PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2022, **24**, 9921

Binding reactions at finite systems†

Ronen Zangi [ab]

A perpetual yearn exists among computational scientists to scale down the size of physical systems, a desire shared as well with experimentalists able to track single molecules. A question then arises whether averages observed at small systems are the same as those observed at large or macroscopic systems. Utilizing statistical-mechanics formulations in ensembles in which the total numbers of particles are fixed, we demonstrate that properties of binding reactions are not homogeneous functions. This means that averages of intensive parameters, such as the concentration of the bound-state, at finite systems are different than those at large systems. The discrepancy increases with decreasing temperature, volume, and to some extent, numbers of particles. As perplexing as it may sound, despite variations in average quantities, extracting the equilibrium constant from systems of different sizes does yield the same value. The reason is that correlations in reactants' concentrations ought to be accounted for in the expression of the equilibrium constant, being negligible at large-scale but significant at small-scale. Similar arguments pertain to the calculations of the reaction rate constants, more specifically, the bimolecular rate of the forward reaction is related to the average of the product (and not to the product of the averages) of the reactants' concentrations. Furthermore, we derive relations aiming to predict the composition only from the equilibrium constant and the system's size. All predictions are validated by Monte-Carlo and molecular dynamics simulations. An important consequence of these findings is that the expression of the equilibrium constant at finite systems is not dictated solely by the chemical equation of the reaction but requires knowledge of the elementary processes involved.

Received 31st December 2021,
Accepted 18th March 2022

DOI: 10.1039/d1cp05984j

rsc.li/pccp

Consider the following association reaction,

$$A + B \rightleftharpoons AB, \tag{1}$$

in which N_A° gas molecules of A and N_B° gas molecules of B are placed in an empty closed container with fixed volume, V, and temperature, T, to reach equilibrium with the bound product, AB. The experiment is then repeated under the same conditions but with λN_A°, λN_B°, and λV instead. Are the concentrations of AB particles in the two experiments equal? In the thermodynamic limit, the answer is yes because intensive and extensive properties are homogeneous zero-order and first-order functions, respectively,[1]

$$X\left(T, \lambda V, \lambda N_A^\circ, \lambda N_B^\circ\right) = \lambda^\alpha X\left(T, V, N_A^\circ, N_B^\circ\right), \tag{2}$$

where $\alpha = 0$ if X is an intensive property and $\alpha = 1$ if X is extensive. However, would eqn (2) hold if we scale down the system to a regime not belonging to the thermodynamic limit (hereafter, referred to as small or finite system), for example that of $N_A^\circ = N_B^\circ = 1$? Currently accepted dogma assumes the validity of eqn (2) for all system sizes,[2–12] provided sufficient statistical data are collected (yet, it is understood that relative magnitudes of fluctuations are inversely proportional to the system size). In this paper we argue that for bimolecular reactions, the homogeneous function character of the system's properties stated in eqn (2) breaks down at finite systems.

Results

I. Statistical mechanical derivation of the equilibrium constant for association

The process in eqn (1) is chosen to be described by the canonical $\left(N_A^\circ, N_B^\circ, V, T\right)$ ensemble, where $N_A^\circ = N_A + N_{AB}$ and $N_B^\circ = N_B + N_{AB}$ are the total numbers of A and B particles. The particle labels are arranged to satisfy $N_A^\circ \leq N_B^\circ$. All three components on both sides of eqn (1) are assumed to be gases with ideal behavior, that means, apart from the reaction described they are not interacting with one another (this also excludes interactions between three or more particles). Upon the formation of one bound AB particle, the potential energy of the system changes by an amount of ε_{AB} (*i.e.*, ε_{AB} is negative).

[a] *POLYMAT & Department of Organic Chemistry I, University of the Basque Country UPV/EHU, Avenida de Tolosa 72, 20018, Donostia-San Sebastián, Spain. E-mail: r.zangi@ikerbasque.org*

[b] *IKERBASQUE, Basque Foundation for Science, Plaza Euskadi 5, 48009 Bilbao, Spain*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1cp05984j

In this model, the energy states of the system due to interparticle interactions can be uniquely mapped onto N_{AB}. Thus, the canonical partition function of the system can be written as,

$$Q = \sum_{i=0}^{N_A^\circ} \frac{q_A^{N_A^\circ - i}}{(N_A^\circ - i)!} \cdot \frac{q_B^{N_B^\circ - i}}{(N_B^\circ - i)!} \cdot \frac{q_{AB}^i}{i!} = \sum_{i=0}^{N_A^\circ} W_{N_A^\circ, N_B^\circ}^i \, q_A^{N_A^\circ - i} \, q_B^{N_B^\circ - i} \, q_{AB}^i, \quad (3)$$

where the summation over index i ($i \equiv N_{AB}$) includes all possible numbers of bound AB particles and thereby all possible energy states. q_A and q_B are single-particle partition functions of unbound A and B particles (hence, obtained by summing only internal energies) and q_{AB} is the paired-particle partition function of one bound AB particle (which includes the Boltzmann factor $e^{-\beta\varepsilon_{AB}}$). These partition functions can be expressed in different forms and they are described in detail in Section SI-1 of the ESI.† Given that the A particles are indistinguishable, and likewise the B particles, $W_{N_A^\circ, N_B^\circ}^i$ in eqn (3),

$$W_{N_A^\circ, N_B^\circ}^i \equiv \frac{1}{(N_A^\circ - i)!(N_B^\circ - i)!i!}, \quad (4)$$

corrects the over-counting when raising the partition functions to the power of the particle numbers. Alternatively, the value of $W_{N_A^\circ, N_B^\circ}^i$ can be obtained by first correcting all particles to be indistinguishable, *i.e.* dividing by $N_A^\circ!N_B^\circ!$, and then multiplying by the number of ways to form from N_A° and N_B° distinguishable particles, i pairs (where the order in each of the formed groups is not important), that is $N_A^\circ!N_B^\circ!/\left[(N_A^\circ - i)!(N_B^\circ - i)!i!\right]$.

The equilibrium constant is defined by,

$$K = e^{-\Delta G^\varnothing / RT}, \quad (5)$$

where $\Delta G^\varnothing$ is the standard Gibbs free energy change of the association reaction. This means, $\Delta G^\varnothing$ is the change in Gibbs energy when one mole of A reacts with one mole of B to form one mole of AB, given that all components are under their standard reference conditions. Here we choose the reference state of component x to be the concentration $c_x^\varnothing$ at temperature T. Measurements of $\Delta G^\varnothing$ are normally not performed on exactly one mole of particles, N_{Avogadro}, but are scaled to correspond to this number. For generality, we set the volume of the macroscopic reference system to $V^\varnothing$ from which the numbers of particles undergoing the association reaction in the reference system can be obtained by $N_x^\varnothing = V^\varnothing c_x^\varnothing$. In writing the partition function of the reference system, $Q^\varnothing$, we can still use eqn (3) but by substituting N_A° with $N_A^\varnothing$ and N_B° with $N_B^\varnothing$. Additionally, because $V^\varnothing$ is not equal to V, the single- and paired-particle partition functions in the reference system, $q_x^\varnothing$, are different than those in our system, q_x. The dependency of these functions on volume is due to the translational partition function and is therefore linear. Thus the following equality,

$$\frac{q_{AB}^\varnothing \Big/ V^\varnothing}{q_A^\varnothing \Big/ V^\varnothing \cdot q_B^\varnothing \Big/ V^\varnothing} = \frac{q_{AB}/V}{q_A/V \cdot q_B/V}, \quad (6)$$

between the particle partition functions in the two systems, having the same temperature, exists and will be used below.

We start by expressing the Gibbs free energy change, $\Delta G_{0\to N_A^\varnothing}$, when $N_A^\varnothing$ particles of A associate with $N_A^\varnothing$ (out of $N_B^\varnothing$) particles of B, *i.e.*, when all components are under their reference conditions. Then, we will obtain $\Delta G^\varnothing$ by scaling $\Delta G_{0\to N_A^\varnothing}$ to the stoichiometric number of moles of the reaction. The corresponding change in Helmholtz free energy, $\Delta F_{0\to N_A^\varnothing}$, can be calculated from the ratio of the probability to find the system in the bound state, p^{AB} (*i.e.*, the fraction of the state $i = N_A^\varnothing$ in the sum of the partition function for the reference system, $Q^\varnothing$), to the probability of the unbound state, p^{A+B} (the fraction of the state $i = 0$). Thus, $\Delta G_{0\to N_A^\varnothing}$ can be written as,

$$\begin{aligned}\Delta G_{0\to N_A^\varnothing} &\equiv G_{i=N_A^\varnothing} - G_{i=0} = \Delta F_{0\to N_A^\varnothing} + V^\varnothing \Delta P_{0\to N_A^\varnothing} \\ &= -k_B T \ln \frac{p^{AB}}{p^{A+B}} + V^\varnothing \Delta P_{0\to N_A^\varnothing} \\ &= -k_B T \ln \left[\frac{\left(q_{AB}^\varnothing\right)^{N_A^\varnothing} \left(q_B^\varnothing\right)^{N_B^\varnothing - N_A^\varnothing}}{N_A^\varnothing! \left(N_B^\varnothing - N_A^\varnothing\right)!} \frac{N_A^\varnothing! N_B^\varnothing!}{\left(q_A^\varnothing\right)^{N_A^\varnothing} \left(q_B^\varnothing\right)^{N_B^\varnothing}} \right] \\ &\quad + V^\varnothing \Delta P_{0\to N_A^\varnothing}, \end{aligned} \quad (7)$$

where $\Delta P_{0\to N_A^\varnothing}$ is the change in the pressure of the system accompanied the reaction. Almost without exception, the reference concentrations are chosen to be the same for all components ($c_x^\varnothing = c^\varnothing$ for all x), thus, eqn (7) reduces to,

$$\Delta G_{0\to N_A^\varnothing} = -N_A^\varnothing k_B T \ln \frac{q_{AB}^\varnothing}{q_A^\varnothing q_B^\varnothing} - k_B T \ln N_A^\varnothing! + V^\varnothing \Delta P_{0\to N_A^\varnothing}. \quad (8)$$

Applying Stirling's approximation to evaluate $\ln N_A^\varnothing!$, thus, requiring $N_A^\varnothing$ to be large, as is always the case for the standard state realized by a macroscopic measurement of $\Delta G^\varnothing$, and subsequently substituting $N_A^\varnothing$ with $V^\varnothing c^\varnothing$ gives,

$$\Delta G_{0\to N_A^\varnothing} = -N_A^\varnothing k_B T \ln \frac{q_{AB}^\varnothing \Big/ V^\varnothing}{q_A^\varnothing \Big/ V^\varnothing \cdot q_B^\varnothing \Big/ V^\varnothing} - N_A k_B T \ln c^\varnothing + N_A^\varnothing k_B T + V^\varnothing \Delta P_{0\to N_A^\varnothing}. \quad (9)$$

