EPJ manuscript No.
(will be inserted by the editor)

Cavity averages for hard spheres in the presence of polydispersity and incomplete data

Michael Schindler and A. C. Maggs

UMR Gulliver 7083 CNRS, ESPCI ParisTech, PSL Research University, 10 rue Vauquelin, 75005 Paris, France

September 3, 2015

Abstract We develop a cavity-based method which allows to extract thermodynamic properties from position information in hard-sphere/disk systems. So far, there are *available-volume* and *free-volume* methods. We add a third one, which we call *available-volume-after-takeout*, and which is shown to be mathematically equivalent to the others. In applications, where data sets are finite, all three methods show limitations, and they do this in different parameter ranges. We illustrate the principal equivalence and the limitations on data from molecular dynamics – In particular, we test robustness against missing data. We have in mind experimental limitations where there is a small polydispersity, say 4% in the particle radii, but individual radii cannot be determined. We observe that, depending on the used method, the errors in such a situation are easily 100% for the pressure and $10\,kT$ for the chemical potentials. Our work is meant as guideline to the experimentalist for choosing the right one of the three methods, in order to keep the outcome of experimental data analysis meaningful.

1 Introduction

Recent years have seen a growing number of experiments which use real-space measurements to extract quantitative information on thermodynamics and on dynamics of colloids [1,2,3,4,5]. Some of the investigated systems are crystalline [6,7,8,9,10,11], some are disordered and more or less dense [12,13,14,15,16,17]. Extracting a full thermodynamic description out of the available data is still a challenging task. Typical data consist of configuration snapshots, for example from video recording, or from confocal microscopy, which provide the positions of all particles at given times. Any thermodynamic description depends on the choice of the disregarded degrees of freedom. In particular, one typically disregards the degrees of freedom of a surrounding fluid and describes it in a more or less effective way. The challenge in this task thus consists in the determination of a set of thermodynamic variables, say pressure and chemical potentials, which is consistent, that is both quantities are calculated from the same data – for example from the particle centers alone, disregarding surrounding fluids.

A number of experiments aim at hard colloids [4,18,14], where the data analysis is the same as for hard-sphere simulations. There, simulations can serve for the development of good data treatment algorithms and for their calibration. On top of the conceptual question how to extract a consistent set of thermodynamic quantities, in real-world application, also the robustness of these algorithms against noise and missing data is important. Usual sources of noise in experiments are limited resolution of position and size of the particles [19], shape variations, and many more. In particular, there seems to be an unavoidable amount of size variations (polydispersity) of around 4% [18,14]. Are the used algorithms for the data analysis robust against this variation?

For hard spheres the determination of pressure and chemical potentials reduces to a geometrical problem. One family of algorithms to calculate them is based on measuring the space where a particle can be inserted into a given configuration. The general scheme is to take a set of configurations, calculate a certain geometrical quantity from each, and average it over the configurations. The different methods vary in the choice of the geometrical quantity and of the configurations. Widom [20] described the *available volume*, that is the volume V_0 where yet another particle can be inserted into a configuration of N particles. Clearly, the available volume may consist of several disconnected regions *(cavities)*. Figure 1a shows an example, where V_0 is the union of the green volumes. The notion of chemical potentials is directly linked to the insertion of another particle into a system. The link between V_0 and the pressure is less obvious. It has been established by Speedy [21] who expressed the pair-distribution function $g(r)$ in terms of the ratio of averages $\langle S_0\rangle/\langle V_0\rangle$, where the average is done over all possible configurations and S_0 is the area of the surface of V_0. This work made it principally possible to calculate the pressure from cavity averages. On a different line of reasoning, Hoover *et al.* [22] expressed the pressure on the basis of the region which a particle can explore when all others remain fixed. This cavity, which is called *free volume*, is another geometrical quantity which can be

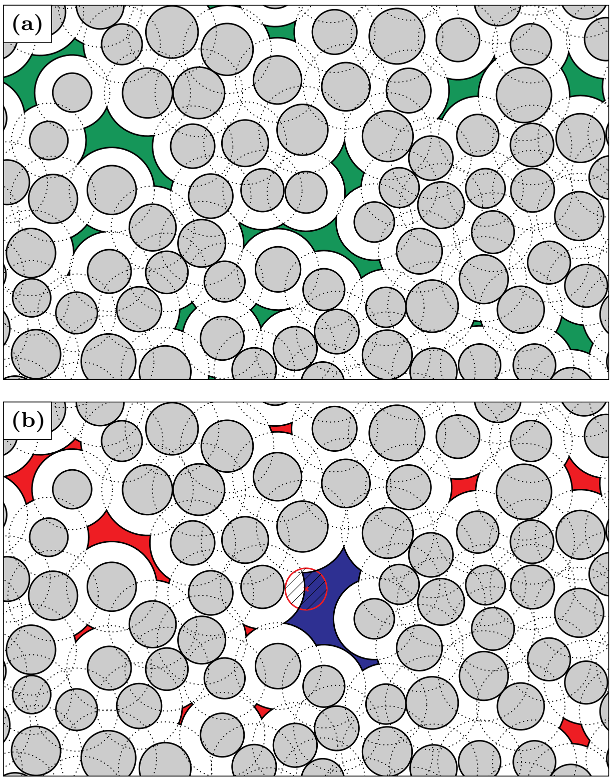

Figure 1. Some cavities in a given polydisperse configuration of hard disks. Dotted circles indicate the excluded volume for the center of the particle in question. **(a)** Available volume (in green) for an additional particle. **(b)** Free volume (blue) of the indicated particle (hatched). The new AVATO average uses the union of the free volume and the remaining cavities (red).

extracted from a given configuration.[1] An example is given in fig. 1b. Notice that if a different particle is chosen, the resulting free volume is different, and the free volumes of two particles may overlap. Clearly, the free volume has advantages over the available volume in dense systems, where adding yet another particle might be impossible. The free volume, however, never vanishes, because we first take out the particle in question, then determine the cavity where we can put it back in. Hoover *et al.* used a dynamical argument for their formula for the pressure. Speedy [25] found the same formula by showing that the above ratio of averages equals an average of a ratio, $\langle S_f/V_f \rangle$, where V_f is the free volume, S_f its boundary. Some years later, Speedy [26,27] found a way to re-derive the pressure result without recourse to the pair distribution function, by only counting configurations and applying elementary thermodynamics. This approach is more precise on how the averages are calculated. Interestingly, in order to get the pressure of N particles, one needs averages over ensembles of $(N-1)$ spheres. This difference is the key for the introduction of free-volume methods, where configurations of one particle less are regarded. Taking this difference seriously allows to provide algorithms which work seamlessly in all cases, being dilute, dense disordered, or crystalline.

If the particles have different sizes, the precise way how to average over cavities changes. Instead of a single available volume, we now have to calculate one for each radius of particle we want to insert. Finally, pressure and chemical potentials are weighted averages over these available volumes. How to calculate this average was shown by Corti & Bowles [28]. They generalised Speedy's work that uses pair-distribution functions to the polydisperse case. They did not use the cleaner configuration counting approach.

In the present paper we enrich the family of cavity-based algorithms by one member. To the methods based on *available-volume* (AV) and on *free-volume* (FV), we add a third one, which we call *available-volume-after-takeout* (AVATO). Its derivation below makes the AVATO method appear as the natural extension of the configuration counting idea, when passing from averages over $(N-1)$ particles to those over N particles – similarly to the free-volume method, but conceptually simpler. It is mathematically equivalent to the other two methods if averages over all configurations are available. As does the FV method, also the AVATO-average is possible in dense configurations. We provide formulae and algorithms for calculating both the pressure and the chemical potentials in all three methods, in the presence of polydispersity. For the two established methods, AV and FV, most of the equations have already been published and implemented [29,30,31,32] – but not for all the combinations of chosen method, pressure, chemical potential and polydispersity. For completeness we give all the equations here.

The second contribution of the present paper is to apply the derived algorithms to numerical data in two dimensions. We reproduce their principal equivalence and show how they start to differ on finite sets of data. We further test their robustness if we throw away information on individual particle radii at small values of polydispersity. The provided comparisons should guide experimentalists working with colloids, such that they can avoid the indicated problems and choose the best algorithm for analysing their data.

The outline of the paper is as follows. In sects. 2.1 and 2.2 we recall the established cavity methods and define some notation, restricting ourselves to the monodisperse case. The AVATO average is derived in sect. 2.3. The expressions for pressure and for chemical potentials in the fully polydisperse case follow in sects. 2.4 and 2.5. Numerical results of all three methods are presented for pressure (sect. 3.2) and for chemical potentials (sect. 3.3). The effect of missing radius information is shown in sect. 3.4.

[1] There is another branch of cavity-based method which goes back to Kirkwood [23] and which has been specialized to hard spheres by Wood [24]. This method also uses the term *free volume*, but it should not be confused with the free-volume method presented here. The essential difference is that in the Wood/Kirkwood method, the cavities are extracted from one averaged snapshot, where the averaging may have destroyed the non-overlapping constraint of hard spheres. Here, cavities are taken from many snapshots which are all compatible with the constraint, and the average is done subsequently.

2 Cavity methods

We now present the established cavity methods and derive our new method in the monodisperse case. We will also define the necessary ensembles and averages. Consider a collection of N spheres/disks of diameter σ in d dimensions, enclosed in a volume V at temperature T. In order to exclude surface effects, we take the volume to be periodic.

