=1}^{B}E_b+\frac{1}{2}\sum_{\substack{b,b'=1\\ b'\neq b}}^{B}E_{b\text{-}b'}, \tag{A1}$$

where E_b is the energy of box b and $E_{b\text{-}b'}$ is the interaction energy between particles in box b and particles in box b'. Since the macroscopic sample is large, the number of boxes near the surface of the sample is negligible in front of the number of boxes in the volume of the sample. We can therefore write, assuming a homogeneous sample divided in B identical boxes of volume V_b,

$$\frac{1}{2}\sum_{\substack{b,b'=1\\ b'\neq b}}^{B}E_{b\text{-}b'}\simeq B\frac{1}{2}\sum_{\substack{b'=1\\ b'\neq\text{primary box } b}}^{B}E_{b\text{-}b'}. \tag{A2}$$

We conclude that the energy per box is

$$E:=\frac{E_{\mathrm{sample}}}{B}=E_b+\frac{1}{2}\sum_{\substack{b'=1\\ b'\neq\text{primary box } b}}^{B}E_{b\text{-}b'}. \tag{A3}$$

All interactions between the simulation box and the periodic images of the simulation box must therefore be accounted for with a factor 1/2 to avoid double counting in the energy of the macroscopic sample. Notice that the ratio of the energy per box divided by the volume of the simulation box, $e:=E/V_b=E_{\mathrm{sample}}/(V_bB)$, does measure correctly the energy density of the macroscopic sample since V_bB is the volume of that sample. This explains the factor 1/2 in Eq. (1) when $(n_x,n_y)\neq(0,0)$.

When the system contains a neutralizing background, the interaction of the primary box b_1 with an image box b_2 is the sum of four interactions energies (which must all be accounted for with a factor 1/2) between four different entities:

$$E_{b_1\text{-}b_2}=E_{c_1\text{-}c_2}+E_{c_1\text{-}\rho_2}+E_{\rho_1\text{-}c_2}+E_{\rho_1\text{-}\rho_2}, \tag{A4}$$

where c_i (ρ_i) stands for the charges (neutralizing background charge density) in box b_i, $i=1,2$. The absence of factor 1/2 in Eq. (6) results therefore from the fact that this energy is actually the sum of the two energies $E_{c_1\text{-}\rho_2}$ and $E_{\rho_1\text{-}c_2}$ which lead to the same value after summation over all boxes. The energies $E_{c_1\text{-}c_2}$ and $E_{\rho_1\text{-}\rho_2}$ appear only once in Eq. (A4), so there is indeed a factor 1/2 in equations such as Eq. (8) or Eq. (7) when $\boldsymbol{r}'\notin V_{b_1}$. When $\boldsymbol{r}'\in V_{b_1}$, the factor 1/2 in Eq. (7) simply prevents double counting within the box, according to the usual relation

$$\sum_i\sum_{j>i}E_{i\text{-}j}=\frac{1}{2}\sum_{\substack{i,j\\ i\neq j}}E_{i\text{-}j}. \tag{A5}$$

APPENDIX B: EFFECTIVE KERNEL FORMULATION

In this appendix, we demonstrate that the energy given by Eq. (9) of a charged system under 2D-PBC can be written as a sum over an effective kernel $\phi_{\mathrm{PBC}}(\boldsymbol{r})$ as in (18), similar to the Ewald sum in 3D-PBC.[36,37] To this aim, we reformulate Eq. (3) as

$$\begin{aligned} E(R) = {} & \frac{1}{2}\sum_{i,j=1}^{N} q_i q_j V(\boldsymbol{r}_i - \boldsymbol{r}_j) + \frac{1}{2}\sum_{i,j=1}^{N} q_i q_j \sum_{(n_x,n_y)\in D(R)} \\ & \times \frac{1}{|n_x\boldsymbol{L}_x + n_y\boldsymbol{L}_y|} + \frac{1}{2}\sum_{i,j=1}^{N} q_i q_j \sum_{(n_x,n_y)\in D(R)} \\ & \times \left(\frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j + n_x\boldsymbol{L}_x + n_y\boldsymbol{L}_y|} - \frac{1}{|n_x\boldsymbol{L}_x + n_y\boldsymbol{L}_y|}\right), \end{aligned} \tag{B1}$$

where

$$V(\boldsymbol{r}) = \begin{cases} 0 & \text{if } \boldsymbol{r} = \boldsymbol{0} \\ 1/|\boldsymbol{r}| & \text{if } \boldsymbol{r} \neq \boldsymbol{0}, \end{cases} \tag{B2}$$

and $D(R)$ denotes all $(n_x,n_y)\in\mathbb{Z}^2$ with $0<(n_xL_x)^2+(n_yL_y)^2 \leq R^2$. The first term is already of the desired form; by approximation of the sum by an integral, one can see that the second term is of the form

$$\frac{1}{2}\sum_{(n_x,n_y)\in D(R)} \frac{1}{|n_x\boldsymbol{L}_x + n_y\boldsymbol{L}_y|} = \frac{\pi R}{L_xL_y} + S + \mathcal{O}(R^{-1}), \tag{B3}$$

where the first term added up over all particles gives the singularity $(\pi Q_{\mathrm{tot}}^2/L_xL_y)R$ that is subtracted in Eq. (9), and the constant term S is the self-energy of a unit charge, which is given by $S=-\mathcal{F}/(2L)$ with $\mathcal{F}\simeq 3.900\,26$ in the case of a square lattice $L_x=L_y=L$ [see Eq. (31)]. Regarding the third term in Eq. (B1), we note that for two vectors $\boldsymbol{r}$ and $\boldsymbol{a}$,

$$\frac{1}{|\boldsymbol{r}+\boldsymbol{a}|} + \frac{1}{|\boldsymbol{r}-\boldsymbol{a}|} - \frac{2}{|\boldsymbol{a}|} = \mathcal{O}(|\boldsymbol{a}|^{-3}), \tag{B4}$$

which means that the last sum over (n_x,n_y) in the third term is in fact absolutely convergent provided the summands for (n_x,n_y) and $(-n_x,-n_y)$ are added in pairs, i.e., as in Eq. (B4) with $\boldsymbol{a}=n_x\boldsymbol{L}_x+n_y\boldsymbol{L}_y$ (note that in 3D-PBC, the analogous sum would be only conditionally convergent). Therefore, Eq. (B1) can be written as

$$\lim_{R\to\infty}\left(E(R) - \frac{\pi Q_{\mathrm{tot}}^2}{L_xL_y}R\right) = \frac{1}{2}\sum_{i,j=1}^{N} q_iq_j\phi_{\mathrm{PBC}}(\boldsymbol{r}_i-\boldsymbol{r}_j), \tag{B5}$$

where

$$\begin{aligned} \phi_{\mathrm{PBC}}(\boldsymbol{r}) \coloneqq {} & V(\boldsymbol{r}) + S + \lim_{R\to\infty}\sum_{(n_x,n_y)\in D(R)} \\ & \times \frac{1}{2}\left(\frac{1}{|\boldsymbol{r}+n_x\boldsymbol{L}_x+n_y\boldsymbol{L}_y|} - \frac{1}{|n_x\boldsymbol{L}_x+n_y\boldsymbol{L}_y|}\right). \end{aligned} \tag{B6}$$

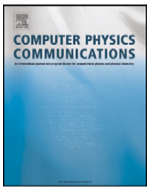
[1] D. E. Parry, Surf. Sci. **49**, 433 (1975).
[2] D. M. Heyes, M. Barber, and J. H. R. Clarke, J. Chem. Soc., Faraday Trans. 2 **73**, 1485 (1977).
[3] J. Hautman and M. L. Klein, Mol. Phys. **75**, 379 (1992).
[4] B. R. A. Nijboer and F. W. de Wette, Physica (Amsterdam) **23**, 309 (1957).
[5] E. R. Smith, Mol. Phys. **65**, 1089 (1988).
[6] M. Kawata, M. Mikami, and U. Nagashima, J. Chem. Phys. **116**, 3430 (2002).
[7] A. Arnold and C. Holm, Comput. Phys. Commun. **148**, 327 (2002).
[8] S. Tyagi, Phys. Rev. E **70**, 066703 (2004).
[9] I.-C. Yeh and M. L. Berkowitz, J. Chem. Phys. **111**, 3155 (1999).
[10] R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles* (IOP, Bristol, 1988).
[11] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).
[12] A. Arnold, J. de Joannis, and C. Holm, J. Chem. Phys. **117**, 2496 (2002).
[13] S. Alireza Ghasemi, A. Neelov, and S. Goedecker, J. Chem. Phys. **127**, 224102 (2007).
[14] G. Hummer, L. R. Pratt, and A. E. Garcia, J. Phys. Chem. **100**, 1206 (1996).
[15] A. Arnold and C. Holm, Adv. Polym. Sci. **185**, 59 (2005).
[16] V. Ballenegger, J. J. Cerda, O. Lenz, and C. Holm, J. Chem. Phys. **128**, 034109 (2008).
[17] G. Hummer, L. R. Pratt, and A. E. Garcia, J. Phys. Chem. **102**, 7885 (1998).
[18] P. H. Hünenberger and J. A. McCammon, J. Chem. Phys. **110**, 1856 (1999).
[19] H. D. Herce, A. E. Garcia, and T. Darden, J. Chem. Phys. **126**, 124106 (2007).
[20] S. Tyagi, A. Arnold, and C. Holm, J. Chem. Phys. **129**, 204102 (2008).
[21] Note that the last term in Eq. (3.10) of Ref. 20 should have a minus sign.
[22] B. R. A. Nijboer and Th. W. Ruijgrok, J. Stat. Phys. **53**, 361 (1988).
[23] H. J. Limbach, A. Arnold, B. A. Mann, and C. Holm, Comput. Phys. Commun. **174**, 704 (2006) (http://espresso.mpg.de).
[24] P. Jungwirth and D. J. Tobias, Chem. Rev. (Washington, D.C.) **106**, 1259 (2006).
[25] E. J. Smith, T. Bryk, and A. D. J. Haymet, J. Chem. Phys. **123**, 034706 (2005).
[26] D. Horinek and R. R. Netz, Phys. Rev. Lett. **99**, 226104 (2007).
[27] J. D. Jackson, *Classical Electrodynamics*, 3rd ed. (Wiley, New York, 1999).
[28] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, J. Phys. Chem. **91**, 6269 (1987).
[29] D. E. Smith and L. X. Dang, J. Chem. Phys. **100**, 3757 (1994).
[30] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. **7**, 306 (2001) http://www.gromacs.org.
[31] G. Iversen, Y. I. Kharkats, and J. Ulstrup, Mol. Phys. **94**, 297 (1998).
[32] It is actually sufficient to introduce only the two first image charges: a charge $q_1=-q(\epsilon_1-\epsilon_2)/(\epsilon_1+\epsilon_2)$ at position $-z$ and a charge $q_2=-q_1$ at position $-(z+h)$. Notice that $q_1+q_2=0$ in accordance with the charge neutrality of the slab.
[33] V. Ballenegger and J.-P. Hansen, J. Chem. Phys. **122**, 114711 (2005).
[34] S. Tyagi, A. Arnold, and C. Holm, J. Chem. Phys. **127**, 154723 (2007).
[35] S. Bogusz, T. E. Cheatham III, and B. R. Brooks, J. Chem. Phys. **108**, 7070 (1998).
[36] S. W. de Leeuw, J. W. Perram, and E. R. Smith, Proc. R. Soc. London, Ser. A **373**, 27 (1980).
[37] G. Hummer, L. R. Pratt, and A. E. Garcia, J. Phys. Chem. **99**, 14188 (1995).

Computer Physics Communications 182 (2011) 1919–1923

Contents lists available at ScienceDirect

Computer Physics Communications

www.elsevier.com/locate/cpc



Removal of spurious self-interactions in particle–mesh methods

V. Ballenegger [a,*], J.J. Cerdà [b], C. Holm [b]

[a] *Institut UTINAM, Université de Franche-Comté, UMR 6213, 16, route de Gray, 25030 Besançon cedex, France*
[b] *Institute for Computational Physics, Universität Stuttgart, Pfaffenwaldring 27, 70569 Stuttgart, Germany*

ARTICLE INFO

Article history:
Received 29 July 2010
Accepted 28 January 2011
Available online 15 February 2011

Keywords:
Particle–mesh methods
Accuracy
Conservation law
SPME
P3M

ABSTRACT

We derive an analytic formula for subtracting the spurious self-forces in particle–mesh methods that use the analytical differentiation scheme, such as the Smooth Particle Mesh Ewald (SPME) method and the Particle–Particle Particle–Mesh (P3M) method with analytical differentiation. The impact of the self-forces on the accuracy of the particle–mesh methods is investigated, and it is shown that subtracting them can improve the accuracy of the calculation for some choices of the method's parameters. It is also suggested to subtract exactly the approximate, mesh-computed, self-energy of each particle, replacing them by the exact value. Subtracting in this way the self-energy and self-force of each particle not only improves the accuracy, but also reduces the violation of momentum and energy conservation in particle–mesh methods with analytical differentiation.

1. Introduction

The computation of long-range forces in a many-particle system is a demanding task that scales a priori as $O(N^2)$ with the number N of particles. Particle–mesh methods reduce the complexity of this problem to $O(N \log N)$ by discretizing the system onto a mesh and by taking advantage of the Fast Fourier Transform (FFT) algorithm to solve the Poisson equation in Fourier space, where it reduces to a simple multiplication by the Green function. For higher efficiency, the pair interaction is moreover decomposed into a short-range and a long-range parts, with the short-range interactions computed in real-space, as in the Ewald method [1]. The Particle–Particle Particle–Mesh (P3M) method [2] and the Smooth Particle Mesh Ewald (SPME) method [3] are two such particle–mesh methods that are widely used to compute Coulomb or gravitational forces in computer simulations.

In particle–mesh methods, various routes have been proposed to compute forces from the mesh-based potential:

- differentiation in real space by finite differences,
- differentiation in real space using the exact gradient of the assignment function used to interpolate the charge density onto/from the mesh,
- differentiation in Fourier space by multiplying the potential by $i\mathbf{k}$.

The second route is referred to as the analytical differentiation (AD) scheme, and it is used in the optimal energy-conserving scheme discussed in Hockney and Eastwood's book [2]. The third route, called the $i\mathbf{k}$-differentiation or force-interpolation scheme, corresponds to calculate the electric (or gravitational) field at each mesh point, and to interpolate this vectorial "force" field back to the particle positions. The finite-differences scheme was favored by Hockney and Eastwood in their implementation of the P3M method, but the other schemes, implemented first in the PME method [4] (for $i\mathbf{k}$-differentiation) and in the SPME method [3] (for AD differentiation), proved to be efficient and were later also implemented in the P3M method [5,6].

Each scheme has its own merits and drawbacks. The $i\mathbf{k}$-differentiation scheme is the most accurate one but also the most computationally expensive one, as it requires 3 inverse FFTs to transform the vectorial electric field mesh back to real space (compared to only one inverse FFT of the potential mesh in the other schemes). The higher accuracy of this scheme permits however the use of a coarser mesh, and accepting the two additional Fourier transforms can be competitive on desktop computers (this route becomes however less attractive in parallel distributed-memory computers because of the global communications required by the FFT's) [7]. As shown by the thorough analysis of particle–mesh methods in Hockney and Eastwood's book, the $i\mathbf{k}$-differentiation and finite-differences schemes conserve momentum but not energy, while analytical differentiation conserves energy (in the limit of small time steps) but not momentum [2, §5.3.3, §5.5 and §7.6]. In the AD scheme, a correction to all forces must be applied to conserve at least the center-of-mass momentum [3]. This correction has unfortunately the collateral effect of breaking the exact conservation of energy (a mass-weighted correction reduces this drawback but does not remove it entirely) [8].

---

[*] Corresponding author. Tel.: +33 3 81 66 64 79.
E-mail address: vincent.ballenegger@univ-fcomte.fr (V. Ballenegger).

0010-4655/$ – see front matter
doi:10.1016/j.cpc.2011.01.026

When a local violation of momentum conservation can be accepted, for example when the system is coupled to a thermostat, analytical differentiation can be the most efficient route since it requires less operations than the other schemes (especially when compared to a finite difference operator of high order). The AD scheme is commonly used in constant-temperature simulations of condensed matter, plasmas and soft matter systems, including large-scale biophysical simulations.

The main drawback of analytical differentiation is of course its violation of momentum conservation, which arises both because the pair force (computed via the mesh) does not obey the action–reaction principle, and because each particle is subjected to a self-force that depends on its position relative to the mesh. It has been proposed recently to subtract the self-forces by using a tabulation of the self-force determined at the beginning of a simulation [9,10]. After a brief reminder of the AD scheme (Section 2), we derive in Section 3 analytical formulae for the self-force and self-energy, for any position of a particle within a mesh cell. These formulae provide a better understanding of the self-interactions, and can be used to subtract them directly without having to first tabulate them. The accuracy gain obtained by subtracting self-forces in the analytically differentiated P3M method and the SPME method is discussed in Section 4.

2. The analytical differentiation scheme

We consider a system made up of N charges q_i at positions $\boldsymbol{r}_i$ in a periodic box of dimensions L_1, L_2, L_3. The box is assumed to be orthorhombic, and its volume is $V = L_1 L_2 L_3$. The particle–mesh methods use a mesh $\mathbb{M}$ of dimensions N_1, N_2, N_3. The lattice spacings are $h_1 = L_1/N_1$, $h_2 = L_2/N_2$, $h_3 = L_3/N_3$ and the grid nodes are $\boldsymbol{r_n} = (h_1 n_1, h_2 n_2, h_3 n_3)$, $n_i = 0 \ldots N_i$. Formally, the energy of the system is approximated on the mesh by

$$E_{\mathrm{PM}} = \frac{1}{2} \sum_{\boldsymbol{r_n}, \boldsymbol{r_{n'}} \in \mathbb{M}} q_{\boldsymbol{r_n}} q_{\boldsymbol{r_{n'}}} G(\boldsymbol{r_n} - \boldsymbol{r_{n'}}) \tag{1}$$

where $G(\boldsymbol{r_n})$ is the lattice Green function in real space for periodic boundary conditions. The charges at each mesh point are computed via an assignment function $W(\boldsymbol{r})$,

$$q_{\boldsymbol{r_n}} = \sum_{i=1}^{N} q_i W(\boldsymbol{r}_i - \boldsymbol{r_n}), \tag{2}$$

where $W(\boldsymbol{r})$ is the Hockney–Eastwood charge assignment function [11,2], equal to a product of three one-dimensional centered B-splines of order P [12]. In the P3M method, the lattice Green function is adjusted so that Eq. (1) gives the best possible discrete approximation to the energy of the original system in continuous space, while the lattice Green function in the SPME method differs from the optimal one [12]. In the analytical differentiation scheme, the forces are obtained by differentiating analytically the approximate energy (1):

$$\begin{aligned} \boldsymbol{F}_i^{\mathrm{PM}} &= -\nabla_i E_{\mathrm{PM}} \\ &= -\sum_{\boldsymbol{r_n}, \boldsymbol{r_{n'}} \in \mathbb{M}} (\nabla_i q_{\boldsymbol{r_n}}) q_{\boldsymbol{r_{n'}}} G(\boldsymbol{r_n} - \boldsymbol{r_{n'}}) \\ &= -q_i \sum_{\boldsymbol{r_n} \in \mathbb{M}} \Phi_{\boldsymbol{r_n}} \nabla_i W(\boldsymbol{r}_i - \boldsymbol{r_n}) \end{aligned} \tag{3}$$

where we used the fact that $G(\boldsymbol{r_n})$ is even and introduced the mesh-based potential

$$\Phi_{\boldsymbol{r_n}} = \sum_{\boldsymbol{r_{n'}} \in \mathbb{M}} q_{\boldsymbol{r_{n'}}} G(\boldsymbol{r_n} - \boldsymbol{r_{n'}}). \tag{4}$$

The particle–mesh force (3) is hence obtained by interpolating the mesh-based potential back to the particle position using the gradient of the charge assignment function. The convolution (4) for the potential mesh is computed in practice as a multiplication in Fourier space by the reciprocal lattice Green function $\widetilde{G}(\boldsymbol{k_n})$, the result of which is transformed back to real space to compute the forces via (3). The expression of the optimal $\widetilde{G}(\boldsymbol{k_n})$ for the analytically differentiated P3M method can be found in Ref. [6], while the corresponding expression for the SPME method is derived in Ref. [13]:

$$\widetilde{G}_{\mathrm{SPME}}(\boldsymbol{k_n}) = \frac{\hat{\phi}(\boldsymbol{k_n})}{(\sum_{\boldsymbol{m} \in \mathbb{Z}^3} \hat{U}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}} \cdot \boldsymbol{m}}))^2} \tag{5}$$

where $\underline{\mathbf{N}} = \mathrm{diag}(N_1, N_2, N_3)$ is a diagonal matrix made with the mesh dimensions and $\hat{\phi}(\boldsymbol{k})$ is the pair interaction between the particles in Fourier space. When the pair interaction is decomposed according to the Ewald method [1,14], $\phi(\boldsymbol{r})$ corresponds to the Coulomb interaction of a point charge interacting with a Gaussian charge distribution of width α^{-1}: $\hat{\phi}(\boldsymbol{k}) = \frac{4\pi}{k^2} \exp(-k^2/(4\alpha^2))$. The complementary interaction $\psi(\boldsymbol{r}) \equiv 1/r - \phi(\boldsymbol{r}) = (1 - \mathrm{erf}(\alpha r))/r$ is short-ranged and computed in real space by direct pair-wise summation. The Ewald screening length α^{-1} is a free parameter that impacts, with an exponential dependence, the accuracy of the method, but not its computational cost; it has to be fine-tuned for getting the best accuracy (see Section 4). The wavevector $\boldsymbol{k_n}$ belongs to the finite reciprocal mesh $\widetilde{\mathbb{M}} = \{\boldsymbol{k_n} = n_1 \frac{2\pi}{L_1} \hat{\boldsymbol{e}}_1 + n_2 \frac{2\pi}{L_2} \hat{\boldsymbol{e}}_2 + n_3 \frac{2\pi}{L_3} \hat{\boldsymbol{e}}_3\}$ where $n_\beta = -N_\beta/2 + 1, -N_\beta/2 + 2, \ldots, N_\beta/2$, $\beta = 1, 2, 3$. The function $\hat{U}(\boldsymbol{k}) = \hat{W}(\boldsymbol{k})/(h_1 h_2 h_3)$ is equal to the Fourier transform of the B-spline assignment function $W(\boldsymbol{r})$ of order P divided by the volume of a mesh cell:

$$\hat{U}(\boldsymbol{k_n}) = \left(\frac{\sin(\pi n_1/N_1)}{\pi n_1/N_1} \frac{\sin(\pi n_2/N_2)}{\pi n_2/N_2} \frac{\sin(\pi n_3/N_3)}{\pi n_3/N_3} \right)^P. \tag{6}$$

3. Expressions for the self-force and self-energy

The self-force is an artefact caused by the analytical differentiation scheme which breaks a symmetry: the forward and backward mapping of the charges onto/from the mesh is not performed using the same charge assignment function, since its gradient is used in the force interpolation (3) [2, p. 151]. Cerutti et al. [9] measured the self-force at various points in a mesh cell, and showed that the self-force, along direction $\beta = 1, 2, 3$, on a charge q located at $\boldsymbol{r}$ is well described by a Fourier sine series

$$F_\beta^{\mathrm{self}}(\boldsymbol{r}) \simeq q^2 \sum_{n=1}^{\infty} a_\beta^{(n)} \sin(n 2\pi s_\beta), \quad s_\beta = \frac{r_\beta}{h_\beta}. \tag{7}$$

Note that the self-force is periodic over distances h_β, as expected. More generally, the self-force can be expanded as

$$F_\beta^{\mathrm{self}}(\boldsymbol{r}) = q^2 \sum_{\boldsymbol{m} \in \mathbb{Z}^3} b_\beta^{(\boldsymbol{m})} \sin\bigl(2\pi (m_1 s_1 + m_2 s_2 + m_3 s_3)\bigr). \tag{8}$$

We prove in Appendix A that, for an orthorhombic simulation cell, the Fourier coefficients $b_\beta^{(\boldsymbol{m})}$ are given by

$$b_\beta^{(\boldsymbol{m})} = \frac{2\pi m_\beta}{h_\beta} \frac{1}{2V} \sum_{\boldsymbol{k_n} \neq 0} \widetilde{G}(\boldsymbol{k_n}) \sum_{\boldsymbol{m}' \in \mathbb{Z}^3} \hat{U}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}} \cdot \boldsymbol{m}'}) \hat{U}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}} \cdot (\boldsymbol{m}'+\boldsymbol{m})}). \tag{9}$$

We note that these coefficients are odd in $\boldsymbol{m}$: $b_\beta^{(-\boldsymbol{m})} = -b_\beta^{(\boldsymbol{m})}$. The coefficients $a_\beta^{(n)}$ in expansion (7) are therefore given by $a_\beta^{(n)} =$

$2b_\beta^{(n\hat{\boldsymbol{e}}_\beta)}$, or, fully explicitly,

$$a_\beta^{(n)} = \frac{2\pi m_\beta}{h_\beta}\frac{1}{V}\sum_{\boldsymbol{k_n}\neq 0}\widetilde{G}(\boldsymbol{k_n})\sum_{\boldsymbol{m}'\in\mathbb{Z}^3}\hat{U}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\cdot\boldsymbol{m}'})\hat{U}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\cdot(\boldsymbol{m}'+\boldsymbol{m})}) \tag{10}$$

where $\boldsymbol{m}$ is fixed to $\boldsymbol{m} = n\hat{\boldsymbol{e}}_\beta$ with $\hat{\boldsymbol{e}}_\beta$ the unit vector in direction β. The general formula (9) for the coefficient $b_\beta^{(\boldsymbol{m})}$ can be used to compute terms beyond the decoupled approximation (7), for example terms varying like $\sin(2\pi(s_1+s_2))$, $\sin(2\pi(s_1+s_3))$ or $\sin(2\pi(s_2+s_3))$.