Now we will evaluate the ratio inside the first logarithm in eqn (9). Given the equality in eqn (6), we can do that using a different system, which is convenient for us to study, at the same temperature but with arbitrary numbers of particles N_A°, N_B° and volume V, thus at arbitrary concentrations, as long as the ideal behavior of the system is maintained. That means, we chose the system for which the partition function in eqn (3) was written for. We begin

by multiplying and dividing this ratio by the term,

$$\sum_{i=0}^{N_A^\circ-1}(i+1)\,W^{i+1}_{N_A^\circ,N_B^\circ}q_A^{N_A^\circ-i}q_B^{N_B^\circ-i}q_{AB}^{i}, \tag{10}$$

and obtain,

$$\begin{aligned}
&V^\varnothing\frac{q_{AB}^\varnothing}{q_A^\varnothing q_B^\varnothing}\\
&=V\frac{q_{AB}}{q_Aq_B}\\
&=V\frac{\sum_{i=0}^{N_A^\circ-1}(i+1)\,W^{i+1}_{N_A^\circ N_B^\circ}q_A^{N_A^\circ-(i+1)}q_B^{N_B^\circ-(i+1)}q_{AB}^{i+1}}{\sum_{i=0}^{N_A^\circ-1}\frac{(i+1)}{[N_A^\circ-(i+1)]![N_B^\circ-(i+1)]!(i+1)!}q_A^{N_A^\circ-i}q_B^{N_B^\circ-i}q_{AB}^{i}}.
\end{aligned} \tag{11}$$

We change the index of the sum in the numerator to $j = i + 1$ and rewrite the factorials in the denominator,

$$V\frac{q_{AB}}{q_Aq_B}=V\frac{\sum_{j=1}^{N_A^\circ}j\,W^{j}_{N_A^\circ,N_B^\circ}q_A^{N_A^\circ-j}q_B^{N_B^\circ-j}q_{AB}^{j}}{\sum_{i=0}^{N_A^\circ-1}(N_A^\circ-i)(N_B^\circ-i)\,W^{i}_{N_A^\circ,N_B^\circ}q_A^{N_A^\circ-i}q_B^{N_B^\circ-i}q_{AB}^{i}}. \tag{12}$$

Without changing the value of the sum in the numerator, we can let index j start from zero. The same is true if we let index i in the denominator end at N_A°. This yields,

$$\begin{aligned}
V\frac{q_{AB}}{q_Aq_B}&=V\frac{\frac{1}{Q}\sum_{j=0}^{N_A^\circ}j\,W^{j}_{N_A^\circ,N_B^\circ}q_A^{N_A^\circ-j}q_B^{N_B^\circ-j}q_{AB}^{j}}{\frac{1}{Q}\sum_{i=0}^{N_A^\circ}(N_A^\circ-i)(N_B^\circ-i)\,W^{i}_{N_A^\circ,N_B^\circ}q_A^{N_A^\circ-i}q_B^{N_B^\circ-i}q_{AB}^{i}}\\
&=V\frac{\langle N_{AB}\rangle}{\langle N_AN_B\rangle}=\frac{\langle c_{AB}\rangle}{\langle c_Ac_B\rangle},
\end{aligned} \tag{13}$$

where the sum in the numerator is identified as the ensemble average of the number of bound particles, $\langle N_{AB}\rangle$, and the sum in the denominator is the average of the product of the numbers of unbound particles, $\langle N_AN_B\rangle$, both in our chosen arbitrary system under equilibrium conditions. Inserting this result into eqn (9) gives,

$$\begin{aligned}
\Delta G_{0\to N_A^\varnothing}=&-N_A^\varnothing k_BT\ln\frac{\langle c_{AB}/c^\varnothing\rangle}{\langle c_A/c^\varnothing\cdot c_B/c^\varnothing\rangle}\\
&+N_A^\varnothing k_BT+V^\varnothing\Delta P_{0\to N_A^\varnothing}.
\end{aligned} \tag{14}$$

For ideal gases, the term $V^\varnothing\Delta P_{0\to N_A^\varnothing}$ equals $-N_A^\varnothing k_BT$, so the last two terms in eqn (14) cancel each other. In addition, the value of $\Delta G^\varnothing$ is reported per mole of stoichiometric coefficients in the chemical equation,

$$\Delta G^\varnothing=\frac{N_{\text{Avogadro}}}{N_A^\varnothing}\Delta G_{0\to N_A^\varnothing}. \tag{15}$$

Considering the definition of K in eqn (5) we obtain,

$$K=\frac{\langle c_{AB}\rangle}{\langle c_Ac_B\rangle}\cdot c^\varnothing=\frac{\langle P_{AB}\rangle}{\langle P_AP_B\rangle}\cdot P^\varnothing, \tag{16}$$

stating the equilibrium constant of binding reactions must include cross correlations in the reactants' concentrations. The second equality relates K to the corresponding ratio of the partial pressures of the different components where $P^\varnothing = c^\varnothing k_BT$ is the standard reference pressure.

Notice that during the entire derivation there were no conditions imposed specifying a finite system. In fact the definition of K implies a reference system with stoichiometric numbers of moles of particles, justifying the application of Stirling's approximation. It is only for convenience that we might use a small system to evaluate the ratio of the partition functions (of single- and paired-particle natures) encountered in eqn (9). Yet, it is specifically in this case that the resulting equilibrium constant expressed in eqn (16) is substantially different from an analogous expression neglecting correlations, K',

$$K'=\frac{\langle c_{AB}\rangle}{\langle c_A\rangle\langle c_B\rangle}\cdot c^\varnothing. \tag{17}$$

It is also essential to note that in statistical mechanics textbooks,[13–15] the equilibrium constant is derived using an ensemble at constant N_A, N_B, N_{AB}, V, T, where the numbers of particles are identified as those at equilibrium upon imposing the macroscopic conditions of chemical equilibrium. Fixing the numbers of particles of all components in the system, and inevitably their corresponding chemical potentials (conjugated parameters), render the description of chemical equilibrium of the reaction macroscopic. Not surprisingly, the resulting equilibrium constant is obtained in its thermodynamic form, $K = c_{AB}c^\varnothing/(c_Ac_B)$, even if along the derivation, relations taken from concepts of statistical mechanics were applied.

II. Computational validation

To test our derivation, we constructed a simple system of Lennard–Jones A and B molecules able to establish the equilibrium binding reaction of eqn (1). Three out of four parameters specifying the system in the canonical ensemble, N_A°, N_B°, and V, were changed systematically at a constant temperature, producing three different series of simulations, R1, R2, and R3. The first two series were subject to three different simulation methods; Monte Carlo (MC), molecular dynamics with a Nosé–Hoover thermostat (MD-NH), and molecular dynamics with a velocity-rescaling thermostat (MD-VR). The third series of simulations (R3) was conducted only by MC. Section SI-2 provides further details on the model system and computational methodologies (ESI†).

Fig. 1 displays the equilibrium constant, K, calculated by eqn (16), as well as, the value of K' defined in eqn (17). Clearly, the inclusion of cross-correlations in the reactants' concentrations is crucial for the equilibrium constant to stay constant at finite systems. In contrast, K' depends on the number of particles and/or volume of the system studied, where its deviation from K increases with decreasing size of the system. For systems large

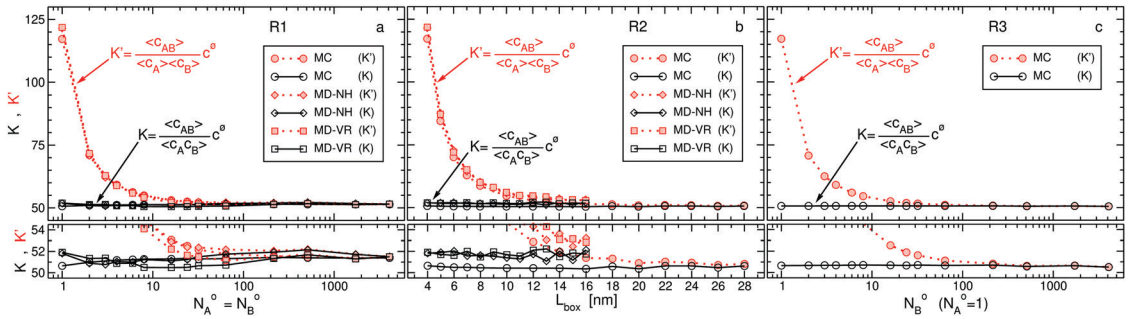

Fig. 1 The equilibrium constant K defined by eqn (16) ($c^\varnothing \equiv 1$ M) for three series of simulations at: (a) constant $c_A^\circ = c_B^\circ = 0.026$ M (R1), (b) constant $N_A^\circ = N_B^\circ = 1$ (R2), and (c) constant $N_A^\circ = 1$ and $c_B^\circ = 0.026$ M (R3), all performed in the canonical ensemble at $T = 300$ K. The value of K' defined by eqn (17) is shown in red for comparison. The lower panels are magnified plots around the value of K. The simulations were performed by three methods: Monte-Carlo (MC), molecular-dynamics with a Nosé–Hoover thermostat (MD-NH), and molecular-dynamics with a velocity-rescaling thermostat (MD-VR). The left-most point in all series represents the same system ($N_A^\circ = N_B^\circ = 1$, $L_{box} = 4$ nm). The estimated errors for the values of K are smaller or about the size of the symbols. Results from simulations at lower and higher temperatures are shown in Fig. SI-3.1 in the ESI.†

enough, where correlations become negligible, K' approaches K. The fact that K is constant for all sizes of the system, even for the smallest system possible, indicates that the law of mass action[16] holds not only for macroscopic but for finite-systems as well, in contrast to arguments found in the literature.[17,18] In section SI-1 (ESI†) we consider even a simpler model system, of single-site reactants, to facilitate easy comparison between the value of K obtained by eqn (16) and analytical/numerical calculations. Excellent agreements, with all three simulations methods, are attained.

As might be expected, the extent of divergence of K' from K is also a function of temperature. To demonstrate this, we conducted additional simulations of R1 series at different temperatures. Fig. 2, as well as Fig. SI-3.1 in Section SI-3 (ESI†), indicate this divergence of K' increases with decreasing temperature (or with increasing $-\varepsilon_{AB}/k_B T$). For example, for $N_A^\circ = N_B^\circ = 1$, K' is larger than K by a factor of 300 at $T = 200$ K, whereas it is nearly equal to K at $T = 1200$ K.

Taking the average of the product of reactants' concentrations, and not the product of their averages, when calculating K has a direct consequence for the condition of equilibrium. Using the relation between the chemical potentials of component x at c_x and at $c^\varnothing$ at the same temperature, $\mu_x = \mu_x^\varnothing + RT\ln(c_x/c^\varnothing)$, and identifying $\Delta G^\varnothing$ with $\mu_{AB}^\varnothing - \mu_A^\varnothing - \mu_B^\varnothing$, it follows from eqn (16) that the condition for equilibrium is,

$$\langle \mu_{AB} \rangle - \langle \mu_A + \mu_B \rangle = 0, \tag{18}$$

and not that expressed by the stoichiometric sum of the average of each component,

$$\langle \mu_{AB} \rangle - \langle \mu_A \rangle - \langle \mu_B \rangle = 0, \tag{19}$$

unless the system is large enough to render the correlations negligible.