2.1 Available volume

Speedy [26] arrived at an expression for the pressure in the monodisperse case, using only counting arguments. His result is the ratio of two averages,

$$\frac{pV}{NkT} = 1 + \frac{\sigma}{2d}\frac{\langle S_0(k)\rangle_{k\in\Omega(N-1)}}{\langle V_0(k)\rangle_{k\in\Omega(N-1)}}. \tag{1}$$

Here, $\Omega(N-1)$ is the set of all possible ways to place $(N-1)$ spheres in the volume, and $\langle\cdot\rangle_{\Omega(N-1)}$ is the average over this set. In order to render the set finite, we may think of space being cut into small pixels of volume ω and observe that the final expression for the pressure does not depend on ω. We denote one such configuration (or "state") by the centers $\mathbf{r}_i$ of the spheres, $k \in \Omega(N-1) : k = \{\mathbf{r}_1, \dots \mathbf{r}_{N-1}\}$. Despite every particle having a unique number, they are not discerned within a state. The *available volume* $V_0(k)$ is then the volume where the center of an additional particle, $\mathbf{r}_N$ in the above case, can be placed. See fig. 1a for an example. The available volume may fall into several disjoint connected regions, called *cavities*. $S_0(k)$ is the boundary of the available volume.

The formula (1) is not directly applicable in data analysis, for two reasons: First, if we have data from an experiment or a simulation with N particles, we never have access to strictly all configurations. We rather take a finite series of K snapshots, $\mathcal{S}(N)$, and average over those. $\mathcal{S}(N)$ represents the data which is really available from an experiment. The number of snapshots, K, introduces a first level of approximation, which becomes exact in the limit $K \to \infty$. We will therefore denote this approximation as an equality. The second problem is that we have snapshots of N particles, not of $(N-1)$ particles. This problem gives rise to the *free-volume* and the *available-volume-after-takeout* methods below. Here, it simply introduces another level of approximation.

Applying eq. (1) to real-world data, we approximate the averages by those which we have, namely over the snapshots $\mathcal{S}(N)$, which contain all particle centers (and radii) at given times,

$$\frac{pV}{NkT} - 1 \approx \frac{\sigma}{2d}\frac{\langle S_0(k)\rangle_{k\in\mathcal{S}(N)}}{\langle V_0(k)\rangle_{k\in\mathcal{S}(N)}}. \tag{2}$$

Equation (2) can be translated into the following algorithm:

```
Loop over the snapshots $k \in \mathcal{S}(N)$:
    Enlarge all sphere diameters by $\sigma$.
    Find the space not covered by any sphere (cavities).
    $V_0(k) \leftarrow$ sum of all cavity volumes.
    $S_0(k) \leftarrow$ sum of all cavity boundaries.
    Accumulate $S_0(k)$ and $V_0(k)$ in independent averages.
End loop over $k$.
```

The enlargement step takes care of the excluded volume of both the present particles and the additionally inserted one, see fig. 1. The cavities are for the *center* of the inserted particle only. The algorithm to find the volumes and boundaries which are not covered by any sphere is discussed in sect. 3.1.

Effect of finite N and K

Generally speaking, in the limit $N \to \infty$ and $K \to \infty$ equation (2) and similar expressions in the following sections become exact. In practice both parameters are finite, and this introduces deviations in the approximations. The precise nature of the deviations is subtle, they depend on the number density N/V and on the quantity that is being averaged. We now try to capture some aspects of the deviations, without being exhaustive.

Let us focus on the individual effect of finite N first, assuming K to cover all possible configurations. This is as if we replaced $\mathcal{S}(N)$ by $\Omega(N)$ in eq. (2). Still, this equation aims at calculating the pressure of a different system than the original one, namely $p(N+1, V, T)$ instead of $p(N, V, T)$. There is a (small) error in the density of the order $1/N$ which translates into an error in the pressure. We can neglect it in the limit of many particles.

Now, if K is finite, everything depends on the concrete configurations contained in $\mathcal{S}(N)$. Above all we need the cavities to be sufficiently numerous to build reliable averages. If the particle density is sufficiently small, this is the case because we can nearly always insert another particle. The density deviation described above is thus the only systematic error and can be controlled by choosing N large enough. The finite number of snapshots introduces additional random noise. Surely, $K \to \infty$ will make this noise disappear, but this does not imply that the rate of convergence is sufficient for practical applications. This point will remain open in the present paper; we would only like to mention that there are examples in the literature [33] where the quantity of interest converges so slowly that advanced extrapolation methods are required.

The influence of finite K and N is more subtle for crystalline systems. It may happen that not a single snapshot in the given set $\mathcal{S}(N)$ allows insertion of another particle. In such a case we cannot calculate the average $\langle\cdot\rangle_{k\in\mathcal{S}(N)}$. (What we said above in the exact limit $K \to \infty$ remains valid, however, because we will find at least one extensible configuration in $\Omega(N)$ if N is large enough for the given density.) In order to see cavities in a monocrystal of N particles, we rely on *spontaneous* fluctuations to make sufficient space for particle insertion. Their rate of appearance is a function of the density N/V and of the total number N

and becomes exponentially small at high densities [27,34]. In practice, where K is limited to a few thousand, we are not astonished to see few or no cavities in the snapshots and to see badly converged averages.

For computer simulations one way out of this problem is to take $\mathcal{S}(N-1)$ from the beginning, that is to simulate a crystal with a vacancy [35]. This guarantees that the number of cavities is around one even in the densest system. In fact, this gives directly an approximation to eq. (1). We will not pursue this idea further in the present paper, because it is rather limited to computer simulations, and we here focus on what can be done with experimental data, taking simulations only as a test ground. Notice however that the number of cavities, $N_c(k)$, is a nice example for a function that does not give the same value when averaged over $\Omega(N)$ and over $\Omega(N-1)$.

2.2 Free volume

A well-established way for the above problem, passing from averages over $N-1$ particles to N particles, but without changing the system, is to introduce the so-called *free-volume* averages [25,27,28,30]. This method allows to measure pressures even in crystalline or nearly jammed systems.

From a configuration $k \in \Omega(N)$ we choose one particle at $\mathbf{r}_i$, take it out, and such construct the reduced state $k\backslash\mathbf{r}_i \in \Omega(N-1)$. The resulting available volume $V_0\big(k\backslash\mathbf{r}_i\big)$ is nonzero. One of its cavities contains the point $\mathbf{r}_i$. This cavity is called the *free volume*, denoted by $V_f\big(k\backslash\mathbf{r}_i\big)$. It is depicted in blue in fig. 1b. In terms of free volumes, eq. (1) reads

$$\frac{pV}{NkT} = 1 + \frac{\sigma}{2d}\left\langle \frac{S_f}{V_f} \right\rangle_{\mathcal{F}(N)}, \tag{3}$$

where the ratio of averages has turned into the average of a ratio. The precise definition of $\mathcal{F}(N)$ requires going into greater detail, which is done in several steps in the remainder of this section. The passage from (1) to (3) has been described several times in the literature [25,27,28,30]. We find it worth summarising these references in order to make the difference to our new method clear, which will be introduced in sect. 2.3.

In a first step, one constructs the set $\mathcal{C}(N-1)$ of all possible cavities which can be obtained from $\Omega(N-1)$. The ratio of averages has not changed at this point, and both are counting averages,

$$\frac{\big\langle S_0(k)\big\rangle_{k\in\Omega(N-1)}}{\big\langle V_0(k)\big\rangle_{k\in\Omega(N-1)}} = \frac{\big\langle S_c(k,l)\big\rangle_{(k,l)\in\mathcal{C}(N-1)}}{\big\langle V_c(k,l)\big\rangle_{(k,l)\in\mathcal{C}(N-1)}}. \tag{4}$$

Here, we labelled every cavity by both the configuration k in which it occurs and a number $l = 1, \ldots, N_c(k)$, where $N_c(k)$ is the number of cavities in this configuration. $V_c(k,l)$ denotes the volume of the cavity, and $S_c(k,l)$ its boundary.

The next step introduces a *probabilistic* element in the averages used in eqs. 4. Instead of averaging over all possible cavities, one chooses cavities at random with a probability proportional to their volume $V_c(k,l)$ [30],

$$P(k,l) := V_c(k,l) \Bigg/ \sum_{(k',l')\in\mathcal{C}(N-1)} V_c(k',l'). \tag{5}$$

The corresponding average of a quantity $f(k,l)$ is denoted by

$$\big\langle f\big\rangle_{\mathcal{P}(N-1)} := \sum_{(k,l)\in\mathcal{C}(N-1)} P(k,l)f(k,l). \tag{6}$$

The ratio of averages in eq. (1) now becomes an average of ratios,

$$\frac{\big\langle S_c(k,l)\big\rangle_{\mathcal{C}(N-1)}}{\big\langle V_c(k,l)\big\rangle_{\mathcal{C}(N-1)}} = \left\langle \frac{S_c(k,l)}{V_c(k,l)} \right\rangle_{\mathcal{P}(N-1)} \tag{7}$$

In a last step one passes from $(N-1)$ to N spheres, which defines the average over $\mathcal{F}(N)$ as the following procedure: One chooses uniformly a configuration $k \in \Omega(N)$ and then again uniformly one of the spheres $i \in \{1, \ldots N\}$. This latter choice can be repeated many times without changing the probability space, such that in the end it is equal to deterministically choosing every sphere once. In order to prove the equivalence of the averages over $\mathcal{F}(N)$ and over $\mathcal{P}(N-1)$, one has to show that the cavity probability $P(k',l)$ ($k' \in \Omega(N-1)$) leads indeed to a homogeneous distribution of configurations $k \in \Omega(N)$, and vice versa.