The self-interaction energy $E_{\mathrm{MS}}(\boldsymbol{r})$ of a charged particle located at $\boldsymbol{r}$ can also be expressed as a Fourier series. As shown in Appendix A, we have

$$E_{\mathrm{MS}}(\boldsymbol{r}) = q^2\sum_{\boldsymbol{m}\in\mathbb{Z}^3} c^{(\boldsymbol{m})}\cos\bigl(2\pi(m_1s_1+m_2s_2+m_3s_3)\bigr) \tag{11}$$

with coefficients given by

$$c^{(\boldsymbol{m})} = -\frac{1}{2V}\sum_{\boldsymbol{k_n}\neq 0}\widetilde{G}(\boldsymbol{k_n})\sum_{\boldsymbol{m}'\in\mathbb{Z}^3}\hat{U}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\cdot\boldsymbol{m}'})\hat{U}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\cdot(\boldsymbol{m}'+\boldsymbol{m})}). \tag{12}$$

$E_{\mathrm{MS}}(\boldsymbol{r})$ is the approximate (mesh-computed) "Madelung self-energy" of a particle (see [15]), defined as the sum of the interaction energy of the particle with itself ("Ewald" self-energy) and with its own periodic images (Madelung energy). Note that $\boldsymbol{F}^{\mathrm{self}}(\boldsymbol{r}) = -\nabla E_{\mathrm{MS}}(\boldsymbol{r})$. The self-force is therefore a consequence of the approximate space-dependent self-energy and of the analytical differentiation scheme.

4. Numerical results and discussion

Our test system is made up of $N = 800$ charged particles (400 carry a positive and 400 a negative unit charge) distributed at random in a cubic simulation box of length $L = 20$ units. Forces are measured in units of $\mathcal{C}^2/\mathcal{L}^2$ where the unit of charge $\mathcal{C}$ and the unit of length $\mathcal{L}$ are arbitrary ($\mathcal{C}$ and $\mathcal{L}$ could be for example the electronic charge and one Ångström, or a solar mass and a parsec in a cosmological simulation). If the unit of length is multiplied by a scale factor $s > 0$ ($\mathcal{L}' = s\mathcal{L}$), forces are multiplied by a factor s^{-2} and the density of the system is multiplied by factor s^{-3} (i.e. it is reduced if $s > 1$). By applying such a scaling to our results, the density of our test system (0.1 particle/$\mathcal{L}^3$) can easily be converted into any other value. The results obtained for our test system are therefore representative for charged systems characterized by a uniform random charge distribution of arbitrary density. We chose the same reference density 0.1 particle/$\mathcal{L}^3$ as in Ref. [12] but work with a system size twice larger for better statistics and for allowing larger real-space cutoffs (note that results for $L = 10$, $N = 100$, mesh size $M = N_1N_2N_3 = 32^3$ and real-space cutoff $r_c = 4$ are equivalent by extensivity to results for $L = 20$, $N = 800$, mesh size $M = 64^3$ and the *same* cutoff $r_c = 4$). We varied the mesh size from $M = 16^3$ to 64^3, the real-space cutoff from $r_c = 2$ to 9 and the spline interpolation order P from 3 to 7.

Fig. 1 shows the accuracy gain brought by subtracting the self-forces in the P3M algorithm with analytical differentiation, as a function of the Ewald splitting parameter α, for various choices of the mesh size and of the real-space cutoff distance r_c. The spline interpolation order is fixed to $P = 5$, but similar curves are found for other values of P. Analogous results are obtained for the SPME method (data not shown). When $r_c = 9$, the optimal accuracy of the P3M method is obtained when $\alpha \simeq 0.3$ for $M = 16^3$ and $\alpha \simeq 0.4$ for $M = 64^3$. The optimal value of α shifts to larger values when the real-space cutoff is reduced. For $r_c = 3$, the optimal value of α is for example $\alpha_{\mathrm{opt}} \simeq 0.83$ for $M = 32^3$ and $P = 5$. With parameters in this range, the accuracy of the forces is improved by about 30% when the self-forces are subtracted.

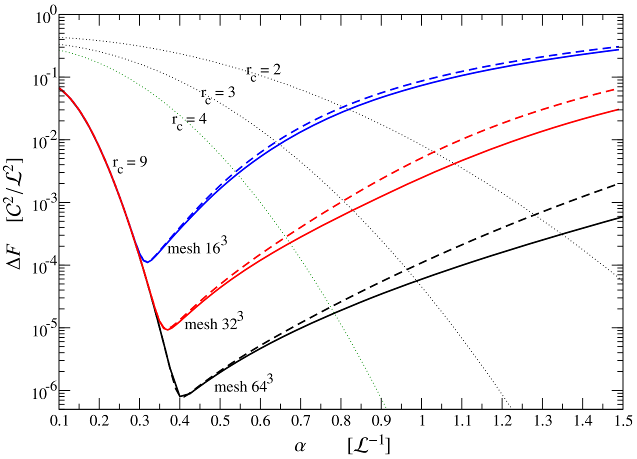


Fig. 1. The root-mean-square accuracy $\Delta F = \sqrt{\frac{1}{N}\sum_i(\boldsymbol{F}_i - \boldsymbol{F}_i^{\mathrm{exact}})^2}$ of the P3M forces (computed with analytical differentiation) is shown for our uniform test system (see text) as a function of the Ewald parameter α for different mesh sizes (from 16^3 to 64^3) and different real-space cutoff distances r_c (from 2 to 9). The spline interpolation order is set to $P = 5$. The accuracies of the particle–particle part of the force calculation for cutoffs $r_c = 2, 3$ and 4 are shown as dotted lines. Solid lines represent the accuracy of the particle–mesh forces obtained when self-forces are subtracted, while dashed lines correspond to the accuracy with self-forces included.

Table 1
Fourier coefficients $b_\beta^{(\boldsymbol{m})}$ for the self-force along direction $\beta = 1$ computed using Eq. (9) for a cubic box of side $L = 20$, mesh size $M = 32^3$, Ewald parameter $\alpha = 0.83$ and spline interpolation order $P = 5$. The self-force $F_{\beta=1}^{\mathrm{self}}(r_1, r_2, r_3)$ does not depend solely on particle coordinate r_1; its dependence on coordinates r_2 and r_3 is described by Fourier coefficients with vector $\boldsymbol{m} = (m_1, m_2, m_3)$ not purely along direction 1.

m_1	m_2	m_3	Fourier coefficient ($\mathcal{L}^{-2}$)
1	0	0	1.706×10^{-3}
2	0	0	1.528×10^{-4}
3	0	0	4.198×10^{-5}
4	0	0	1.722×10^{-5}
...			
1	±1	0	1.960×10^{-6}
1	0	±1	1.960×10^{-6}
2	±1	0	1.682×10^{-7}

When using a large cutoff $r_c \gtrsim 9$ (and hence a small α), no gain in accuracy is obtained by subtracting the self-forces: the dominant source of error is then not the self-forces, but errors in the interparticle forces. It is therefore worthy to subtract the self-forces only when one uses a relatively small cutoff in the simulation. The choice of the cut-off is dictated in part by the particular system under study (i.e. whether there are other short-range interactions to be computed in the pair-wise summation in real space) and by the respective timings, for various combination of parameters r_c, M and P, of the particle–particle part and of the particle–mesh part of the P3M (or SPME) method. Decreasing the cut-off makes indeed the particle–particle part of the algorithm faster, but the particle–mesh calculation has to be able to deliver the required accuracy sufficiently quickly. At the end, the cutoff is determined from the optimal set of parameters $\{r_c, M, P\}$ that provides the prescribed accuracy with the shortest possible computational time. Small cut-offs are usually present in simulations of dense (charged) systems.

We list in Table 1 the first few Fourier coefficients $b_\beta^{(\boldsymbol{m})}$ for the self-force in direction $\beta = 1$, calculated with Eq. (9) for parameter values in the region where the subtraction of the self-force brings a substantial increase in accuracy. Thanks to the rapid decay of the coefficients, the Fourier series of the self-force can be truncated essentially after the first two terms, in agreement with the findings of Ref. [9]. The self-force $F_{\beta=1}^{\mathrm{self}}(r_1, r_2, r_3)$ does not depend solely on particle coordinate r_1. Its dependence on the other coordinates is described by Fourier coefficients with vector $\boldsymbol{m} = (m_1, m_2, m_3)$ not purely along direction 1. As these terms are found to be smaller than the uncoupled terms $a_\beta^{(n)}$ for $n \leqslant 4$, they can safely be neglected, unless one wants to calculate the self-force to a very high precision.

Though particle–mesh algorithms with analytical differentiation don't conserve momentum, they do conserve energy in the limit of small time steps. The subsequent correction of forces to enforce conservation of the center-of-mass momentum breaks down however this exact conservation. To preserve it as well as possible, one has to subtract not only the self-forces, but also the interaction energy $E_{\mathrm{MS}}(\boldsymbol{r})$ of each particle with itself and with its periodic images, which is given by formula (11). The exact Madelung self-energy of a particle (which is a constant) can then be added back so that particles have the right self-energy associated to the periodic boundary conditions. In this way, the energy (1) of the discretized system is modified in a way consistent with the subtraction of the self-forces to maintain the equality $\boldsymbol{F}_i = -\nabla_i E_{\mathrm{PM}}$. A correction of this kind has been introduced in Ref. [16] to remove, *on average*, the bias of the P3M energies that is caused by the approximate (mesh computed) self-interactions. Correcting exactly these self-interactions at the level of each particle will improve the accuracy of the computed energies beyond the average correction of Ref. [16].

In summary, we have shown that the accuracy of particle–mesh calculations can be enhanced, especially for small cut-offs, by removing (or correcting in the case of the energy) the self-interaction of each particle. Our analytic expressions for the self-interactions, namely the self-force associated to the analytical differentiation scheme and the self-energy, can be evaluated during a simulation at virtually no computational cost. A further benefit of subtracting the self-interactions is that it reduces the amount of non-conservation of momentum (and energy) in particle–mesh methods with analytical differentiation.

Appendix A. Proof of formulae (8)–(9) for the self-force

Let $\boldsymbol{F}(\boldsymbol{r}_1, \boldsymbol{r}_2)$ be the particle–mesh force felt by a test particle with unit charge at $\boldsymbol{r}_1$ due to a particle with unit charge at $\boldsymbol{r}_2$ when analytical differentiation is used. That function can be expressed as a Fourier series in both variables

$$\boldsymbol{F}(\boldsymbol{r}_1, \boldsymbol{r}_2) = \frac{1}{V^2} \sum_{\boldsymbol{k}_1, \boldsymbol{k}_2} \hat{\boldsymbol{F}}(\boldsymbol{k}_1, \boldsymbol{k}_2)\, e^{i(\boldsymbol{k}_1 \cdot \boldsymbol{r}_1 + \boldsymbol{k}_2 \cdot \boldsymbol{r}_2)} \tag{A.1}$$

where the Fourier coefficients are given by [10]

$$\hat{\boldsymbol{F}}(\boldsymbol{k}_1, \boldsymbol{k}_2) = -i\boldsymbol{k}_1 V \hat{U}(\boldsymbol{k}_1) \hat{U}(\boldsymbol{k}_2) \widetilde{G}\big(\boldsymbol{k}_1^{\mathrm{G}}\big) \sum_{\boldsymbol{m} \in \mathbb{Z}^3} \delta_{\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_{\underline{\mathbf{N}} \cdot \boldsymbol{m}}}. \tag{A.2}$$

Vector $\boldsymbol{k}_1^{\mathrm{G}}$ is defined as in Ref. [10] – that is as vector $\boldsymbol{k}_1$ folded back into the first Brillouin zone $\widetilde{\mathbb{M}}$; matrix $\underline{\mathbf{N}}$ and vector $\boldsymbol{k}_{\boldsymbol{n}}$ (here for $\boldsymbol{n} = \underline{\mathbf{N}} \cdot \boldsymbol{m}$) are defined in Section 2 of the main text. The self-force is obtained by setting $\boldsymbol{r}_1 = \boldsymbol{r}_2$ in (A.1):

$$\begin{aligned}
&\boldsymbol{F}^{\mathrm{self}}(\boldsymbol{r}) \\
&\quad = \boldsymbol{F}(\boldsymbol{r}, \boldsymbol{r}) \\
&\quad = \frac{1}{V} \sum_{\boldsymbol{k}_1, \boldsymbol{k}_2} (-i\boldsymbol{k}_1) \hat{U}(\boldsymbol{k}_1) \hat{U}(\boldsymbol{k}_2) \widetilde{G}\big(\boldsymbol{k}_1^{\mathrm{G}}\big) \sum_{\boldsymbol{m}} \delta_{\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_{\underline{\mathbf{N}} \cdot \boldsymbol{m}}} e^{i(\boldsymbol{k}_1 + \boldsymbol{k}_2) \cdot \boldsymbol{r}} \\
&\quad = \frac{1}{V} \sum_{\boldsymbol{k}_1} (-i\boldsymbol{k}_1) \widetilde{G}\big(\boldsymbol{k}_1^{\mathrm{G}}\big) \hat{U}(\boldsymbol{k}_1) \sum_{\boldsymbol{m}} \hat{U}(\boldsymbol{k}_1 + \boldsymbol{k}_{\underline{\mathbf{N}} \cdot \boldsymbol{m}}) e^{-i\boldsymbol{k}_{\underline{\mathbf{N}} \cdot \boldsymbol{m}} \cdot \boldsymbol{r}} \\
&\quad = -\frac{1}{V} \sum_{\boldsymbol{k}_1} \boldsymbol{k}_1 \widetilde{G}\big(\boldsymbol{k}_1^{\mathrm{G}}\big) \hat{U}(\boldsymbol{k}_1) \sum_{\boldsymbol{m}} \hat{U}(\boldsymbol{k}_1 + \boldsymbol{k}_{\underline{\mathbf{N}} \cdot \boldsymbol{m}}) \sin(\boldsymbol{k}_{\underline{\mathbf{N}} \cdot \boldsymbol{m}} \cdot \boldsymbol{r})
\end{aligned} \tag{A.3}$$

where the last equality holds because $\widetilde{G}(\boldsymbol{k})$ and $\hat{U}(\boldsymbol{k})$ are even functions of vector $\boldsymbol{k}$. Since $\boldsymbol{k}_{\underline{\mathbf{N}} \cdot \boldsymbol{m}} \cdot \boldsymbol{r} = 2\pi(m_1 s_1 + m_2 s_2 + m_3 s_3)$ with $s_\beta = r_\beta / h_\beta$, the self-force is periodic over distances h_β and we can write

$$\boldsymbol{F}^{\mathrm{self}}(\boldsymbol{r}) = \sum_{\boldsymbol{m}} \boldsymbol{b}^{(\boldsymbol{m})} \sin\big(2\pi(m_1 s_1 + m_2 s_2 + m_3 s_3)\big) \tag{A.4}$$

where

$$\boldsymbol{b}^{(\boldsymbol{m})} \equiv -\frac{1}{V} \sum_{\boldsymbol{k}_1} \boldsymbol{k}_1 \widetilde{G}\big(\boldsymbol{k}_1^{\mathrm{G}}\big) \hat{U}(\boldsymbol{k}_1) \hat{U}(\boldsymbol{k}_1 + \boldsymbol{k}_{\underline{\mathbf{N}} \cdot \boldsymbol{m}}). \tag{A.5}$$

Decomposing the wavevector $\boldsymbol{k}_1$ as $\boldsymbol{k}_1 = \boldsymbol{k}_{\boldsymbol{n}} + \boldsymbol{k}_{\underline{\mathbf{N}} \cdot \boldsymbol{m}'}$ with $\boldsymbol{k}_{\boldsymbol{n}} = \boldsymbol{k}_1^{\mathrm{G}}$ the folded position into the first Brillouin zone $\widetilde{\mathbb{M}}$ and $\boldsymbol{m}' \in \mathbb{Z}^3$, we find that the term with $\boldsymbol{k}_{\boldsymbol{n}}$ gives a contribution to the self-force that vanishes by symmetry, since $\boldsymbol{k}_{\boldsymbol{n}}$ is odd while $G(\boldsymbol{k}_{\boldsymbol{n}})$ is even and

$$f^{(\boldsymbol{m})}(\boldsymbol{k}_{\boldsymbol{n}}) \equiv \sum_{\boldsymbol{m}' \in \mathbb{Z}^3} \hat{U}(\boldsymbol{k}_{\boldsymbol{n} + \underline{\mathbf{N}} \cdot \boldsymbol{m}'}) \hat{U}(\boldsymbol{k}_{\boldsymbol{n} + \underline{\mathbf{N}} \cdot \boldsymbol{m}' + \underline{\mathbf{N}} \cdot \boldsymbol{m}}) \tag{A.6}$$

is an even function of $\boldsymbol{k}_{\boldsymbol{n}}$ for any fixed vector $\boldsymbol{m}$. This result corresponds to the absence of self-force in the case of $i\boldsymbol{k}$-differentiation [in that scheme the particle–mesh pair force is indeed given by (A.2) with $\boldsymbol{k}_1$ replaced by $\boldsymbol{k}_1^{\mathrm{G}}$]. Eq. (A.5) reduces therefore to

$$\boldsymbol{b}^{(\boldsymbol{m})} = -\frac{1}{V} \sum_{\boldsymbol{k}_{\boldsymbol{n}} \in \widetilde{\mathbb{M}}} \widetilde{G}(\boldsymbol{k}_{\boldsymbol{n}}) \sum_{\boldsymbol{m}' \in \mathbb{Z}^3} \boldsymbol{k}_{\underline{\mathbf{N}} \cdot \boldsymbol{m}'} \hat{U}(\boldsymbol{k}_{\boldsymbol{n} + \underline{\mathbf{N}} \cdot \boldsymbol{m}'}) \hat{U}(\boldsymbol{k}_{\boldsymbol{n} + \underline{\mathbf{N}} \cdot \boldsymbol{m}' + \underline{\mathbf{N}} \cdot \boldsymbol{m}}). \tag{A.7}$$

Since $\boldsymbol{b}^{(\boldsymbol{m})} = -\boldsymbol{b}^{(-\boldsymbol{m})}$, we can rewrite the previous expression as

$$\boldsymbol{b}^{(\boldsymbol{m})} = \frac{\boldsymbol{b}^{(\boldsymbol{m})} - \boldsymbol{b}^{(-\boldsymbol{m})}}{2} = \boldsymbol{k}_{\underline{\mathbf{N}} \cdot \boldsymbol{m}} \frac{1}{2V} \sum_{\boldsymbol{k}_{\boldsymbol{n}}} \widetilde{G}(\boldsymbol{k}_{\boldsymbol{n}}) f^{(\boldsymbol{m})}(\boldsymbol{k}_{\boldsymbol{n}}) \tag{A.8}$$

where symmetries were used to factor out the vector $\boldsymbol{k}_{\underline{\mathbf{N}} \cdot \boldsymbol{m}}$. This formula for coefficient $\boldsymbol{b}^{(\boldsymbol{m})}$ is equivalent to expression (9) given in the main text.

Formulae (11)–(12) for the Madelung self-energy $E_{\mathrm{MS}}(\boldsymbol{r})$ of a particle can be obtained by integrating term by term the Fourier series (A.4) for the self-force. Alternatively, it can also be deduced directly by noticing that $E_{\mathrm{MS}}(\boldsymbol{r})$ is given by the first line of Eq. (A.3) in which factor $(-i\boldsymbol{k}_1)$ is replaced by $\frac{1}{2}$.

References

[1] P.P. Ewald, Ann. Phys. 369 (1921) 253–287.
[2] R.W. Hockney, J.W. Eastwood, Computer Simulations Using Particles, McGraw–Hill, New York, 1981.
[3] U. Essmann, L. Perera, M. Berkowitz, T. Darden, H. Lee, L. Pedersen, J. Chem. Phys. 103 (1995) 8577–8593.
[4] T. Darden, D. York, L. Pedersen, J. Chem. Phys. 98 (1993) 10089–10092.
[5] R. Ferrell, E. Bertschinger, J. Mod. Phys. 5 (1994) 933.
[6] H. Stern, K. Calkins, J. Chem. Phys. 128 (2008) 214106.
[7] H. Wang, F. Dommert, C. Holm, J. Chem. Phys. 133 (2010) 034117.
[8] R. Skeel, D. Hardy, J. Phillips, J. Comput. Phys. 225 (2007) 1–5.

[9] D. Cerutti, R. Duke, T. Darden, T. Lybrand, J. Chem. Theory Comput. 5 (2009) 2322–2338.
[10] A. Neelov, C. Holm, J. Chem. Phys. 132 (2010) 234103.
[11] J.W. Eastwood, Computational Methods in Classical and Quantum Physics, Advance Publications Limited, 1976, pp. 206–228.
[12] M. Deserno, C. Holm, J. Chem. Phys. 109 (1998) 7678.
[13] V. Ballenegger, J.J. Cerdà, C. Holm, A simple error estimate for the SPME method, J. Chem. Theory Comput., submitted for publication..
[14] D. Frenkel, B. Smit, Understanding Molecular Simulation, 2nd ed., Academic Press, San Diego, 2002.
[15] J.J. Cerdà, V. Ballenegger, O. Lenz, C. Holm, J. Chem. Phys. 129 (2008) 234104.
[16] V. Ballenegger, J.J. Cerdà, O. Lenz, C. Holm, J. Chem. Phys. 128 (2008) 034109.

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

How to Convert SPME to P3M: Influence Functions and Error Estimates

V. Ballenegger,*[†] J. J. Cerdà,[‡] and C. Holm[§]

[†]Institut UTINAM, Université de Franche-Comté, UMR 6213, 16, route de Gray, 25030 Besançon cedex, France

[‡]Instituto de Física Interdisciplinar y Sistemas Complejos, IFISC (CSIC-UIB), Universitat de les Illes Balears, E-07122 Palma de Mallorca, Spain

[§]Institut für Computer Physik, Universität Stuttgart, Pfaffenwaldring 27, 70569 Stuttgart, Germany

ABSTRACT: We demonstrate explicitly how the two seemingly different particle mesh Ewald methods, the smooth particle mesh Ewald (SPME) and the particle–particle particle mesh (P3M), can be mathematically transformed into each other. This allows us in particular to convert the error estimate of the P3M method in the energy-conserving scheme (also known as "P3M with analytic differentiation") into an error estimate for the SPME method, via a simple change of the lattice Green function. Our error estimate is valid for any values of the SPME parameters (mesh size, spline interpolation order, Ewald splitting parameter, real-space cutoff distance), including odd orders of splines. The problem with the self-forces is avoided thanks to an analytical formula that allows to subtract them directly within the particle mesh calculation. Plots of the accuracy of the SPME forces are provided for a wide range of parameter values. The main use of the error estimate is to allow a simulation program to scan quickly the multidimensional parameter space to find the best set of parameters to achieve a target accuracy at the smallest computational cost. As a byproduct, we show how a SPME code can be transformed into a P3M version by changing a few lines of code. We demonstrate also that the P3M lattice Green function can be approximated by a closed form expression, computable on-the-fly, that provides essentially the same accuracy as the full function.

1. INTRODUCTION

Particle mesh methods, such as particle–particle particle mesh (P3M)[1] and smooth particle mesh Ewald (SPME),[2] are efficient methods to compute long-range Coulomb (or gravitational) forces in simulations of many-particle systems.[3,4] By using fast Fourier transform (FFT) routines to solve the Poisson equation in Fourier space, they convert the task of computing $O(N^2)$ pairwise interactions into a task that scales as $O(N \log N)$ with the number of particles. To improve the accuracy, the pair interaction is moreover decomposed into a short-ranged component that decays rapidly in real space and a smooth long-ranged component that decays rapidly in Fourier space.[5] The short-ranged interactions are computed via a double loop over particle–particle interactions, while the long-range components are computed by particle–mesh mapping of the density (charge or mass), convolution of the mesh-based density with the appropriate Green function, and mesh–particle mapping of the resulting potential.

Several variants of P3M algorithms exist, associated to different choices of the splitting function used to decompose the interaction into a short- and a long-ranged part and to different choices of the differentiation scheme used to compute forces from the mesh-based potential. The differentiation schemes fall in two classes: energy- and momentum-conserving schemes.[1,6] In momentum conserving schemes, the potential is differentiated numerically (using finite differences in real space or by multiplication by $i\mathbf{k}$ in Fourier space) and then interpolated to the particle positions in continuous space. In the energy conserving scheme, forces are obtained from the exact gradient of the potential energy of the system discretized on the grid. That latter scheme conserves energy but not momentum.[1,7] The accuracy of the P3M method has been studied in depth for different choices of ingredients in the method.[1,6,8−10] Thanks to that analysis, the sources of errors in P3M algorithms are fully understood, and the accuracy of the method has been optimized via the use of error-minimizing lattice Green functions, that help reducing truncation and aliasing errors. A further benefit of that analysis is that the error of the method is under full control: An analytical formula predicts the accuracy as function of the various "free" (user defined) parameters of the method. Using this a priori error estimate, a simulation program can determine automatically the set of P3M parameters (mesh size, interpolation order, Ewald splitting parameter, real-space cutoff) that is optimal, in the sense that it achieves the prescribed accuracy at a minimal computational cost.

The purpose of this paper is to show how the SPME method can be transformed into a P3M method and vice versa, and that the error analysis of the P3M method can be adapted to understand and predict the errors of the SPME method.[2] We focus in this paper on SPME, as it supersedes particle mesh Ewald (PME)[26] in practice, and it is implemented in several mainstream molecular dynamics simulation packages, such as GROMACS,[11] AMBER,[12] and NAMD.[13] Whereas the derivation of SPME and P3M are different, the result of the SPME derivation is a special case of the more general P3M method, in which the splitting function is the error function. In the SPME approach, the Green function for the system discretized on the lattice differs from the continuous-space interaction by contributions inherited from the employed exponential

Received: March 15, 2011
Published: January 24, 2012

Euler spline approximation. This contrasts with the P3M approach, where the lattice Green function is optimized to ensure that the mesh calculation give results as close as possible, in a least-squares sense, to those of a continuous-space calculation. A further difference between the two approaches lies in the choice of the differentiation scheme. P3M is often implemented with a momentum-conserving differentiation scheme (the original implementation[14] used finite differences, several later implementations employ ik differentiation), while the SPME method uses normally the "analytical differentiation" (AD) scheme, which is an energy-conserving scheme introduced in the 1970s.[6,15] Since the AD scheme provides better accuracy in less operations than the finite differences scheme, it can be a cost-effective solution provided that the small local violation of momentum conservation inherent to that scheme can be accepted in the simulations. The optimization of the lattice Green function of a P3M method that uses the AD scheme can be found in refs 6 and 16. In this work, we give the link between the different notation and terminology used in the P3M and SPME methods, building upon the presentation and comparison of particle mesh methods of ref 9. By viewing the SPME method as a P3M method that uses analytical differentiation and a lattice Green function that differs from the optimal P3M error-minimizing one, we show how the a priori error estimates derived for P3M can be converted into an error estimate for the SPME method.