The statistical thermodynamics expression of the equilibrium constant (eqn (16)) can also be rationalized from dynamics. At equilibrium, the average (over replica or over time) net change in the product's and reactants' concentrations is zero, thus we have,

$$\left\langle \frac{dc_{AB}}{dt} \right\rangle = \left\langle -\frac{dc_A}{dt} \right\rangle = \left\langle -\frac{dc_B}{dt} \right\rangle = \langle k_{fw} c_A c_B - k_{bw} c_{AB} \rangle = 0, \tag{20}$$

where k_{fw} and k_{bw} are the rate constants of the forward and backward reactions, respectively. The backward reaction is a simple unimolecular process, while the forward reaction is a bimolecular process and its rate is proportional to the collision probability between A and B. In turn, this collision probability at each point in time is proportional to the product of the corresponding instantaneous concentrations. That is, averaging the rate of the forward reaction in finite systems requires the cross-correlations of the two reactants' concentrations. By defining K as the ratio between forward and backward rate constants, and rendering its value dimensionless *via* $c^\varnothing$, we recover eqn (16). We calculated k_{fw} and k_{bw} from the MD simulations (section SI-2, ESI†) and the results corroborating k_{fw} at finite systems must include correlations between c_A and c_B (Fig. 3). Clearly, ignoring these correlations will produce rate constants that depend on concentrations (Fig. 3b) as

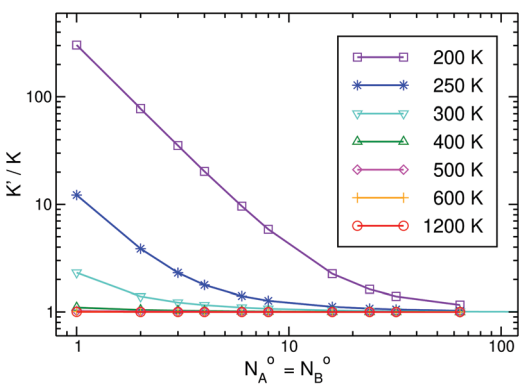

Fig. 2 The ratio of the equilibrium constant in which correlations between the reactant concentrations are ignored to that in which they are accounted for, K'/K, from MC R1 series of simulations at different temperatures.

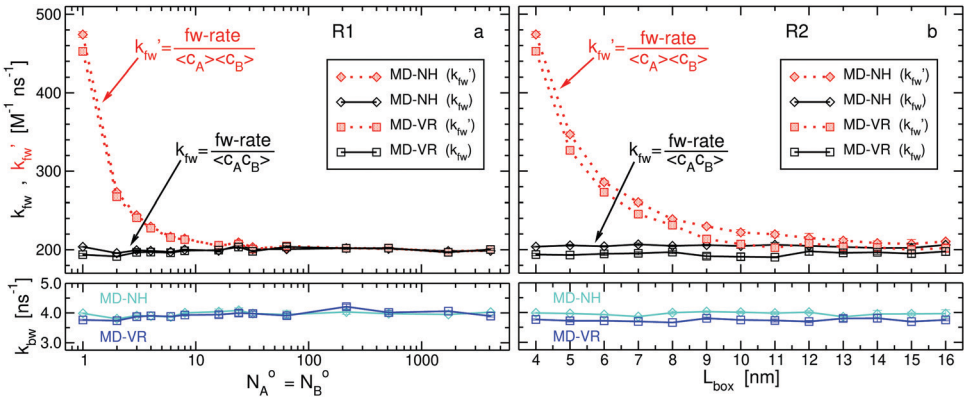

Fig. 3 The rate constants of the binding reaction for simulation series R1 (a) and R2 (b) obtained from molecular dynamics simulations. The top panels show the rate constant in the forward direction, k_{fw}, whereas the lower panels show the rate constant in the backward direction, k_{bw}. For comparison we also present k'_{fw} calculated by uncorrelated reactants' concentrations.

evidenced when analyzing single-molecule fluorescence binding experiments.[19]

In constructing R1 series of simulations, we multiplied all extensive parameters specifying the system by the same factor, exactly as described by eqn (2). This means, intensive properties are expected to have the same average values for all system sizes if the system's properties were homogeneous in character. However, Fig. 4 demonstrates that this is not the case at finite systems. In particular, the concentrations of the bound state, as well as the inter-particle energy per particle, exhibit increasing divergence from a horizontal line as the number of particles decreases. We also plot the radial distribution function between a and b sites. Again eqn (2) predicts overlapping curves for all system sizes, however, different distributions are obtained where the maxima describing the bound state for small-sized systems are higher in accordance with their larger concentrations. It is worth mentioning that these changes in the average properties of finite systems are not emerging from artefacts due to neglected concentration fluctuations in small simulations[20] or application of periodic boundary conditions in finite simulation boxes,[17,21–23] but are a consequence of incompatibility between two-body interactions and linear scaling.

III. Calculating concentrations from fluctuations

It is well known that fluctuations are related to susceptibilities. In our system, the incessant transitions at equilibrium between reactants and product force the number of particles of each component to fluctuate. We now show that the composition of the system (particle numbers, or concentrations) can be determined only from the magnitudes of these fluctuations.

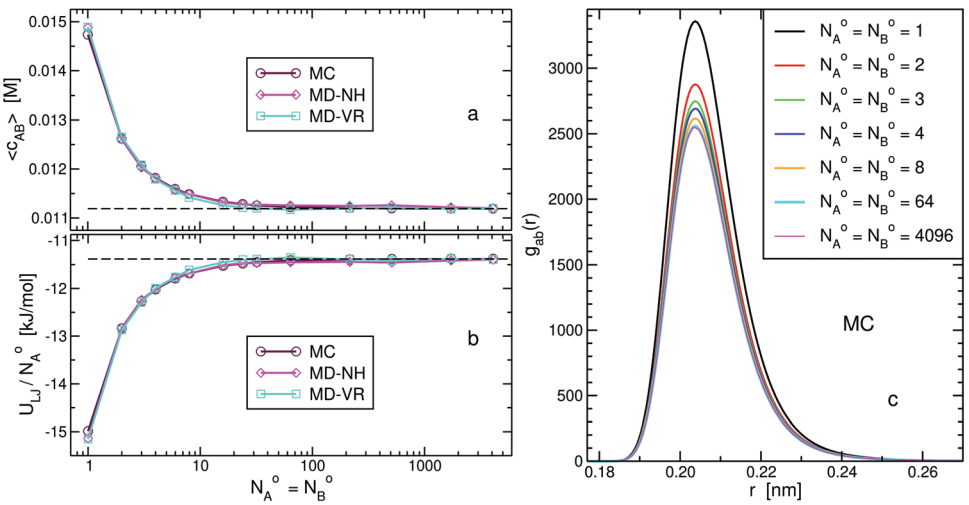

Fig. 4 Results exhibiting the inhomogeneous character of properties of bimolecular reactions upon scaling-down system size. The analyses were performed on R1, *i.e.* the series of simulations generated by scaling all extensive parameters specifying the system $\left(N_A^\circ = N_B^\circ, V\right)$ by the same factor. (a) The concentration of bound molecules, $\langle c_{AB} \rangle$, (b) the inter-particle energy per particle, and (c) the radial distribution function between a and b sites for different system sizes. (a) and (b) are almost perfect mirror-image of each other, and the estimated errors are smaller than the size of the symbols. In (c), only results from MC simulations are shown, however, very similar figures are obtained for MD-NH and MD-VR. If average quantities of the system were homogeneous functions, the data points in (a) and (b) would follow the horizontal dashed line, and the pair-distribution functions in (c) would collapse on the curve of the largest system.

Adopting the notation of Lebowitz *et al.*,[24] we define the cross fluctuations between quantities ζ and η as,

$$L(\zeta,\eta) = \langle\zeta\eta\rangle - \langle\zeta\rangle\langle\eta\rangle, \tag{21}$$

and their relative magnitude by,

$$l(\zeta,\eta) = \frac{L(\zeta,\eta)}{\langle\zeta\rangle\langle\eta\rangle}. \tag{22}$$

We now look at the following difference,

$$l(N_{AB},N_{AB}) - l(N_{AB},N_AN_B) = \frac{1}{\langle N_{AB}\rangle}\left[\frac{\langle {N_{AB}}^2\rangle}{\langle N_{AB}\rangle} - \frac{\langle N_{AB}N_AN_B\rangle}{\langle N_AN_B\rangle}\right], \tag{23}$$

and concentrate on evaluating the term inside the square brackets. We start by evaluating the term,

$$\frac{\langle {N_{AB}}^2\rangle}{\langle N_{AB}\rangle} = \frac{\frac{1}{Q}\sum_{i=0}^{N_A^\circ} i^2 W^i_{N_A^\circ,N_B^\circ} q_A^{N_A^\circ-i} q_B^{N_B^\circ-i} q_{AB}^i}{\frac{1}{Q}\sum_{i=0}^{N_A^\circ} i W^i_{N_A^\circ,N_B^\circ} q_A^{N_A^\circ-i} q_B^{N_B^\circ-i} q_{AB}^i} = \frac{\sum_{j=0}^{N_A^\circ-1}(j+1)^2 W^{j+1}_{N_A^\circ,N_B^\circ} q_A^{N_A^\circ-j} q_B^{N_B^\circ-j} q_{AB}^j}{\sum_{j=0}^{N_A^\circ-1}(j+1) W^{j+1}_{N_A^\circ,N_B^\circ} q_A^{N_A^\circ-j} q_B^{N_B^\circ-j} q_{AB}^j}, \tag{24}$$

where we skipped the terms corresponding to $i = 0$ and changed the index of the summation to $j = i - 1$. In the second equality we also multiplied and divided the ratio by q_Aq_B/q_{AB}. Similarly, we can express the second term inside the square brackets in eqn (23) by,

$$\frac{\langle N_{AB}N_AN_B\rangle}{\langle N_AN_B\rangle} = \frac{\frac{1}{Q}\sum_{i=0}^{N_A^\circ} i\left(N_A^\circ-i\right)\left(N_B^\circ-i\right) W^i_{N_A^\circ,N_B^\circ} q_A^{N_A^\circ-i} q_B^{N_B^\circ-i} q_{AB}^i}{\frac{1}{Q}\sum_{i=0}^{N_A^\circ} \frac{\left(N_A^\circ-i\right)\left(N_B^\circ-i\right)}{\left(N_A^\circ-i\right)!\left(N_B^\circ-i\right)!i!} q_A^{N_A^\circ-i} q_B^{N_B^\circ-i} q_{AB}^i} = \frac{\sum_{i=0}^{N_A^\circ-1} i(i+1)\ W^{i+1}_{N_A^\circ,N_A^\circ} q_A^{N_A^\circ-i} q_B^{N_B^\circ-i} q_{AB}^i}{\sum_{i=0}^{N_A^\circ-1}(i+1)\ W^{i+1}_{N_A^\circ,N_B^\circ} q_A^{N_A^\circ-i} q_B^{N_B^\circ-i} q_{AB}^i}. \tag{25}$$

The second equality is realized by letting index i in the sum end at $N_A^\circ - 1$ and rewriting the coefficients of the single/paired-particle partition functions in terms of $W^{i+1}_{N_A^\circ,N_B^\circ}$. Note that the denominators of eqn (24) and (25) are the same, so the difference of the two terms inside the square brackets in eqn (23) can be easily evaluated,

$$\frac{\langle {N_{AB}}^2\rangle}{\langle N_{AB}\rangle} - \frac{\langle N_{AB}N_AN_B\rangle}{\langle N_AN_B\rangle} = \frac{\sum_{i=0}^{N_A^\circ-1}(i+1) W^{i+1}_{N_A^\circ,N_B^\circ} q_A^{N_A^\circ-i} q_B^{N_B^\circ-i} q_{AB}^i}{\sum_{i=0}^{N_A^\circ-1}(i+1) W^{i+1}_{N_A^\circ,N_B^\circ} q_A^{N_A^\circ-i} q_B^{N_B^\circ-i} q_{AB}^i} = 1, \tag{26}$$

which actually reduces to one. This means eqn (23) becomes,

$$l(N_{AB},N_{AB}) - l(N_{AB},N_AN_B) = \frac{1}{\langle N_{AB}\rangle}, \tag{27}$$

from which the average concentration of the bound AB particles can be expressed by,

$$\langle c_{AB}\rangle = \frac{1}{[l(N_{AB},N_{AB}) - l(N_{AB},N_AN_B)]V}. \tag{28}$$

Eqn (27) can also be derived by a more conventional procedure, *i.e.* by partially differentiating the partition function of the system with respect to temperature. However in this case we need to assume K is given by eqn (16) (see section SI-4, ESI†).