On the algorithmic level, what eq. (3) means in the analysis of a series of snapshots is the following:

```
Loop over the snapshots $k \in \mathcal{S}(N)$:
  Loop over all particles $i$:
    Take out particle $i$.
    Increase all other diameters by $\sigma_i$.
    Find the space not covered by any sphere (cavities).
    Identify the cavity which contains the center $\mathbf{r}_i$.
    $V_f\big(k\backslash\mathbf{r}_i\big) \leftarrow$ volume of this cavity.
    $S_f\big(k\backslash\mathbf{r}_i\big) \leftarrow$ boundary of this cavity.
    Accumulate $S_f/V_f$ over both loops.
  End loop over $i$.
End loop over $k$.
```

or as a formula,

$$\frac{pV}{NkT} - 1 = \frac{\sigma}{2d}\left\langle \frac{1}{N}\sum_{i=1}^{N} \frac{S_f\big(k\backslash\mathbf{r}_i\big)}{V_f\big(k\backslash\mathbf{r}_i\big)} \right\rangle_{k\in\mathcal{S}(N)} \tag{8}$$

We write eq. (8) as an identity, which is strictly valid only in the limit of an infinite number of snapshots.

2.3 Available volume after take-out

We will now develop a third, new method which turns out to be an alternative to the available-volume and free-volume averages.

Let us start with a naive algorithmic question: In the above algorithm, why not take the whole available volume of the reduced state instead of only one of its cavities? In the configuration of fig. 1b we would take the union of the blue and the red cavities. We call this union volume the *available volume after take-out* (AVATO), denoted by $V_0\bigl(k\backslash\mathbf{r}_i\bigr)$. An average over this volume inherits the advantage from the FV-average that even at high densities there is always at least one cavity to calculate. With the AV-average it shares the advantage that there are more than one cavity, possibly many, which contribute when the system is not dense. This improves the stability of the averages. In this sense the AVATO-average combines the best of two worlds.

Using this AVATO-average, the pressure is calculated by

$$\frac{pV}{NkT} = 1 + \frac{\sigma}{2d}\left\langle \frac{1}{N}\sum_{i=1}^{N}\frac{S_0\bigl(k\backslash\mathbf{r}_i\bigr)}{V_0\bigl(k\backslash\mathbf{r}_i\bigr)}\right\rangle_{k\in\Omega(N)} \tag{9}$$

The answer to the above question is affirmative: Equation (9) is entirely equivalent to eq. (1). Even better, this equivalence is based only on the counting of states, no additional probability is required. Let us start with all configurations $k' \in \Omega(N-1)$. If a state k' is extensible, it can be the result of taking one particle out of a configuration $k \in \Omega(N)$. In fact, there are exactly $V_0(k')/\omega$ different such states k which lead to k'. If a state k' is inextensible, we have $V_0(k') = 0$. In the reverse direction, starting with all configurations $k \in \Omega(N)$, for each of them there are exactly N ways to produce an extensible state k'. The such constructed mapping between states k and k' is far from being one-to-one, but the whole $\Omega(N)$ is mapped to the extensible states, a subset of $\Omega(N-1)$. This means that for any function $f : \Omega(N-1) \to \mathbb{R}$ we have the equality

$$\sum_{k\in\Omega(N)}\sum_{i=1}^{N} f\bigl(k\backslash\mathbf{r}_i\bigr) = \sum_{k'\in\Omega(N-1)}\frac{V_0(k')}{\omega}\,f(k'). \tag{10}$$

Every extensible state is counted an equal number on both sides of the equation. Notice that the weighting $V_0(k')/\omega$ automatically eliminates the inextensible states from the sum on the right-hand side. We may therefore sum over all states $\Omega(N-1)$. Notice further that when we specialize eq. (10) to the constant function $f(k') = 1$, we obtain Speedy's eq. (5) for the number of configurations [26]. The equivalence of eqs. (1) and (9) is now shown by writing out the averages as sums and then using eq. (10) twice, with $f(k') = S_0(k')/V_0(k')$ and with $f(k') = 1$.

When applied to a finite set of snapshots, only the type of average changes, as compared to eq. (9),

$$\frac{pV}{NkT} - 1 = \frac{\sigma}{2d}\left\langle \frac{1}{N}\sum_{i=1}^{N}\frac{S_0\bigl(k\backslash\mathbf{r}_i\bigr)}{V_0\bigl(k\backslash\mathbf{r}_i\bigr)}\right\rangle_{k\in\mathcal{S}(N)} \tag{11}$$

Again, this approximation becomes exact in the limit of infinite snapshots. The corresponding algorithm is nearly identical to the one for free volumes, only the averaged quantity is different,

```
Loop over the snapshots $k \in \mathcal{S}(N)$:
  Loop over all particles $i$:
    Take out particle $i$.
    Increase all other diameters by $\sigma_i$.
    Find the space not covered by any sphere (cavities).
    $V_0\bigl(k\backslash\mathbf{r}_i\bigr) \leftarrow$ sum of all cavity volumes.
    $S_0\bigl(k\backslash\mathbf{r}_i\bigr) \leftarrow$ sum of all cavity boundaries.
    Accumulate $S_0/V_0$ over both loops.
  End loop over $i$.
End loop over $k$.
```

2.4 Polydisperse case: pressure

For simplicity of the notation, we derive the expressions for a binary mixture only, the generalisation to multicomponent systems is evident from the structure of the formulae. The mixture contains N_A spheres of diameter σ_A and N_B spheres of a different diameter σ_B. In a configuration

$$k = \Bigl(\bigl\{\mathbf{r}_1^A, \dots \mathbf{r}_{N_A}^A\bigr\}, \bigl\{\mathbf{r}_1^B, \dots \mathbf{r}_{N_B}^B\bigr\}\Bigr) \tag{12}$$

all the A-particles can be interchanged among themselves without changing the state, and equally the B-particles. The set of all possible states is denoted by $\Omega(N_A, N_B)$. When we allow particles to have different sizes, also the available volume starts to depend on the (type of) particle. We will write V_0^α for the volume where we can insert the center of another particle *of type* α, and accordingly for $S_0^\alpha, V_f^\alpha, \dots$

In order to generalize eq. (1) to the polydisperse case, we can follow the same arguments as in Ref. [26]. This is a straightforward task, but it has not yet been published. The result is similar to Corti & Bowles' eq. (59), only that the $(N-1)$ averages show up explicitly,

$$\frac{pV}{N\,kT} - 1 = \frac{1}{2dN}\sum_{\alpha=A,B} N_\alpha\sigma_\alpha \frac{\bigl\langle S_0^\alpha\bigr\rangle_{\Omega(N_\alpha-1,N.)}}{\bigl\langle V_0^\alpha\bigr\rangle_{\Omega(N_\alpha-1,N.)}}. \tag{13}$$

The notation $\Omega(N_\alpha-1, N.)$ stands for $\Omega(N_A-1, N_B)$ or for $\Omega(N_A, N_B-1)$, respectively. The algorithm now includes a loop over the particle types,

```
Loop over the snapshots $k \in \mathcal{S}(N)$:
  Loop over the particle types $\alpha$:
    Enlarge all sphere diameters by $\sigma_\alpha$.
    Find the space not covered by any sphere (cavities).
    $V_0^\alpha(k) \leftarrow$ sum of all cavity volumes.
    $S_0^\alpha(k) \leftarrow$ sum of all cavity boundaries.
    Accumulate $S_0^\alpha(k)$ and $V_0^\alpha(k)$ in averages.
  End loop over $\alpha$.
End loop over $k$.
```

The algorithm evaluates the average in the available-volume approximation for the pressure:

$$\frac{pV}{N\,kT}-1\approx\frac{1}{2dN}\sum_{\alpha=A,B}N_\alpha\sigma_\alpha\frac{\big\langle S_0^\alpha(k)\big\rangle_{k\in\mathcal{S}(N_A,N_B)}}{\big\langle V_0^\alpha(k)\big\rangle_{k\in\mathcal{S}(N_A,N_B)}}\qquad(p/\text{AV})$$

From the available-volume average in eq. (13) we can follow the same arguments as in sect. 2.2 to obtain the free-volume average,

$$\frac{pV}{N\,kT}-1=\frac{1}{2dN}\sum_{\alpha=A,B}N_\alpha\sigma_\alpha\left\langle\frac{S_f^\alpha}{V_f^\alpha}\right\rangle_{\mathcal{F}^\alpha(N_A,N_B)}\qquad(14)$$

which generalises eq. (3) to the polydisperse case. This is Corti & Bowles' eq. (75). In practice, the free volumes are different geometrical objects for each i and each α, such that it is simpler to write one sum over all particles instead of separate sums over types and particles. We had this sum already in eq. (8), which now becomes

$$\frac{pV}{NkT}-1=\frac{1}{2dN}\sum_{i=1}^{N}\sigma_i\left\langle\frac{S_f^{\alpha_i}(k\backslash\mathbf{r}_i)}{V_f^{\alpha_i}(k\backslash\mathbf{r}_i)}\right\rangle_{k\in\mathcal{S}(N_A,N_B)}\qquad(p/\text{FV})$$

Here, σ_i and α_i are diameter and type of particle i, respectively.