The error estimate for the SPME method introduced in this paper can be used to tune easily the parameters of the method [size of the mesh in each direction (M_1, M_2, M_3), charge interpolation order (P), Ewald splitting parameter (α), and real-space cutoff distance (r_c)] to maximize efficiency and hence save valuable computer and user time. Compared to a SPME error estimate derived recently,[17] our error estimate is more general and simpler to use. It is valid for any interpolation order, and it is also given by an analytic formula which can be evaluated quickly. Contrary to the error estimate derived in ref 17, our approach does not require an expensive calculation of the root-mean-square (rms) error arising from self-forces: We subtract these self-forces directly within the particle mesh calculation using a formula we derived in ref 10.

The paper is organized as follows. In Section 2, we recall the main formulas of the P3M method in the general case of a triclinic cell, with a special emphasis on the analytical differentiation scheme and the definition and meaning of the lattice Green function. This provides the background needed in Section 3 to derive the mathematical link between the two methods, which is embodied in the lattice Green function (eq 3.7) of the SPME method. In Section 4, we give the error estimate for the SPME method and expose also how the SPME algorithm can be modified in practice to use the optimized lattice Green function derived within the P3M approach. The numerical results presented is Section 5 demonstrate the accuracy and usefulness of the error estimate. We show that the P3M lattice Green function optimized for computing the energy, for which a closed form expression exists, provides virtually the same accuracy as the true optimal lattice Green function. For cutoffs and accuracies typically employed in atomistic simulations (e.g., 9 Å cutoff and rms force error of ~10^{-4}), the increase in accuracy brought by modifying SPME to use the optimal P3M lattice Green function turns out to be small. Section 6 contains a summary of the findings and concluding remarks.

2. SUMMARY OF P3M METHOD

We consider an overall neutral system of point charges q_i at locations $\boldsymbol{r}_i$ (i = 1, ..., N) in a unit cell of volume V defined by edge vectors $\boldsymbol{a}_\beta$, β = 1, 2, 3. The charges interact via Coulomb's law, and periodic boundary conditions are applied, i.e., the cell is replicated along the directions of the edge vectors $\boldsymbol{a}_\beta$'s, which need not to be orthogonal. The objective is to compute efficiently the energy E of the system

$$E = \frac{1}{2} \sum_{\boldsymbol{n}\in\mathbb{Z}^3} \sum_{i,j=1}^{N}{}' \frac{q_i q_j}{|\boldsymbol{r}_i - \boldsymbol{r}_j + \underline{\mathbf{L}}\cdot\boldsymbol{n}|} \tag{2.1}$$

and the forces acting on each particle, which are defined by $\boldsymbol{F}_i = -\nabla_i E$. In eq 2.1, $\underline{\mathbf{L}} = \mathrm{diag}(|\boldsymbol{a}_1|, |\boldsymbol{a}_2|, |\boldsymbol{a}_3|)$ is a diagonal matrix formed with the lengths of the primitive cell translation vectors, and the sum over vectors $\boldsymbol{n}$ (which have integer components) account for the periodic boundary conditions. The prime indicates the omission of the $i = j$ term when $\boldsymbol{n} = 0$. The sum is only conditionally convergent, and we assume here the usual spherical summation order with conducting boundary conditions.[18,19]

The P3M method is based on three main ideas: (i) decomposition of the pair interaction into a short-range interaction (calculated in real space) and a smooth long-range interaction with a fast decay in Fourier space; (ii) discretization of the system onto a mesh to take advantage of the FFT algorithm to solve, in Fourier space, the Poisson equation associated to the long-range forces (this equation reduces to a simple multiplication by a lattice Green function); and (iii) use of a lattice Green function that optimizes the accuracy of the calculation.

2.1. Decomposition of the Pair Interaction.

The splitting of the interaction into a short- and a long-range part is performed in P3M using the same principle as in the Ewald method,[5,20] namely by introducing a compensating charge distribution around each particle. In the Ewald method, this screening distribution is a Gaussian of width $1/(2\alpha)$, which leads to the following decomposition of the interaction:

$$\frac{1}{r} = \psi(r) + \varphi(r) = \frac{1 - \mathrm{erf}(\alpha r)}{r} + \frac{\mathrm{erf}(\alpha r)}{r} \tag{2.2}$$

where $\mathrm{erf}(x)$ is the error function and the free parameter α controls the relative importance of both contributions. In that decomposition, the long-range part $\phi(r) = \mathrm{erf}(\alpha r)/r$ corresponds to the interaction of a point charge with a Gaussian charge distribution of width $1/(2\alpha)$, while the short-range part $\psi(r) = (1-\mathrm{erf}(\alpha r))/r$ corresponds to the interaction of a point charge with a point charge screened by the Gaussian charge distribution. In the P3M method, several shapes for the screening charge distribution have been considered, among which the Gaussian shape (called S3) corresponds to the above Ewald splitting.[21,22] Since the various screening distributions gave similar accuracies,[1,23] it is assumed here that P3M is implemented with standard Gaussian screening (for other choices, see e.g. ref 24). Since the potential $\psi(r)$ is short ranged, ψ can be computed easily by direct summation in real space. The long-range interaction ϕ is sometimes called the "reciprocal-space interaction", because interactions ϕ between the particles are most easily computed in reciprocal (Fourier) space. The Fourier transform of that interaction:

$$\hat{\varphi}(\boldsymbol{k}) = \int \varphi(\boldsymbol{r}) e^{-i\boldsymbol{k}\cdot\boldsymbol{r}} d^3\boldsymbol{r} = \frac{4\pi}{k^2} \exp(-k^2/(4\alpha^2)) \tag{2.3}$$

has a fast decay with k. Upon rewriting the long-range interactions between particles in Fourier space (Ewald method), the energy of the system can be written as the sum of the real-space energy $E(r) = 1/2\sum_{n\in\mathbb{Z}^3} \sum_{i,j=1}^{N}{}' q_i q_j \psi(|\boldsymbol{r}_i - \boldsymbol{r}_j + \underline{\mathbf{L}}\cdot\boldsymbol{n}|)$, the reciprocal-space energy

$$E^{(k)} = \frac{1}{2}\frac{1}{V}\sum_{\boldsymbol{k}\neq 0} \hat{\varphi}(\boldsymbol{k})|\hat{\rho}(\boldsymbol{k})|^2 \tag{2.4}$$

and the self-energy $E^{(s)} = -(\alpha/\sqrt{\pi})\sum_i q_i^2$. In eq 2.4, the structure factor

$$\hat{\rho}(\boldsymbol{k}) = \sum_{i=1}^{N} q_i e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_i} \tag{2.5}$$

is the Fourier transform of the charge density $\rho(\boldsymbol{r}) = \sum_i q_i \delta(\boldsymbol{r} - \boldsymbol{r}_i)$. The sum over $\boldsymbol{k}$ runs over all reciprocal lattice points (except $\boldsymbol{k} = 0$): The wave vector $\boldsymbol{k}$ is of the form

$$\boldsymbol{k}_{\boldsymbol{n}} = 2\pi(n_1\boldsymbol{a}_1^* + n_2\boldsymbol{a}_2^* + n_3\boldsymbol{a}_3^*) \tag{2.6}$$

with $\boldsymbol{n} = (n_1, n_2, n_3) \in \mathbb{Z}^3$ a triplet of integers, and the reciprocal lattice vectors $\boldsymbol{a}_\alpha^*$ are defined by the relations $\boldsymbol{a}_\alpha^*\cdot\boldsymbol{a}_\beta = \delta_{\alpha\beta}$ (the Kronecker δ), for α, β = 1, 2, 3. The reciprocal lattice vectors are thus given by $\boldsymbol{a}_1^* = (1/V)(\boldsymbol{a}_2 \times \boldsymbol{a}_3)$ with cyclic permutations for $\boldsymbol{a}_2^*$ and $\boldsymbol{a}_3^*$, and $V = \boldsymbol{a}_1\cdot(\boldsymbol{a}_2 \times \boldsymbol{a}_3)$. Notice that the self-energy term $E^{(s)}$ subtracts the Ewald self-energy $\phi(0)$ of each particle included in the reciprocal energy $E^{(k)}$.

2.2. Mesh Calculation of the Reciprocal Energy. The computation of eq 2.4 requires a priori $O(N^2)$ operations (or $O(N^{3/2})$ if the cutoff used in the real-space calculation depends on the Ewald splitting parameter α in a specific way),[25] because eq 2.5 is a sum of N terms and needs to be evaluated for a number of wave vectors that grows linearly with the volume of the simulation box and hence with the number of particles. The basic idea in particle mesh methods is to reduce the computational cost to $O(N \log N)$ by replacing the continuous Fourier transform $\hat{\rho}(\boldsymbol{k})$ in eq 2.4 by a FFT. To this end, the system is discretized onto a grid with fixed lattice spacing, that is, each direction $\boldsymbol{a}_\alpha$ of the unit cell is divided into N_α mesh points. The edges of a mesh cell have lengths $h_1 = |\boldsymbol{a}_1|/N_1$, $h_2 = |\boldsymbol{a}_2|/N_2$, and $h_3 = |\boldsymbol{a}_3|/N_3$. The charge $q_{\boldsymbol{r}_n}$ at each mesh point $\boldsymbol{r}_n$ is computed using an assignment function $W(\boldsymbol{r})$:

$$q_{\boldsymbol{r}_n} \equiv \sum_i q_i W(\boldsymbol{r}_n - \boldsymbol{r}_i) \tag{2.7}$$

where periodic boundary conditions are implicitly assumed in eq 2.7. If analytical differentiation is used, $W(\boldsymbol{r})$ must be differentiable. An assignment function that distributes a charge to its P nearest mesh points (separately in each dimension) is called a P order assignment function. It is usually chosen to be of the product form

$$W^{(P)}(\boldsymbol{r}) = w(s_1)w(s_2)w(s_3) \tag{2.8}$$

where $w(s)$ is an even function with finite support $[-P/2, P/2]$ and

$$\boldsymbol{r} = \frac{s_1}{N_1}\boldsymbol{a}_1 + \frac{s_2}{N_2}\boldsymbol{a}_2 + \frac{s_3}{N_3}\boldsymbol{a}_3 \tag{2.9}$$

where $s_\alpha = N_\alpha \boldsymbol{a}_\alpha^*\cdot\boldsymbol{r} \in [0, N_\alpha]$ is the coordinate of $\boldsymbol{r}$ along direction $\boldsymbol{a}_\alpha$ scaled by the number of mesh points N_α. Hockney and Eastwood's choice for $w(s)$ is a B-spline of order P. More details on charge assignment can be found in refs [1,9,16,21,23] We can introduce the mesh-based charge density by setting

$$\rho_M(\boldsymbol{r}_n) = \frac{N_1N_2N_3}{V} q_{\boldsymbol{r}_n} \tag{2.10}$$

where $\boldsymbol{r}_n \in \mathbb{M}$ and $\mathbb{M} = \{\boldsymbol{r}_n = (n_1/N_1)\boldsymbol{a}_1 + (n_2/N_2)\boldsymbol{a}_2 + (n_3/N_3)\boldsymbol{a}_3 \colon n_\alpha = 0, 1, ..., N_\alpha - 1, \alpha = 1, 2, 3\}$ denotes the set of all mesh points. Notice that $\rho_M(\boldsymbol{r}_n)$ can be written as a convolution of the (periodic) charge density $\rho(\boldsymbol{r})$ with the assignment function:

$$\rho_M(\boldsymbol{r}) = \frac{N_1N_2N_3}{V}\int_{\mathbb{R}^3} \rho(\boldsymbol{r}')W(\boldsymbol{r} - \boldsymbol{r}')\mathrm{d}^3\boldsymbol{r}' \tag{2.11}$$

Once the charges are assigned to the mesh, the finite Fourier transform of the mesh-based charge density

$$\tilde{\rho}_M(\boldsymbol{k}) = \mathrm{FFT}[\rho_M] = \frac{V}{N_1N_2N_3}\sum_{\boldsymbol{r}_n\in\mathbb{M}} \rho_M(\boldsymbol{r}_n)e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_n} \tag{2.12}$$

may be computed via the FFT algorithm (we denote finite Fourier transforms by a tilde instead of a hat to distinguish them from usual Fourier transforms). Notice that the volume element $\mathrm{d}^3\boldsymbol{r}$ of the usual transform takes here the form of the volume of a mesh cell $V/(N_1N_2N_3)$. $\tilde{\rho}_M(\boldsymbol{k})$ is defined for wave vectors in the (finite) reciprocal lattice

$$\begin{aligned}\tilde{\mathbb{M}} = \{&\boldsymbol{k}_{\boldsymbol{n}} = 2\pi(n_1\boldsymbol{a}_1^* + n_2\boldsymbol{a}_2^* + n_3\boldsymbol{a}_3^*) \colon n_\alpha \in \mathbb{Z},\\ &|n_\alpha| \leq (N_\alpha - 1)/2,\ \alpha = 1, 2, 3\}\end{aligned} \tag{2.13}$$

In the P3M method, the reciprocal Ewald energy eq 2.4 is approximated by the mesh formula

$$E_{\mathrm{P3M}}^{(k)} = \frac{1}{2}\frac{1}{V}\sum_{\substack{\boldsymbol{k}\neq 0\\ \boldsymbol{k}\in\tilde{\mathbb{M}}}} \tilde{G}(\boldsymbol{k})|\tilde{\rho}_M(\boldsymbol{k})|^2 \tag{2.14}$$

The reciprocal interaction $\hat{\varphi}(\boldsymbol{k})$ is thus replaced by a lattice Green function (or "influence function") $\tilde{G}(\boldsymbol{k})$, which will be chosen, as suggested by Eastwood,[6] in a way that optimizes the accuracy of the mesh calculation. For later purposes, it is useful to note that eq 2.4 can be expressed in real-space as

$$E^{(k)} = \frac{1}{2}\sum_{i,j=1}^{N} q_i q_j G_{\mathrm{cont.}}(\boldsymbol{r}_i - \boldsymbol{r}_j) \tag{2.15}$$

where $G_{\mathrm{cont.}}(\boldsymbol{r}) = V^{-1}\sum_{k\neq 0} \exp(i\boldsymbol{k}\cdot\boldsymbol{r})\hat{\varphi}(\boldsymbol{k})$ is the continuous-space Green function of the system under periodic boundary conditions. If that Green function reduces to $\hat{\varphi}(\boldsymbol{k})$ in Fourier space $[\hat{G}_{\mathrm{cont.}}(\boldsymbol{k}) = \hat{\varphi}(\boldsymbol{k})$ for $\boldsymbol{k}\in\tilde{\mathbb{M}}]$, it differs from $\phi(\boldsymbol{r})$ in real space, since it is given by the inverse Fourier series (rather than full) transform of $\hat{\varphi}(\boldsymbol{k})$. That difference accounts precisely for the periodic boundary conditions that are embedded in the Green function. When the system is discretized onto the mesh, the energy is computed using eq 2.14 or equivalently in real space by

$$E_{\mathrm{P3M}}^{(k)} = \frac{1}{2}\sum_{\boldsymbol{r}_n,\boldsymbol{r}_m\in\mathbb{M}} q_{\boldsymbol{r}_n} q_{\boldsymbol{r}_m} G(\boldsymbol{r}_n - \boldsymbol{r}_m) \tag{2.16}$$

where $G(\boldsymbol{r}_n) = \mathrm{FFT}^{-1}[\tilde{G}](\boldsymbol{r}_n)$ is the lattice Green function in real space. The reciprocal-space energy (eq 2.15) can also be written as $E^{(k)} = (1/2)\sum_i q_i \Phi^{(k)}(\boldsymbol{r}_i)$ in term of the (reciprocal space) electrostatic potential at point r defined by

$$\Phi^{(k)}(\boldsymbol{r}) = \int_V \rho(\boldsymbol{r}')G_{\text{cont.}}(\boldsymbol{r} - \boldsymbol{r}')\mathrm{d}^3\boldsymbol{r}' \tag{2.17}$$

In Fourier space, this convolution becomes the product $\hat{\Phi}^{(k)}(\boldsymbol{k}) = \hat{\rho}(\boldsymbol{k})\hat{G}_{\text{cont.}}(\boldsymbol{k})$, which can be approximated on the mesh by

$$\hat{\Phi}^{(k)}(\boldsymbol{k}) \simeq \tilde{\rho}_M(\boldsymbol{k})\tilde{G}(\boldsymbol{k}) \tag{2.18}$$

2.3. Calculation of Forces. Once the potential $\Phi^{(k)}$ is known, the forces acting on the particles are obtained by differentiating the energy. This step can be performed in several ways:

(i) finite-differences scheme: differentiation by computing finite differences in real space

(ii) analytical differentiation (AD) scheme: forces are obtained by differentiating exactly the mesh-based energy (eq 2.16):

$$\begin{aligned}\boldsymbol{F}_i^{(k)} &= -\nabla_i E_{\text{P3M}}^{(k)} \\ &= -\sum_{\boldsymbol{r_n},\boldsymbol{r_n}'\in\mathbb{M}} (\nabla_i q_{\boldsymbol{r_n}})q_{\boldsymbol{r_n}'}G(\boldsymbol{r_n} - \boldsymbol{r_n}') \\ &= -q_i \sum_{\boldsymbol{r_n}\in\mathbb{M}} \Phi^{(k)}(\boldsymbol{r_n})\nabla_i W(\boldsymbol{r}_i - \boldsymbol{r}_n)\end{aligned} \tag{2.19}$$

where we used that $G(\boldsymbol{r}_n)$ is even and $\Phi^{(k)}(\boldsymbol{r}_n) \equiv \sum_{\boldsymbol{r}_{n'}\in\mathbb{M}} q_{\boldsymbol{r}_{n'}}G(\boldsymbol{r}_n - \boldsymbol{r}_{n'})$ is the potential at mesh point $\boldsymbol{r}_n$. Forces are therefore calculated in this scheme by interpolating the mesh-based potential back to a particle position using the gradient $\nabla W(\boldsymbol{r})$ of the charge assignment function.

(iii) *ik*-differentiation scheme: differentiation in Fourier space by multiplying the potential by *ik*.

In the first two schemes, one computes the inverse FFT of the (scalar) electrostatic potential (eq 2.18) to get $\Phi^{(k)}(\boldsymbol{r}_n)$ in real space and calculate then the forces either by using finite differences or by using the exact analytic gradient $\nabla W(\boldsymbol{r})$. It should be emphasized that the analytical differentiation scheme amounts to computing the exact gradient of an approximate energy (the energy, eq 2.16, of the discretized system), while the *ik*-differentiation scheme corresponds to approximating the formula for the force (i.e., the gradient of the exact energy, eq 2.15) as well as possible on a mesh.

The *ik*-differentiation scheme, also known as force-interpolation scheme, is the most accurate one, but it requires to compute the inverse FFT of a vectorial field. The force-interpolation scheme amounts to calculating the electric (or gravitational) field at each mesh point and to interpolating this vectorial "force" field back to the particle positions. In the context of particle mesh methods, force interpolation was introduced first in the PME method[26] and then in P3M by Bertschwinger et al.[8] The *ik*-differentiation and finite-differences schemes are known as "momentum conserving schemes", as they conserve momentum but not the energy.[1]

The forces obtained with analytical differentiation are conservative (since they are the exact gradient of an energy), so the AD scheme is an energy-conserving scheme. In Lewis' original energy-conserving scheme,[15] analytical differentiation is used together with a potential that is solved by using a specific finite-difference form of Poisson's equation, but Langdon showed that almost any potential solver can be used.[7] The AD scheme conserves the energy (in the limit of small time steps) but not momentum.[1,7] When using this scheme in a simulation, a correction must be applied to the forces to enforce at least the conservation of the center-of-mass momentum.[2] That correction has unfortunately the collateral effect of breaking the exact conservation of energy.

For a more thorough discussion of the merits and drawbacks of each scheme and comparison of accuracies, see refs 1, 3, and 9. P3M is often associated with finite differences because Hockney and Eastwood advocated the use of this scheme in their book[1] (that scheme turned out to be the most effective at that time where memory constraints were strong and where the computations were done at low force accuracy [interpolation order 1 or 2] due to the small computing power). We will refer to a P3M method with *ik*-differentiation as "P3M-ik" and to a P3M method with analytical differentiation as "P3M-AD".

2.4. Optimal Lattice Green Functions. In the simplest implementation of P3M, the lattice Green function would be taken to be equal to the continuous-space Green function:

$$\tilde{G}(\boldsymbol{k_n}) = \hat{G}_{\text{cont.}}(\boldsymbol{k_n}) = \hat{\varphi}(\boldsymbol{k_n}) \tag{2.20}$$

But this simple Poisson solver gives poor accuracy, because it neglects the reshaping effect due to the charge assignment onto the mesh, which amounts to giving a finite size to the particles. The mesh-based charge density can indeed be seen as resulting from the sampling, at grid points, of a continuous smoothed charge density obtained by replacing each point charge by a cloud of charge density $W(\boldsymbol{r})$.[1] We stress that this reshaping effect is due purely to charge assignment, and it is not related to the Ewald Gaussian screening contained in the reciprocal interaction $\hat{\varphi}(\boldsymbol{k})$. Because of this assignment (or interpolation) smoothing, the high-frequency harmonics of the mesh-based charge density $\tilde{\rho}_M(\boldsymbol{k})$ are damped compared to those of $\hat{\rho}(\boldsymbol{k})$. In P3M, the lattice Green function is adjusted to compensate for this damping and also to reduce other errors arising from truncation and aliasing.

Since eq 2.11 is a convolution, the (continuous) Fourier transform of $\rho_M(\boldsymbol{r})$ is given by the product $\tilde{\rho}_M(\boldsymbol{k}) = \hat{U}(\boldsymbol{k})\hat{\rho}(\boldsymbol{k})$, where $\hat{U}(\boldsymbol{k})$ is equal to the Fourier transform $\hat{W}(\boldsymbol{k})$ of the charge assignment function divided by the volume of a mesh cell. For a B-spline assignment function of order P, one has

$$\begin{aligned}\hat{U}(\boldsymbol{k}) &\equiv \frac{N_1N_2N_3}{V}\hat{W}(\boldsymbol{k}) \\ &= \left(\frac{\sin(\pi n_1/N_1)}{\pi n_1/N_1}\right)^P \left(\frac{\sin(\pi n_2/N_2)}{\pi n_2/N_2}\right)^P \left(\frac{\sin(\pi n_3/N_3)}{\pi n_3/N_3}\right)^P\end{aligned} \tag{2.21}$$

If the harmonic content of function $\hat{\rho}_M(\boldsymbol{k})$ is negligible beyond the Nyquist frequency π/h_β (for each direction $\beta = 1, 2, 3$), no information is lost by the sampling on the mesh, and the discrete Fourier transform yields the exact spectrum: $\tilde{\rho}_M(\boldsymbol{k}) = \hat{\rho}_M(\boldsymbol{k}) = \hat{U}(\boldsymbol{k})\hat{\rho}(\boldsymbol{k})$. From this relation, and by comparing eqs 2.14 to 2.4, we find that the accuracy of the Poisson solver can be increased by setting the lattice Green function to

$$\tilde{G}(\boldsymbol{k_n}) = \frac{\hat{\varphi}(\boldsymbol{k_n})}{\hat{U}^2(\boldsymbol{k_n})} \tag{2.22}$$

That function contains a "sharpening" factor $1/\hat{U}^2(\boldsymbol{k}_n)$ that compensates for the smoothing introduced by the mapping of the charges onto the mesh and by the back interpolation of the mesh results to continuous space. The use of the adjusted Green function, eq 2.22, in place of eq 2.20 improves substantially the accuracy of the particle mesh calculation.

A key idea in P3M is to consider $\tilde{G}(\boldsymbol{k}_n)$ as an adjustable function, that has to be optimized such that the results of the mesh calculation are as close as possible to the results of the original continuous problem. The optimal P3M lattice Green function is the one that minimizes the rms error of the mesh calculation. The optimal lattice Green function includes corrections beyond the previous sharpening factor. The optimization of the lattice Green function enhances the accuracy at no computational cost because: (i) if one disregards caching issues, $\tilde{G}(\boldsymbol{k}_n)$ is a fixed quantity (apart in constant pressure simulations) that can be tabulated; and (ii) the lattice Green function optimized for computing the energy, which can be computed on-the-fly since it is known in closed analytical form, provides in practice virtually the same accuracy as the true optimal lattice Green function associated to the P3M-AD and P3M-ik force calculations (see Section 5).

To derive the optimal P3M lattice Green function, one needs the exact relation between the structure factors $\tilde{\rho}_M(\boldsymbol{k})$ and $\hat{\rho}(\boldsymbol{k})$, which can be obtained using the sampling and convolution theorems (see ref 1 or eq 5.2 of ref 27):

$$\tilde{\rho}_M(\boldsymbol{k}) = \sum_{\boldsymbol{m}\in\mathbb{Z}^3} \hat{U}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\hat{\rho}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}), \qquad \boldsymbol{k} = \boldsymbol{k}_{\boldsymbol{n}} \in \tilde{\mathbb{M}} \tag{2.23}$$

where $\underline{\mathbf{N}} = \mathrm{diag}(N_1, N_2, N_3)$ is a diagonal matrix formed with the numbers of mesh points in each dimension. In agreement with eq 2.6, the vector

$$\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}} = 2\pi((n_1 + m_1N_1)\boldsymbol{a}_1^* + (n_2 + m_2N_2)\boldsymbol{a}_2^* + (n_3 + m_3N_3)\boldsymbol{a}_3^*) = \boldsymbol{k}_{\boldsymbol{n}} + \boldsymbol{k}_{\underline{\mathbf{N}}\boldsymbol{m}} \tag{2.24}$$

is equal to the vector $\boldsymbol{k}_n$ shifted outside the first Brillouin zone by reciprocal lattice vector $\boldsymbol{k}_{\underline{\mathbf{N}}\boldsymbol{m}}$. Equation 2.23 embodies in a simple formula the effects on the spectrum induced by the charge assignment and the sampling of the charge density on a mesh. The sum over integer triplets $\boldsymbol{m}$ shows that spurious contributions from the high-frequency spectrum of $\hat{U}(\boldsymbol{k})\hat{\rho}(\boldsymbol{k})$ are introduced into the first Brillouin zone $\tilde{\mathbb{M}}$. These unwanted copies of the other Brillouin zones into the first one are known as aliasing errors.[1] To reduce aliasing errors, the charge assignment function $\hat{U}(\boldsymbol{k})$ must decay quickly in k-space (it must also decay to zero over a short distance in real-space to avoid having to assign the charges to a too large number of grid points). If aliasing ($\boldsymbol{m} \neq 0$) contributions are neglected in eq 2.23, one recovers the relation $\tilde{\rho}_M(\boldsymbol{k}) = \hat{U}(\boldsymbol{k})\hat{\rho}(\boldsymbol{k})$ that leads to the modified lattice Green function, eq 2.22.