It is interesting to comment that although $l(N_{AB}, N_{AB})$ is necessarily positive, the relative fluctuations in $l(N_{AB}, N_AN_B)$ measure correlations between two quantities that are anti-correlated and hence always negative. Thus, the quantity inside the square brackets of the denominator in eqn (28) is a summation of two positive terms with magnitude that reduces with increasing system size. For large systems, this reduction is proportional to the reciprocal of the volume so that $\langle c_{AB}\rangle$ approaches a constant.

The relation in eqn (27) was tested for all simulations performed. In Fig. 5a we plot the results at $T = 300$ K, and in Fig. SI-3.2 (ESI†) the results of the R1 series at different temperatures. All data points, independent of temperature, fall on the same straight line as predicted. The correlation coefficients of the linear regressions turned-out perfect, within the accuracy of the analysis software, likely because comparison is made between two quantities calculated from the same simulation allowing elimination of certain errors.

IV. Calculating concentrations from K

A drawback of eqn (27) or eqn (28) is when the system simulated or studied experimentally is not of the same size as the target system. In this case, relative fluctuations of the target system are needed in order to compute composition. Thus it would be more practical if we can determine the concentrations from K and the parameters defining the target system.

We start by rewriting the expression of K,

$$\langle N_{AB}\rangle = \frac{K}{Vc^\varnothing}\langle N_AN_B\rangle = \frac{K}{Vc^\varnothing}\left\langle\left(N_A^\circ - N_{AB}\right)\left(N_B^\circ - N_{AB}\right)\right\rangle = \frac{K}{Vc^\varnothing}\left[N_A^\circ N_B^\circ - \left(N_A^\circ + N_B^\circ\right)\langle N_{AB}\rangle + L(N_{AB},N_{AB}) + \langle N_{AB}\rangle^2\right], \tag{29}$$

and solve the quadratic equation to obtain,

$$\langle c_{AB}\rangle = \frac{\left(c_A^\circ + c_B^\circ + \frac{c^\varnothing}{K}\right) - \sqrt{\left(c_A^\circ + c_B^\circ + \frac{c^\varnothing}{K}\right)^2 - 4[l(N_{AB},N_{AB})+1]c_A^\circ c_B^\circ}}{2[l(N_{AB},N_{AB})+1]}. \tag{30}$$

If we performed simulations/experiments at a finite size and wish to know the concentrations in the thermodynamic limit

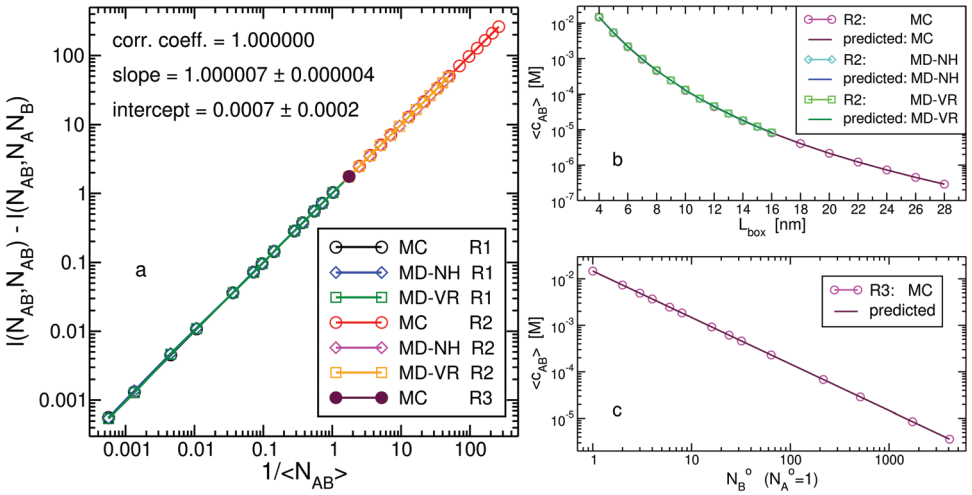

Fig. 5 (a) A relation between two relative fluctuations and the reciprocal of average number of bound particles. All simulation results (displayed for T = 300 K, here, for other temperatures see Fig. S3.2, ESI†) fall on a linear line crossing the origin with a slope of one as described in eqn (27). The results obtained from linear regression (using xmgrace) of all data points are indicated. All points of R3 have the same x, y values. (b) The concentration of bound particles as a function of box length for R2 series. The results obtained from simulations are shown along predictions based on the value of the equilibrium constant (eqn (32)). (c) Same as (b) but for R3 series of simulations, in which case, the concentration is plotted as a function of the total number of B particles.

$\left(V, N_A^\circ \to \infty\right)$, we simply set $l(N_{AB}, N_{AB}) = 0$ and recover general chemistry textbook results,[25]

$$\langle c_{AB}\rangle_\infty = \frac{1}{2}\left[c_A^\circ + c_B^\circ + \frac{c^\varnothing}{K} - \sqrt{\left(c_B^\circ - c_A^\circ\right)^2 + \frac{2c^\varnothing\left(c_A^\circ + c_B^\circ\right)}{K} + \frac{{c^\varnothing}^2}{K^2}}\right], \tag{31}$$

and because $L(N_{AB}, N_{AB}) = L(N_A, N_B)$, we can substitute the correlated reactants' concentrations appearing in the expression for K with the uncorrelated concentrations, *i.e.* $K' \to K$.

Another important case, especially for simulation studies and single-molecule experiments, is that of $N_A^\circ = 1$ (where $N_B^\circ \geq N_A^\circ$). Here it is easy to show that the relation, $\langle N_A N_B\rangle = N_B^\circ\left(N_A^\circ - \langle N_{AB}\rangle\right)$, is satisfied which leads to,

$$\langle c_{AB}\rangle_{N_A^\circ=1} = \frac{N_A^\circ N_B^\circ K}{V\left(Vc^\varnothing + N_B^\circ K\right)}. \tag{32}$$

In addition, the relative fluctuations obey,

$$l(N_{AB}, N_{AB})_{N_A^\circ=1} = \frac{Vc^\varnothing}{KN_B^\circ}, \tag{33}$$

by noting that in this case $(N_A^\circ{=}1)$, $l(N_{AB}, N_A N_B) = 1$ and $\langle {N_{AB}}^2\rangle = \langle N_{AB}\rangle$. R2 and R3 series of simulations fall within this special case, therefore, in Fig. 5b and c we predict $\langle c_{AB}\rangle$ as stated by eqn (32), and in Fig. SI-3.3 (ESI†) we predict $l(N_{AB}, N_{AB})$ according to eqn (33). In both cases, the agreement is almost perfect.

In contrast to the thermodynamic limit and systems with $N_A^\circ = 1$, predicting $\langle c_{AB}\rangle$ from the value of K for other finite systems, thus with $N_A^\circ > 1$, is less simple because of the difficulty of predicting $l(N_{AB}, N_{AB})$. Obviously, the magnitude of $l(N_{AB}, N_{AB})$ for $N_A^\circ > 1$ must be smaller than that for $N_A^\circ = 1$. A plausible guess can be that it is inversely proportional to the system size. We therefore express $l(N_{AB}, N_{AB})$ for $N_A^\circ > 1$ by scaling the corresponding value at $N_A^\circ{=}1$ according to,

$$\begin{aligned} l(N_{AB}, N_{AB})_{N_A^\circ \geq 1} &= l(N_{AB}, N_{AB})_{N_A^\circ=1} \cdot \frac{1}{\left(N_A^\circ\right)^\lambda} \\ &= \frac{Vc^\varnothing}{KN_B^\circ} \cdot \frac{1}{\left(N_A^\circ\right)^\lambda}, \end{aligned} \tag{34}$$

where $0 \leq \lambda \leq 1$.

When λ = 0, eqn (34) reduces to eqn (33), whereas for the thermodynamic limit it turns out from the simulations that λ = 1. Empirically we find λ can be approximated by,

$$\lambda \simeq \frac{1}{1 + K/\left(Vc^\varnothing \ln N_B^\circ\right)}. \tag{35}$$

This approximation is investigated in Fig. SI-3.4 (ESI†) for the R1 series of simulations. Very good agreement with the simulation data is obtained where the accuracy of prediction increases with temperature. Armed with the ability to estimate $l(N_{AB}, N_{AB})$, we proceed to predict the concentrations *via* eqn (30) in Fig. SI-3.5 (ESI†). Although not perfect at lower temperatures, the approximation yields satisfactory agreement with concentrations observed in the simulations. Due to the asymmetric roles of N_A° and N_B° in eqn (34), we examined the approximation also on another series of simulations, R4, in which N_A° and N_B° are not equal (Section SI-2, ESI†). Here the accuracy of the prediction, shown in Fig. SI-3.6, (ESI†) is even better. Moreover, we scale $g_{ab}(r)$ obtained at finite systems to the corresponding distribution of a macroscopic system, as shown in Fig. SI-5.1 and discussed in Section SI-5 (ESI†).

Discussion

An example

We now exemplify and discuss the calculation of K for the association reaction in eqn (1) for the smallest system possible, $N_A^\circ = N_B^\circ = 1$. In this case, there are only two possible macroscopic states in the system, one corresponding to a bound AB particle and the other to unbound $A + B$ particles. Suppose the fraction of independent configurations in which the bound state is observed is f^{AB} (thus, the fraction of the unbound state is $f^{A+B} = 1 - f^{AB}$). Applying the expression of K with uncorrelated reactants' concentrations defined in eqn (17) yields,

$$K' = \frac{f^{AB}/V}{\left[(1-f^{AB})/V\right]^2} \cdot c^{\varnothing}. \tag{36}$$

Although this is currently the most employed expression in the literature,[3–12] it provides erroneous results at finite systems as demonstrated throughout the manuscript. This is because correlations in reactants' concentrations, which must be taken into account, are augmented as the system size is decreased. Yet, for this system it is possible to calculate K from the ratio of f^{AB} to f^{A+B}. However application of the plain ratio,

$$K'' = \frac{f^{AB}}{1-f^{AB}}, \tag{37}$$

which is equal to the ratio of probabilities to find this particular system in the bound and unbound states, does not correspond to K. The reason is that this ratio is size-dependent. There are many more possible microstates for the unbound state than for the bound state, and scaling with system-size follows different power-laws for the two states. In our derivation (eqn (7)), this is expressed in the corresponding translational partition functions; the number of possible states is proportional to the volume for the bound particles whereas it is proportional to the square of the volume for the unbound particles. It is only when q_A, q_B, and q_{AB} are, each, divided by V that the ratio becomes size-independent as argued in eqn (6). Thus normalizations of the probabilities (or the number of configurations) in eqn (37) by a factor of V and V^2, to obtain probability densities, are necessary, albeit the introduction of a dimension of volume to the ratio. It is therefore for these cases, *i.e.* when the number of particles on the reactant side is not equal to that on the product side, that a reference to a standard system is necessary to render K dimensionless. In eqn (9) the standard concentration, $c^{\varnothing}$, emerged from the term $N_A^{\varnothing}!$ that was not canceled-out in the ratio of probabilities of the two states (in eqn (7)). Dividing the probabilities in eqn (37) by the normalization factors and eliminating the dimension of the ratio by $c^{\varnothing}$ gives,

$$K = \frac{f^{AB}V}{1-f^{AB}} \cdot c^{\varnothing}, \tag{38}$$

an expression identical to that we would have obtained had we used eqn (16). Obviously this simple direct counting of configurations of the two states to obtain K, or even just the free energy difference ΔG for the studied system,[26] can only work for $N_A^\circ = N_B^\circ = 1$. The reason is, that in this case there are only two macroscopic states which represent the two states for which the standard Gibbs free energy change is calculated for.