Finally, the arguments of sect. (2.3) hold for every particle type individually. For each α we obtain an equation such as (10),

$$\sum_{k\in\Omega(N_A,N_B)}\sum_{i=1}^{N_\alpha}f^\alpha(k\backslash\mathbf{r}_i^\alpha)=\sum_{k'\in\Omega(N_\alpha-1,N.)}\frac{V_0^\alpha(k')}{\omega}\,f^\alpha(k'),\qquad(15)$$

where now also the averaged quantity depends on α. Using the AVATO average, the pressure evaluation from data then turns out to be

$$\frac{pV}{NkT}-1=\frac{1}{2dN}\sum_{i=1}^{N}\sigma_i\left\langle\frac{S_0^{\alpha_i}(k\backslash\mathbf{r}_i)}{V_0^{\alpha_i}(k\backslash\mathbf{r}_i)}\right\rangle_{k\in\mathcal{S}(N_A,N_B)}\qquad(p/\text{AVATO})$$

The generalisation of the analytical formula (9) to the polydisperse case is analogous, only that the average is done over $\Omega(N_A,N_B)$.

2.5 Polydisperse case: chemical potentials

In continuum thermodynamics, the chemical potentials μ_A,μ_B are said to be derivatives of the suitable thermodynamic potentials with respect to N_A,N_B. As these latter take discrete values, we have to choose either the upper or the lower difference. In agreement with Ref. [27] we choose the lower one, because it leads to averages over configurations of $(N{-}1)$ particles, consistent with those in eq. (1). In a *canonical* ensemble we have the lower difference of Helmholtz' free energies,

$$\begin{aligned}\mu_A(T,V,N_A,N_B):=\;&+F(T,N_A,N_B,V,\sigma_A,\sigma_B)\\&-F(T,N_A{-}1,N_B,V,\sigma_A,\sigma_B)\end{aligned}\qquad(16)$$

In the *microcanonical* ensemble we have differences of entropies,

$$\begin{aligned}\frac{\mu_A}{T}(E,V,N_A,N_B):=\;&-S(E,N_A,N_B,V,\sigma_A,\sigma_B)\\&+S(E,N_A{-}1,N_B,V,\sigma_A,\sigma_B).\end{aligned}\qquad(17)$$

The cavity methods analyse only the configuration integral and disregard the kinetic part of the phase space. We thus require the phase space integral to factorise into two parts – which is the case both for the canonical partition function (always), and in the considered hard-sphere case also for the microcanonical integral. If we denote as $\varphi_{\text{kin}},\varphi_{\text{conf}}$ the two factors of the phase space integral (synonymously for the canonical and microcanonical cases), the differences of eqs. (16) and (17) become (written for particle type A),

$$\begin{aligned}\frac{\mu_A}{kT}&=\ln\left(\frac{\varphi_{\text{kin}}(N_A{-}1,N_B)}{\varphi_{\text{kin}}(N_A,N_B)}\,\frac{\varphi_{\text{conf}}(N_A{-}1,N_B)}{\varphi_{\text{conf}}(N_A,N_B)}\right)\\&=\ln\left(\frac{\lambda_A^d}{\ell^d}\,\frac{|\Omega(N_A{-}1,N_B)|\ell^d}{|\Omega(N_A,N_B)|\omega}\right)\\&=\ln\frac{\lambda_A^d\,N_A}{\big\langle V_0^A\big\rangle_{\Omega(N_A-1,N_B)}}.\end{aligned}\qquad(18)$$

Here, ℓ denotes an arbitrary length scale. The constants λ_α are the thermal de-Broglie wavelengths, given in the canonical case by $h/\sqrt{2\pi\,kT\,m_\alpha}$, and in the microcanonical case by the indicated ratio of φ_{kin}. The λ_α only set the origin of the energy axis and play no role in the following. Equation (18) is Corti & Bowles' eq. (48), only that here the average over $(N{-}1)$ particles becomes explicit. In the last step we made use of their relation (30).

The appearance of the $\Omega(N{-}1)$-average in eq. (18) gives rise to free-volume and AVATO averages, similar to the treatment of the pressure in the above sections – with the important difference, however, that here we do not treat a relative property (a ratio such as S/V), but an absolute one. This will hinder us from eliminating the average $\langle\cdot\rangle_{\Omega(N-1)}$ entirely [2]. Concerning the average over $\Omega(N{-}1)$, we have the same problem as for the pressure, that we cannot calculate eq. (18) from data snapshots. The best we can do for the moment is to use the statistical averages we can compute, arriving at

$$\frac{\mu_\alpha}{kT}-\ln\lambda_\alpha^d\approx\ln\frac{N_\alpha}{\big\langle V_0^\alpha(k)\big\rangle_{k\in\mathcal{S}(N_A,N_B)}}.\qquad(\mu/\text{AV})$$

This approximation aims at evaluating $\mu(N{+}1)$ instead of $\mu(N)$, as we saw already for the pressure in eq. (2). Again, the approximation breaks down for systems so dense that one cannot insert another particle.

The free-volume equivalent of eq. (18) is

$$\frac{\mu_A}{kT}-\ln\lambda_A^d=\ln\frac{N_A\big\langle 1/V_f^A\big\rangle_{\mathcal{F}^A(N_A,N_B)}}{\big\langle N_c^A(k)\big\rangle_{k\in\Omega(N_A-1,N_B)}}\qquad(19)$$

[2] This *caveat* also leads to the problem mentioned by Sastry *et al.* [30], that *"the chemical potential cannot be determined from free-volume information alone"*, we need also the number of cavities.

This is Sastry's eq. (9) [30], generalised to the polydisperse case. Notice the average number of cavities in the denominator, which still uses the counting average. When we want to turn this equation into an algorithm, again, we have to replace $\langle N_c^A(k)\rangle_{k\in\Omega(N_A-1,N_B)}$ by something we can compute, for example $\langle N_c^A(k)\rangle_{k\in\mathcal{S}(N_A,N_B)}$, without knowing the error we make. Thus, in data analysis we will use

$$\frac{\mu_\alpha}{kT} - \ln\lambda_\alpha^d \approx \ln\frac{\sum_{i=1}^{N_\alpha}\left\langle \frac{1}{V_f^\alpha(k\backslash\mathbf{r}_i^\alpha)}\right\rangle_{k\in\mathcal{S}(N_A,N_B)}}{\langle N_c^\alpha\rangle_{\mathcal{S}(N_A,N_B)}} \qquad (\mu/\text{FV})$$

Upon increasing density, we will be limited by the denominator, just as in eq. (μ/AV). As soon as most of the states are not extensible anymore, the average number of cavities vanishes.

Finally, let us calculate the chemical potentials in terms of AVATO averages. Here, the difference between extensible and inextensible states is essential. We therefore introduce the notation $\Omega^+(N_A-1,N_B)$ for those states in $\Omega(N_A-1,N_B)$ which are extensible by a particle of type A. Equation (15) does not work with quantities such as $f(k') = 1/V_0^\alpha(k')$ which are infinite for inextensible states. They would formally annihilate the weighting $V_0^\alpha(k')$ which is necessary to discard inextensible states from the sum. Using the restricted summation repairs this problem,

$$\sum_{k\in\Omega(N_A,N_B)}\sum_{i=1}^{N_A}\frac{1}{V_0^A(k\backslash\mathbf{r}_i^A)} = \sum_{k'\in\Omega^+(N_A-1,N_B)} 1/\omega = \frac{|\Omega^+(N_A-1,N_B)|}{\omega}. \qquad (20)$$

The take-out average of $f^A = 1/V_0^A$ becomes

$$\left\langle\frac{1}{V_0^A}\right\rangle_{\mathcal{T}^A(N_A,N_B)} := \frac{1}{N_A}\sum_{i=1}^{N_A}\left\langle\frac{1}{V_0^A(k\backslash\mathbf{r}_i^A)}\right\rangle_{k\in\Omega(N_A,N_B)} = \frac{|\Omega^+(N_A-1,N_B)|}{\sum_{k'\in\Omega^+(N_A-1,N_B)} V_0^A(k')} = \frac{1}{\gamma_A}\,\frac{1}{\langle V_0^A\rangle_{\Omega(N_A-1,N_B)}}. \qquad (21)$$

In the last step we introduced the ratio of the total number of states and the number of extensible states,

$$\gamma_A := \frac{|\Omega(N_A-1,N_B)|}{|\Omega^+(N_A-1,N_B)|}. \qquad (22)$$

The chemical potentials are then written as

$$\frac{\mu_\alpha}{kT} - \ln\lambda_\alpha^d = \ln N_\alpha\gamma_\alpha\left\langle\frac{1}{V_0^\alpha}\right\rangle_{\mathcal{T}^\alpha(N_A,N_B)} \qquad (23)$$

The occurrence of the ratio γ_α in this equation is unfortunate. We will not be able to compute it from N-particle snapshots. In dilute systems, where we can insert another particle in practically all configurations, this factor is close to unity and does not interfere. In dense systems, however, we do not know in advance into how many configurations we can insert a particle, knowing only that the factor will be larger than one, diverging for the largest possible density. The following approximation therefore *underestimates* the true chemical potential:

$$\frac{\mu_\alpha}{kT} - \ln\lambda_\alpha^d \lessapprox \ln\sum_{i=1}^{N_\alpha}\left\langle\frac{1}{V_0^\alpha(k\backslash\mathbf{r}_i^\alpha)}\right\rangle_{k\in\mathcal{S}(N_A,N_B)} \qquad (\mu/\text{AVATO-A})$$