Formula 2.23 can be used to predict analytically, for any positions of two particles within the simulation box, the error of the P3M pair interaction (energy, force, etc.) compared to the exact continuous-space interaction. By minimizing the rms error of this pair interaction, one finds the optimal P3M lattice Green function. The optimal lattice Green function for computing the energy reads[27]

$$\tilde{G}_{\mathrm{opt}}(\boldsymbol{k}_{\boldsymbol{n}}) = \frac{\sum_{\boldsymbol{m}} \hat{\varphi}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})}{(\sum_{\boldsymbol{m}} \hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}))^2} \simeq \hat{\varphi}(\boldsymbol{k}_{\boldsymbol{n}}) \frac{\hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}})}{(\sum_{\boldsymbol{m}} \hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}))^2} \tag{2.25}$$

where the last expression holds thanks to the fast decay of $\hat{\varphi}(\boldsymbol{k})$. The sum over $\boldsymbol{m}$ in the numerator originates from the minimization of truncation errors, which are small if $\hat{\varphi}(\boldsymbol{k})$ decays rapidly, and the sum over $\boldsymbol{m}$ in the denominator arises from aliasing errors. As can be seen by comparing the optimal lattice Green function, eq 2.25, to the simpler result, eq 2.22, the "sharpening" factor $1/\hat{U}^2(\boldsymbol{k})$ is somewhat damped in eq 2.25 to compensate partially for aliasing errors. The denominator in eq 2.25 can actually be calculated analytically. For the first few spline interpolation orders, one has for instance:[23]

$$\sum_{m\in\mathbb{Z}} \hat{U}^2(2\pi(n + Nm)) = \begin{cases} 1 - 2z^2/3 & \text{for } P = 2 \\ 1 - z^2 + 2z^4/15 & \text{for } P = 3 \\ 1 - 4z^2/3 + 2z^4/5 + 4z^6/315 & \text{for } P = 4 \\ 1 - 5z^2/3 + 7z^4/9 - 17z^6/189 & \text{for } P = 5 \\ \quad + 2z^8/2835 & \end{cases} \tag{2.26}$$

where $z = \sin\,(\pi n/N)$ and the full function is

$$\sum_{\boldsymbol{m}} \hat{U}^2(\mathbf{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}) = \prod_{j=1,2,3} \sum_{m\in\mathbb{Z}} \hat{U}^2(2\pi(n_j + N_jm)) \tag{2.27}$$

This provides a closed form expression for the lattice Green function, eq 2.25, which is most useful for computing this function on-the-fly.

Expressions for the optimal lattice Green function to compute forces, both with finite differences or *ik*-differentiation, are given in refs 1 and 8. The optimal influence function for the P3M-AD algorithm reads[6,16]

$$\tilde{G}_{\mathrm{opt}}(\boldsymbol{k}_{\boldsymbol{n}}) = \frac{\sum_{\boldsymbol{m}} \hat{\varphi}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\boldsymbol{k}^2_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}\hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})}{(\sum_{\boldsymbol{m}} \hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}))\sum_{\boldsymbol{m}} \hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\boldsymbol{k}^2_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}} \simeq \hat{\varphi}(\boldsymbol{k}_{\boldsymbol{n}}) \frac{\hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}})\boldsymbol{k}^2_{\boldsymbol{n}}}{(\sum_{\boldsymbol{m}} \hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}))\sum_{\boldsymbol{m}} \hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\boldsymbol{k}^2_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}} \tag{2.28}$$

In all earlier expressions, the sum over $\boldsymbol{m} \in \mathbb{Z}^3$ converges quickly. The lattice Green functions can therefore be evaluated in practice by truncating the sum at $|m_\beta| = 2$ in each direction ($\beta = 1, 2, 3$).

3. LINK BETWEEN SPME AND P3M: THE SPME LATTICE GREEN FUNCTION

The SPME method corresponds to a specific form of particle–particle particle mesh algorithm. It uses the traditional Ewald splitting, eq 2.2, of the interaction and an FFT-based Poisson solver to compute the reciprocal-space potential (contrary to P3M, the lattice Green function of the Poisson solver is not optimized in SPME). The SPME method differs from the standard P3M method mainly by the choice of the analytical differentiation scheme and by the point of view taken when discretizing the system onto a mesh (which leads to the SPME lattice Green function).

The combination of an FFT-based Poisson solver with the analytical differentiation (energy-conserving) scheme corresponds to the P3M-AD algorithm introduced in the previous section. In a comparison of schemes, Eastwood found that the P3M-AD method was not competitive and discarded the method, but his analysis was limited to one-dimensional problems.[6] In three dimensions, the use of the AD scheme results in a very fast and efficient algorithm, since it requires fewer FFT's than *ik*-differentiation (the main drawback of the method lies in local violations of momentum conservation). That choice of ingredients makes the SPME method very effective.

In the SPME methods, the discretization of eqs 2.4 and 2.5 onto a mesh is viewed alternatively as interpolation of the exponential factor $\exp(i\boldsymbol{k}\cdot\boldsymbol{r}_i)$ from its values at nearby grid points. This is done using Lagrange interpolation in the PME method, while SPME uses exponential Euler splines. The latter approximation amounts to setting

$$\hat{\rho}(\boldsymbol{k}) \approx b(\boldsymbol{k})\tilde{\rho}_M(\boldsymbol{k}) \tag{3.1}$$

and hence to calculate

$$E^{(k)}_{\mathrm{SPME}} = \frac{1}{2}\frac{1}{V}\sum_{\substack{\boldsymbol{k}\neq 0\\ \boldsymbol{k}\in\widetilde{\mathbb{M}}}} \hat{\varphi}(\boldsymbol{k})|\hat{b}(\boldsymbol{k})|^2|\tilde{\rho}_M(\boldsymbol{k})|^2 \tag{3.2}$$

where $\tilde{\rho}_M(\boldsymbol{k})$ is computed using a B-spline of order P as assignment function, which is identical, up to a shift of $P/2$, to the charge assignment function introduced in P3M:

$$M^{(P)}(s) = w(s - P/2) \tag{3.3}$$

The shift of the charge mesh in SPME by $P/2$ compared to eq 2.7 is of course irrelevant since the system obeys periodic boundary conditions. In eq 3.2, $b(\boldsymbol{k})$ is given by $b(\boldsymbol{k}_n) = b(n_1)b(n_2)b(n_3)$, where for even P (see ref 2 for the treatment of odd P values)

$$b(n_\alpha) = \frac{e^{2\pi i P n_\alpha/N_\alpha}}{\sum\limits_{l=1}^{P-1} M^{(P)}(l)e^{2\pi i n_\alpha l/N_\alpha}} \tag{3.4}$$

The SPME formula, eq 3.2, for the energy is identical to the P3M formula, eq 2.14, with specific choices of the lattice influence function and of the charge assignment function. In the SPME algorithm, the exponential Euler spline approximation gives thus rise to a modified lattice Green function

$$\tilde{G}_{\mathrm{SPME}}(\boldsymbol{k}) = \hat{\varphi}(\boldsymbol{k})|b(\boldsymbol{k})|^2 \tag{3.5}$$

That expression for $\tilde{G}_{\mathrm{SPME}}(\boldsymbol{k})$ corrects and generalizes eq 28 of ref 9 to general triclinic unit cells. We note that the denominator in $b(\boldsymbol{k})$ is a FFT of $M^{(P)}(s)$ and is hence given by an aliasing sum:

$$\begin{aligned} b(\boldsymbol{k}_n) &= \frac{e^{i\boldsymbol{k}_n\cdot\left(\frac{\boldsymbol{a}_1}{N_1}+\frac{\boldsymbol{a}_2}{N_2}+\frac{\boldsymbol{a}_3}{N_3}\right)P}}{\frac{N_1N_2N_3}{V}\sum_{\boldsymbol{m}}\hat{M}(-\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})} \\ &= \frac{e^{i\boldsymbol{k}_n\cdot\left(\frac{\boldsymbol{a}_1}{N_1}+\frac{\boldsymbol{a}_2}{N_2}+\frac{\boldsymbol{a}_3}{N_3}\right)P/2}}{\sum_{\boldsymbol{m}}\hat{U}(k_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})} \end{aligned} \tag{3.6}$$

where we used eqs 3.3, 2.8, and 2.21 and the symmetry $\hat{U}(-\boldsymbol{k}) = \hat{U}(\boldsymbol{k})$ in the second equality. The lattice Green function of SPME can eventually be rewritten as

$$\tilde{G}_{\mathrm{SPME}}(\boldsymbol{k}_n) = \frac{\hat{\varphi}(\boldsymbol{k}_n)}{\left(\sum_{\boldsymbol{m}}\hat{U}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\right)^2} \tag{3.7}$$

in terms of the harmonics $\hat{U}(\boldsymbol{k})$ of the spline assignment function defined in P3M (eq 2.21). Comparing eq 3.7 to the optimal P3M influence function for the energy, eq 2.25, we see that if aliasing terms $\boldsymbol{m} \neq 0$ are neglected, both functions reduce to the same lattice Green function, eq 2.22, containing the sharpening factor $1/\hat{U}^2(\boldsymbol{k})$.

We note finally that Eastwood's idea of minimizing rms errors of the mesh calculation can also be implemented in a SPME approach, and this has been done in the case of a SPME method with *ik*-differentiation.[19,28] The lattice Green function of the resulting optimized SPME-ik algorithm should then be equivalent to the optimal P3M-ik lattice Green function (both methods were found in ref 19 to have a similar accuracy). Contrary to the P3M approach, the SPME route does however not enlighten the connection between the adjusted lattice Green function and the errors that originate from interpolation smoothing and aliasing.

4. FROM P3M TO SPME AND BACK: CONVERTING ALGORITHMS AND ERROR ESTIMATES

To convert a SPME algorithm into a P3M algorithm, or vice versa, with the same differentiation scheme (that is SPME-ik ↔ P3M-ik or SPME ↔ P3M-AD), it suffices to change the influence function in the program. Since P3M is not always implemented with Gaussian screening, one might furthermore have to make sure that the real-space part of the P3M algorithm does compute the real-space energy associated to the Ewald splitting, eq 2.2.

The conversion SPME → P3M-AD is done for example by replacing eqs 3.7 by 2.28, which is done in practice by replacing the SPME array $|b(\boldsymbol{k}_n)|^2 = (1/(\sum_m\hat{U}(\boldsymbol{k}_{n+\underline{\mathrm{N}}m}))^2$, where $\hat{U}(\boldsymbol{k})$ is defined in eq 2.21 by the array

$$|b_{\mathrm{opt}}(\boldsymbol{k}_n)|^2 = \frac{\hat{U}^2(\boldsymbol{k}_n)\boldsymbol{k}_n^2}{\left(\sum_{\boldsymbol{m}}\hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\right)\sum_{\boldsymbol{m}}\hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\boldsymbol{k}^2_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}} \tag{4.1}$$

$$\simeq \frac{\hat{U}^2(\boldsymbol{k}_n)}{\left(\sum_{\boldsymbol{m}}\hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\right)^2} \tag{4.2}$$

The last expression, eq 4.2, corresponds to using the lattice Green function, eq 2.25, optimized for the energy instead of eq 2.28. Numerical tests have shown that the use of this simpler lattice Green function, which does not need to be tabulated since it can be written in closed form using eq 2.26, does not lead to any noticeable loss of accuracy.

Using the array $|b_{\mathrm{opt}}(\boldsymbol{k}_n)|^2$ in a SPME algorithm will ensure that the FFT Poisson solver is optimized, which will improve somewhat the accuracy at no computational cost. As mentioned in the previous section, an optimized lattice Green function for SPME in the case of the *ik*-differentiation scheme has been derived within the SPME approach, resulting in an algorithm that is basically identical to the P3M-ik algorithm.[19,28]

Similarly, one can convert the error estimate derived for the P3M-AD method into an error estimate for the SPME method. The mean-square error of the forces

$$(\Delta F)^2 \equiv \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{F}_i - \boldsymbol{F}_i^{\mathrm{exact}})^2 \tag{4.3}$$

can be written as the sum of a real- and a reciprocal-space contributions:

$$(\Delta F)^2 = (\Delta F^{(r)})^2 + (\Delta F^{(k)})^2 \tag{4.4}$$

if both errors are assumed to be statistically independent. These errors can be predicted exactly for random uniformly distributed charge distributions, as function of the system's parameters (number N of charged particles, valencies $\{q_i\}$, volume V) and of the method's parameters: Ewald splitting parameter α, real-space cutoff r_c, mesh size (N_1, N_2, N_3) and interpolation order P of the charge assignment function. The accuracy of the real-space part of the calculation is controlled by a formula derived by Kolafa and Perram:[29]

$$\Delta F^{(r)} = \frac{2\mathrm{e}^{-\alpha^2 r_c^2}}{\sqrt{r_c N V}}\sum_{i=1}^{N} q_i^2 \tag{4.5}$$

The accuracy of the reciprocal-space part of the calculation is controlled by[9]

$$\Delta F^{(k)} = V^{-1}\sqrt{\frac{Q}{N}}\sum_{i=1}^{N} q_i^2 \tag{4.6}$$

where

$$Q \equiv \int_V\int_V (F_{\mathrm{P3M}}(\boldsymbol{r}_1, \boldsymbol{r}_2) - F_{\mathrm{exact}}^{(k)}(\boldsymbol{r}_1, \boldsymbol{r}_2))^2\,\mathrm{d}^3\boldsymbol{r}_1\mathrm{d}^3\boldsymbol{r}_2 \tag{4.7}$$

measures the error in the P3M force $F_{\mathrm{P3M}}(\boldsymbol{r}_1, \boldsymbol{r}_2)$ on a test unit charge at $\boldsymbol{r}_1$ due to a unit charge at $\boldsymbol{r}_2$, for all possible positions of both particles in the simulation box $[F_{\mathrm{exact}}^{(k)}(\boldsymbol{r}_1, \boldsymbol{r}_2)]$ is the exact force which can be calculated with an Ewald sum]. Rewriting eq 4.7 in Fourier space and using the explicit expressions for the approximate and the exact forces, Hockney and Eastwood showed that error Q is given by a quadratic polynomial in the influence function

$$Q = \sum_{\substack{\boldsymbol{k}\neq 0\\ \boldsymbol{k}\in\tilde{\mathbb{M}}}} [A(\boldsymbol{k})\tilde{G}^2(\boldsymbol{k}) - 2B(\boldsymbol{k})\tilde{G}(\boldsymbol{k})] + C \tag{4.8}$$

where C is the constant defined by

$$C = \sum_{\boldsymbol{k}\neq 0}\hat{\varphi}(\boldsymbol{k})\boldsymbol{k}^2 = \sum_{\substack{\boldsymbol{k_n}\neq 0\\ \boldsymbol{k_n}\in\tilde{\mathbb{M}}}}\sum_{\boldsymbol{m}}\hat{\varphi}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}^2 \tag{4.9}$$

and expressions for $A(\boldsymbol{k})$ and $B(\boldsymbol{k})$ for both *ik* and analytical differentiation schemes can be found in refs 8 and 16 and are recalled below for convenience:

$$A(\boldsymbol{k_n}) = \begin{cases} (\sum_{\boldsymbol{m}}\hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}))(\sum_{\boldsymbol{m}'}\hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}'})\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}'}^2) & \text{for AD scheme}\\ (\sum_{\boldsymbol{m}}\hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}))^2\boldsymbol{k}_{\boldsymbol{n}}^2 & \text{for } i\boldsymbol{k} \text{ scheme}\end{cases} \tag{4.10}$$

and

$$B(\boldsymbol{k_n}) = \begin{cases} \sum_{\boldsymbol{m}}\hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\hat{\varphi}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}}^2 & \text{for AD scheme}\\ \sum_{\boldsymbol{m}}\hat{U}^2(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\hat{\varphi}(\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}})\boldsymbol{k}_{\boldsymbol{n}}\cdot\boldsymbol{k}_{\boldsymbol{n}+\underline{\mathbf{N}}\boldsymbol{m}} & \text{for } i\boldsymbol{k} \text{ scheme}\end{cases} \tag{4.11}$$

The optimal lattice Green function is the one for which error Q is minimal. It is easy to see that minimizing eq 4.8 yields $\tilde{G}_{\mathrm{opt}}(\boldsymbol{k}) = B(\boldsymbol{k})/A(\boldsymbol{k})$. The numerator in eq 2.28 is therefore the expression of $B(\boldsymbol{k})$ in the case of analytical differentiation, while the denominator is $A(\boldsymbol{k})$.

Formula 4.6 together with eq 4.8 predicts the accuracy of the particle mesh calculation as function of the system's and method's parameters. It applies not only to the P3M, P3M-ik, and P3M-AD methods [in which case Q reduces to its minimal value $Q_{\min} = C - \sum_{\boldsymbol{k}\neq 0} B^2(\boldsymbol{k})/A(\boldsymbol{k})$], but also to the PME and SPME methods. In the case of the SPME method, we simply substitute the SPME influence function, eq 3.7, in eq 4.8 and use the expressions for $A(\boldsymbol{k})$ and $B(\boldsymbol{k})$ that correspond to analytical differentiation.

When using the analytical differentiation scheme (i.e., the SPME or P3M-AD method), each particle is subjected to a self-force that depends on the particle position relative to a mesh cell.[1,10,30] It is important to note that the contribution of the self-forces to the rms error $\Delta F^{(k)}$ is not included in the above error estimate. The rms error due to self-interactions can also be predicted analytically, as has been done for example in the case of the P3M error estimate for the energy[27] (which includes contributions due to self-energies) and for self-energies and self-torques in a dipolar P3M algorithm.[31] Instead of including the contribution of self-interactions into the error estimate, it is actually better to simply subtract the self-interactions directly within the particle mesh calculation. This can be done easily thanks to the analytical formulas for the self-interactions (self-energy and self-force in the case of the SPME and P3M-AD algorithms) derived in ref 10. Alternatively, the self-forces can also be measured and tabulated before the start of a simulation, as first suggested by Cerutti et al.[30] The analysis of ref 10 shows that the subtraction of self-forces in the P3M-AD or SPME algorithm improves the accuracy of the calculation by a factor between 0% and 30%, depending on the values of the method's parameters. All numerical results presented in the next section apply to a SPME algorithm in which the self-forces are subtracted.

5. NUMERICAL RESULTS

We choose the same test system as in ref 9. In the smallest setup, it consists in $N = 100$ particles randomly distributed within a cubic box of length $L = 10$, half of them carry a positive, the other half a negative unit charge. Forces are measured in units of $C^2/\mathcal{L}^2$, where the unit of charge C and the unit of length $\mathcal{L}$ are arbitrary (C and $\mathcal{L}$ could be, for example, the electronic charge and one nanometer or a solar mass and a parsec in a cosmological simulation). By appropriately choosing the unit $\mathcal{L}$, the density of the system, $n = N/L^3 = 0.1\ \mathcal{L}^{-3}$, can correspond to any desired physical density. Once the unit $\mathcal{L}$ is chosen, results for another density n' can moreover be deduced from the results for density n by simply scaling the length $\mathcal{L}$ by a factor $s = (n/n')^{1/3}$. Under the scaling $\mathcal{L} \rightarrow s\mathcal{L}$, the density is indeed multiplied by $s^{-3} = n'/n$, while the forces are multiplied by s^{-2}, the Ewald parameter α is multiplied by s^{-1}, and the mesh spacing h and the real-space cutoff r_c are multiplied by s. The numerical results presented here for the test system exhibit therefore the accuracy of the SPME method for homogeneous (and uncorrelated) systems of arbitrary density. Since the charge valencies $\{q_i\}$ appear moreover only via the simple factor $\sum_i q_i^2$ in the error estimate, eq 4.6, the results can be converted straightforwardly to a system with other charge valencies. An example showing how to predict the accuracy of the SPME forces in a simulation of water from the accuracy curves of the test system will be given below (Section 5.3.3).

By extensivity, identical results would be obtained in larger systems, for instance $N = 100\,000$ particles in a box of length $L = 100$, provided the size of the mesh is increased at the same time proportionally (here 10 times). No new effect is expected indeed to appear when increasing the box size because all wavelengths are uncoupled (the sole characteristic length in this test system is the mean interparticle distance). We checked this point explicitly by performing calculations in larger simulation boxes (all at density $0.1\ \mathcal{L}^{-3}$), in particular $N = 800$ particles in a box $L = 20$ and $N = 12\,500$ in a box $L = 50$. We found that the small setup ($L = 10$) has already reached the thermodynamic limit, as far as the accuracy of forces is concerned, for almost all values of the Ewald splitting parameter α, except for very small ones; a slight system size dependence is observed when $L < 20$ and α is in region $r_c > 13$ in Figure 1. To show curves corresponding to the infinitely extended system even in that region, we opted to use a box with $N = 800$ particles and $L = 20$.

5.1. Predicted versus Actual Errors. We demonstrate the accuracy of the error estimate, eq 4.6, by comparing predicted errors versus actual errors measured on the test system with $L = 20$ defined previously. Exact reference forces are computed by a well converged standard Ewald sum. For better statistics, the measured errors were calculated by averaging the rms error obtained for 10 different random configurations of the charges. We recall that, in our implementation of SPME, the self-forces are subtracted within the particle mesh calculation. We varied the mesh size from 16^3 to 128^3, the real-space cutoff from $r_c = 1$ to 13 and the spline order P from 2 to 7.

As displayed in Figures 1 and 2, we observe excellent agreement between the actual errors and those predicted by eqs 4.5 (real-space error estimate) and 4.6–4.11 (reciprocal-space error estimate). The curves in Figure 1 hold for a B-spline interpolation order $P = 4$, a value that is often a good compromise between accuracy and speed of calculation. Figure 2 shows, in the case of a mesh of size $M = 16$, how the accuracy depends on order P. Increasing P improves of course the accuracy, but this comes at an increased cost for the steps of mapping charges onto the mesh and back-interpolating results from the mesh; these operations scale indeed as $O(P^3)$ with respect to the spline order.

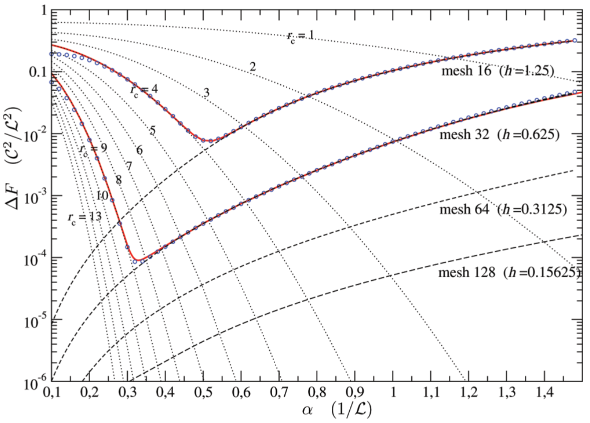


Figure 1. Actual and estimated errors of the SPME method as function the Ewald splitting parameter α. The system has a uniform density equal to 0.1 charged particles per $\mathcal{L}^3$ (800 particles in box $L = 20$). Dotted lines are the estimated errors for the direct part of the Ewald summation, for real-space cutoffs (r_c) varying from 1 to 13 (in units of $\mathcal{L}$). Dashed lines are the error estimate for the reciprocal-space part of the SPME method with self-forces subtracted. The B-spline interpolation order is set to 4, and various mesh sizes are considered: 16^3, 32^3, 64^3, and 128^3. Red lines show the estimate for the total error for two choices of parameters: mesh size 16^3 with $r_c = 4$ and mesh size 32^3 with $r_c = 9$. Actual measured errors for these two sets of parameters are shown as circles.

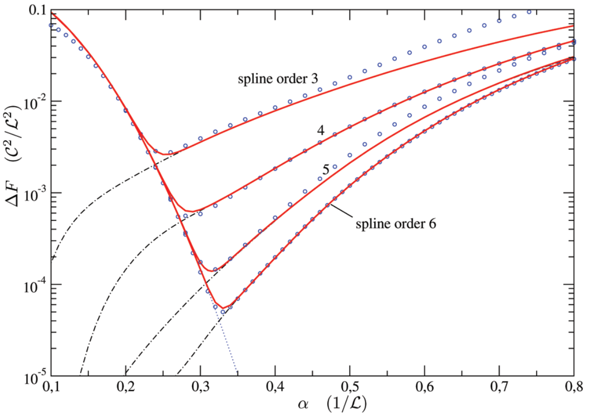


Figure 2. Actual and estimated errors of the SPME method as function of α. The system is the same as in Figure 1. The cutoff in real space is set to $r_c = 9\ \mathcal{L}$ and the mesh size is 16^3 (lattice spacing $h = L/M = 1.25\ \mathcal{L}$). The B-spline order is varied from 3 to 6. Dot-dashed lines show the error estimate for the reciprocal-space part of the SPME calculation. Red lines are the estimated total errors. Measured errors are shown as circles for interpolation orders 3–6.

Since the analytical error estimate is built by uniformly averaging the squared error of the pair interaction over all possible positions of the two particles within the simulation cell, this perfect agreement when comparing to uncorrelated systems (charges located at random) is expected by construction. Real systems have of course correlations, but we will show in Section 5.3, on several examples, that the uncorrelated error estimate still predicts reasonably well the accuracy in "real" systems with correlations. The uniform uncorrelated test system, for which the accuracy is fully understood and under control, serves thus as a useful generic guide on the accuracy. Since it contains no other characteristic length scale apart from the mean interparticle distance, that system exhibits conveniently how the accuracy of a method to compute long-range forces depends intrinsically on its parameters.

Figure 1 displays how the accuracy increases with respect to the number M of mesh points per direction and with respect to the magnitude of the real-space cutoff. At small values of the Ewald splitting parameter α, errors in the direct space part of the calculation dominate, while reciprocal-space errors dominate when α is large. For a given mesh size M and a given cutoff r_c, the parameter α affects only the accuracy and not the computational cost of the calculation. The value of α must therefore be chosen to be at, or at least close to, the point where the error reaches a minimum. The location of that point of highest accuracy is given in good approximation by the condition that the real- and reciprocal-space errors have the same magnitude: $\Delta F^{(r)} = \Delta F^{(k)}$. That equation, which defines the optimal value of α for a given M and r_c, can easily be solved numerically. From Figure 1, we see that to reach an accuracy of, say, 10^{-4} (in units of $\mathcal{C}^2/\mathcal{L}^2$), one can use for example a mesh of size 32^3 in association with a real-space cutoff $r_c = 9$ (and $\alpha \simeq 0.32$) or a mesh 64^3 with a real-space cutoff $r_c = 5$ (and $\alpha \simeq 0.58$). The smaller real-space cutoff makes the direct space part of the calculation faster, but the increased size of the mesh makes the reciprocal space part of the calculation slower. The best choice between the previous two sets of parameters is naturally the one that takes, overall, the least computational time. Note that if particles interact not only via Coulomb's law (or the force of gravity) but also via other short-range forces, it might be necessary to keep the real-space cutoff r_c above a minimum value to ensure that these other short-range forces are computed within the required accuracy. This constraint on the cutoff may be relaxed in some cases since short-range forces that derive from a central potential can be computed using a P3M computation with a Green function adjusted to get the right force.