The difference with unimolecular processes

It is important to emphasize that the arguments presented in this paper are pertinent to bimolecular reactions or two-body properties. Consider the chemical equation,

$$A + B \rightleftharpoons 2C, \tag{39}$$

representing, for example, the recombination of hydronium and hydroxide ions to form two water molecules. As it is a bimolecular process, the equilibrium constant is expressed as,

$$K = \frac{\langle c_C{}^2 \rangle}{\langle c_A c_B \rangle}. \tag{40}$$

Another process that can also be represented by exactly the same chemical equation is, for example, the transitions between different conformations of a peptide, where A, B, and C denote α-helix, β-sheet, and random-coil. In this case, eqn (39) is actually a sum of two chemical equations ($A \rightleftharpoons C$ and $B \rightleftharpoons C$) in which α-helix and β-sheet, separately, form equilibrium with the coil conformation. The transitions between the different conformations are unimolecular in nature and the expression of K in eqn (40) is not appropriate. As no correlations exist between the α-helix and β-sheet conformations, the equilibrium constant should be computed by,

$$K = \frac{\langle c_C \rangle^2}{\langle c_A \rangle \langle c_B \rangle}, \tag{41}$$

that is, the product of the equilibrium constants of the two unimolecular reactions. The outcome of these two examples contradicts the principle upon which chemists view chemical equilibrium, that is, K is dictated only by the chemical equation of the reaction, irrespective of its nature. This is indeed true as long as the system is macroscopic or large enough. However, at finite systems, the expression of the equilibrium constant of two reactions with the same chemical equation can be different. The distinction emerges because different averaging applies for calculating K depending on the order of the elementary process(es) involved, as is the case when determining rate constants.

Magnitude of the two-body correlations

As pointed out above, the magnitude of correlations between the reactants, which can be represented also by the deviation of the K'/K ratio from 1, is influenced by temperature (or by the 'reduced' temperature, $k_B T/\varepsilon_{AB}$). Fig. 2 shows this clearly, yet it indicates that the correlations are affected as well by the numbers of particles and/or volume, because in R1 series when $N_A^\circ = N_B^\circ$ increases, V increases by the same factor to keep $c_A^\circ = c_B^\circ$ constant. In the R2 series, the volume is the only parameter changing and from Fig. 1b it is evident that it affects the magnitude of correlations. Given the well-known expression of fluctuations in the number of particles in the grand-canonical ensemble, it is tempting to assume that the correlations in our system would decay inversely with the number of

particles. In Fig. SI-3.7a (ESI†) we plot K'/K for R4 series where all simulations have the same T, V, and N_B° and only N_A° was increased from 1 to 8. However, this increase in the value of N_A° did not have an effect on the ratio of K'/K. In contrast, the value of N_B° does influences the correlations. This can be seen in Fig. SI-3.7b (ESI†) where we compare the R1 and R2 series. In both series, $N_A^\circ = N_B^\circ$, however in R2 these numbers equal 1 for all simulations whereas in R1 they vary. The curves, plotted as a function of V, indicate that the value of K'/K at fixed V is lower for R1 where the numbers of particles are larger, however, the decay is much smaller (less strong) than $1/N_B^\circ$.

Conclusions

In this paper we demonstrated that equilibrium constants, as well as rate constants, of binding reactions at finite systems must include correlations between reactants' concentrations. That being the case, equilibrium is achieved when the average chemical potential of the bound product is equal to the average of the sum, and not to the sum of the averages, of the chemical potentials of the unbound reactants. This point has never been considered in the literature, likely because the working assumption followed an outcome presented in statistical mechanics textbooks based on an ensemble, claimed here inappropriate, which leads only to the well-known expression applicable for macroscopic systems. Instead, a different derivation is offered in which the constructed ensemble fixes only the total number of each particle-type in the system. This allows the numbers of reactants and product(s) of the reaction to experience fluctuations, with magnitude dictated by the parameters specifying the system. Accordingly, the resulting expression of K provides information on how to perform averaging over the ensemble utilized. A key step in the derivation is the evaluation of the ratio $Vq_{AB}/(q_Aq_B)$. By applying a sequence of algebraic operations, we showed this ratio to be equal to $\langle c_{AB}\rangle/\langle c_Ac_B\rangle$, where the brackets indicate ensemble average under equilibrium conditions. This inclusion of correlations in calculating the equilibrium constant, can produce values that differ by few orders of magnitude compared with those neglecting them. Because correlations become less important with increasing system size, for macroscopic systems, the statistical mechanical expression of K reduces to that obtained from thermodynamics.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

I would like to thank Zohar Nussinov for stimulating and insightful discussion. Technical and human support of the computer cluster provided by IZO-SGI SGIker of UPV/EHU and the European fundings, ERDF and ESF, are greatly acknowledged.

References

1 H. B. Callen, *Thermodynamics and an Introduction to Thermostatistics*, John Wiley & Sons, New York, NY, 1985.
2 M. K. Gilson, J. A. Given, B. L. Bush and J. A. McCammon, *Biophys. J.*, 1997, **72**, 1047–1069.
3 P. H. Hünenberger, J. K. Granwehr, J.-N. Aebischer, N. Ghoneim, E. Haselbach and W. F. van Gunsteren, *J. Am. Chem. Soc.*, 1997, **119**, 7533–7544.
4 H. Luo and K. Sharp, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 10399–10404.
5 Y. Zhang and J. A. McCammon, *J. Chem. Phys.*, 2003, **118**, 1821–1827.
6 E. Psachoulia, P. W. Fowler, P. J. Bond and M. S. P. Sansom, *Biochemistry*, 2008, **47**, 10503–10512.
7 Y. Deng and B. Roux, *J. Phys. Chem. B*, 2009, **113**, 2234–2246.
8 R. Skorpa, J.-M. Simon, D. Bedeaux and S. Kjelstrup, *Phys. Chem. Chem. Phys.*, 2014, **16**, 1227–1237.
9 M. De Vivo, M. Masetti, G. Bottegoni and A. Cavalli, *J. Med. Chem.*, 2016, **59**, 4035–4061.
10 J. J. Montalvo-Acosta and M. Cecchini, *Mol. Inform.*, 2016, **35**, 555–567.
11 L. A. Patel and J. T. Kindt, *J. Chem. Theory Comput.*, 2017, **13**, 1023–1033.
12 E. Duboué-Dijon and J. Hénin, *J. Chem. Phys.*, 2021, **154**, 204101.
13 D. A. McQuarrie, *Statistical Thermodynamics*, University Science Books, Mill Valley, CA, 1973.
14 D. Chandler, *Introduction to Modern Statistical Mechanics*, Oxford University Press, New York, NY, 1987.
15 H. Gould and J. Tobochnik, *Statistical and Thermal Physics: With Computer Applications*, Princeton University Press, Princeton, NJ, 2010.
16 P. Waage and C. M. Guldberg, *Avh. - Nor. Vidensk.-Akad., 1: Mat.-Naturvidensk. Kl.*, 1864, 35–45.
17 D. H. De Jong, L. V. Schäfer, A. H. De Vries, S. J. Marrink, H. J. C. Berendsen and H. Grubmüller, *J. Comp. Chem.*, 2011, **32**, 1919–1928.
18 X. Zhang, L. A. Patel, O. Beckwith, R. Schneider, C. J. Weeden and J. T. Kindt, *J. Chem. Theory Comput.*, 2017, **13**, 5195–5206.
19 J. Yang and J. E. Pearson, *J. Chem. Phys.*, 2012, **136**, 244506.
20 T. E. Ouldridge, A. A. Louis and J. P. K. Doye, *J. Phys.: Condens. Matter*, 2010, **22**, 104102.
21 R. Cortes-Huerto, K. Kremer and R. Potestio, *J. Chem. Phys.*, 2016, **145**, 141103.
22 N. Dawass, P. Krüger, S. K. Schnell, D. Bedeaux, S. Kjelstrup, J. M. Simon and T. J. H. Vlugt, *Mol. Sim.*, 2018, **44**, 599–612.
23 R. Hall, T. Dixon and A. Dickson, *Front. Mol. Biosci.*, 2020, **7**, 106.
24 J. L. Lebowitz, J. K. Percus and L. Verlet, *Phys. Rev.*, 1967, **153**, 250–254.
25 D. W. Oxtoby and N. H. Nachtrieb, *Principles of Modern Chemistry*, Saunders College Publishing, Orlando, FL, 3rd edn, 1996.
26 W. F. van Gunsteren, X. Daura and A. Mark, *Helv. Chim. Acta*, 2002, **85**, 3113–3129.

Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supplementary Information: Binding Reactions at Finite Systems

Ronen Zangi*[1,2]

[1] *POLYMAT & Department of Organic Chemistry I, University of the Basque Country UPV/EHU, Avenida de Tolosa 72, 20018, Donostia-San Sebastián, Spain*

[2] *IKERBASQUE, Basque Foundation for Science, Plaza Euskadi 5, 48009 Bilbao, Spain*

March 4, 2022

*r.zangi@ikerbasque.org

SI-1 Comparisons with Analytical/Numerical Methods

We now compare the value of the equilibrium constant using Eq. 16 to two well-known analytical expressions derived from evaluations of the single-particle, $q_{_A}$ and $q_{_B}$, and pair-particle, $q_{_{AB}}$, partition functions. In the first method, these partition functions are evaluated by integration over the coordinates of the particles, whereas in the second method, $q_{_{AB}}$ is obtained by integrations over the coordinates and momenta of the center-of-mass and relative motions of the bound state.

For the purpose of comparisons, we choose a finite, $N_{_A}^{\circ} = N_{_B}^{\circ} = 1$, model system at $c_{_A}^{\circ} = c_{_B}^{\circ} = 0.00462963\ molecule/nm^3$ (corresponding to $L_{box} = 6.0\ nm$) in which the reference expressions (see below) can be easily calculated analytically or numerically if we describe A and B as single-site particles*, $A \equiv a$ and $B \equiv b$. We also modified the well-depth of the Lennard-Jones potential to $\epsilon_{_{AB}}^{LJ} = 22.15\ kJ/mol$ so that its magnitude is similar to the effective attraction between A and B in the simulations with diatomic monomers described above. All other simulation parameters are unchanged.