We may try to find different expressions, on the search to avoid the problem with the γ_α in eq. (23). For example, we may average N_c^A/V_0^A, which is a relative property, thus γ_A will not intervene. Its take-out average is

$$\left\langle\frac{N_c^A}{V_0^A}\right\rangle_{\mathcal{T}^A(N_A,N_B)} = \frac{\langle N_c^A\rangle_{\Omega(N_A-1,N_B)}}{\langle V_0^A\rangle_{\Omega(N_A-1,N_B)}} = \left\langle\frac{1}{V_f^A}\right\rangle_{\mathcal{F}^A(N_A,N_B)}, \qquad (24)$$

we find a quantity which we have seen already, namely the free-volume average of $1/V_f^A$. We can thus express eq. (19) in yet another way, namely

$$\frac{\mu_A}{kT} - \ln\lambda_A^d = \ln\frac{N_A\langle N_c^A/V_0^A\rangle_{\mathcal{T}^A(N_A,N_B)}}{\langle N_c^A\rangle_{\Omega(N_A-1,N_B)}} \qquad (25)$$

Of course, this expression suffers from the same limitations concerning the denominator as does eq. (19). It is nevertheless interesting to see how we can again replace an average of free volumes by an AVATO average. In data analysis, we will use the following approximation for the denominator in eq. (25),

$$\frac{\mu_\alpha}{kT} - \ln\lambda_\alpha^d \approx \ln N_\alpha\frac{\langle N_c^\alpha/V_0^\alpha\rangle_{\mathcal{T}^A(N_A,N_B)}}{\langle N_c^\alpha\rangle_{\mathcal{S}(N_A,N_B)}} = \ln\frac{\sum_{i=1}^{N_\alpha}\left\langle\frac{N_c^\alpha(k\backslash\mathbf{r}_i^\alpha)}{V_0^\alpha(k\backslash\mathbf{r}_i^\alpha)}\right\rangle_{k\in\mathcal{S}(N_A,N_B)}}{\langle N_c^\alpha\rangle_{\mathcal{S}(N_A,N_B)}}. \qquad (\mu/\text{AVATO-B})$$

3 Numerical method and results

We now proceed with the comparison of equations (p/AV), (p/FV), and (p/AVATO) for the pressure, and of the equations (μ/AV), (μ/FV), (μ/AVATO-A), (μ/AVATO-B) for the chemical potentials by applying them to data. Some of these equations are approximations to the real thermodynamic quantity, so we hope to get from their comparison some information about the quality of these approximations. In order to avoid at best other sources of error, we constrain ourselves to snapshots which come from simulations, where we are limited only by the numerical precision in the centers and radii of the particles.

3.1 Numerical methods

To generate snapshots, we used event-driven molecular dynamics [36] with elastic collisions, such that the statistical ensemble is the microcanonical one – without the need for a thermostat. In addition to energy conservation, also the total linear momentum is conserved during the simulation. All our simulations use periodic boundary conditions to avoid possible inaccuracies from cutting cavities at boundaries.

The snapshots contain $N = 2150$ circular disks in two dimensions ($d = 2$) which have all identical mass $m_i = 1$. The diameters of the disks are chosen randomly from a Normal distribution with mean 1 and a prescribed standard deviation, where outliers beyond three standard deviations were discarded. In different simulations we used different standard deviations, starting with 0% (monodisperse) and increasing in steps of 1.5% until 9%. Simulation units are fixed by the average particle diameter, the particle mass and the temperature ($kT = 1$). At the beginning of a simulation, we shrunk the particles drastically and arranged them on a hexagonal grid, with random velocities chosen according to a Normal distribution – subject to the two constraints of zero total momentum and prescribed total energy. In a first phase of the simulation run, the particles, being in a gas phase, were successively grown to their individually prescribed target diameter. This was done without creating additional collisions, such that the total energy and momentum remained unaffected. After all diameters were attained, the system was allowed to relax to its thermodynamic equilibrium in a second phase of the simulation run. The equilibrium was either a gas state, at surface fractions $\phi < 0.70$, or a crystalline state, at $\phi \gtrsim 0.72$, see footnote.[3] Polydispersity shifted the transition to higher values: For example, at 6% we found the crystalline phase at $\phi \gtrsim 0.78$, with disclination defects, and at 9% polydispersity we did not find a crystalline structure anymore. Finally, the simulation run was continued, and snapshots were recorded periodically in time. The period was chosen such that on average every particle collided five times between subsequent snapshots.

For the pressure, which is at the same time a thermodynamic and a mechanical quantity, we have a reference value at our disposition. During simulation runs, also the mechanical definition of pressure, that is the *volume-averaged linear momentum current* was recorded as a time average. It comprises a kinetic part and a virial part [37,38]. This pressure, which converges very quickly because it is calculated from all collisions, will serve as a reference in the comparison of the different cavity results. We verified that this numerical reference pressure coincides with the known virial expansion [39,40] and found deviations smaller than 0.1% up to $\phi = 0.35$.

For measuring the volume and boundary of the cavity in all three methods, we implemented the algorithm of Ref. [29], see also Refs. [30,31]. We adapted it to the periodic case, in two dimensions, for configurations where every particle can in principle have a different diameter. The general idea of the algorithm is to cut space into triangles, such that for each of them only a single disk is to be considered. The triangles in question are pieces of the Voronoi cells which come from a *radical-plane construction*, see fig. 2 for an example. The Voronoi diagram (also called power diagram) is the dual of a *weighted Delaunay* triangulation (also called *regular triangulation*), where the weights are the squares of the extended radii [41,42]. For the calculation of the regular triangulation, we relied on the robust and efficient C++ library CGAL [43,44]. Their algorithm scales as $N \log N$ in the number of particles (instead of N^2 as in Ref. [31]).

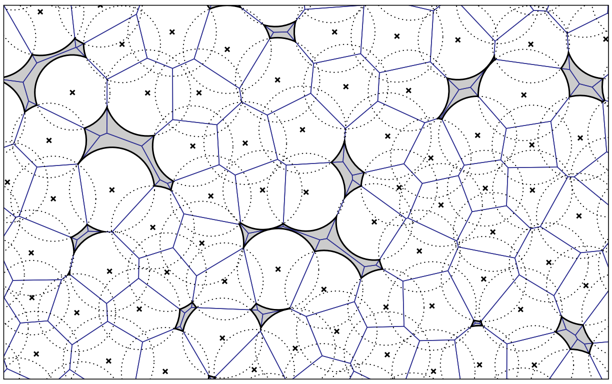

Figure 2. The Voronoi cells (blue) which served to calculate the cavities in fig. 1a. Dotted circles are the particles, extended by the radius of the particle for which the cavities are calculated.

The periodic version of the regular triangulation in CGAL is about to be developed, so that we had to implement this part ourselves, combining what CGAL offers for periodic and for regular nonperiodic triangulations [45,46]. For the unweighted Delaunay triangulation, it has been proved [47] that the periodic triangulation can be extracted from 3×3 periodic copies of the initial input. We found that this result also holds for weighted Delaunay triangulations.

The source code of the program to calculate the cavities will be made publicly available [48].

3.2 Precision of the methods, pressure

The result of the data analysis for the pressure, according to eqs. (p/AV), (p/FV), and (p/AVATO) are plotted as solid lines in fig. 3. We used 200 snapshots. Generally, all three cavity methods coincide with the reference pressure within the linewidth, only the AV average shows extreme errors in the crystalline phase. This behaviour is expected because snapshots which allow insertion of an (N+1)st particle are very rare in the dense phase, resulting in an average over a single cavity in the worst case.

In order to quantify the differences, the top-left insets of fig. 3 show relative errors with respect to the reference pressure. Generally speaking, the error is less than 1%, with largest errors in the dense gaseous phase, close to the

[3] We neither discuss the hexatic phase here, nor the precise location and nature of the phase transitions.