Figures 1 and 2 demonstrate that the error estimate is accurate for most possible values of the parameters (r_c, M, P, α) of the SPME method, and specifically it is very accurate near the parameter set yielding minimal error. Such plots, or the formulas 4.5 and 4.6 themselves, can be used in association with timings measurements to determine optimal working parameters for the SPME method, i.e., parameters that yield a target accuracy at a minimal computational cost. The open-source soft-matter simulation package ESPResSo[32] contains such an automatic tuning routine for the parameters of a P3M method.

5.2. SPME with Optimized Lattice Green Function.

The FFT Poisson solver is not fully optimized in the SPME method with analytical differentiation, since the SPME lattice Green function is determined by the exponential Euler spline interpolation employed in the charge mapping onto the mesh, whereas this function (which may be tabulated in the algorithm) can actually be adjusted to improve the accuracy of the method, as done in the P3M approach. We recall that if P3M is implemented with analytical differentiation (P3M-AD algorithm), then the SPME and P3M-AD methods differ, at a computational level, only by the choice of the lattice Green function. As shown in Section 4, the lattice Green function in a SPME algorithm can be changed to the error-minimizing lattice Green function of P3M-AD by setting the tabulated array $|b(\boldsymbol{k})|^2$ to expression 4.1 instead of its usual SPME expression.

Figure 3 shows that modifying the lattice Green function to use the P3M-AD optimal lattice Green function improves the accuracy only marginally. This may be traced back to the smallness of the aliasing ($\boldsymbol{m} \neq 0$) terms in the lattice Green functions. Neglecting entirely all aliasing terms, the SPME and P3M lattice Green functions both reduce to the same expression, namely the influence function 2.22 which contains a sharpening factor $1/\hat{U}^2(\boldsymbol{k}_n)$ that compensates for interpolation smoothing. Equation 2.22 may be called the $\boldsymbol{m} = 0$ approximation for the lattice Green function, though we stress that it can be derived trivially without any reference to the aliasing theorem (see reasoning before eq 2.22). If one uses eq 2.22 as the lattice Green function, the resulting accuracy is close to those of the SPME and P3M-AD methods (in-between the two curves or slightly less accurate than SPME depending on the value of α).

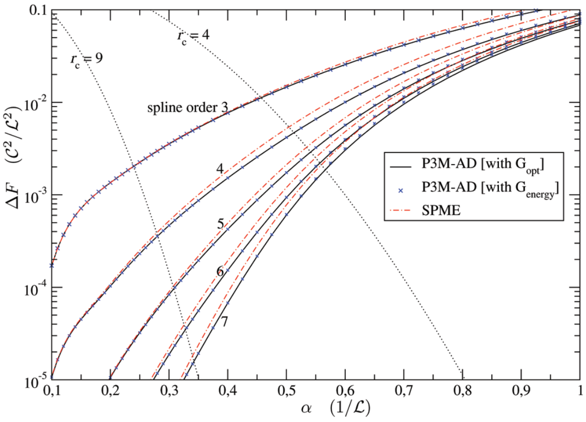


Figure 3. Comparison of reciprocal-space rms errors forces as function of α for three mesh methods: standard SPME (dashed lines), P3M-AD [or equivalently SPME with optimized lattice Green function] (solid lines), and P3M-AD with the closed form approximation 2.25 as the lattice Green function (crosses). Self-forces are subtracted within the particle mesh calculations. The system is the same as in Figure 2. The mesh size is 16^3 (lattice spacing $h = 1.25\ \mathcal{L}$), and the B-spline order is varied from 3 to 7. The real-space error estimate is shown for $r_c = 9$ and 4 $\mathcal{L}$ (dotted lines).

Interestingly, the lattice Green function optimized for the energy, eq 2.25, provides essentially the same accuracy as the true optimal P3M-AD lattice Green function, as evidenced in Figure 3. This is a useful result because eq 2.25 is known in closed analytical form and can be computed on-the-fly in simulations. The small deviations at high values of α are due to truncation errors; these unimportant deviations can be removed by using the full expression for the P3M lattice Green function optimized for the energy, i.e., keeping the aliasing sum in the numerator of eq 2.25. The fact that P3M-AD is very slightly more accurate than SPME demonstrates that the aliasing sum that optimizes the accuracy is indeed the one that appears in eq 2.25 $[(\sum_m \hat{U}^2(\boldsymbol{k}_{n+\underline{N}m}))^2/\hat{U}^2(\boldsymbol{k}_n)]$ and not the one $[(\sum_m \hat{U}(\boldsymbol{k}_{n+\underline{N}m}))^2]$ that appears in SPME.

From the practical point of view, the fact that the SPME lattice Green function is very close to the true optimal lattice Green function confirms the effectiveness of the SPME approach in the standard case where the charge density is discretized on a single mesh. The accuracy gain provided by the true optimal lattice Green function increases when using coarser meshes. If changing the lattice Green function in a standard SPME algorithm may seem unnecessary, it can be expected that it can provide substantial improvement in the accuracy in the case of the interlaced (also known as staggered)

SPME algorithm,[30] since interlaced P3M-AD can be more than 10 times more accurate than interlaced SPME.[33]

5.3. Applications to Correlated Systems. The a priori error estimate introduced in Section 4 is expected to predict only approximately the accuracy in correlated systems, since it assumes all interionic distances to be equiprobable [recall the uniform weighting in definition, eq 4.7, of the mean-square error Q], whereas in reality some distances are more likely than others because of correlations. We observed, however, on several examples of real systems that the uncorrelated error estimate still predicts reasonably well the error even in systems with strong correlations, in agreement with the findings of ref 9. We present comparisons of predicted versus measured errors for three different correlated condensed matter systems: an ionic melt, an ionic crystal, and a liquid water system.

5.3.1. System 1: Ionic Melt. We consider the same NaCl melt as in ref 34. The system consists of a cubic box of size L = 25.557 Å containing 512 particles (256 Na^+ and 256 Cl^- ions) that interact via the Fumi−Tosi potential:

$$v_{ij} = A_{ij}\exp(B(\sigma_{ij} - r)) - \frac{C_{ij}}{r^6} - \frac{D_{ij}}{r^8} + \frac{q_i q_j}{r} \tag{5.1}$$

Parameters A_{ij}, B, σ_{ij}, C_{ij}, and D_{ij} are set to the values given in Table 1 of ref 34. Short-range forces are truncated at r_c = 9 Å, while the long-range Coulomb forces are computed with the P3M method. A short simulation of duration 100 ps was run, at temperature fixed to 1078 K using a Langevin thermostat, to get 10 independent representative configurations of the system. Exact reference forces for these 10 configurations were computed using a well-converged Ewald summation.

Figure 4a shows that the actual rms errors of the Coulombic forces, computed using the SPME method for the 10 configurations, are somewhat smaller (by a factor between 1 and 2) than the errors expected from the error estimate for a random system. The presence of correlations in this system reduces hence the rms errors with respect to an uncorrelated system. This can be understood as arising from the screening of the electric field by the counterion atmosphere: The field (and hence the forces) decays faster than in the uncorrelated system where no such screening occurs. It can moreover be seen that the optimal value of α predicted by the error estimate (for a given mesh size and spline interpolation order) corresponds also essentially to the optimal value for the ionic melt system. As the actual accuracy is higher than the expected accuracy, it is safe, for this particular system, to use the uncorrelated error estimate when tuning the parameters of the SPME method.

As pointed out earlier, Figure 1 contains results on the accuracy of forces in systems of arbitrary density. The density of the test system (0.1 charged particle per $\mathcal{L}^3$) can be made equal to the density of the ionic melt system (0.03 charged particle per Å^3) by setting the unit of length to $\mathcal{L}$ = 1.48 Å. Since the employed mesh spacing is $h = 25.557/16 \simeq 1.6$ Å = 1.08 $\mathcal{L}$, the expected accuracy of the mesh calculation in the ionic melt system (solid line in Figure 4a) lies in between the curves for mesh spacing h = 1.25 and 0.625 in Figure 1. One may take also the opposite path, i.e., convert the accuracy curves from Figure 1 into accuracy curves for Figure 4a, which implies switching to units e^2/Å^2 and adjusting the density to that of the ionic melt. This is achieved by setting $\mathcal{L}$ = 1 Å and by converting the curves in Figure 1 via the scaling $\mathcal{L} \rightarrow s\mathcal{L}$ with s = 1.48, as explained at the beginning of Section 5.

5.3.2. System 2: NaCl Crystal. We consider a NaCl crystal, which is a system where the charges are very strongly correlated. We take again the same parameters for the simulation box and for the Fumi−Tosi interactions as in ref 34. The crystal is made up of 512 particles in a box of size 22.728 Å, which is equivalent to 4 × 4 × 4 units cells with lattice parameter 5.682 Å (the equilibrium value at ambient pressure). Exact reference forces were computed for 10 independent configurations extracted from a 100 ps long simulation of the crystal at 300 K.

Figure 4b shows the comparison between the predicted and the actual errors for SPME forces computed with a mesh of size 16^3. One observes again that the mesh calculation provides more accurate results than expected on the basis of a random system, at least for the region of α values of interest. This lessening of the rms errors in the correlated system is due to partial cancellations of the forces (at zero temperature, all forces would vanish exactly by symmetry).

5.3.3. System 3: SPC/E Water. Since aqueous simulation boxes are ubiquitous in chemical physics and biophysical simulations, we check how the error estimate performs in the case of a water simulation box. We take the water box provided with the GROMACS molecular dynamics simulation package.[11] It contains,

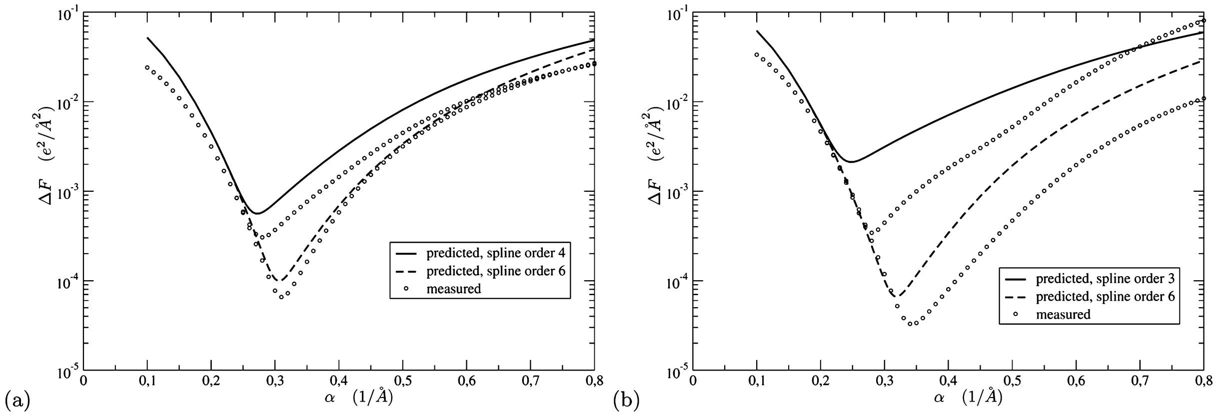


Figure 4. Predicted (solid and dashed lines) and measured (circles) rms accuracy of Coulombic forces as function of α in a simulation box containing 256 Na^+ and 256 Cl^- ions for two different state points: (a) a liquid state (temperature T = 1078 K, box size L = 25.557 Å) and (b) a crystalline state (T = 300 K, L = 22.728 Å). The SPME method is used with a real-space cutoff r_c = 9 Å and a mesh size $M = 16^3$; the used values of the B-spline interpolation order are indicated in the legend box.

in a box of size $L = 18.6206$ Å, 216 water molecules defined by the SPC/E water model.[35] There are therefore in total 648 charged sites in the box: 216 negative sites (O) with charge −0.8476 *e* and 432 positive sites (H) with charge 0.4238 *e*. In this system, the charges are strongly correlated within the molecules (rigid model) and between molecules via the hydrogen-bond network.

The considered water system has a density of 0.1 charged particle/Å^3. Setting the unit of length $\mathcal{L}$ in Figures 1 and 2 to 1 Å, the plots exhibit directly the expected accuracy [in units $e^2/\text{Å}^2 = 1.389 \times 10^4$ kJ/(mol·nm)] for a system with that density. The parameter file provided with GROMACS for this water simulation box specifies, either directly or indirectly, the use of the following SPME parameters: real-space cutoff $r_c = 9$ Å, mesh size $M = 16^3$, interpolation order $P = 4$, and Ewald parameter $\alpha = 0.347$ Å. The employed lattice spacing is therefore $h \simeq 1.16$ Å. Since the lattice spacing used in Figure 2 is slightly coarser (1.25 Å), the reciprocal-space errors will be slightly smaller in the water simulation than in Figure 2.

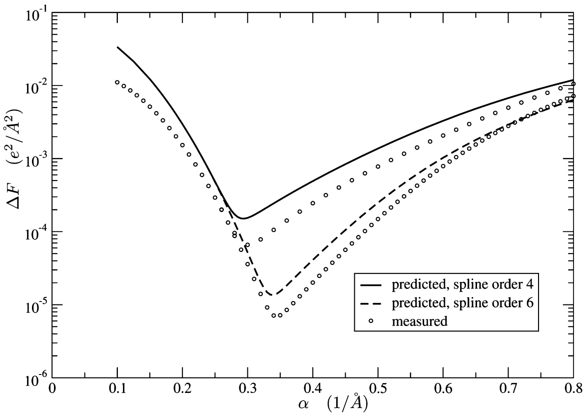


Figure 5. Predicted (solid and dashed lines) and measured (circles) rms accuracy of the forces as function of α, in a simulation box of size $L = 18.6206$ Å containing 216 SPC water molecules. The parameters for the SPME method are: real-space cutoff $r_c = 9$ Å, mesh size $M = 16^3$, and two values are considered for the B-spline interpolation order: $P = 4$ and 6.

The predicted and measured errors for the water simulation box are shown in Figure 5. In that figure, the predicted accuracies have been recalculated using the error estimate, eq 4.4 , in the case of the water simulation box of size $L = 18.6206$ Å. Compared to the rms errors in Figure 2, the reciprocal-space errors are reduced by a factor 0.359 due to the different charge valencies factor ($\sum_i q_i^2$):

$$\frac{216\cdot(-0.8476^2 + 2\cdot 0.4238^2)}{648} = \frac{232.8}{648} \simeq 0.359 \qquad (5.2)$$

One can also observe a further reduction of the reciprocal-space errors by about 25% due to the use of a slightly smaller mesh spacing. The error estimate, valid in principle only for uniform and uncorrelated systems, predicts the actual accuracy of the Coulombic forces in the water system within a factor of about 3.[17] As the errors are overestimated, the error estimate can be used safely in that system. It is not surprising to see a decrease of the rms errors because charges are associated into neutral molecules, which leads to partial cancellations between the forces acting on the charges (at large distances, the force between two molecules reduces to a dipole−dipole interaction).

The optimal values of α predicted by the error estimate ($\alpha_{\text{opt}} \simeq 0.29$ in the case $P = 4$) are also the optimal values for the water system. Figure 5 shows incidentally that value $\alpha = 0.347$ Å specified in the GROMACS parameter file is slightly too large when used with the order $P = 4$: The rms error of the forces would be reduced by almost a factor of 2 if the optimal value of α were used. In the case $P = 6$, the specified value is optimal.

The previous three examples show that correlations in real homogeneous systems tend to reduce the rms errors, with respect to those expected for an uncorrelated system, because of screening effects. If this seems to be the general tendency, it cannot be excluded that correlations can lead to a build up of errors in some specific systems, especially if the system is both strongly correlated and strongly inhomogeneous. When using the error estimate, together with timings measurements, to find an efficient set of particle mesh parameters for a given "real" system, one has to keep in mind that the real accuracy might differ from the prescribed accuracy (which is guaranteed to be satisfied only in the absence of correlations). It is therefore recommended to perform once (or periodically if correlations build up during a simulation) an explicit check of the real accuracy. If the real accuracy turns out to be higher than the expected accuracy, as in the previous examples, the parameters are safe to use. In the opposite case, other parameters must be found, for instance by specifying a higher target accuracy in the tuning routines based on the error estimate.

6. CONCLUSIONS

We exploited the theoretical link between the SPME and P3M-AD methods to demonstrate that the P3M-AD error estimate can be adapted straightforwardly to predict the accuracy of the SPME method as well. On the theoretical side, the existence of the link between these two methods, embodied in expression 3.7 of the SPME lattice Green function, allows to understand the origin of errors in the SPME method using the same powerful concepts and tools (Fourier analysis, sampling theorem, harmonics aliasing) introduced in the error analysis and optimization of the P3M method.

We demonstrated that the a priori error estimate [eqs 4.6 and 4.7] is pinpoint accurate for random uncorrelated systems. It is valid for any values of the SPME parameters, including odd B-spline interpolations orders, contrary to the error estimate derived in ref 17. Our approach avoids also the expensive calculation of the contribution of the self-forces to the rms errors, since we subtract the self-forces within the particle mesh algorithm using the analytical formula derived in ref 10.

The plots in Figures 1−3 can be used to check the accuracy of a SPME calculation and to find the optimal value of the Ewald splitting parameter, for almost any values of the parameters. The figures cover a large range of real-space cutoffs (r_c from 1 to 13) of mesh spacings (from 0.156 to 1.25), and the B-spline order is varied from 2 to 6. A simple scaling of the values can be introduced to deduce from the provided plots the accuracy of the SPME forces in a system of arbitrary density, containing charges with arbitrary valencies. An example of such a scaling is given to predict the accuracy of the forces in a box containing water molecules at normal conditions. In agreement with previous findings, the error estimate is found to predict quite well the accuracy, even though the charges in this system are correlated. The main use of the a priori error estimate is to allow a simulation program to determine automatically the best set of parameters to achieve a target accuracy at a minimal computational cost. Optimal parameters determined in this way depend in general on implementation details of the algorithm and on the machine architecture over which the program is running. An example of an automatic tuning routine for the parameters of the P3M method can be found in the open-source soft-matter simulation package ESPResSo.[32]

We showed moreover that by changing the lattice Green function in a SPME algorithm in favor of the optimized lattice Green function of P3M (which can be either tabulated or computed on-the-fly thanks to a closed form approximation that does not lead to any accuracy loss), the accuracy of the algorithm can be improved marginally. If this change might seem unnecessary for the standard SPME algorithm, it can be expected that it can provide a substantial improvement in accuracy in the case of the interlaced (also known as staggered) SPME algorithm.[30] Our derived error estimate for SPME can moreover be generalized easily to the case of the interlaced SPME method, since the effect of interlacing can be straightforwardly introduced into the a priori error estimate 4.7 (see refs 1 and 33). A detailed analysis of the accuracy and performance of the interlaced SPME and interlaced P3M algorithms will be the subject of a future work.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: vincent.ballenegger@univ-fcomte.fr

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Hockney, R. W.; Eastwood, J. W. *Computer Simulation Using Particles*; IOP: London, U.K., 1988.

(2) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577.

(3) Sagui, C.; Darden, T. A. Molecular dynamics simulations of biomolecules: Long-range electrostatic effects. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 155−179.

(4) Arnold, A.; Holm, C. In Efficient methods to compute long range interactions for soft matter systems. *Advanced Computer Simulation Approaches for Soft Matter Sciences II*; Holm, C., Kremer, K., Eds.; Advances in Polymer Sciences; Springer: Berlin, Germany, 2005; Vol. *II*; pp 59−109.

(5) Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. Phys.* **1921**, *369*, 253−287.

(6) Eastwood, J. W. Optimal particle-mesh algorithms. *J. Comput. Phys.* **1975**, *18*, 1−20.

(7) Langdon, A. B. "Energy-conserving" plasma simulation algorithms. *J. Comput. Phys.* **1973**, *12*, 247−268.

(8) Ferrel, R.; Bertschinger, E. Particle-mesh methods on the connection machine. *Int. J. Mod. Phys. C* **1994**, *5*, 933.

(9) Deserno, M.; Holm, C. How to mesh up Ewald sums. I. A theoretical and numerical comparison of various particle mesh routines. *J. Chem. Phys.* **1998**, *109*, 7678.

(10) Ballenegger, V.; Cerda, J. J.; Holm, C. Removal of spurious self-interactions in particle-mesh methods. *Comput. Phys. Commun.* **2011**, *182*, 1919−1923.

(11) Lindahl, E.; Hess, B.; van der Spoel, D. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. Mol. Mod.* **2001**, *7*, 306−317.

(12) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. The amber biomolecular simulation programs. *J. Computat. Chem.* **2005**, *26*, 1668−1688.

(13) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−1802.

(14) Eastwood, J. W.; Hockney, R. W.; Lawrence, D. N. P3M3DP−the three-dimensional periodic particle-particle/ particle-mesh program. *Comput. Phys. Commun.* **1980**, *19*, 215−261.

(15) Lewis, H. R. Energy-conserving numerical approximations for Vlasov plasmas. *J. Comput. Phys.* **1970**, *6*, 136.

(16) Stern, H. A.; Calkins, K. G. On mesh-based Ewald methods: Optimal parameters for two differentiation schemes. *J. Chem. Phys.* **2008**, *128*, 214106.

(17) Wang, H.; Dommert, F.; Holm, C. Optimizing working parameters of the smooth particle mesh Ewald algorithm in terms of accuracy and efficiency. *J. Chem. Phys.* **2010**, *133*, 034117.

(18) de Leeuw, S. W.; Perram, J. W.; Smith, E. R. Simulation of electrostatic systems in periodic boundary conditions. I. lattice sums and dielectric constants. *Proc. R. Soc. Lond., A* **1980**, *373*, 27−56.

(19) Darden, T. A.; Toukmaji, A.; Pedersen, L. G. Long-range electrostatic effects in biomolecular simulations. *J. Chim. Phys.* **1997**, *94*, 1346−1364.

(20) Frenkel, D.; Smit, B. *Understanding Molecular Simulation*, 2nd ed.; Academic Press: San Diego, CA, 2002.

(21) Eastwood, J. W. Optimal P3M Algorithms For Molecular Dynamics Simulations. In *Computational methods in classical and quantum physics*; Hooper, M., Ed.; Advance Publications Limited: Dublin, Ireland, 1976; pp 206−228.

(22) Luty, B. A.; Tironi, I. G.; van Gunsteren, W. F. Lattice-sum methods for calculating electrostatic interactions in molecular simulations. *J. Chem. Phys.* **1995**, *103*, 3014−21.

(23) Pollock, E. L.; Glosli, J. Comments on P3M, FMM, and the Ewald method for large periodic Coulombic systems. *Comput. Phys. Commun.* **1996**, *95*, 93−110.

(24) Huenenberger, P. H. Optimal charger-shaping functions for the particle-particle-particle-mesh (P3M) method for computing electrostatic interactions in molecular simulations. *J. Chem. Phys.* **2000**, *113*, 10464−10476.

(25) Perram, J. W.; Petersen, H. G.; de Leeuw, S. W. An algorithm for the simulation of condensed matter which grows as the 3/2 power of the number of particles. *Mol. Phys.* **1988**, *65*, 875−893.

(26) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald - an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089.

(27) Ballenegger, V.; Cerdà, J. J.; Lenz, O.; Holm, C. The optimal P3M algorithm for computing electrostatic energies in periodic systems. *J. Chem. Phys.* **2008**, *128*, 034109.

(28) Sagui, C.; Darden, T. A. *P3M and PME: A comparison of the two methods. In Simulation and Theory of Electrostatic Interactions in Solution;* Pratt, L. R., Hummer, G., Eds.; American Institute of Physics: Melville, NY, 1999; ISBN 1-56396-906-8.

(29) Kolafa, J.; Perram, J. W. Cutoff errors in the Ewald summation formulae for point charge systems. *Mol. Simul.* **1992**, *9*, 351−368.

(30) Cerutti, D.; Duke, R. E.; Darden, T. A.; Lybrand, T. P. Staggered mesh Ewald: An extension of the smooth particle-mesh Ewald method adding great versatility. *J. Chem. Theory Comput.* **2009**, *5*, 2322−2338.

(31) Cerdà, J. J.; Ballenegger, V.; Lenz, O.; Holm, C. P3M algorithm for dipolar interactions. *J. Chem. Phys.* **2008**, *129*, 234104.

(32) Limbach, H. J.; Arnold, A.; Mann, B. A.; Holm, C. ESPResSo − an extensible simulation package for research on soft matter systems. *Comput. Phys. Commun.* **2006**, *174*, 704−727.

(33) Neelov, A.; Holm, C. Interlaced P3M algorithm with analytical and ik-differentiation. *J. Chem. Phys.* **2010**, *132*, 234103.

(34) Anwar, J.; Frenkel, D.; Noro, M. G. Calculation of the melting point of NaCl by molecular simulation. *J. Chem. Phys.* **203**, *118*, 728−735.

(35) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **1987**, *91*, 6269−6271.

ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect

Chemical Physics Letters 432 (2006) 78–83

CHEMICAL PHYSICS LETTERS

www.elsevier.com/locate/cplett

Molecular dynamics study of diffusion of formaldehyde in ice

V. Ballenegger [a,*], S. Picaud [a], C. Toubin [b]

[a] Laboratoire de Physique Moléculaire, UMR CNRS 6624, Faculté des Sciences, La Bouloie, Université de Franche-Comté, 25030 Besançon Cedex, France
[b] Laboratoire de Physique des Lasers Atomes et Molécules, UMR CNRS 8523, Centre d'Etudes et de Recherches Lasers et Applications, Université des Sciences et Technologies de Lille, Bat. P5, 59655 Villeneuve d'Ascq Cedex, France

Received 23 August 2006; in final form 2 October 2006
Available online 13 October 2006

Abstract

We report a Molecular Dynamics simulation study of the diffusion process of formaldehyde (CH_2O) in proton-disordered ice Ih at atmospheric pressure, in the temperature range 200–273 K. CH_2O molecules diffuse in ice predominantly by jumping between B sites (bond-breaking mechanism), but substitutional diffusion can also be observed. At 260 K, the diffusion constant is predicted to be 4×10^{-7} cm^2/s with the TIP4P–Ew water model, and 3×10^{-7} cm^2/s with the TIP4P/Ice water model.