MC simulations of 10^{12} trial moves, with same relevant characteristics as described above, were performed to yield an acceptance-ratio of 0.44 and an equilibrium constant, calculated by Eq. 16, of 51.09. In addition two MD simulations, using Nosé-Hoover and velocity-rescaling thermostats, were ran for $48\ \mu s$ and $160\ \mu s$ resulting with a value of K of 50.93 and 52.20, respectively (see Table SI-1.2).

I. K from Integration over Particle's Coordinates

If $\mathcal{T}$ and $\mathcal{U}$ are the kinetic and potential parts of the Hamiltonian, the pair-particle partition function can be written as,

$$q_{_{AB}}(\vec{p_{_A}}, \vec{p_{_B}}, \vec{r_{_A}}, \vec{r_{_B}}) = \frac{1}{h^6} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{-\beta \mathcal{T}(\vec{p_{_A}}, \vec{p_{_B}})} d\vec{p_{_A}} d\vec{p_{_B}} \int_{\vec{r_{_A}}} d\vec{r_{_A}} \int_{0}^{r_c} e^{-\beta \mathcal{U}(r)} d\vec{r} \quad , \quad \text{(SI-1.1)}$$

where h is Planck's constant and r_c the cutoff distance defining the bound state. The integrals over the momenta of each particle are of three dimensions, as is the integral over $\vec{r_{_A}}$, i.e. over all

*The potential danger of products other than the AB bound state is now removed by restricting the simulations to a system with $N_{_A}^{\circ} = N_{_B}^{\circ} = 1$.

possible coordinates of particle A, which yields V. In Eq. SI-1.1 we assumed the potential energy of the system depends only on the relative distance between A and B particles, $r = |\vec{r_A} - \vec{r_B}|$. Further assuming $\mathcal{U}$ vanishes for $r > r_c$, the single-particle partition function, q_A, is

$$q_A(r) = \frac{1}{h^3}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\beta\mathcal{T}(\vec{p_A})} d\vec{p_A} \int_{\vec{r_A}} d\vec{r_A} = \frac{V}{h^3}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\beta\mathcal{T}(\vec{p_A})} d\vec{p_A} \quad , \quad \text{(SI-1.2)}$$

and a corresponding expression holds for q_B. In the ratio for K, the integrals over momenta cancel-out and we are left with,

$$K = \frac{q_{AB}V}{q_A \cdot q_B} \cdot c^{\varnothing} = c^{\varnothing}\int_0^{r_c} e^{-\beta\mathcal{U}(r)} d\vec{r} = c^{\varnothing}\int_0^{r_c} e^{-\beta\mathcal{U}(r)} 4\pi r^2 dr \quad . \quad \text{(SI-1.3)}$$

If there had been other degrees of freedom in the system, integratable at fixed values of r, then instead of $\mathcal{U}(r)$ we would have had $w(r)$, the potential of the averaged force acting between A and B due to those other degrees of freedom[1]. Thus, the need for additional simulations to calculate the potential of mean force is avoided here because A and B are mono-atomic particles, and we can solve Eq. SI-1.3 numerically using the Lennard-Jones potential described above. This gives $K = 51.04$.

II. K from a Molecular Partition Function

The Hamiltonian of the pair-particle partition function can also be written in terms of generalized coordinates and momenta describing translation of the center-of-mass, as well as, rotations and vibrations of the bound AB state. If the rotational and vibrational modes are decoupled, the expression of K becomes,

$$K = \frac{q_{\text{trans}(AB)}\ q_{\text{rot}}\ q_{\text{vib}}\ e^{-\beta\epsilon_{AB}}}{q_{\text{trans}(A)}\ q_{\text{trans}(B)}} V c^{\varnothing} \quad , \quad \text{(SI-1.4)}$$

where ϵ_{AB} equals $-\epsilon_{AB}^{LJ}/N_{\text{Avogadro}}$ set above. We use textbooks[2] results for the translational and high-temperature (rigid-rotor) rotational partition functions. These are,

$$q_{\text{trans}} = \left(\frac{2\pi m k_B T}{h^2}\right)^{3/2} V \quad , \quad \text{(SI-1.5)}$$

and,

$$q_{\text{rot}} = \frac{8\pi^2 I k_B T}{h^2} \quad , \quad \text{(SI-1.6)}$$

where m is the mass of the translating body, $I = \mu R_{eq}^2$ is the moment of inertia with μ the reduced mass and $R_{eq} = 2^{1/6}\sigma_{_{AB}} = 0.2020\ nm$, the equilibrium distance between A and B particles in the bound state. We also assume high-temperatures for the vibrational partition function, arising from the oscillatory motion around the minimum of the LJ potential, and perform numerical integration instead of discrete summation. Here, the Hamiltonian includes a one-dimensional kinetic term of a body with a reduced mass μ and the Lennard-Jones potential is shifted by ϵ_{AB}^{LJ} so its minimum is at zero energy. We therefore have,

$$q_{\text{vib}} = \frac{1}{h}\int_{-\infty}^{\infty} e^{-\beta p^2/2\mu} d\vec{p} \int_0^{r_c} e^{-\beta[U_{LJ}(r)+\epsilon_{AB}^{LJ}]} dr = \left(\frac{2\pi\mu k_{_B} T}{h^2}\right)^{1/2} \int_0^{r_c} e^{-\beta[U_{LJ}(r)+\epsilon_{AB}^{LJ}]} dr \quad . \tag{SI-1.7}$$

We calculate these different elements of the molecular partition function for the system introduced above and present the results in Table SI-1.1. Inserting these values in Eq. SI-1.4 we obtain

Table SI-1.1: The value of different elements in the molecular partition function of a diatomic gas, along with the corresponding monoatomic partition functions and the Boltzmann's factor, necessary to compute the equilibrium constant in Eq. SI-1.4, at $T = 300\ K$. The quantities q_{trans}/V are given in units of m^{-3}.

$q_{\text{trans}}(AB)/V$	q_{rot}	q_{vib}	$e^{-\beta\epsilon_{AB}}$	$q_{\text{trans}}(A)/V = q_{\text{trans}}(B)/V$
$8.734 \cdot 10^{31}$	252.4	0.4854	7187	$3.088 \cdot 10^{31}$

$K = 48.57$.

In Table SI-1.2 we summarize the results obtained from the simulations, as well as, from the two analytical/numerical methods. The agreement of the MC and MD-NH simulations with the numerical evaluation of Eq. SI-1.3 is excellent. The result of MD-VR is slightly less good where it converges to a different value than that determined by Eq. SI-1.3, nevertheless the discrepancy of $0.056\ kJ/mol$ in $\Delta G^{\varnothing}$ is rather small. A mild discrepancy, relative to the other four results, is also observed when we evaluate K by the molecular partition function using Eq. SI-1.4 with a magnitude that translates to $0.12 - 0.18\ kJ/mol$ for the value of $\Delta G^{\varnothing}$. This is not surprising given the assumptions made in Eq. SI-1.4 and is likely to be the least accurate method. The most questionable assumption is the neglect of coupling between vibrational and rotational degrees of

Table SI-1.2: Comparison between values of the equilibrium constant K computed by five different methods, for the reaction described in Eq. 1 using the simplified system of single-site monomers detailed in this section. Simulations utilizing Monte-Carlo (MC) and two Molecular Dynamics, one with a Nosé-Hoover (MD-NH) and one with a velocity-rescaling (MD-VR) thermostats, methods were performed. In these simulations, K was obtained by calculating the ratio between the product and correlated-reactants concentrations according to Eq. 16. The analytical/numerical calculations were based on integration of the particles coordinates (Eq. SI-1.3), as well as on partition functions describing relative motions of a diatomic molecule (Eq. SI-1.4). In addition to the values of K, we also provide (in kJ/mol) the corresponding change in the standard Gibbs energy, $\Delta G^{\varnothing}$, using the definition in Eq. 5.

	Simulations (Eq. 16)			Analytical/Numerical Evaluations	
	MC	MD-NH	MD-VR	Eq. SI-1.3	Eq. SI-1.4
K	51.09 ± 0.06	50.93 ± 0.25	52.20 ± 0.13	51.04	48.57
$\Delta G^{\varnothing}$	-9.812 ± 0.003	-9.804 ± 0.012	-9.865 ± 0.006	-9.809	-9.686

freedom in a system of bound particles held together by an *intermolecular* potential that, for a rigid-rotor approximation, is rather soft.

Comparing the Radial Distribution Functions

We write the total partition function of the system of one A and one B particles based on the way we defined q_{AB} in Eq. SI-1.1 but with an upper bound of the integral over r that includes all possible values of the relative distances between A and B,

$$Q(\vec{p_A}, \vec{p_B}, \vec{r_A}, \vec{r_B}) = \frac{1}{h^6} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{-\beta \mathcal{T}(\vec{p_A}, \vec{p_B})} d\vec{p_A} d\vec{p_B} \int_{\vec{r_A}} d\vec{r_A} \int_0^{r_{box}} e^{-\beta \mathcal{U}(r)} 4\pi r^2 dr \quad . \tag{SI-1.8}$$

Note that computer simulations often use rectangular-shaped boxes which are not so convenient to integrate by a spherically symmetric coordinate system. This is easy to solve if we recall our assumption that $\mathcal{U}(r)$ vanishes for $r > r_c$, because for these values of r the integrand is 1 and we are integrating only the relative spatial coordinates. Thus all we need to do is to perform the

integration from 0 to r_c and add the remaining volume element, $V - 4\pi r_c^3/3$. Alternatively, we can integrate from 0 to r_{box}, as indicated in Eq. SI-1.8, where we set $r_{box} = (3V/4\pi)^{1/3}$, i.e., substituting the rectangular box with a sphere of the same volume.

The probability density of finding particle B at a distance r from particle A is,

$$P(r) = \frac{\frac{1}{h^6}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{-\beta\mathcal{T}(\vec{p_A},\vec{p_B})}d\vec{p_A}d\vec{p_B}\ V\ e^{-\beta\mathcal{U}(r)}4\pi r^2}{Q} \quad , \tag{SI-1.9}$$

whereas for a random distribution this probability density is,

$$P_{\text{random}}(r) = \frac{4\pi r^2}{V} \quad . \tag{SI-1.10}$$

The radial distribution function is exactly the ratio between these two probabilities,

$$g_{AB}(r)_{N_A^\circ=N_B^\circ=1} = \frac{P(r)}{P_{\text{random}}(r)} = \frac{V}{\int_0^{r_{box}} e^{-\beta\mathcal{U}(r)}4\pi r^2 dr}\ e^{-\beta\mathcal{U}(r)} \quad , \tag{SI-1.11}$$

which can be easily solved numerically given the above mentioned LJ potential. In Fig. SI-1.1 we compare this numerical result to the three different simulations. The MC simulations produce almost identical radial distribution function to that obtained from Eq. SI-1.11. The agreement of MD-NH is again excellent, however, MD-VR displays small but significant discrepancies in line with the slight overestimation of K (Table SI-1.2).