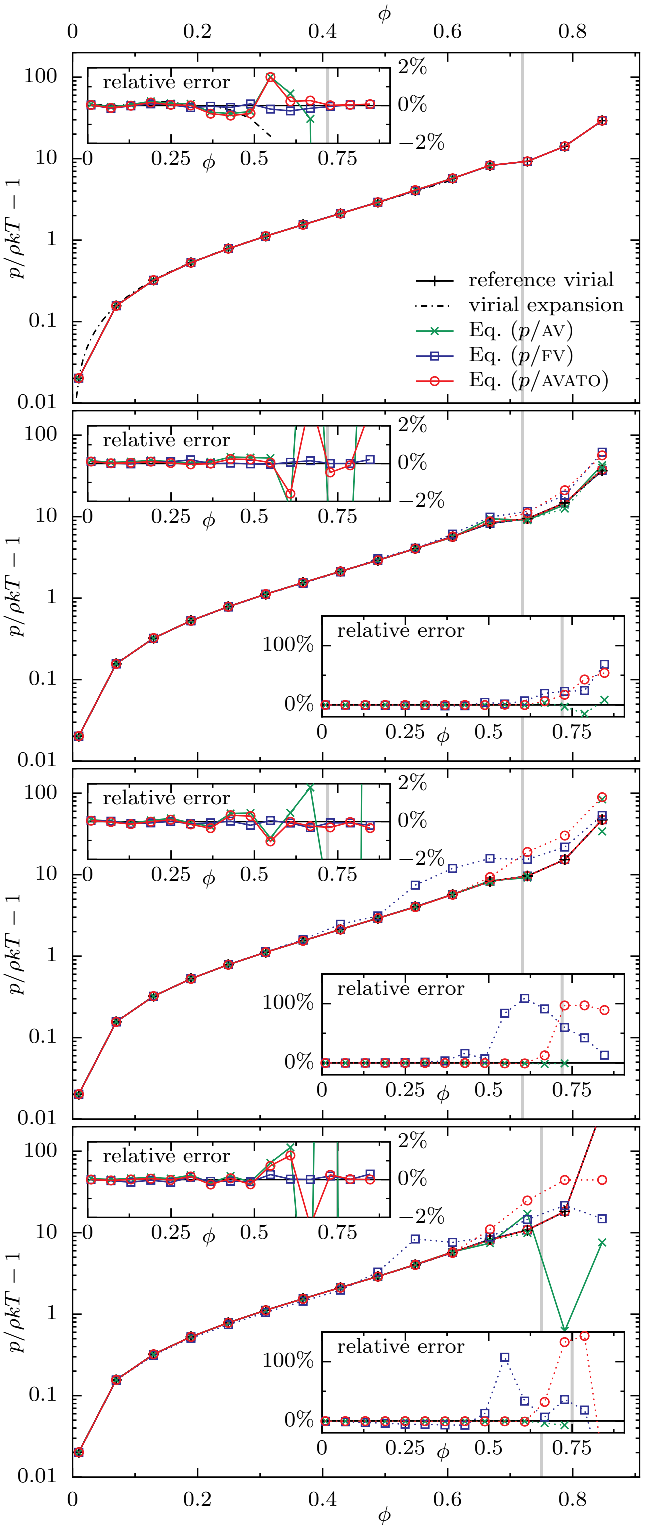

Figure 3. Pressure extracted from molecular-dynamics snapshots. From top to bottom, polydispersities are 0%, 1.5%, 3%, and 6%. **Solid lines:** the true radii are used; number of snapshots is $K = 200$. **Dotted lines:** disks are intentionally treated as if they were monodisperse; $K = 1700$. Vertical gray lines are guides to the eye to separate liquid from crystalline states (without formal definition).

transition to more ordered phases. There, the AVATO average shows slightly larger fluctuations than the FV average. Far in the crystalline phase the errors of the FV and the AVATO averages then drop again to very small values. The errors of the AV average grow strongly at around $\phi \gtrsim 0.65$, again because the number of extensible snapshots decreases strongly.

3.3 Chemical potentials

For the chemical potentials, which are not mechanical properties, we do not have reference data at our disposition, and we can only plot the cavity averages. Since a truly polydisperse system has as many particle classes as particles, instead of all the chemical potentials we plot the free enthalpy, also known as Gibbs free energy,

$$\frac{G}{kT} = \sum_{\alpha=1}^{M} N_\alpha \frac{\mu_\alpha}{kT}. \qquad (26)$$

Here, M denotes the number of classes; for 0% polydispersity we have $M = 1$, otherwise $M = N$. In order to make free enthalpies with different numbers of classes comparable, we had to add a term $\ln M$. The result of the data analysis according to eqs. (μ/AV), (μ/FV), $(\mu/\text{AVATO-A})$, and $(\mu/\text{AVATO-B})$ are plotted as solid lines in fig. 4. We used the same 200 snapshots as for the pressure.

In the figure, we see good agreement between the four averages, with an exception of the FV average which shows too low energies at moderately small surface fractions, $0.07 \lesssim \phi \lesssim 0.2$. We do not have a good explanation for this deviation at the moment, but we observe that it is related to the number of classes, thus stems from the denominator in eq. (μ/FV). The deviation disappears when using $M = 1$ in the monodisperse case in the top panel of fig. 4. It reappeared in a test where we forced $M = N$ on the same data. However, the same denominator is found also in the AVATO average in eq. $(\mu/\text{AVATO-B})$, where it does not cause a deviation. We tested also a bi-disperse system and found pronounced fluctuations in the FV average – to smaller and larger values – just at those values of ϕ where it exhibits the too low energies in fig. 4. We thus conclude that eq. (μ/FV) is extremely sensible in this ϕ-region and converges more slowly than elsewhere.

Furthermore, only one of the four methods, namely eq. $(\mu/\text{AVATO-A})$ gives values beyond the gaseous phase. The limitation for the three others comes of course again from their denominators, which are AV averages and thus give no data if the snapshots are inextensible. The fourth method, eq. $(\mu/\text{AVATO-A})$ is known to underestimate systematically the chemical potentials, so that one cannot rely on its results. We can check the consistency of pressure and chemical potentials using the Gibbs–Duhem relation $\sum_\alpha N_\alpha d\mu_\alpha = V dp - S dT$, where the last term vanishes because we worked at constant temperature. For the monodisperse case we rewrite the relation in two ways which avoid numerical differentiation,

$$\frac{dp}{d\mu} = \rho, \quad \text{and} \quad \frac{d(\rho\mu - p)}{d\rho} = \mu, \qquad (27)$$

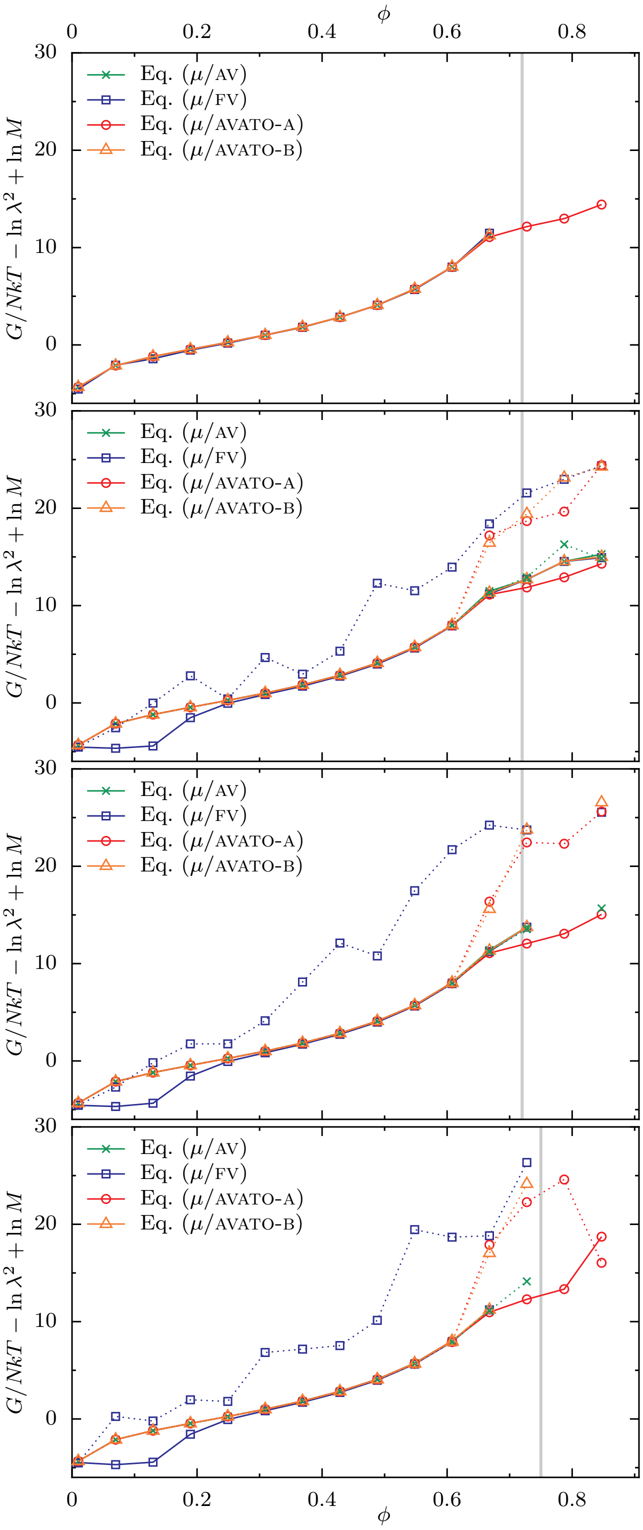

Figure 4. Free enthalpy per particle, extracted from molecular dynamics snapshots. From top to bottom, polydispersities are 0%, 1.5%, 3%, and 6%. **Solid lines:** the true radii are used; number of snapshots is $K = 200$. **Dotted lines:** disks are intentionally treated as if they were monodisperse; $K = 1700$.