---

1. Introduction

Interest in the interactions between trace gases and ice surfaces has been stimulated in recent years by the recognition of the crucial role that ice surfaces can play in catalytic ozone destruction and in partitioning of photochemically active trace gases between air and ice [1]. More recently, the role of the snowpack in the atmospheric chemistry of gases such as formaldehyde (CH_2O) and NO_x above snow-covered surfaces has also been evidenced and it appears now essential to take into account the trace gases/snow interactions in the atmospheric studies above polar regions [2–4].

However, understanding the possible exchange of such trace gases between snow and the lower atmosphere requires the knowledge of their incorporation, diffusion, and release mechanisms in snow crystals. Recent studies have in particular focused on the formaldehyde molecule [5–11], because its photolysis produces oxidizing radicals that enhance the oxidizing capacity of the atmosphere in polar regions, where the other sources of these radicals are reduced. Although the location of CH_2O in snow appeared controversial from the conclusions of these studies, the diffusion coefficient of CH_2O in ice is certainly one among several key parameters that need to be known to quantify the contribution of the snowpack to the atmospheric chemistry of CH_2O [11].

Unfortunately, diffusion measurements of trace gases in ice are scarce due to experimental challenges. Conflicting results have been reported for HCl, with diffusion coefficients ranging from $D \simeq 10^{-5}$ cm^2/s to $D \simeq 10^{-13}$ cm^2/s at $T = 185$ K [12–17]. This wide range of diffusion coefficients has been attributed to different ice preparation conditions, species concentrations, variable defects and grain boundaries, and trapping phenomena [17]. Notice that a strong acid such as HCl certainly dissociates at the above temperatures, influencing thereby the diffusion mechanism. A partially oxidized organic compound such as formaldehyde does not dissociate in ice, but the experimental studies of its diffusion are still very rare. To our knowledge, only one value can be found in the literature, namely $D \simeq 8 \times 10^{-11}$ cm^2/s at 258 K [11]. A similar value has been reported for the diffusion of methanol in ice, but at a much lower temperature ($D \simeq 2 \times 10^{-11}$ cm^2/s at 185 K) [18].

In the present Letter, we use Molecular Dynamics (MD) simulations to characterize the diffusion mechanisms of formaldehyde in ice at a molecular level, and to provide an accurate value of the diffusion coefficient

---

[*] Corresponding author. Fax: +33 3 81 66 64 75.
E-mail address: vincent.ballenegger@univ-fcomte.fr (V. Ballenegger).

0009-2614/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cplett.2006.10.014

of CH_2O in a perfect 3D proton-disordered ice Ih crystal at atmospheric pressure, for temperatures in the range 200–273 K. Indeed, such simulations have been recently proved to be an accurate tool for investigating the diffusion mechanisms of small guest particles in ice, such as He, H_2O, CO_2, N_2, O_2 and CH_4 [19–22].

2. Computational method

2.1. Intermolecular potentials

Many effective pair potentials have been devised to simulate water at liquid densities. We decided to use the TIP4P model of Jorgensen et al. [23], as reparametrized by Horn et al. [24] for use with Ewald summations, because this four-site model is a good compromise between accuracy and speed of computation. The melting temperature of TIP4P–Ew ice at atmospheric pressure is 244 ± 3 K [25]. Although higher than the melting temperatures of the SPC/E and standard TIP4P models, this value is still significantly below the experimental result 273 K. For comparison purposes, we performed also the same simulations with the new TIP4P/Ice model [26], which was designed to reproduce correctly the melting temperature of ice Ih.

For consistency with the chosen water model, we treat the formaldehyde molecule as a rigid body having only translational and rotational degrees of freedom. The interaction sites of a CH_2O molecule are taken from the OPLS force field, and are summarized in Table 1 of Ref. [27]. The cross interaction between a formaldehyde and a water molecule is defined by using the usual Lorentz–Berthelot combining rules (geometric mean of $C^{(6)}$ and $C^{(12)}$ van der Waals coefficients) for each site–site interaction pair. The combination of OPLS and TIP4P parameters has been validated in Ref. [27].

As the models used in this study are slightly different from the original TIP4P model, we carried out energy minimization calculations for the CH_2O–H_2O dimer, and compared with quantum chemistry calculations (B3LYP/6-31+G(d,p) and MP2/6-311+G(d,p)). Both models, TIP4P–Ew and TIP4P/Ice, lead to the same geometry with comparable bond lengths and bond angles. However this configuration is slightly distorted when compared to the ab initio or TIP4P dimer configurations. From an energetic point of view, very similar binding energies are obtained with TIP4P–Ew (−229 meV) compared to TIP4P (−221 meV) and ab initio values (−220 meV (B3LYP) and −228 meV (MP2)), whereas the TIP4P/Ice model tends to overestimate the CH_2O–H_2O interaction (−245 meV). Since the energy discrepancies (more important in the case of the TIP4P/Ice) do not alter significantly the binding properties of CH_2O with H_2O, both models should reproduce the CH_2O–H_2O and CH_2O/ice interaction within reasonable accuracy.

Table 1
Diffusion constant of formaldehyde in ice, computed with the TIP4P–Ew water model (a) and with the TIP4P/Ice water model (b)

T (K)	D (10^{-7} cm^2/s)
(a) TIP4P–Ew	
210	0.44
220	0.78
230	1.45
235	1.80
240	1.92
245	2.27
250	2.37
(b) TIP4P/Ice	
230	1.00
240	1.53
250	2.73
260	2.31
270	3.29
275	3.71

2.2. Simulation details of ice Ih

The system consists in an orthorhombic simulation cell, with edges $a \simeq 2.69$, $b \simeq 3.14$ and $c \simeq 2.89$ (in nm units), containing 768 water molecules for one CH_2O molecule. In the starting configuration, the water molecules are arranged according to the crystallographic structure of hexagonal ice Ih [28]. The protons are oriented randomly, under the restriction that the Bernal–Fowler ice rules are respected [28] and that the system carries no net dipole moment. The simulation cell accommodates a lattice made up of 8 layers, each containing 6×8 hexagonal rings, that are stacked vertically (i.e. along the z- or c-axis). The formaldehyde molecule is placed initially at the centre of an interstitial cavity.

The MD simulations were performed using GROMACS [29], in the NpT ensemble. The temperature and pressure are kept constant by coupling the system to a Berendsen thermostat and barostat. Periodic boundary conditions are applied in all three directions, and long-range electrostatic interactions are calculated using Ewald summations (Ewald splitting parameter $\alpha = 3.47$ nm^{-1}). The cut-off for Lennard-Jones and real-space coulombic interactions was set to 0.9 nm. The Smooth Particle Mesh Ewald method was used to compute reciprocal electrostatic forces and energies; the mesh had a size of $5 \times 6 \times 6$ cells, and cubic splines were used.

The equations of motion were integrated using the leapfrog algorithm with a time step of 2 fs. Each production run lasted at least 100 ns. This very long simulation time is necessary to gather enough statistics on the diffusion of the CH_2O molecule, which is a slow process at low temperatures.

2.3. Determination of diffusion constant

In an isotropic medium, the diffusion constant D of a molecule is related to its mean square displacement (MSD) according to Einstein's formula

$$D = \lim_{t\to\infty} \frac{\langle(\mathbf{r}(t) - \mathbf{r}(0))^2\rangle}{6t} \tag{1}$$

($\mathbf{r}(t)$ is the position of the centre of mass of the molecule). Admittedly, ice is anisotropic, but we observed in our simulations that the diffusion of formaldehyde parallel and perpendicular to the *c*-axis are of the same order of magnitude. We analyse therefore the simulation data by using the isotropic formula.

The ensemble average in the MSD cannot be performed here as usual by averaging over different molecules (since there is only one formaldehyde molecule in the system). The MSD must therefore be computed by averaging the squared displacement measured in several independent trajectories. (This averaging is important because the molecule can travel back and forth and end up very close to its starting position at the end of a simulation). In practice, we use a very long trajectory (duration $T_{\text{sim}} \simeq 100$ ns) that we regard as being composed of N 'independent' trajectories of duration $\Delta t = T_{\text{sim}}/N$. The ensemble average is then approximated by

$$\langle(\mathbf{r}(t) - \mathbf{r}(0))^2\rangle \simeq \frac{\sum_{i=0}^{N-1} \Theta(T_{\text{sim}} - t - t_i)(\mathbf{r}(t+t_i) - \mathbf{r}(t_i))^2}{\sum_{i=0}^{N-1} \Theta(T_{\text{sim}} - t - t_i)} \tag{2}$$

where $t_i = i\Delta t$ is the starting time of the ith segment of the trajectory, and Θ is the Heaviside function. At short times, $0 \leqslant t \leqslant \Delta t$, the MSD is thus obtained as an average of the displacements observed in the N segments, while obviously less statistics is available at longer times. Fig. 1 contains an example of the above 'ensemble' averaging for the MSD of a typical trajectory. We use very small 'restart' times $\Delta t = T_{\text{sim}}/N$ to ensure optimal exploitation of the data ($\Delta t = 200$ ps is small enough for $T_{\text{sim}} = 100$ ns).

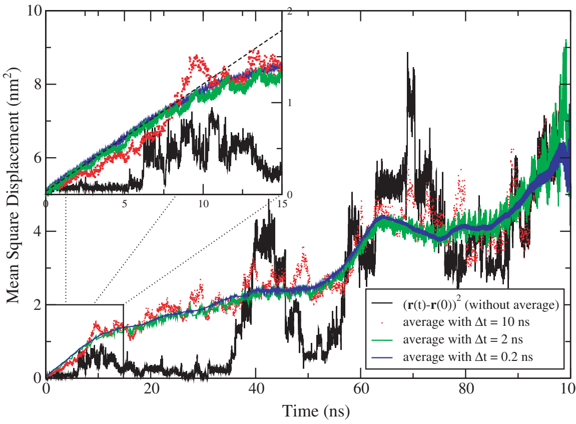


Fig. 1. Mean square displacement of a formaldehyde molecule in proton-disordered ice at 235 K. The ensemble average (2) is computed for three increasingly small 'restart' times Δt. The diffusion coefficient is obtained from the slope of a linear fit of the data for times $t \leqslant 7$ ns, as shown in the inset (dashed line). Linear behaviour is not observed at larger times because of insufficient statistics.

3. Results and discussion

3.1. Diffusion constant

The diffusion of formaldehyde in ice was simulated at constant atmospheric pressure for various temperatures in the range 210–275 K using two different water models: TIP4P–Ew and TIP4P/Ice (see previous section for simulation parameters). The diffusion constant is measured in each run using formulae (1) and (2), and Table 1 summarizes the results.

The highest temperatures chosen (250 K for TIP4P–Ew and 275 K for TIP4P/Ice) are slightly above the melting temperature of these models, but the ice crystal did not melt on the timescale of the simulation (100 ns). Such metastable superheated phases do not exist experimentally, but are known to occur in simulations of bulk systems [25,30], because simulation times are much shorter than the typical time needed to cross the activation energy barrier for melting.

Fig. 2 represents the evolution of the diffusion constant as a function of the (inverse) temperature for the two simulated models. The error bars are rough estimates of the uncertainties based on the convergence of D: they correspond to the maximum difference between D and the values of the diffusion constants which would be obtained by considering only the first, resp. the last, 50 ns of the simulation. For both models, the points clearly follow a linear relationship, showing that the diffusion process obeys, not surprisingly, to Arrhenius law

$$D = D_0 \exp(-E_{\text{diff}}/kT) \tag{3}$$

where E_{diff} is the activation energy of the diffusion process.

For the TIP4P–Ew model, we get $E_{\text{diff}} \simeq 20.2$ kJ/mol and $D_0 \simeq 5 \times 10^{-3}$ cm^2/s, while $E_{\text{diff}} \simeq 14.3$ kJ/mol and $D_0 \simeq 2 \times 10^{-4}$ cm^2/s for the TIP4P/Ice model. In the range 230–250 K, the diffusion constant predicted by both models agree within simulation uncertainties. With the results of the TIP4P/Ice model at higher temperatures

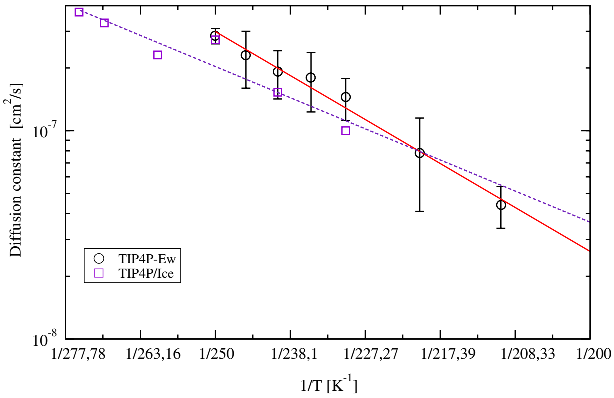


Fig. 2. Arrhenius plot of the diffusion constant of formaldehyde in ice Ih for two different water models.

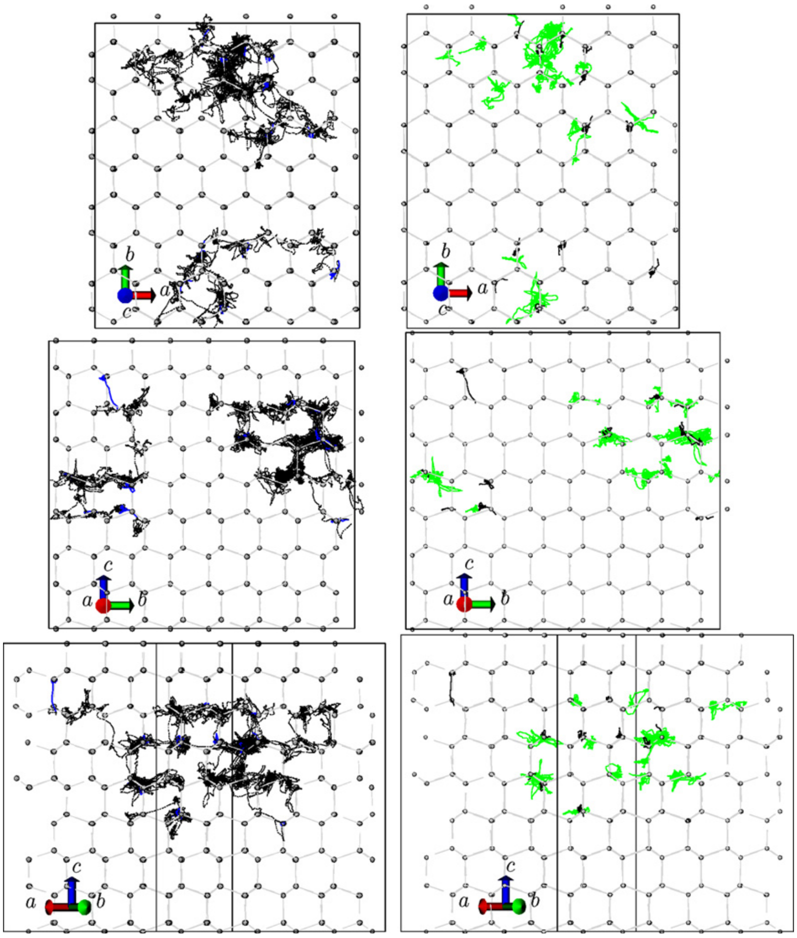


Fig. 3. Left column: Trajectory of the formaldehyde molecule (oxygen atom) during the first 70 ns of the simulation at 250 K (TIP4P–Ew water). The path is shown as a dotted line projected onto the (0001), $(2\bar{1}\bar{1}0)$ and $(1\bar{2}10)$ planes. Right column: some stable sites in the trajectory. The position of the oxygen atom is shown as a black line when the potential energy the formaldehyde molecule is at its low value. Gray lines represent the centre of mass of the molecule in stable sites with high potential energy.

(where statistics is quite good), we see however that the activation energy of the latter model is about 30% lower. The difference between the two models can be made smaller if we rescale the TIP4P results to interpret the value at the melting point of the model (245 K), as corresponding effectively to the experimental melting temperature (273 K), while leaving unchanged the result at the lowest simulated temperature (210 K). If predicting an accurate value of the activation energy is not easy, the simulations give however theoretical estimates for the order of magnitude of the diffusion constant in the whole temperature range 200–273 K.

3.2. Stable sites and diffusion mechanism

Fig. 3 shows the diffusion path followed by the formaldehyde molecule, or more precisely its oxygen atom, in the simulation at 250 K (in this section, all references are to the simulations performed with the TIP4P–Ew model). The trajectory of the O atom is projected onto the (0001), $(2\bar{1}\bar{1}0)$ and $(1\bar{2}10)$ planes[1], and has been smoothed, by a running average over a 20 ps window, to reduce thermal noise. It has been shown by previous authors that a small apolar guest molecule, such as He, Ne, Ar or H_2, diffuse by jumping between Tu (trigonal uncapped) interstitial sites, without distorting the ice lattice [19,21]. The Tu site is located on the axis of the open hexagonal channels along the *c*-axis, and can accomodate larger molecules than the Tc interstitial site (see Fig. 4). It has been shown moreover that large apolar molecules, such as O_2, N_2 and CH_4, diffuse in ice by jumping between so-called B (bond) sites, by the bond-breaking mechanism [19,22]. The *B* site is located at the centre of an O–O bond in the basal plane (the centre of an O–O bond parallel to the *c*-axis corresponds to a similar, but not crystallographically equivalent, Bc site). In the bond-breaking mechanism, the dominant

---

[1] In the Miller–Bravais notation appropriate for hexagonal structures, the third index is redundunt [$i = -h - k$ in $(hkil)$] and is used to show the equivalent directions by cyclic permutations of the indices.

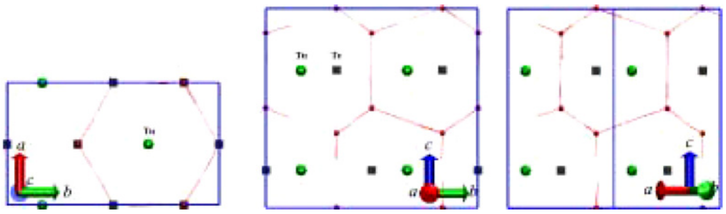


Fig. 4. Orthorhombic unit cell for ice Ih ($a = 4.519$ Å, $b = a\sqrt{3}$, $c = 7.357$ Å at 253 K). The locations of the Tu and Tc interstitial sites are shown by little spheres and cubes respectively. The Tu sites are located on the axis of the open hexagonal channels along the *c*-axis, midway between two layers. The smaller Tc (trigonal capped) interstitial sites are located at the same height, midway between oxygen atoms that are aligned vertically (i.e. along the *c*-axis) and that are not adjacent in a hexagonal ring.

factor governing the diffusion is the repulsive steric interactions between the guest and the H_2O molecules. These interactions distort the structure of the ice lattice, breaking an O–O bond to accomodate the guest molecule in a B site. The two water molecules of the former O–O bond are displaced from their original lattice positions, and are pushed towards them because of the remaining hydrogen bonds with the surrounding H_2O molecules.

A comparison of the diffusion path of CH_2O (Fig. 3) and the positions of the interstitial sites (Fig. 4) shows that this molecule does not diffuse via interstitial sites: the Tu and Tc sites do not accomodate the large CH_2O molecule. This finding is consistent with the results obtained for N_2, O_2, CH_4 and CO_2 [19,22].

Identifying the stable sites of CH_2O in the ice lattice is a difficult task. It is done both by analysing the time evolution of the potential energy $E_f(t)$ of the formaldehyde molecule, and by monitoring the positions $\mathbf{r}_i(t)$ of different sites ($i =$ centre of mass, oxygen, carbon) of the molecule along the trajectory. $E_f(t)$ is plotted in Fig. 5 for the diffusion path represented in Fig. 3 (a running average has been applied to reduce thermal noise). This quantity takes either a low value of about -52 kJ/mol, or a higher value of about -37 kJ/mol. The lower value corresponds to configurations with in average two hydrogen bonds between the oxygen atom and the neighbouring water molecules. Although these configurations minimize $E_f(t)$, they occur only 24% of the time at the considered temperature (250 K). The higher energy sites are favoured because they lead to a lower total energy for the system (CH_2O and lattice). The analysis of configurations $\mathbf{r}_i(t)$ characterized by the lower values of E_f shows that they correspond to $\mathbf{r}_O$ positions at a lattice point normally occupied by the oxygen atom of a water molecule (see Fig. 3). These sites corresponds therefore to the formaldehyde replacing a water molecule, which has been pushed out from its lattice position. As a consequence, jumps between such sites correspond to the so-called 'substitutional' diffusion mechanism. Configurations with high energies E_f correspond by contrast to the formaldehyde occupying a B site, as shown by the position $\mathbf{r}_{CM}(t)$ of its centre of mass (see Fig. 3). Formaldehyde diffuses therefore predominantly by jumping between B sites by the bond-breaking mechanism, though substitutional diffusion does also occur.

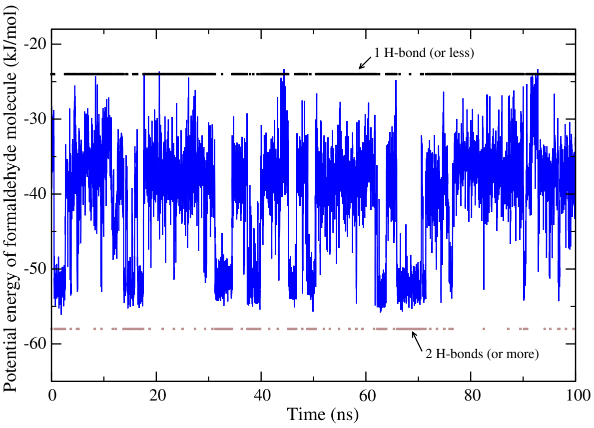


Fig. 5. Potential energy of a diffusing formaldehyde molecule at 250 K. The discontinuous black and grey horizontal lines indicate the time during which the formaldehyde forms one, resp. more than one, hydrogen bond (in average) with the surrounding water molecules.

4. Conclusion

In this study, long simulations (100 ns) were used to elucidate the diffusion mechanism of formaldehyde in ice Ih, and to get accurate values for the diffusion constant of this compound. In the considered temperature range 200–273 K, we found that CH_2O diffuses predominantly by jumping between B sites, though substitutional diffusion does also happen. Because formaldehyde can accept two hydrogen bonds with surrounding water molecules, its diffusion path is more subtle than the pure bond-breaking mechanism that characterizes the diffusion of large apolar molecules.

The time over which the formaldehyde molecule jumps between stable sites is about 20 ps. Since this is one order of magnitude larger than the characteristic time of molecular motion (1 ps), free energy calculations can be used to calculate precisely the jump probability between B sites (using Transition State Theory [31]). The simulation times used in this study are however long enough to allow relatively accurate estimates of the diffusion constant. This quantity obeys Arrhenius law, as expected for an activated process. Using the activation energies reported after Eq. (3), we find that the diffusion constant of CH_2O at 260 K is $D \simeq 4 \times 10^{-7}$ cm^2/s with the TIP4P–Ew water model, and $D \simeq 3 \times 10^{-7}\ \mathrm{cm}^2/\mathrm{s}$ with the TIP4P/Ice water model. These values are much larger than the experimental result of Perrier et al. [11], who measured $D \simeq 8 \times 10^{-11}\ \mathrm{cm}^2/\mathrm{s}$ at 258 K in a macroscopic ice sample. Theory and experiment are known to agree in experimentally well controlled materials, such as in semiconductors and zeolites [32,33]. The present discrepancy may be due to trapping phenomena in the ice crystal, which inevitably contains defects and grain boundaries. Further experimental work on diffusion of CH_2O (and air molecules O_2, N_2, CO_2, CH_4) would be most welcome, especially if it probes the diffusion on a molecular scale in a single crystal.

Acknowledgements

The various snapshots of the simulation cell have been produced with the program VMD (Visual Molecular Dynamics) [34]. The authors gratefully acknowledge the Programme National de Chimie Atmosphérique for financial support.

References

[1] J.H. Seinfeld, S.N. Pandis, Atmospheric Chemistry and Physics, Wiley, New York, 1998.
[2] A.L. Sumner, P.B. Shepson, Nature 398 (1999) 230.
[3] R.E. Honrath, M.C. Peterson, S. Guo, J.E. Dibbs, P.B. Shepson, B. Campbell, Geophys. Res. Lett. 26 (1999) 695.
[4] A.E. Jones, R. Weller, P.S. Anderson, H.W. Jacobi, E.W. Wolff, O. Schrems, H. Miller, Geophys. Res. Lett. 28 (2001) 1499.
[5] M.A. Hutterli, R. Rothlisberger, R.C. Bales, Geophys. Res. Lett. 26 (1999) 1691.
[6] M.A. Hutterli, R.C. Bales, J.R. McConnell, R.W. Stewart, Geophys. Res. Lett. 29 (2002) 1029.
[7] H.W. Jacobi et al., Atmos. Environ. 36 (2002) 2619.
[8] A.K. Winkler, N.S. Holmes, J.N. Crowley, Phys. Chem. Chem. Phys. 4 (2002) 5270.
[9] S. Perrier, S. Houdier, F. Dominé, A. Cabanes, L. Legagneux, A.L. Sumner, P.B. Shepson, Atmos. Environ. 36 (2002) 2695.
[10] J.F. Burkhart, M.A. Hutterli, R.C. Bales, Atmos. Environ. 36 (2002) 2157.
[11] S. Perrier, P. Sassin, F. Dominé, Can. J. Phys. 81 (2003) 319.
[12] M.J. Molina, T.L. Tso, L.T. Molina, F.C. Wang, Science 238 (1987) 1253.
[13] E.W. Wolff, R. Mulvaney, K. Oates, Geophys. Res. Lett. 16 (1989) 487.
[14] F. Dominé, E. Thibert, E. Silvente, M. Legrand, J.L. Jaffrezo, J. Atmos. Chem. 21 (1995) 165.
[15] E. Thibert, F. Dominé, J. Phys. Chem. B 101 (1997) 3554.
[16] A.B. Horn, J. Sully, J. Chem. Soc., Faraday Trans. 93 (1997) 2741.
[17] F.E. Livingston, S.M. George, J. Phys. Chem. A 105 (2001) 5155.
[18] F.E. Livingston, J.A. Smith, S.M. George, J. Phys. Chem. A 106 (2002) 6309.
[19] T. Ikeda-Fukazawa, S. Horikawa, T. Hondoh, K. Kawamura, J. Chem. Phys. 117 (2002) 3886.
[20] A. Demurov, R. Radhakrishnan, B.L. Trout, J. Chem. Phys. 116 (2002) 702.
[21] T. Ikeda-Fukazawa, K. Kawamura, T. Hondoh, Chem. Phys. Lett. 385 (2004) 467.
[22] T. Ikeda-Fukazawa, K. Kawamura, T. Hondoh, Mol. Sim. 30 (2004) 973.
[23] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, J. Chem. Phys. 79 (2005) 926.
[24] H.W. Horn, W.C. Swope, J.W. Pitera, J.D. Madura, T.J. Dick, G.L. Hura, T. Head-Gordon, J. Chem. Phys. 120 (2004) 9665.
[25] R. García Fernández, J.L.F. Abascal, C. Vega, J. Chem. Phys. 124 (2006), art. no 144506.
[26] J.L.F. Abascal, E. Sanz, R. García Fernández, C. Vega, J. Chem. Phys. 122 (2005), art. no 234511.
[27] B. Collignon, S. Picaud, Chem. Phys. Lett. 393 (2004) 457.
[28] V.F. Petrenko, R.W. Whitworth, Physics of Ice, Oxford University Press, 1999.
[29] E. Lindahl, B. Hess, D. van der Spoel, J. Mol. Mod. 7 (2001) 306–317. http://www.gromacs.org.
[30] S.C. Gay, E.J. Smith, A.D.J. Haymet, J. Chem. Phys. 116 (2002) 8876.
[31] B.H. Mahan, J. Chem. Educ. 51 (1974) 709.
[32] V Gusakov, J. Phys.: Condens. Matter 17 (2005) S2285.
[33] M. Kawano, B. Vessal, C.R.A. Catlow, J. Chem. Soc., Chem. Commun. (1992) 879.
[34] W. Humphrey, A. Dalke, K. Schulten, J. Mol. Graphics 14 (1996) 33.