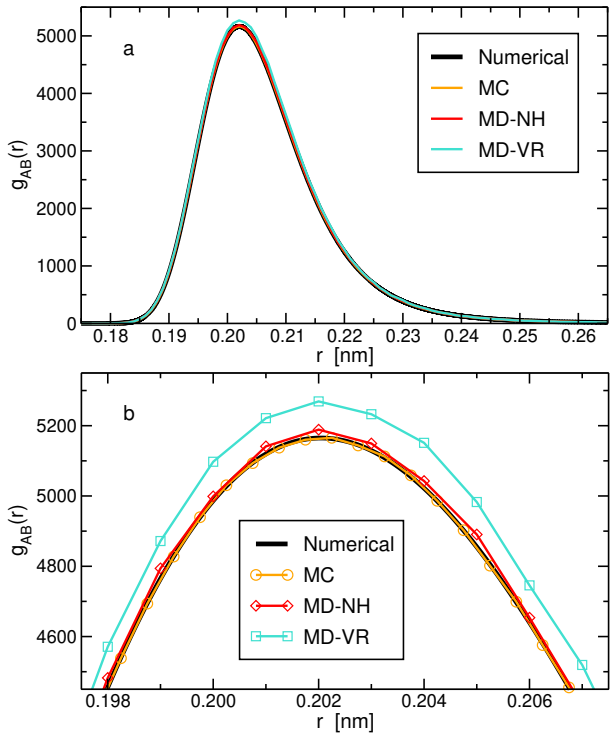

Figure SI-1.1: The radial distribution function between A and B in the single-site monomers model calculated numerically using Eq. SI-1.11, as well as, from the trajectories of the MC, MD-NH, and MD-VR simulations (a). In (b) we magnified a section around the maximum, representing the bound state, and added symbols to the plots of the simulations.

SI-2 Computational Details

The model system consists of two types of molecules where each molecule is represented by two sites, $A \equiv ah$ and $B \equiv bh$, 'covalently' bonded with a bond-length of $0.15\ nm$ as shown schematically in Fig. SI-2.1. The role of the h atoms is to prevent any clustering of the molecules, apart

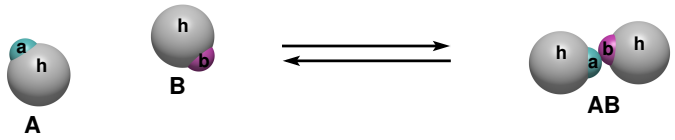

Figure SI-2.1: Simulation model for the association process between A and B molecules to produce the bound state AB. Molecule A and molecule B consist of, uncharged LJ, $a-h$ and $b-h$ atom-sites, respectively. The interaction between a and b is strongly attractive, whereas other intermolecular interactions are repulsive (see Table SI-2.1). Within each molecule, the intramolecular distance between the two atom-sites, having a value of $0.15\ nm$, is either fixed (Monte-Carlo simulations) or held together by a harmonic potential (molecular-dynamics simulations).

from product formation. All atom-sites have zero charge, $q_a = q_b = q_h = 0.0\ e$, and their intermolecular interactions are described by Lennard-Jones (LJ) potentials truncated at a distance of $2.0\ nm$. The different possible σ and ϵ parameters are specified in Table SI-2.1, yielding essentially repulsive interactions between all sites except for a strong attraction between the a and b atoms. This model results in a two-state system of unbound, $A + B$, and bound, AB, gas particles. Based on the location of the first minimum of $g_{ab}(r)$ (see Fig. SI-5.1b), the bound state is defined for $r_{ab} < 0.4\ nm$. We did not encounter any product other than this bound, AB, state in all frames of all simulations.

Periodic boundary conditions were applied along all three Cartesian axes. The total number of A molecules is denoted by $N_A^\circ = N_A + N_{AB}$ and that of B molecules by $N_B^\circ = N_B + N_{AB}$. Three main series of simulations were designed. In the first, labeled R1, we changed the value of $N_A^\circ = N_B^\circ$ from 1 to 4096, and concomitantly, the volume of the cubic simulation box, V, keeping the concentrations, $c_A^\circ = N_A^\circ/V$ and $c_B^\circ = N_B^\circ/V$, constant at $0.015625\ molecules/nm^3$

Table SI-2.1: Lennard-Jones parameters between all atom sites for a system with $A(ah)$ and $B(bh)$ molecules.

	σ $[nm]$	ϵ $[kJ/mol]$
$a\cdots a$	1.00	0.1
$b\cdots b$	1.00	0.1
$h\cdots h$	0.50	0.1
$a\cdots h$	0.35	0.1
$b\cdots h$	0.35	0.1
$a\cdots b$	0.18	30.0

($\sim 0.026\ M$). In the second series of simulations, R2, we considered only one molecule of A, $N_A^\circ = 1$, and one molecule of B, $N_B^\circ = 1$, and increased V by increasing the length of the cubic box from $L_{box} = 4.0\ nm$ to $L_{box} = 28.0\ nm$. The third series of simulations, R3, consisted of asymmetrical concentrations of the A and B molecules, in which $N_A^\circ = 1$ is fixed whereas N_B° varied from 1 to 4096, coupled to changes of V to satisfy $c_B^\circ = 0.015625\ molecules/nm^3$. In order to further examine the validity of the approximation to predict composition from K at finite systems where $N_A^\circ > 1$ (see below) a fourth series of simulations, R4, also with asymmetrical concentrations, was conducted. In this case, $N_B^\circ = 8$ and $V = 512\ nm^3$ (i.e., $c_B^\circ = 0.015625\ molecules/nm^3$) were kept constant whereas N_A° varied from 1 to 8.

All four series of simulations were performed by the Monte-Carlo (MC) technique (coded in-house in double-precision) where the canonical ensemble emerges naturally from the generated configurations[3,4]. The Metropolis acceptance criteria[5] was applied to either accept or reject trial moves. Each trial move is composed of randomly selecting one A and one B molecules which are then displaced, in each of the three Cartesian-axes, and rotated around each of the two axes perpendicular to the molecular axis. The displacements and rotations are performed, as rigid bodies, on each of the molecules separately. Their magnitudes and directions were determined randomly from a uniform distribution with maximum values of $0.4\ nm$ for displacements along each of the Cartesian-axes, 0.1 for $\cos\theta$ when rotating around angle θ ($0 \leq \theta \leq \pi$), and $0.314\ rad$ for rotations around angle ϕ ($0 \leq \phi \leq 2\pi$). These trial moves resulted in acceptance-ratios that varied from

0.313, for the system with the largest $N_A^\circ = N_B^\circ$ in R1, to 0.996, for the system with the largest L_{box} in R2. The number of trial moves applied for each simulation was inversely proportional to the size of the system. More specifically, the equilibration and data collection stages ranged from 10^4 and $1.4 \cdot 10^{12}$ moves, respectively, for the smallest system of $N_A^\circ = N_B^\circ = 1$, to 10^9 and $1.5 \cdot 10^{10}$ moves for the largest system of $N_A^\circ = N_B^\circ = 4096$.

Unless stated differently, the simulations were carried out at $T = 300\ K$. Nonetheless, we also performed the R1 series of simulations at temperatures of $200,\ 250,\ 400,\ 500,\ 600$, and $1200\ K$. Here $N_A^\circ = N_B^\circ$ ranged from 1 to 64, and the number of trial moves for data collection at the lowest two temperatures, $1.5 \cdot 10^{11}$, was three times larger than that at the highest four temperatures. Note that for the system $N_A^\circ = N_B^\circ = 1$, the average number of bound particles at the lowest temperature $(200\ K)$ is 0.997 whereas at the highest temperature $(1200\ K)$ it is 0.003, spanning a wide range of values for the equilibrium constant. This system $(N_A^\circ = N_B^\circ = 1)$ at $1200\ K$ exhibited the largest acceptance-ratio of 0.995 whereas the smallest acceptance-ratio, 0.0034, was recorded at $200\ K$ for the largest four systems. Larger systems at lower temperatures are more difficult to equilibrate and reach convergence. Nevertheless, the results presented here are converged as demonstrated in Fig. SI-2.2 and Table SI-2.2 for the most challenging system.

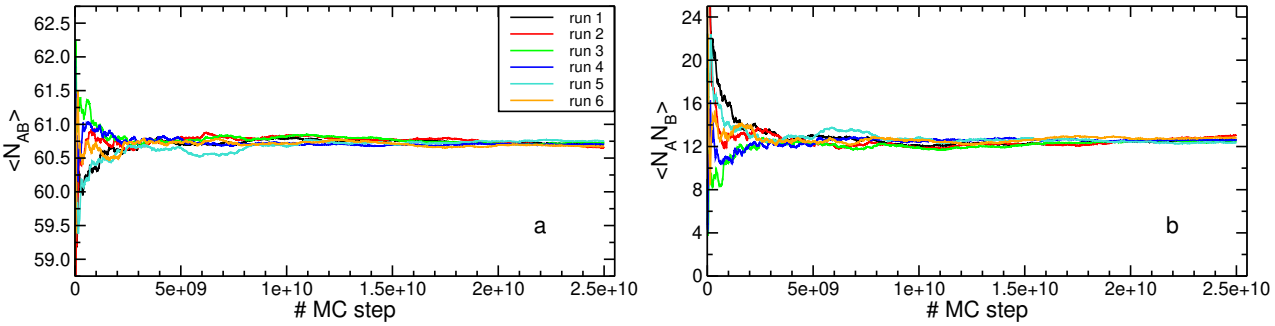

Figure SI-2.2: (a) The average number of bound particles as a function of MC steps for the system of $N_A^\circ = N_B^\circ = 64$ (R1 series) at $T = 200\ K$. Six curves corresponding to six different runs are shown where each spans $2.5 \cdot 10^{10}$ MC steps plotted every $5 \cdot 10^5$ steps. (b) The same as (a) but for the average of the product of the number of unbound A and unbound B particles.

We also attempted simulations at $T = 150\ K$, however, with the number of trial-moves specified

Table SI-2.2: Results obtained from six independent simulations, each with $2.5{\cdot}10^{10}$ MC trial-moves, for the system $N_A^\circ = N_B^\circ = 64$ of R1 series at $T = 200\ K$. The table presents the average number of bound particles, the average of the product between unbound A and unbound B particles, and the variance $\sigma^2 = L(N_{AB}, N_{AB}) = L(N_A, N_B)$ defined in Eq. 21. Unlike Fig. SI-2.2, here all averages are calculated over all MC steps.

Simulation #	$\langle N_{AB} \rangle$	$\langle N_A N_B \rangle$	σ^2
1	60.70	12.60	1.74
2	60.65	13.05	1.83
3	60.73	12.45	1.73
4	60.70	12.59	1.72
5	60.76	12.35	1.83
6	60.67	12.82	1.76

above convergence was not attained and therefore the results were not considered.

Besides MC, we also performed molecular-dynamics (MD) simulations for the R1 and R2 series utilizing the software package GROMACS version 4.6.5[6] (single-precision). A time step of $0.002\ ps$ was employed to integrate the equations of motion and a mass of $10.0\ amu$ was assigned to all atom-sites. The $a-h$ and $b-h$ 'covalent' bonds were represented by a harmonic potential with bond-length of $0.15\ nm$ and force-constant of $2 \cdot 10^5\ kJ/(mol \cdot nm^2)$. A temperature of $300.0\ K$ was maintained by applying either the Nosé-Hoover[7,8] (MD-NH) or the velocity-rescaling[9] (MD-VR) thermostats. In the first, the equations of motion were propagated by the velocity-Verlet algorithm in which the kinetic energy is determined by the average of the two half-steps (see the Gromacs manual). Due to systems with very few degrees of freedom, we applied 10 chained Nose-Hoover thermostats[10] and the coupling strength determining the friction coefficient was set to 0.1. In simulations with the second thermostat, the leap-frog algorithm was used for integrating the equations of motion and the particles' velocities were scaled with a coupling-time of $0.1\ ps$. Note that for the systems described here, MD simulations were less efficient than MC, and therefore, we applied them only to R1 and R2 (up to a box length of $L_{box} = 16.0\ nm$) series of simulations.