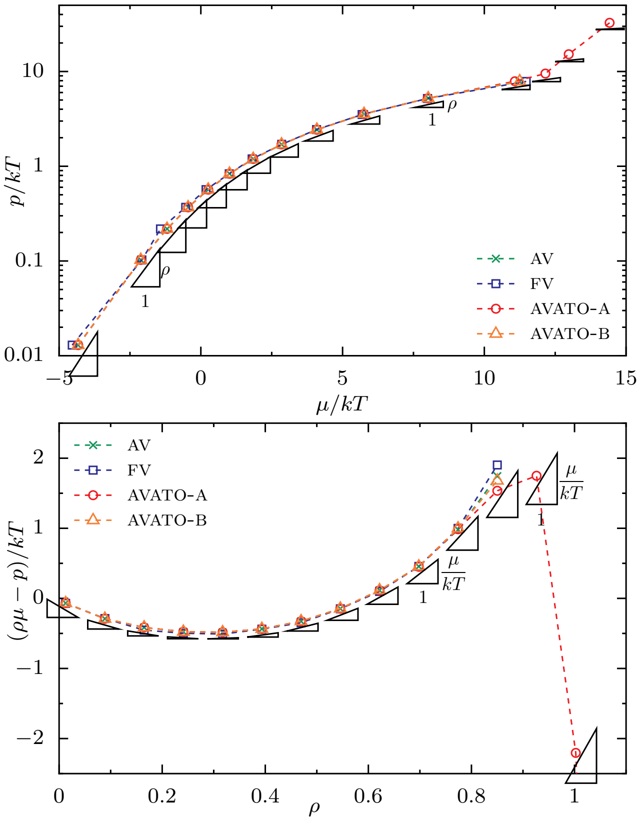

Figure 5. Test of the Gibbs–Duhem relation on monodisperse data. We used the shortcut μ/kT instead of $G/NkT - \ln\lambda^2 + \ln M$. The little triangles indicate the expected local derivatives according to eqs. (27).

where $\rho = N/V$ denotes the number density. Both variants of the equation are plotted in fig. 5. The little triangles indicate the expected derivative, that is the right-hand side of the above equations. By eye, these derivatives seem to coincide well with the general slope of the curves; only in the crystalline part slope and triangles do not agree at all. This was expected and proves that the values of eq. (μ/AVATO-A) beyond the gaseous phase are not physical.

3.4 Missing radius information

Above, we have shown that the three averages, AV, FV, and AVATO give similar answers when applied to numerical data. This is in fact not surprising, since their equivalence has been established analytically. The agreement of the above data is thus a test for the implementation and for the convergence.

We go now a step further and investigate the robustness of the three averages against imprecise snapshots. We will here consider only one source of imprecision, that is lack of the individual radii of the disks/spheres. This is a typical experimental situation, where one has a global idea of the radius standard deviation, but where the resolution is insufficient to reliably determine the radius of each particle [18]. The smallest experimental radius standard deviation is about 4% [14], so our question here is whether already this small polydispersity changes the reliability of results extracted with one or the other cavity method.

On our numerical data this task can be done quite easily. It suffices to throw away the information of individual radii. The algorithm which determines the cavities works as well on freely invented radii as on the true ones. One could for example attribute completely new radii according to a given distribution, or shuffle the old ones among all particles. Generally, we think that the details of radius attribution does not play a major role. We therefore follow a conceptually simpler route, that is we treat the polydisperse system as if it were monodisperse and attribute a unity diameter to every particle. The results for the pressure and for the free enthalpy are plotted as dotted lines in figs. 3 and 4.

For the pressure, we find extremely large deviations from the reference pressure, which grow with the surface fraction ϕ. The FV average is the first to deviate, it shows errors of 100% already far in the gas phase for $\phi \gtrsim 0.5$. One has to keep the polydispersity as tiny as 1.5% to keep the errors of the FV average below 20%. The AVATO average is more stable, it works up to $\phi \approx 0.7$. Remarkably both the FV and the AVATO averages are not robust against missing radius information in the crystalline phase, not even for 1.5% polydispersity – Remember that their errors were negligible when the radii were all correctly taken into account. The third method, the AV average works best up to the point where it fails also for the full radii information. Generally speaking, also the fluctuations in the curves is larger than in the full-radii case, although we took many more snapshots into account.

For the chemical potentials, fig. 4 shows a complete failure of the FV method, even at rather small densities, and even at tiny polydispersity. Errors are generally about 5–10 kT, and the curves fluctuate strongly. The AVATO average gives the correct result, until it deviates also at high densities, $\phi \gtrsim 0.66$. The AV average, again, gives the correct result in the ϕ range where it works.

In the search for an explanation why the FV method, and to some extend also the AVATO method, are not robust against missing particle radii, we remark that many configuration snapshots are *incompatible* with a modified radius attribution: Particles may overlap, as can be seen in the right-hand column of fig. 6. This has severe implications for the cavities, especially for the free volumes. Figures 6a',b' show cases where a cavity, originally the one which was the free volume of the particle, is now missed by the particle center. This cavity is not counted in the FV average, but it is counted in the AVATO average. The smaller a cavity, the more easily is it missed. However, especially the small cavities contribute strongly to the chemical potentials, where $1/V$ is averaged. In the pressure, which averages S/V, small cavities also contribute more strongly than large ones, but the scaling is weaker. For very small cavities, a second effect occurs, they may completely disappear, as shown in last row of fig. 6. In total, it is qualitatively understandable why we see more errors in dense than in dilute systems, and why the errors are larger in chemical potentials than in pressures, and why FV averages are more affected than the others.

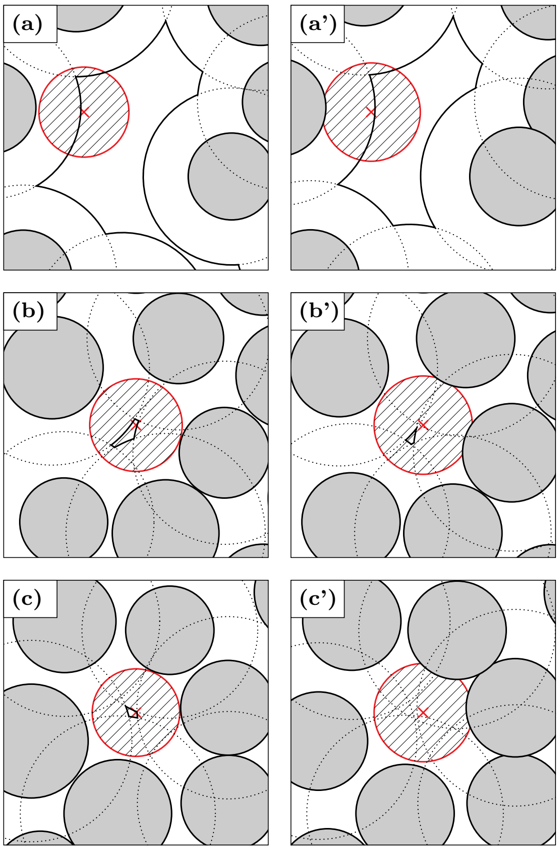

Figure 6. Examples of what happens to the free-volume cavity in case of missing radius information. **Left** with the true radii, **right** with artificially identical radii. **(a', b')** The particle center is not in "the" free-volume cavity. **(c')** The cavity has vanished completely.

We tested a number of variants of the original algorithms. Concerning missed free volumes, we relaxed the criterion a bit and counted as "free volume" also cavities with a distance up to 0.05 from the particle center. The cavities in figs. 6a',b' would have been counted as "free volumes". As another variant, we modified the normalisation of the averages. In the original algorithm, it reads $\langle f\rangle_{\mathcal{S}(N)} = \frac{1}{K}\sum_{k\in\mathcal{S}(N)} f(k)$. Instead of the total number of snapshots, K, we now divided by the number of snapshots which gave a non-vanishing f, thus discarding those where cavities may have been missed. Finally, we modified the number of particle "classes" (M) for the chemical potentials, grouping the particles with similar radii together. None of the above variants, nor combinations of them improved the dotted curves in figs. 3 and 4.

4 Conclusions

We investigated three geometrical methods to extract the pressure and the chemical potentials from hard-sphere positions. All three methods are cavity-based averages and differ in the precise way how the average is done and what quantity is averaged. The theoretical derivation of these methods all start with an average over all possibilities to place $(N-1)$ particles in a given volume. The FV and AVATO methods consistently turn this average into one over all possibilities to place N particles in the volume, whereas the AV average stays with $(N-1)$ particles. The differences in how the average is done and what quantity is averaged implies that the three methods differ in their applicability to real-world data sets, that is in their convergence rate, and in their response to noise.

As could be expected, there are hardly any problems in *dilute, gas-like* systems. Such systems easily explore their phase space, and it is easy to insert another particle. Consequently, there is negligible difference in averages over N and over $(N-1)$ particles, and these averages tend to converge well. We encountered, however, an exception to this general trend, namely the combination FV average/chemical potentials in fig. 4. With the same exception, we found that the methods are robust to variations in the particle radii in the dilute regime, say for $\phi \lesssim 0.5$. This conclusion is in agreement with a previous experimental/numerical study [49] done in three dimensions. In this reference, the AV average was calculated by counting pixels of the data snapshots, a procedure which introduced errors of at least 5–10% in the individual particle radii, the system being minimally polydisperse. Further, the precise position of the particles was limited by the Brownian diffusion time scale. It appears that in the dilute regime, the AV average – together with the additional data treatment described in the reference – was nevertheless able to give correct results for pressure and chemical potential.

Problems really start when systems become dense. At *very high densities*, say $\phi \gtrsim 0.75$, exploring the phase space takes a forbiddingly long time, which, together with the uncertainty whether an average over N and over $(N-1)$ particles are equivalent, renders the AV average unusable. Here, the FV and AVATO methods propose a valuable alternative – for calculating the pressure. Indeed, we found excellent agreement of these two methods with the reference pressure in fig. 3. The chemical potentials, however, are still out of reach, for the same reasons as for the AV average.