Available online at www.sciencedirect.com
ScienceDirect
Planetary and Space Science 56 (2008) 1607–1617
Planetary and Space Science
www.elsevier.com/locate/pss

A theoretical investigation into the trapping of noble gases by clathrates on Titan

C. Thomas*, S. Picaud, O. Mousis, V. Ballenegger

Institut UTINAM, Université de Franche-Comté, CNRS/INSU, UMR 6213, 25030 Besançon Cedex, France

Received 23 April 2008; accepted 23 April 2008
Available online 6 May 2008

Abstract

In this paper, we use a statistical thermodynamic approach to quantify the efficiency with which clathrates on the surface of Titan trap noble gases. We consider different values of the Ar, Kr, Xe, CH_4, C_2H_6 and N_2 abundances in the gas phase that may be representative of Titan's early atmosphere. We discuss the effect of the various parameters that are chosen to represent the interactions between the guest species and the ice cage in our calculations. We also discuss the results of varying the size of the clathrate cages. We show that the trapping efficiency of clathrates is high enough to significantly decrease the atmospheric concentrations of Xe and, to a lesser extent, of Kr, irrespective of the initial gas phase composition, provided that these clathrates are abundant enough on the surface of Titan. In contrast, we find that Ar is poorly trapped in clathrates and, as a consequence, that the atmospheric abundance of argon should remain almost constant. We conclude that the mechanism of trapping noble gases via clathration can explain the deficiency in primordial Xe and Kr observed in Titan's atmosphere by *Huygens*, but that this mechanism is not sufficient to explain the deficiency in Ar.

Keywords: Titan; Clathrates; Atmosphere; Noble gases; Cassini–Huygens

1. Introduction

Saturn's largest satellite, Titan, has a thick atmosphere, primarily consisting of nitrogen, with a few percent of methane (Niemann et al., 2005). An unexpected feature of this atmosphere is that no primordial noble gases, other than Ar, were detected by the gas chromatograph mass spectrometer (GCMS) aboard the *Huygens* probe during its descent on January 14, 2005. The detected Ar includes primordial ^{36}Ar (the main isotope) and the radiogenic isotope ^{40}Ar, which is a decay product of ^{40}K (Niemann et al., 2005). The other primordial noble gases ^{38}Ar, Kr and Xe, were not detected by the GCMS instrument, yielding upper limits of 10^{-8} for their mole fractions in the gas phase. Furthermore, the value of the ^{36}Ar/^{14}N ratio is about six orders of magnitude lower than the solar value, indicating that the amount of ^{36}Ar is surprisingly low within Titan's atmosphere (Niemann et al., 2005). These observations seem to be at odds with the idea that noble gases are widespread in the bodies of the solar system. Indeed, these elements have been measured in situ in the atmospheres of the Earth, Mars and Venus, as well as in meteorites (Owen et al., 1992). The abundances of Ar, Kr and Xe were also measured to be oversolar by the *Galileo* probe in the atmosphere of Jupiter (Owen et al., 1999).

In order to explain the observed deficiency of primordial noble gases in Titan's atmosphere, Osegovic and Max (2005) proposed that the noble gases within the atmosphere could be trapped in clathrates located on the surface of Titan. The authors calculated the composition of clathrates on the surface of Titan using the program CSMHYD (developed by Sloan, 1998) and showed that such crystalline ice structures may act as a sink for Xe. The facts that the code used by Osegovic and Max (2005) is not suitable below about 140 K for gas mixtures of interest, and that the authors did not explicitly calculate the trapping efficiencies of Ar and Kr in clathrates on the surface of Titan led Thomas et al. (2007) to reconsider their results. In particular, Thomas et al. (2007) performed more accurate calculations of the trapping of

---

*Corresponding author. Tel.: +33 3 81 66 64 78; fax: +33 3 81 66 64 75.
E-mail address: caroline.thomas@univ-fcomte.fr (C. Thomas).

0032-0633/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.pss.2008.04.009

noble gases in clathrates using a statistical thermodynamic model based on experimental data and on the original work of van der Waals and Platteeuw (1959). On this basis, Thomas et al. (2007) showed that Xe and Kr could have been progressively absorbed in clathrates located at the surface of Titan during its thermal history, in contrast with Ar, which is poorly trapped in these crystalline structures. They then concluded that their calculations are only partly consistent with the *Huygens* probe measurements, since the presence of clathrates on the surface of Titan cannot explain the primordial Ar deficiency in its atmosphere.

In this paper, we aim to extend the work of Thomas et al. (2007) by considering the clathration of a more plausible gas mixture representative of the composition of Titan's atmosphere, in which we include Ar, Kr and Xe all together. Indeed, Osegovic and Max (2005) and Thomas et al. (2007) both considered atmospheric compositions containing only one noble gas at a time. As we shall see, the competition between the various species strongly affects the efficiency with which they are trapped in clathrates. In addition, because the incorporation conditions of guest species in clathrates depend on the structural characteristics of the crystalline network and on the intermolecular potentials, we examine the influence of the size of cages and of the interaction potential parameters on our calculations. The study of the influence of these parameters is motivated by two distinct facts. First, it is known that the size of the clathrate cages depends on the temperature (Shpakov et al., 1998; Belosludov et al., 2003; Takeya et al., 2006), and secondly, several different sets of Kihara parameters are available in the literature (Parrish and Prausnitz, 1972; Diaz Peña et al., 1982; Jager, 2001). As a result, this study of sensitivity to parameters allows us to better quantify the accuracy of the results.

In Section 2, we describe the statistical model used to calculate the composition of the clathrates. This hybrid model is based on the work of van der Waals and Platteeuw (1959), and on available experimental data. We also compare the dissociation pressure of clathrates obtained from our model with that obtained from the CSMHYD program (Sloan, 1998). CSMHYD uses a more sophisticated and rigorous approach, but is limited to carrying out calculations above about 140 K (Sloan, 1998). In Section 3, we investigate the sensitivity of our model to various parameters, and also examine the influence of these parameters on the predicted clathrate composition. In Section 4, the statistical approach developed in Section 2 is used to calculate the relative abundances of guests trapped in clathrates that may exist on the surface of Titan. Several hypotheses for the abundance of noble gases in the atmosphere of Titan are tested. Section 5 is devoted to the summary and discussion of our results.

2. Theoretical background

To calculate the relative abundance of guest species incorporated in a clathrate from a coexisting gas of specified composition at given temperature and pressure, we follow the formalism developed by Lunine and Stevenson (1985), which is based on the statistical mechanics approach of van der Waals and Platteeuw (1959). Such an approach relies on four key assumptions: the host molecules contribution to the free energy is independent of the clathrate occupancy (this assumption implies in particular that the guest species do not distort the cages), the cages are singly occupied, there are no interactions between guest species in neighboring cages, and classical statistics is valid, i.e., quantum effects are negligible (Sloan, 1998).

In this formalism, the occupancy fraction of a guest species G for a given type t ($t =$ small or large) of cage, and for a given type of clathrate structure (I or II) can be written as

$$y_{G,t} = \frac{C_{G,t} P_G}{1 + \sum_J C_{J,t} P_J}, \tag{1}$$

where $C_{G,t}$ is the Langmuir constant of guest species G in the cage of type t, and P_G is the partial pressure of guest species G. Note that this assumes that the sample behaves as an ideal gas. The partial pressure is given by $P_G = x_G \times P$, with x_G the molar fraction of guest species G in the initial gas phase, and P the total pressure. The sum, $\sum_J$, in the denominator runs over all species J which are present in the initial gas phase.

The Langmuir constants indicate the strength of the interaction between each guest species and each type of cage. This interaction can be accurately described, to a first approximation, on the basis of the spherically averaged Kihara potential $w_G(r)$ between the guest species G and the water molecules forming the surrounding cage (McKoy and Sinanoğlu, 1963), written as

$$\begin{aligned} w_G(r) = 2z\varepsilon_G & \frac{\sigma_G^{12}}{R_c^{11} r}\left(\delta_G^{(10)}(r) + \frac{a_G}{R_c}\delta_G^{(11)}(r)\right) \\ & - \frac{\sigma_G^6}{R_c^5 r}\left(\delta_G^{(4)}(r) + \frac{a_G}{R_c}\delta_G^{(5)}(r)\right), \end{aligned} \tag{2}$$

where R_c represents the radius of the cavity assumed to be spherical. z is the coordination number of the cell and r the distance from the guest molecule to the cavity center. The parameters R_c and z depend on the structure of the clathrate and on the type of the cage (small or large), and are given in Table 1. The functions $\delta_G^{(N)}(r)$ are defined as

$$\delta_G^{(N)}(r) = \frac{1}{N}\left[\left(1 - \frac{r}{R_c} - \frac{a_G}{R_c}\right)^{-N} - \left(1 + \frac{r}{R_c} - \frac{a_G}{R_c}\right)^{-N}\right], \tag{3}$$

where a_G, σ_G and ε_G are the Kihara parameters for the interactions between guest species and water. The choice of the Kihara parameters for the guest species considered in the present study is discussed in the next section. The parameters chosen for our calculations are given in Table 2.

Table 1
Parameters for the cavities

Clathrate structure	I		II	
Cavity type	Small	Large	Small	Large
R_c (Å)	3.975	4.300	3.910	4.730
b	2	6	16	8
z	20	24	20	28

R_c is the radius of the cavity (values taken from Parrish and Prausnitz, 1972). b represents the number of small (b_s) or large (b_ℓ) cages per unit cell for a given structure of clathrate (I or II), z is the coordination number in a cavity.

Table 2
Three different sets of parameters for the Kihara potential

Ref.	Molecule	σ (Å)	ε/k_B (K)	a (Å)
(a)	CH_4	3.2398	153.17	0.300
	C_2H_6	3.2941	174.97	0.400
	N_2	3.2199	127.95	0.350
	Xe	3.1906	201.34	0.280
	Ar	2.9434	170.50	0.184
	Kr	2.9739	198.34	0.230
(b)	CH_4	3.019	205.66	0.313
	C_2H_6	3.038	399.07	0.485
	N_2	2.728	145.45	0.385
	Xe	3.268	302.49	0.307
	Ar	2.829	155.30	0.226
	Kr	3.094	212.70	0.224
(c)	CH_4	3.1514	154.88	0.3834
	C_2H_6	3.2422	189.08	0.5651
	N_2	3.0224	127.67	0.3526
	Xe	3.3215	192.95	0.2357
	Ar	–	–	–
	Kr	–	–	–

σ is the Lennard-Jones diameter, ε is the depth of the potential well, and a is the radius of the impenetrable core. These parameters come from the papers of (a) Parrish and Prausnitz (1972), (b) Diaz Peña et al. (1982) and (c) Jager (2001).

The Langmuir constants are then determined by integrating the Kihara potential within the cage as

$$C_{G,t} = \frac{4\pi}{k_B T} \int_0^{R_c} \exp\left(-\frac{w_G(r)}{k_B T}\right) r^2 \, dr, \tag{4}$$

where T represents the temperature and k_B the Boltzmann constant.

Finally, the relative abundance f_G of a guest species G in a clathrate (of structure I or II) is defined as the ratio of the average number of guest molecules of species G in the clathrate over the average total number of enclathrated molecules :

$$f_G = \frac{b_s y_{G,s} + b_\ell y_{G,\ell}}{b_s \sum_J y_{J,s} + b_\ell \sum_J y_{J,\ell}}, \tag{5}$$

where the sums in the denominator run over all species present in the system, and b_s and b_ℓ are the number of small and large cages per unit cell, respectively. Note that the relative abundances of guest species incorporated in a clathrate can differ strongly from the composition of the coexisting gas phase because each molecular species has a different affinity with the clathrate.

The calculations are performed at temperature and pressure conditions at which the multiple guest clathrates are formed. The corresponding temperature and pressure values ($T = T_{\mathrm{mix}}^{\mathrm{diss}}$ and $P = P_{\mathrm{mix}}^{\mathrm{diss}}$) can be read from the dissociation curve of the multiple guest clathrates.

In the present study, the dissociation pressure is determined from available experimental data and from a combination rule due to Lipenkov and Istomin (2001). Thus, the dissociation pressure $P_{\mathrm{mix}}^{\mathrm{diss}}$ of a multiple guest clathrate is calculated from the dissociation pressure P_G^{diss} of a pure clathrate of guest species G as

$$P_{\mathrm{mix}}^{\mathrm{diss}} = \left(\sum_G \frac{x_G}{P_G^{\mathrm{diss}}} \right)^{-1}, \tag{6}$$

where x_G is the molar fraction of species G in the gas phase.

The dissociation pressure P_G^{diss} is derived from laboratory measurements and follows an Arrhenius law (Miller, 1961):

$$\log(P_G^{\mathrm{diss}}) = A + \frac{B}{T}, \tag{7}$$

where P_G^{diss} and T are expressed in Pa and K, respectively. The constants A and B used in the present study have been fitted to the experimental data given by Lunine and Stevenson (1985) and by Sloan (1998) and are listed in Thomas et al. (2007).

The present approach differs from that proposed by Sloan (1998) in the CSMHYD program, in which the dissociation pressure of the multiple guest clathrate is calculated in an iterative way by requiring that the chemical potential in the clathrate phase is equal to that in the gas phase. The determination of this equilibrium requires knowledge of the thermodynamics of an empty hydrate, such as the chemical potential, enthalpy and volume difference between ice (chosen as a reference state) and the empty hydrate. The experimental data available at standard conditions ($T = 273.15\,\mathrm{K}, P = 1\,\mathrm{atm}$) allow the CSMHYD program to calculate chemical potentials, and hence dissociation curves, as long as the temperature and pressure is not too far from the reference point. This method fails to converge at temperatures below 140 K for the clathrates considered in this study. Our approach avoids this problem, because it uses experimentally determined dissociation curves, which are valid down to low temperatures.

As an illustration, Fig. 1 shows a comparison between the dissociation curve obtained from our approach (full line) and that calculated using the CSMHYD program (crosses) for a multiple guest clathrate corresponding to an initial gas phase composition of 4.9% CH_4, 0.1% C_2H_6 and 95% N_2. For this comparison, both calculations have been performed with the same set of parameters (Kihara parameters and cage geometries given by Sloan, 1998).

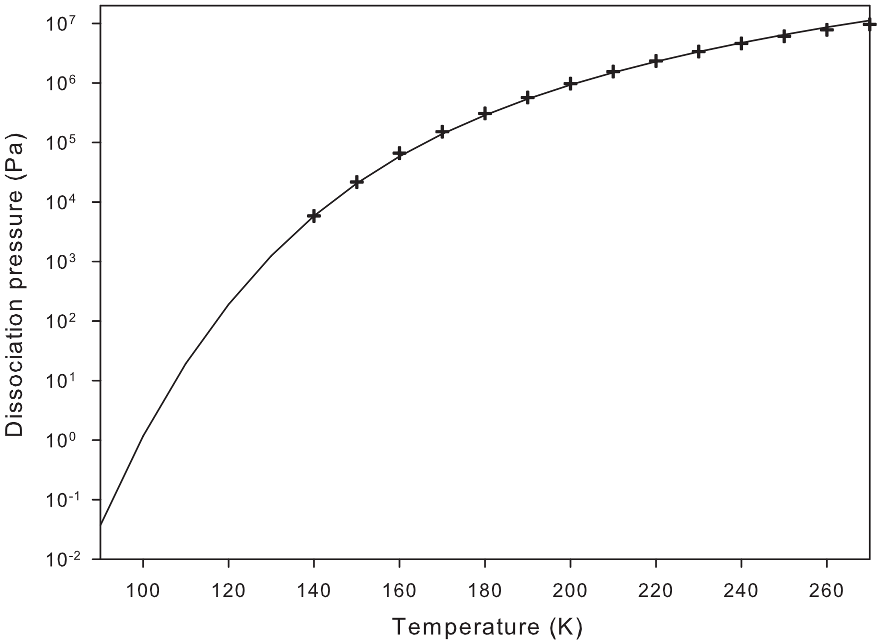


Fig. 1. Dissociation curves pressure as a function of temperature for a multiple guest clathrate corresponding to an initial gas phase composition of 4.9% of CH_4, 0.1% of C_2H_6 and 95% of N_2. The calculations have been performed using either the approach discussed in the present paper (full line) or the CSMHYD program proposed by Sloan (1998) (crosses).

Both methods give very similar results in the range where the CSMHYD program converges.

3. Sensitivity to parameters

The present calculations of the relative abundances of guest species incorporated in a clathrate depend on the structural characteristics of this clathrate (size of the cages for example) and also on the parameters of the Kihara potential. It is thus useful to assess the influence of these structural characteristics and potential parameters on the calculations, in order to better quantify the accuracy of the results.

3.1. Discussion of the Kihara parameters

Papadimitriou et al. (2007) have recently illustrated the sensitivity of clathrate equilibrium calculations to Kihara parameters values, in the case of methane and propane clathrates. Indeed, by perturbing in the range ± (1–10%) the σ and ε Kihara parameters originally given in Sloan (1998), they have demonstrated that these parameters have a significant effect on the values of the Langmuir constants and on the dissociation curves. It appears thus of fundamental importance to assess the accuracy of the Kihara parameters which are used in the studies of clathrates, for example by comparing the theoretical results with available experimental data. However, because in the present paper we are interested in the calculations of the relative abundances of noble gases in clathrates on Titan which are not experimentally available, our choice of Kihara parameters has been based on the literature, only.

For the guest species considered here, i.e., CH_4, C_2H_6, N_2, Ar, Kr, and Xe, as far as we know, there are only two full sets of Kihara parameters in the literature (Parrish and Prausnitz, 1972; Diaz Peña et al., 1982) which have been used in the context of clathrate studies. These parameters are unfortunately very different (see Table 2). The parameters given by Parrish and Prausnitz (1972) have been obtained by comparing calculated chemical potentials based on the structural data of the clathrates cages given by von Stackelberg and Müller (1954) with experimental results based on clathrate dissociation pressure data (Parrish and Prausnitz, 1972). The parameters given by Diaz Peña et al. (1982) have been fitted on experimentally measured interaction virial coefficients for binary mixtures. They have been recently used by Iro et al. (2003) to quantify the trapping by clathrates of gases contained in volatiles observed in comets. The corresponding calculations were performed by using the clathrate cage parameters given by Sparks et al. (1999). These two sets of Kihara parameters (Parrish and Prausnitz, 1972; Diaz Peña et al., 1982) can be partly compared to those recently given for CH_4, C_2H_6, N_2, and Xe, only, in the Ph.D. work of Jager (2001).

Table 2 shows that the σ and ε parameters used by Jager (2001) are quite close to those given by Parrish and Prausnitz (1972) and, as a consequence, the relative abundances f_G (with $G = CH_4$, C_2H_6, N_2, or Xe) we have calculated with these two sets of parameters are of the same order of magnitude, and behave similarly with temperature. By contrast, the relative abundances calculated with the parameters given by Diaz Peña et al. (1982) are very different from those calculated with the two other sets of Kihara parameters, as expected from the conclusions of Papadimitriou et al. (2007).

Because the potential and structural parameters given by Parrish and Prausnitz (1972) (Table 1) have been self-consistently determined on experimentally measured clathrates properties, and also because they give results similar to those obtained when using Jager's (2001) parameters for a reduced set of species, we choose here the Parrish and Prausnitz's parameters for the rest of our study.

3.2. *Influence of the size of the cages*

In the present paper, we have chosen the potential and structural parameters given by Parrish and Prausnitz (1972) (Table 1). However, it has been shown that the size of the cages can vary as a function of the temperature (thermal expansion or contraction) and also of the size of the guest species (Shpakov et al., 1998; Takeya et al., 2006; Belosludov et al., 2003; Hester et al., 2007). Indeed, the structural parameter R_c increases with temperature and with the size of the guest species, whereas it decreases when small guest species are encaged. For example, the lattice constant of the methane hydrate is increased by 0.3% between 83 and 173 K (Takeya et al., 2006).

In order to quantify the influence of variations of the size of the cages on the relative abundances calculated for the multiple guest clathrate considered in the present study, we have modified by ± (1–5%) the R_c values given in Table 1. These variations are compatible with typical thermal expansion or contraction in the temperature range 90–270 K (Shpakov et al., 1998; Takeya et al., 2006; Belosludov et al., 2003; Hester et al., 2007).

Fig. 2 shows the evolution with temperature of f_{Ar}, f_{Kr} and f_{Xe} (calculated from Eq. (5)) in a multiple guest clathrate, for both structures I and II, and for variations of the size of the cages equal to ±1% and ±5%. The calculations have been performed for an initial gas phase containing CH_4, C_2H_6, N_2, Ar, Kr and Xe. The gas phase abundance of CH_4 (4.92%) has been taken from Niemann et al. (2005), whereas the values for C_2H_6 (0.1%), Ar (0.1%), Kr (0.1%) and Xe (0.1%) are based on our previous study (Thomas et al., 2007). The relative abundance of N_2 (94.68%) has been determined accordingly.

Our results show that a small variation (±1%) of the size of the cages has only a very small effect on the trapping of noble gases in the corresponding multiple guest clathrate, irrespective of the temperature (Fig. 2). In particular, the behavior (i.e., increase or decrease) with temperature is not affected by small variations of the size of the cages. A similar small effect is also observed for an expansion of the cage by 5%. By contrast, a large contraction of the cages (R_c decreased by 5%) leads to strong modifications of the relative abundances f_G ($G = \mathrm{Ar, Kr, Xe}$) which can vary by several orders of magnitude.

The evolution of the relative abundances of all guests in the clathrates is also given in Fig. 3 as a function of the sizes of the cages, for given pressure and temperature. The corresponding calculations have been performed at a pressure $P = 1.5$ bar (i.e., the present atmospheric pressure at the surface of Titan), and at the corresponding temperature given by the dissociation curve, i.e., $T_{\mathrm{mix}}^{\mathrm{diss}} = 176$ K (see below). Fig. 3 clearly shows that, for the given P and T conditions, the contractions of the cages have a larger effect than the expansions on the relative abundances calculated in clathrates and that these effects are also strongly dependent on the interaction parameters between the guest species and the cages. As a consequence, the relative abundances in clathrates of the noble gases considered in the present paper appear to be much more dependent on the size of the cages than those calculated for CH_4, C_2H_6 and N_2. This feature can be related to the values of the ε parameters of the Kihara potential (see Table 2) which are larger for the noble gases than for the other species. Moreover, the relative abundances of the smallest species (i.e., the noble gases which have the smallest values of the σ Kihara parameters) are increased when decreasing the sizes of the cages.

To summarize, the results given in Figs. 2 and 3 show that thermal variations of the cages need to be taken into account if these variations are greater than a few percents. Unfortunately, these variations with temperature are often not known, except for a small number of specific systems, such as the methane clathrate between 83 and 173 K (Takeya et al., 2006) for which the variations of the cages have been shown to be very small. In a first approximation, such thermal variations will thus be neglected in our calculations of the f_G evolution on Titan, as a function of the temperature.

4. Trapping of noble gases by clathrates on Titan

The statistical approach outlined in Section 2 is used to calculate the relative abundances of CH_4, C_2H_6, N_2, Ar, Kr and Xe in a multiple guest clathrate (structures I and II), as a function of the temperature. As discussed above, the interactions between the guests and the surrounding cages have been calculated by using the Kihara potential with the parameters given by Parrish and Prausnitz (1972), and by disregarding the possible influence of the thermal variations of the cages.

The initial gas phase abundance of CH_4 (4.92%) is taken from Niemann et al. (2005), whereas three different sets of initial abundances are considered for N_2, C_2H_6, Ar, Kr, and Xe in the atmosphere of Titan. The first set of values (hereafter case 1) is derived from the atmospheric composition considered by Osegovic and Max (2005) and was also used in our previous study (Thomas et al., 2007). The second set of values is calculated under the assumption that each ratio of noble gas to methane gas in the

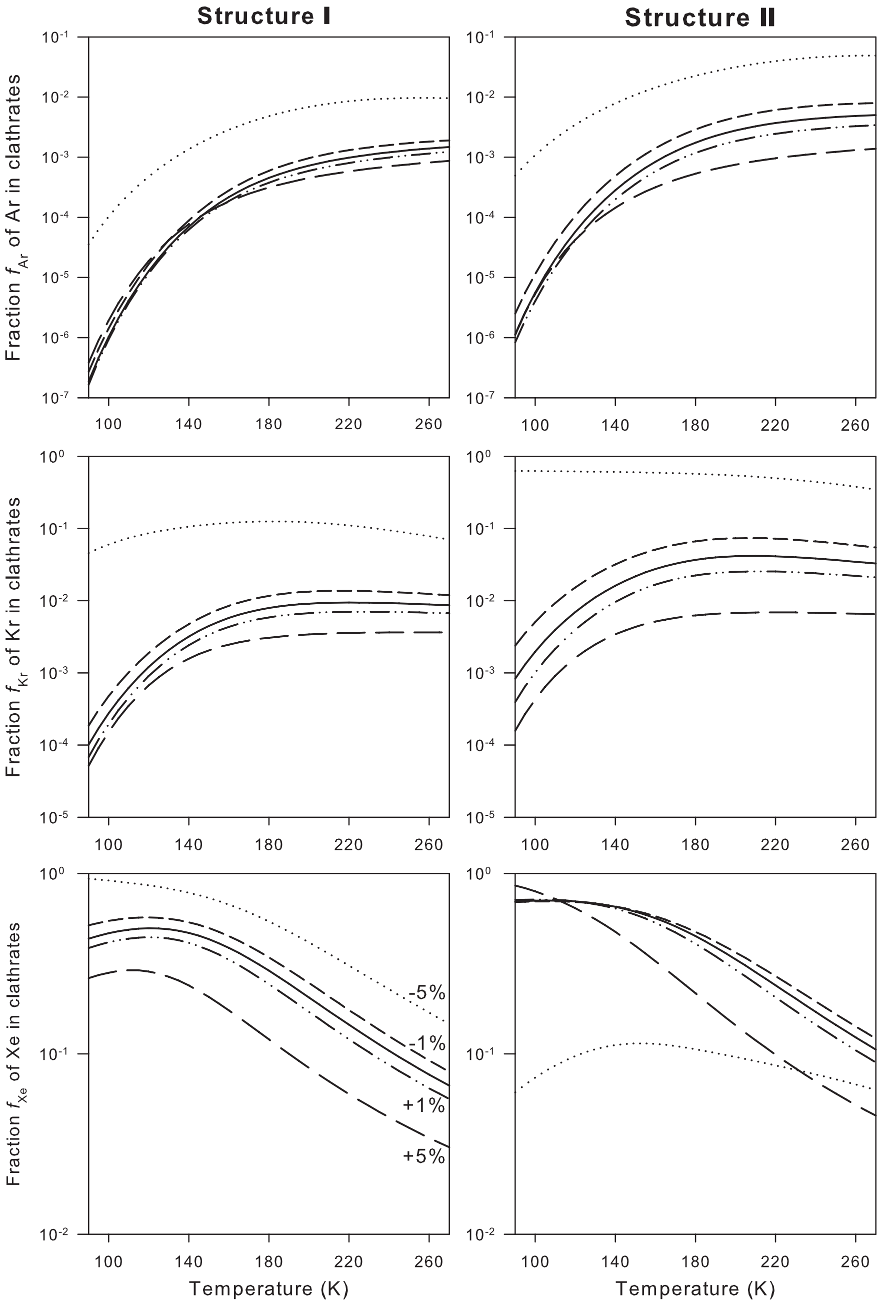


Fig. 2. Relative abundances of Ar, Kr, and Xe in clathrates as a function of temperature for structures I and II. The solid lines represent the results obtained with the parameters of the cages given in Table 1. The dash-dot-dotted and long dashed lines correspond to calculations performed with size of the cages increased by, respectively 1% and 5%. The medium dashed and dotted lines are results obtained with size of the cages decreased, respectively, by 1% and 5%.

atmosphere of Titan corresponds to the solar abundance (Lodders, 2003) with all carbon postulated to be in the form of methane (hereafter case 2). The third set of values is calculated under the assumption that each noble gas to methane gas phase ratio derives from the value calculated by Alibert and Mousis (2007) for planetesimals produced in the feeding zone of Saturn and ultimately accreted by the forming Titan (hereafter case 3). In each case, the relative abundance of N_2 (approximately 95%) has been determined such as $x_{CH_4} + x_{Ar} + x_{Kr} + x_{Xe} + x_{N_2} + x_{C_2H_6} = 1$. The initial gas phase abundances for the three cases are given in Table 3.

Fig. 4 shows that the dissociation curves calculated for the multiple guest clathrates that form in the three considered atmospheres exhibit a similar behavior, although for a given temperature, the dissociation pressure can vary by two orders of magnitude from one case to another one (especially at low temperatures). However, for

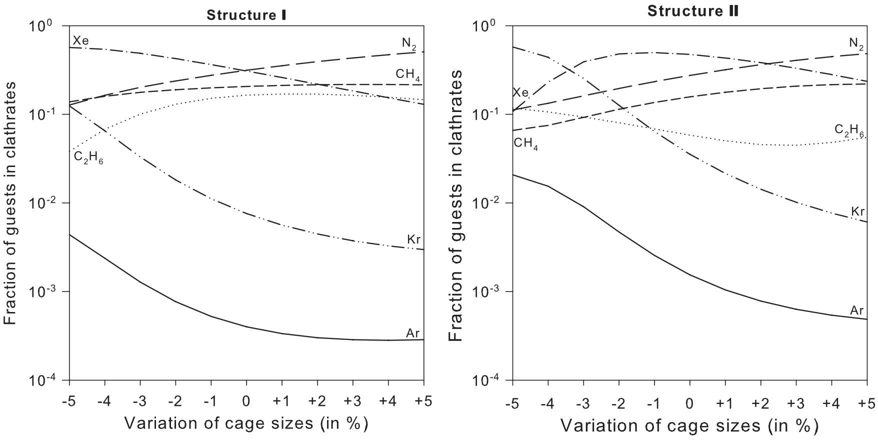


Fig. 3. Relative abundances of guests in clathrates as a function of the cage sizes. These results have been calculated at $P = 1.5$ bar, corresponding to a dissociation temperature $T = 176$ K.

Table 3
Three initial gas phase abundances corresponding to the systems considered by Osegovic and Max (2005) (case 1), the solar nebula (case 2), and Saturnian planetesimals (case 3)

Molecule	Molar fractions (%)		
	Case 1	Case 2	Case 3
Ar	0.1	7.1264×10^{-2}	2.10506
Kr	0.1	3.44×10^{-5}	1.38×10^{-3}
Xe	0.1	3.8×10^{-6}	1.6×10^{-4}
CH_4	4.92	4.92	4.92
N_2	94.68	95	92
C_2H_6	0.1	8.6978×10^{-3}	0.9734

a pressure corresponding to the present pressure at the surface of Titan (i.e., $P = 1.5$ bar), the dissociation temperatures given by Fig. 4 for the three cases are within a 20 K range, with corresponding values $T_{\text{mix}}^{\text{diss}} = 176, 167$, and 185 K, for cases 1, 2 and 3, respectively. These results indicate that the influence of the initial abundances in the gas phase is quite weak on the stability of the corresponding multiple guest clathrate.

Then, we have calculated the variations with temperature of the relative abundances f_G in the multiple guest clathrate considered in the present study, in each case.

Fig. 5 shows the corresponding results calculated for a multiple guest clathrate of structure I or II. For case 1, this figure shows that the relative abundances of Ar, Kr, CH_4 and N_2 decrease when the formation temperature of the clathrate decreases, in contrast with the relative abundances of Xe and C_2H_6 which slightly increase when the temperature decreases. This indicates that the efficiency of the trapping by multiple guest clathrates decreases with temperature for Ar and Kr, whereas it slightly increases for Xe. This result differs from that obtained in our previous study (Thomas et al., 2007) in which the trapping of both Xe and Kr was found to increase when the temperature decreases. However, in this previous study, we performed the calculations by assuming the presence of only one noble gas in the initial gas phase, the two others being excluded. The difference obtained in the present study for a gas phase containing the three noble gases, indicates that there is a strong competition between the trapping of Xe, Kr, and Ar, when considering that they can be trapped all together. Similar conclusions are obtained when considering the two other cases, although the trapping of Xe is found to increase much more than in case 1 when the temperature decreases. Also, in cases 2 and 3, the trapping of Kr appears almost constant in the whole range of temperatures considered in the present study. Note that the absolute values of the relative abundances are very different for the three cases due to the different compositions of the initial gas phase.

As a consequence, it is much more useful to compare the efficiency of the trapping mechanism in each case by calculating the abundance ratios for the three considered noble gases Ar, Kr and Xe. Such a ratio is defined as in our previous paper (Thomas et al., 2007), i.e., as the ratio between the relative abundance f_G of a given noble gas in the multiple guest clathrate and its initial gas phase abundance x_G (see Table 3). The ratios calculated for Xe, Kr, and Ar in the three cases considered here are given in Table 4. These ratios have been calculated at the particular point on the dissociation curves (Fig. 4) corresponding to the present atmospheric pressure at the ground level of Titan (i.e., $P = 1.5$ bar and $T = T_{\text{mix}}^{\text{diss}}$).

Table 4 shows that for this particular point of the dissociation curve, the relative abundances of Xe and Kr trapped in multiple guest clathrates are much higher than in the initial gas phase, irrespective of the initial gas phase composition. By contrast, the relative abundance of Ar is similar in gas phase and in the multiple guest clathrate. These results indicate that the efficiency of the trapping by clathrate may be high enough to significantly decrease the atmospheric concentrations of Xe and, to a lesser extent, of Kr, irrespective of the initial gas phase composition, provided that clathrates are abundant enough at the

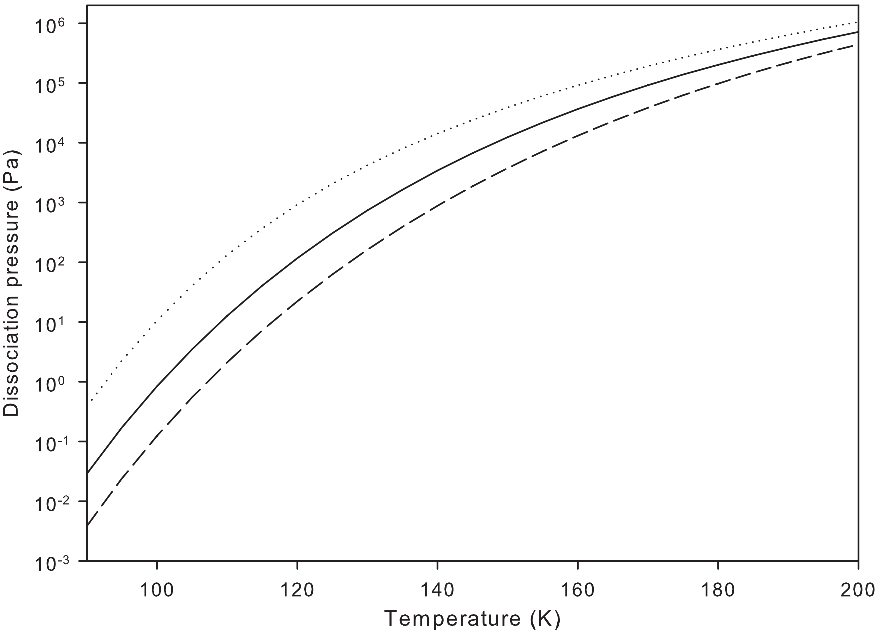


Fig. 4. Dissociation pressures of multiple guest clathrates, as a function of temperature for the three cases considered in the present study (see text): case 1 (solid line), case 2 (dotted line), and case 3 (dashed line).

surface of Titan. On the contrary, with an abundance ratio close to 1, Ar is poorly trapped in clathrates and the Ar atmospheric abundance consequently should remain almost constant.

5. Summary and discussion

We have extended the work of Thomas et al. (2007) by considering the clathration of a gas mixture presumably representing the composition of Titan's atmosphere, where Ar, Kr and Xe are included all together. In this context, we have developed a hybrid statistical model derived from the works of van der Waals and Platteeuw (1959) and Lipenkov and Istomin (2001), and using available experimental data to constrain the clathrates composition. Because it has been shown that clathrates equilibrium calculations are very sensitive to the guest species–cage interaction potential, we have compared different sets of potential parameters existing in the literature. Our calculations were performed using the parameters calculated by Parrish and Prausnitz (1972), because these parameters form a consistent set for our application to clathrates on Titan. We have also assessed the influence of the thermal variations of the size of the cages to better quantify the accuracy of the composition prediction. We show that these variations need to be taken into account if they are greater than a few percents. We have then considered several initial gas phase compositions, including different sets of noble gases abundances, that may be representative of Titan's early atmosphere. We finally show that the trapping efficiency of clathrates is high enough to significantly decrease the atmospheric concentrations of Xe and, in a lesser extent, of Kr, irrespective of the initial gas phase composition, provided that these clathrates are abundant enough at the surface of Titan. On the contrary, with an abundance ratio close to 1, Ar is poorly trapped in clathrates and its atmospheric abundance should remain consequently almost constant. Despite the fact that we consider simultaneously three noble gases in the gas phase composition, in contrast with Thomas et al. (2007), we obtain the same conclusions: the noble gases trapping mechanism via clathration can explain the deficiency in primordial Xe and Kr observed by *Huygens* in Titan's atmosphere, but not that in Ar.

We note that, even if the Visible and Infrared Mapping Spectrometer (VIMS) onboard *Cassini* was able to see the surface unimpeded, the bulk composition of Titan's crust is still unknown. Hence, the presence of clathrates on the surface of Titan is difficult to quantify.

Thomas et al. (2007) calculated that the total sink of Xe or Kr in clathrates would represent a layer at the surface of Titan whose equivalent thickness would not exceed ~50 cm. The sink of these noble gases in clathrates requires the presence of available crystalline water ice on the surface or in the near subsurface of Titan. If an open porosity exists within the top few hundreds meters in the icy mantle of Titan, by analogy with the terrestrial icy polar caps, the amount of available water ice in contact with the atmosphere of Titan would thus increase and help the formation of clathrates inside the pores. Moreover, in presence of methane clathrate on the surface of Titan, diffusive exchange of noble gases with methane might occur in the cavities, thus favoring their trapping in clathrates. One must also note that the efficiency of the noble gases

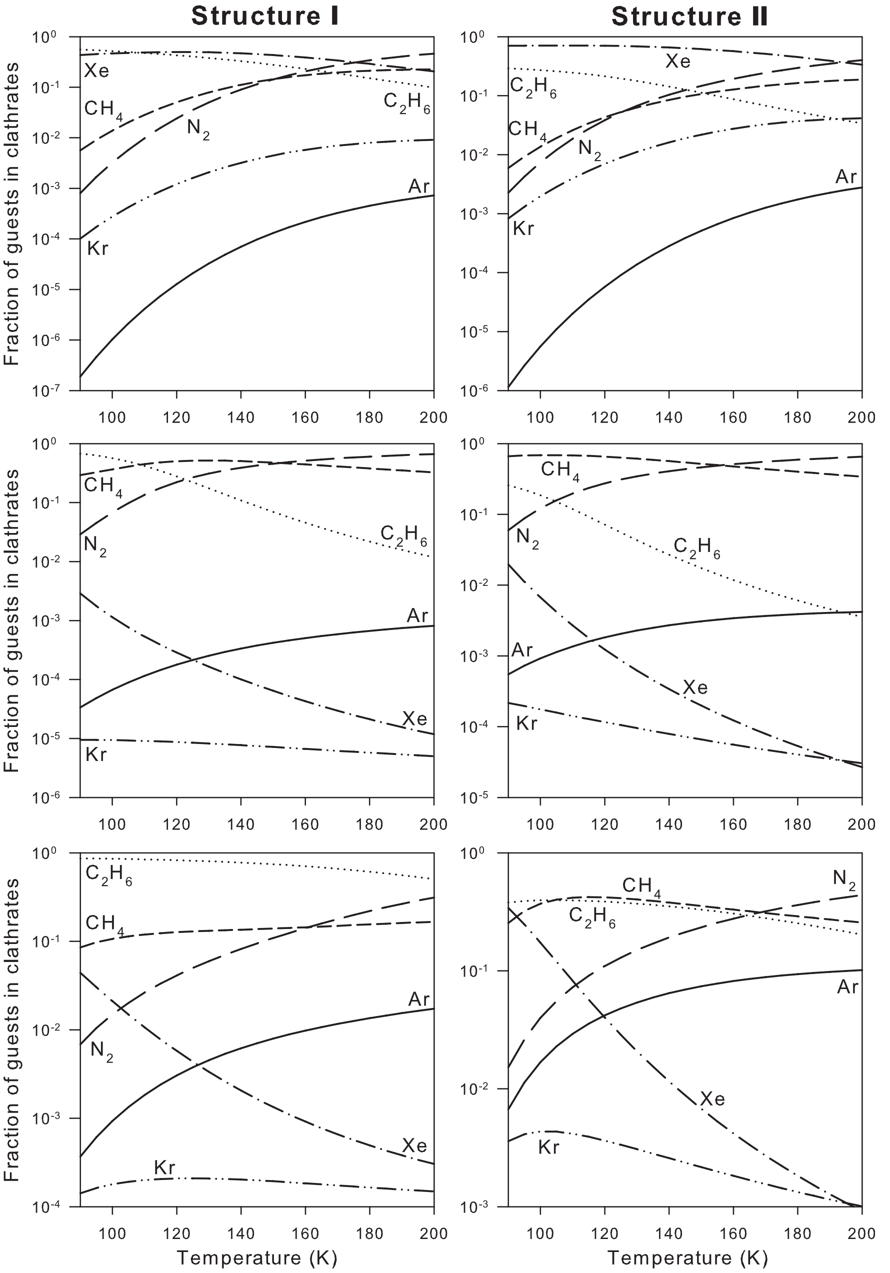


Fig. 5. Fractions of guests in clathrates, as a function of temperature for structures I and II; (top) case 1, (middle) case 2, (bottom) case 3.

trapping by clathrates on Titan can be limited by the very slow (and poorly known) kinetics at these low temperatures and the availability of water ice to clathration. It is then difficult to estimate the timescale needed to remove the proposed quantities of noble gases from the atmosphere of Titan.

To explain the deficiency in Ar in a way consistent with the present results, we can invoke the Titan's formation scenario proposed by Alibert and Mousis (2007) and Mousis et al. (2007). According to this scenario, the lack of CO in the atmosphere of Titan can be explained if Titan was formed from planetesimals that have been partially vaporized in the Saturn's subnebula. The vaporization temperature in the Saturn's subnebula (~50 K) needed to explain the loss of CO in planetesimals ultimately accreted by Titan is also high enough to imply the sublimation of Ar and, in a lower extent, that of Kr (see e.g. Fig. 9 of Alibert and Mousis, 2007). Indeed, Kr can also be partially trapped in methane clathrates formed in the solar nebula (Mousis et al., 2007). On the other hand, Xe still remains trapped in planetesimals because its incorporation occurs in conditions close to those required for the methane clathration in the nebula (Alibert and Mousis, 2007; Mousis et al., 2007).

Table 4
Abundance ratios of noble gas in clathrates to noble gas in the initial gas phase for Ar, Kr and Xe

Case	Initial molar fraction in gas	Abundance ratio structure I	Abundance ratio structure II
	Ar gas		
1	0.1×10^{-2}	0.4	1.5
2	7.1264×10^{-4}	0.8	5
3	2.10506×10^{-2}	0.7	4.6
	Kr gas		
1	0.1×10^{-2}	7.6	35.3
2	3.44×10^{-7}	18.3	143.7
3	1.38×10^{-5}	11.6	89.5
	Xe gas		
1	0.1×10^{-2}	308	473
2	3.8×10^{-8}	863.4	2356
3	1.6×10^{-6}	269.2	948.7

These ratios are calculated at $P = 1.5$ bar, and at the corresponding temperature on the dissociation curves (see Fig. 4). This temperature is equal to 176 K for case 1, 167 K for case 2, and 185 K for case 3.

It is important to mention that the composition of Titan's today atmosphere is almost certainly different from that in past. In particular, the nitrogen isotopes in Titan's atmosphere suggest significant mass loss over time. Moreover, we do not know the outgassing history of methane. In particular, if the methane outgassing is recent (Tobie et al., 2006), the lack of infrared opacity prior to that era must result in freezing out of nitrogen. It is likely that the composition, pressure and temperatures in Titan's atmosphere have differed significantly in the past, which will affect the composition and formation efficiency of clathrates on the surface.

Finally, we note that there remains the possibility that the noble gas abundances are telling a story that entirely differs from the scenario we propose, in which neither molecular nitrogen nor noble gases were initially accreted in clathrates (Atreya et al., 2006, 2007). In this context, Titan would have formed from solids produced at such high temperatures that they would have accreted nitrogen essentially as ammonia hydrate. Planetesimals formed in such conditions would be directly impoverished in noble gases since their trapping in clathrates require lower temperature and pressure conditions (Atreya et al., 2006, 2007).

Acknowledgements

O.M. acknowledges the support of the French "Centre National d'Etudes Spatiales" (CNES). Support from the PID program "Origines des Planètes et de la Vie" of the CNRS is also gratefully acknowledged. We also acknowledge Dr. Jonathan Horner for a careful reading of the manuscript.

References

Alibert, Y., Mousis, O., 2007. Formation of Titan in Saturn's subnebula: constraints from Huygens probe measurements. Astron. Astrophys. 465, 1051–1060.

Atreya, S.K., Adams, E.Y., Niemann, H.B., Demick-Montelara, J.E., Owen, T.C., Fulchignoni, M., Ferri, F., Wilson, E.H., 2006. Titan's methane cycle. Planet. Space Sci. 54, 1177–1187.

Atreya, S.K., Matson, D.L., Castillo-Rogez, J., Johnson, T.V., Adams, E.Y., Lunine, J.I., 2007. Photochemical origin of nitrogen on Titan and Enceladus. AGU Fall Meeting Abstracts 3.

Belosludov, V.R., Inerbaev, T.M., Subbotin, O.S., Belosludov, R.V., Kudoh, J., Kawazoe, Y., 2003. Thermal expansion and lattice distortion of clathrate hydrates of cubic structures I and II. J. Supramolecular Chem. 2 (4–5), 453–458.

Diaz Peña, M., Pando, C., Renuncio, J.A.R., 1982. Combination rules for intermolecular potential parameters. I. Rules based on approximations for the long-range dispersion energy. J. Chem. Phys. 76, 325–332.

Hester, K.C., Huo, Z., Ballard, A.L., Koh, C.A., Miller, K.T., Sloan, E.D., 2007. Thermal expansivity for sI and sII clathrate hydrates. J. Phys. Chem. B 111, 8830–8835.

Iro, N., Gautier, D., Hersant, F., Bockelée-Morvan, D., Lunine, J.I., 2003. An interpretation of the nitrogen deficiency in comets. Icarus 161, 511–532.

Jager, M., 2001. High pressure studies of hydrate phase inhibition using Raman spectroscopy. Ph.D. Thesis.

Lipenkov, V.Ya., Istomin, V.A., 2001. On the stability of air clathrate-hydrate crystals in subglacial Lake Vostok, Antarctica. Materialy Glyatsiol. Issled. 91, 129–133.

Lodders, K., 2003. Solar system abundances and condensation temperatures of the elements. Astrophys. J. 591, 1220–1247.

Lunine, J.I., Stevenson, D.J., 1985. Thermodynamics of clathrate hydrate at low and high pressures with application to the outer solar system. Astrophys. J. Suppl. Ser. 58, 493–531.

McKoy, V., Sinanoğlu, O., 1963. Theory of dissociation pressures of some gas hydrates. J. Chem. Phys. 38 (12), 2946–2956.

Miller, S.L., 1961. The occurrence of gas hydrates in the solar system. Proc. Natl. Acad. Sci. 47 (11), 1798–1808.

Mousis, O., Lunine, J.I., Thomas, C., Alibert, Y., 2007. Constraints on the origin of Titan from Huygens probe measurements. AAS/Division for Planetary Sciences Meeting Abstracts 39, #44.09.

Niemann, H.B., et al., 2005. The abundances of constituents of Titan's atmosphere from the GCMS instrument on the Huygens probe. Nature 438, 779–784.

Osegovic, J.P., Max, M.D., 2005. Compound clathrate hydrate on Titan's surface. J. Geophys. Res. (Planets) 110, 8004.

Owen, T., Bar-Nun, A., Kleinfeld, I., 1992. Possible cometary origin of heavy noble gases in the atmospheres of Venus, Earth, and Mars. Nature 358, 43–46.

Owen, T., Mahaffy, P., Niemann, H.B., Atreya, S., Donahue, T., Bar-Nun, A., de Pater, I., 1999. A low-temperature origin for the planetesimals that formed Jupiter. Nature 402, 269–270.

Papadimitriou, N.I., Tsimpanogiannis, I.N., Yiotis, A.G., Steriotis, T.A., Stubos, A.K., 2007. On the use of the Kihara potential for hydrate equilibrium calculations. In: Kuhs, W. (Ed.), Physics and Chemistry of Ice. Proceedings of the 11th International Conference on the Physics and Chemistry of Ice, vol. 311, pp. 475–482.

Parrish, W.R., Prausnitz, J.M., 1972. Dissociation pressures of gas hydrates formed by gas mixtures. Ind. Eng. Chem. Process Design Dev. 11 (1), 26–35 (Erratum: Parrish, W.R., Prausnitz, J.M., 1972. Industrial and Engineering Chemistry: Process Design and Development 11(3), 462).

Shpakov, V.P., Tse, J.S., Tulk, C.A., Kvamme, B., Belosludov, V.R., 1998. Elastic moduli calculation and instability in structure I methane clathrate hydrate. Chem. Phys. Lett. 282 (2), 107–114.

Sloan Jr., E.D., 1998. Clathrate Hydrates of Natural Gases. Marcel Dekker, New York.

Sparks, K.A., Tester, J.W., Cao, Z., Trout, B.L., 1999. Configurational properties of water clathrates: Monte Carlo and multidimensional integration versus the Lennard-Jones and Devonshire approximation. J. Phys. Chem. B 103 (30), 6300–6308.

Takeya, S., Kida, M., Minami, H., Sakagami, H., Hachikubo, A., Takahashi, N., et al., 2006. Structure and thermal expansion of natural gas clathrate hydrates. Chem. Eng. Sci. 61 (8), 2670–2674.

Thomas, C., Mousis, O., Ballenegger, V., Picaud, S., 2007. Clathrate hydrates as a sink of noble gases in Titan's atmosphere. Astron. Astrophys. 474, L17–L20.

Tobie, G., Lunine, J.I., Sotin, C., 2006. Episodic outgassing as the origin of atmospheric methane on Titan. Nature 440, 61–64.

van der Waals, J.H., Platteeuw, J.C., 1959. Clathrate solutions. In: Advances in Chemical Physics, vol. 2. Interscience, New York, pp. 1–57.

von Stackelberg, M., Müller, H.R., 1954. Feste Gashydrate II. Struktur und Raumchemie. Elektrochemie 58, 25–39.