Between dilute and very dense, there is a *dense, but not extremely dense* regime, say $0.5 \lesssim \phi \lesssim 0.75$. Depending on polydispersity and initial preparation, such systems can be either (rather dilute) crystalline, or disordered glass-like, or disordered with a very long relaxation time for global order. These systems are in the main focus of many experimental studies on colloids. In particular, one would like to really compare the "free energies" of crystalline and disordered systems, and one thus requires both the pressure and the chemical potentials. The question is of course, *how far can we push the cavity-based methods? Are they suitable to provide these informations?* Despite the fact that this question is not directly the subject of the present paper, we think that we provide some elements to its possible answer. First, a close look on figs. 3 and 4 reveals that the AV method can indeed give results close to the crystalline phase, or even within. Its applicability thus has not a strict boundary but rather gradually decreases together with the uncertainty whether an average over N and over $(N-1)$ particles are equivalent. With the AVATO average, we here provide a second method which is at least conceptually consistent. Comparison of FV and AVATO results therefore reveals possible problems with noise. The most direct information of our study on the above question is the importance of high precision: An experimental method shall provide individual radii with an error below 1% in order to allow proper extraction of thermodynamic information on dense systems (see fig. 3) – a challenging, but maybe not unreachable task.

Surely, every experiment has different sources of noise which are more or less pertinent. We here provide an example for only one source of noise, or rather of missing information, which was motivated by an existing experimental situation [50]. Experimentalists who are interested in calibrating their data analysis are invited to use our cavity algorithm, which will be made available as open source [48]. For those who do not want to undertake heavy calibration, we propose the following rules of thumb: (i) Avoid the FV average. It is generally less robust against wrong radius information than the other two methods. It has problems with chemical potentials at moderately low densities. (ii) If the density is below $\phi \lesssim 0.5$, AVATO averages and AV averages give the same result. AVATO averages are conceptually cleaner, but depending on the number of classes, AV averages might be less expensive to calculate. (iii) If your system is dense disordered or crystalline, use more than one method and compare. (iv) If you want to implement only one method, stay with AVATO.

As a final remark, we would like to say that the algorithm is not restricted to thermal systems. The physical interpretation, however, starts from the assumption that all configurations are realised with the same probability. It could therefore be interesting to apply the algorithms also to systems which are not at equilibrium and where equivalence of all ensembles is not guaranteed [51].

We thank Daniel Bonn and Rojman Zargar for interesting us for the free-volume methods.

References

1. V.J. Anderson, H.N.W. Lekkerkerker, Nature **416**, 811 (2002)
2. A. van Blaaderen, P. Wiltzius, Science **270**, 1177 (1995)
3. W.K. Kegel, A. van Blaaderen, Science **14**, 290 (2000)
4. C.P. Royall, W.C.K. Poon, E.R. Weeks, Soft Matter **9**, 17 (2013)
5. P.J. Yunker, K. Chen, M.D. Gratale, M.A. Lohr, T. Still, A.G. Yodh, Rep. Prog. Phys. **77**, 056601 (2014)
6. K. Zahn, A. Wille, G. Maret, S. Sengupta, P. Nielaba, Phys. Rev. Lett. **90**, 155506 (2003)

7. T. Still, C.P. Goodrich, K. Chen, P.J. Yunker, S. Schoenholz, A.J. Liu, A.G. Yodh, Phys. Rev. E **89**, 012301 (2014)
8. B. Liu, T.H. Besseling, M. Hermes, A.F. Demirörs, A. Imhof, A. van Blaaderen, Nature Comm. **5**, 3092 (2014)
9. J. Taffs, S.R. Williams, H. Tanaka, C.P. Royall, Soft Matter **9**, 297 (2013)
10. A. Ghosh, R. Mari, V.K. Chikkadi, P. Schall, A.C. Maggs, D. Bonn, Physica A **390**, 3061 (2011)
11. J. Baumgartl, R.P.A. Dullens, M. Dijkstra, R. Roth, C. Bechinger, Phys. Rev. Lett. **98**, 198303 (2007)
12. R. Dreyfus, Y. Xu, T. Still, L.A. Hough, A.G. Yodh, S. Torquato, Phys. Rev. E **91**, 012302 (2015)
13. R. Zargar, J. Russo, P. Schall, H. Tanaka, D. Bonn, Europhys. Lett. **108**, 38002 (2014)
14. R. Zargar, B. Nienhuis, P. Schall, D. Bonn, Phys. Rev. Lett. **110**, 258301 (2013)
15. C.L. Klix, F. Ebert, F. Weysser, M. Fuchs, G. Maret, P. Keim, Phys. Rev. Lett. **109**, 178301 (2012)
16. V. Chikkadi, P. Schall, Phys. Rev. E **85**, 031402 (2012)
17. A. Ghosh, V. Chikkadi, P. Schall, D. Bonn, Phys. Rev. Lett. **107**, 188303 (2011)
18. H.L. Schöpe, O. Marnette, W. van Megen, G. Bryant, Langmuir **23**, 11534 (2007)
19. W.C.K. Poon, E.R. Weeks, C.P. Royall, Soft Matter **8**, 21 (2012)
20. B. Widom, J. Chem. Phys. **39**, 2808 (1963)
21. R.J. Speedy, J. Chem. Soc., Faraday Trans. 2 **76**, 693 (1980)
22. W.G. Hoover, W.T. Ashurst, R. Grover, J. Chem. Phys. **57**, 1259 (1972)
23. J.G. Kirkwood, J. Chem. Phys. **18**, 380 (1950)
24. W.W. Wood, J. Chem. Phys. **20**, 1334 (1952)
25. R.J. Speedy, J. Chem. Soc., Faraday Trans. 2 **77**, 329 (1981)
26. R.J. Speedy, J. Phys. Chem. **92**, 2016 (1988)
27. R.J. Speedy, H. Reiss, Molecular Physics **72**, 999 (1991)
28. D.S. Corti, R.K. Bowles, Molecular Physics **96**, 1623 (1999)
29. S. Sastry, D.S. Corti, P.G. Debenedetti, F.H. Stillinger, Phys. Rev. E **56**, 5524 (1997)
30. S. Sastry, T.M. Truskett, P.G. Debenedetti, S. Torquato, F.H. Stillinger, Molecular Physics **95**, 289 (1998)
31. M. Maiti, A. Lakshminarayanan, S. Sastry, Eur. Phys. J. E **36**, 5 (2013)
32. M. Maiti, S. Sastry, J. Chem. Phys. **141**, 044510 (2014)
33. K. Chen, T. Still, S. Schoenholz, K.B. Aptowicz, M. Schindler, A.C. Maggs, A.J. Liu, A.G. Yodh, Phys. Rev. E **88**, 022315 (7) (2013)
34. R.K. Bowles, R.J. Speedy, Molecular Physics **83**, 113 (1994)
35. R.J. Speedy, H. Reiss, Molecular Physics **72**, 1015 (1991)
36. D.C. Rapaport, *The Art of Molecular Dynamics Simulation*, 2nd edn. (Cambridge University Press, Cambridge, UK, 2004)
37. D. Forster, *Hydrodynamic fluctuations, broken symmetry, and correlation functions* (Addison-Wesley, 1990)
38. M. Schindler, Chem. Phys. **375**, 327 (2010)
39. *SklogWiki*, `http://www.sklogwiki.org/SklogWiki/index.php/Hard_sphere:_virial_coefficients` (23. July 2015)
40. N. Clisby, B.M. McCoy, J. Stat. Phys. **122**, 15 (2006)
41. B.J. Gellatly, J.L. Finney, Journal of Non-Crystalline Solids **50**, 313 (1982)
42. M. Caroli, P.M.M. de Castro, S. Loriot, O. Rouiller, M. Teillaud, C. Wormser, *Robust and Efficient Delaunay Triangulations of Points on or Close to a Sphere*, in *9th International Symposium on Experimental Algorithms* (2010), Vol. 6049 of *Lecture Notes in Computer Science*, pp. 462–473
43. *CGAL 4.6, Computational Geometry Algorithms Library*, `http://www.cgal.org` (2015)
44. M. Caroli, M. Teillaud, *Computing 3D Periodic Triangulations*, in *Proceedings 17th European Symposium on Algorithms* (2009), Vol. 5757 of *Lecture Notes in Computer Science*, pp. 59–70
45. M. Yvinec, in *CGAL User and Reference Manual* (CGAL Editorial Board, 2015), 4.6 edn., `http://doc.cgal.org/4.6/Manual/packages.html#PkgTriangulation2Summary`
46. N. Kruithof, in *CGAL User and Reference Manual* (CGAL Editorial Board, 2015), 4.6 edn., `http://doc.cgal.org/4.6/Manual/packages.html#PkgPeriodic2Triangulation2Summary`
47. N.P. Dolbilin, D.H. Huson, Periodica Mathematica Hungarica **34**, 57 (1997)
48. `http://www.pct.espci.fr/~michael/`
49. R.P.A. Dullens, W.K. Kegel, D.G.A.L. Aarts, Oil & Gas Science and Technology – Rev. IFP **3**, 295 (2008)
50. R. Zargar, D. Bonn, private communication
51. K. Chen, G. Ellenbroek, Z. Zhang, D. Chen, P. Yunker, S. Henkes, C. Brito, O. Dauchot, W. Sarloos, A. Liu et al., Phys. Rev. Lett. **105**, 025501 (2010)