Equilibration time of at least $1\ \mu s$ was conducted prior to data collection for each system. For R1, the time period for collecting data ranged from $224\ \mu s$ for the smallest system to $3.84\ \mu s$ for the largest system. For R2, data was collected for $224\ \mu s$.

In order to analyze the dynamics of the forward and backward reactions we performed the R1 and R2 series of simulations by the MD-NH and MD-VR techniques again. However, this time the trajectories were saved more frequently; from a frequency of every 200 steps for $N_A^\circ = 1$ to a frequency of every step for $N_A^\circ = N_B^\circ \geq 16$. These frequencies corresponded to, approximately, the lowest frequencies for which trial calculations of the rate constants were not affected upon an increase of the trajectory-saving frequency. At the same time, the duration of trajectories were also smaller than those described above and ranged from $24\ \mu s$ for the smallest system to $12\ ns$ for the largest system. To keep the size of the trajectories manageable, each run was split into few shorter runs. The rates of the forward and backward reactions were calculated by counting the number of transitions per period of time divided by V. A transition between the two states is identified when the distance r_{ab} crossed the cutoff-value of $0.4\ nm$ plus, or minus, a distance of $0.1\ nm$ on either side of the cutoff (i.e., $0.3\ nm$ for an unbound-to-bound transition and $0.5\ nm$ for the opposite transition) to avoid counting return-trajectories originating from transient species in the proximity of the transition state.

SI-3 Supplementary Figures & Tables

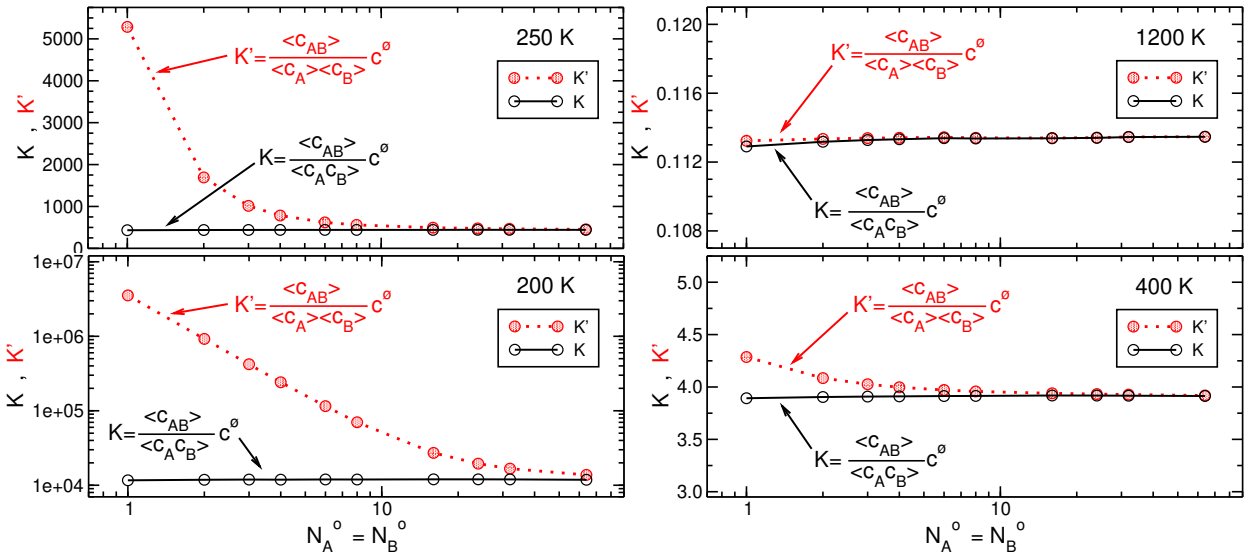

Figure SI-3.1: The equilibrium constant K defined by Eq. 16, as well as the value of K' defined by Eq. 17, from Monte-Carlo R1 series of simulations (i.e., constant $c_{_A}^{\circ} = c_{_B}^{\circ} = 0.026\ M$) at four different temperatures. Here the number of particles, $N_A^{\circ} = N_B^{\circ}$, ranges from 1 to 64. Note the scales of the y-axis are substantially different for the different temperatures and at the lowest temperature, $T = 200\ K$, is not linear but logarithmic.

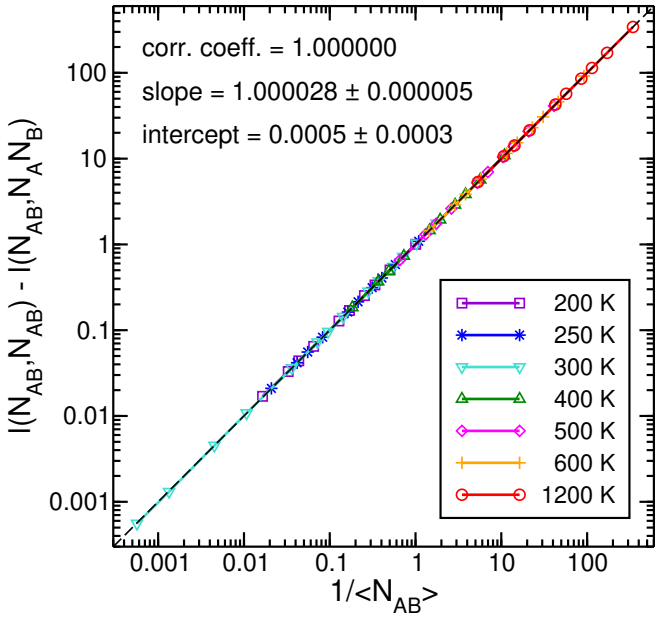

Figure SI-3.2: The difference of the relative correlations, $l(N_{AB}, N_{AB}) - l(N_{AB}, N_A N_B)$, as a function of the reciprocal average of bound AB particles for MC R1 series of simulations at different temperatures. The results at $T = 300\ K$ displayed in Fig. 5a are included here as well as a reference. Linear regression results (obtained by xmgrace) of all data points are indicated. The dashed black line is a $y = x$ line, plotted as a reference for perfect predictions.

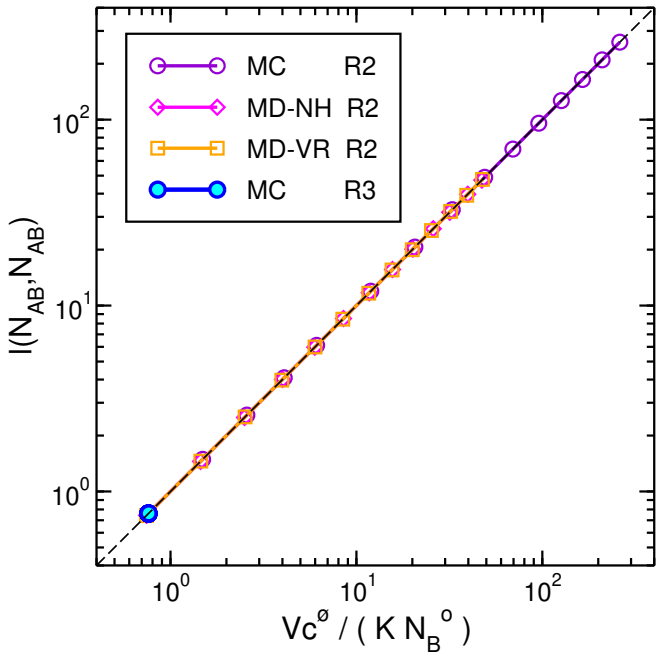

Figure SI-3.3: The expression for predicting relative fluctuations in the number of bound particles, $Vc^{\varnothing}/(KN_B^{\circ})$, for the case $N_A^{\circ} = 1$ ($N_B^{\circ} \geq N_A^{\circ}$) as described in Eq. 33, plotted against the fluctuations themselves for R2 ($N_A^{\circ} = N_B^{\circ} = 1$) and R3 ($N_A^{\circ} = 1$, $c_B^{\circ} = 0.026\ M$) series of simulations. Note that because in R3 series, the ratio V/N_B° is constant, all points in this series have the same value. Linear regression results are presented in Table SI-3.1 below.

Table SI-3.1: Linear-regression analyses (performed by xmgrace) of the predictions of the values of $l(N_{AB}, N_{AB})$ shown in Fig. SI-3.3 above for R2 series of simulations using three different simulation methods.

	Correlation coef.	Slope	Intercept
MC	1.000000	$1.000014 \pm 8 \cdot 10^{-6}$	-0.0001 ± 0.0009
MD-NH	1.000000	$1.00006 \pm 3 \cdot 10^{-5}$	0.0007 ± 0.0006
MD-VR	1.000000	$1.00011 \pm 3 \cdot 10^{-5}$	-0.0006 ± 0.0007

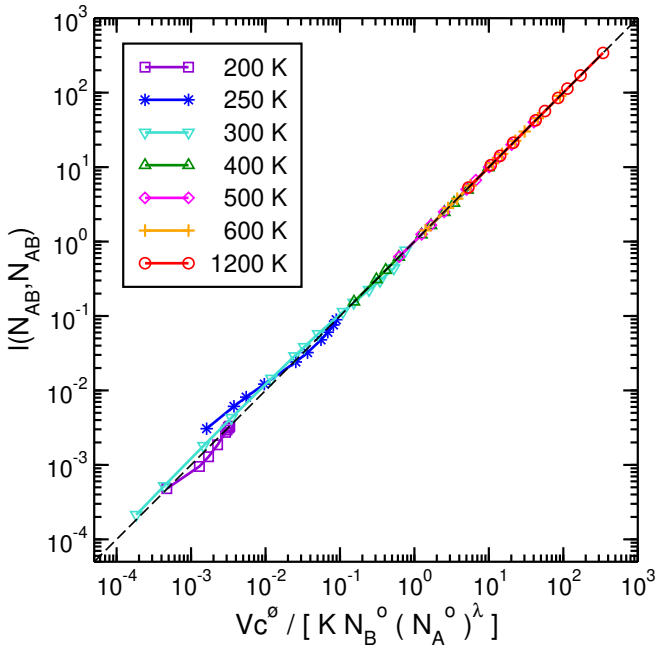

Figure SI-3.4: Approximation results obtained for MC R1 series of simulations at different temperatures. The graph displays the relative fluctuations, $l(N_{AB}, N_{AB})$, as a function of predicted values given by Eq. 34 with $\lambda = [1 + K/(Vc^{\varnothing} \ln N_B^{\circ})]^{-1}$. Linear regression results are presented in Table SI-3.2 below.

Table SI-3.2: Linear-regression analyses of the predictions of the values of $l(N_{AB}, N_{AB})$ shown in Fig. SI-3.4 above at each temperature.

	Correlation coef.	Slope	Intercept
200	0.9882321	1.07 ± 0.06	-0.0003 ± 0.0002
250	0.9948747	0.91 ± 0.03	0.001 ± 0.002
300	0.993245	0.93 ± 0.04	0.001 ± 0.013
400	0.9997421	0.992 ± 0.008	-0.02 ± 0.03
500	0.9999818	0.998 ± 0.002	-0.02 ± 0.03
600	0.9999963	0.999 ± 0.001	-0.03 ± 0.03
1200	0.9999997	0.9998 ± 0.0003	-0.03 ± 0.03

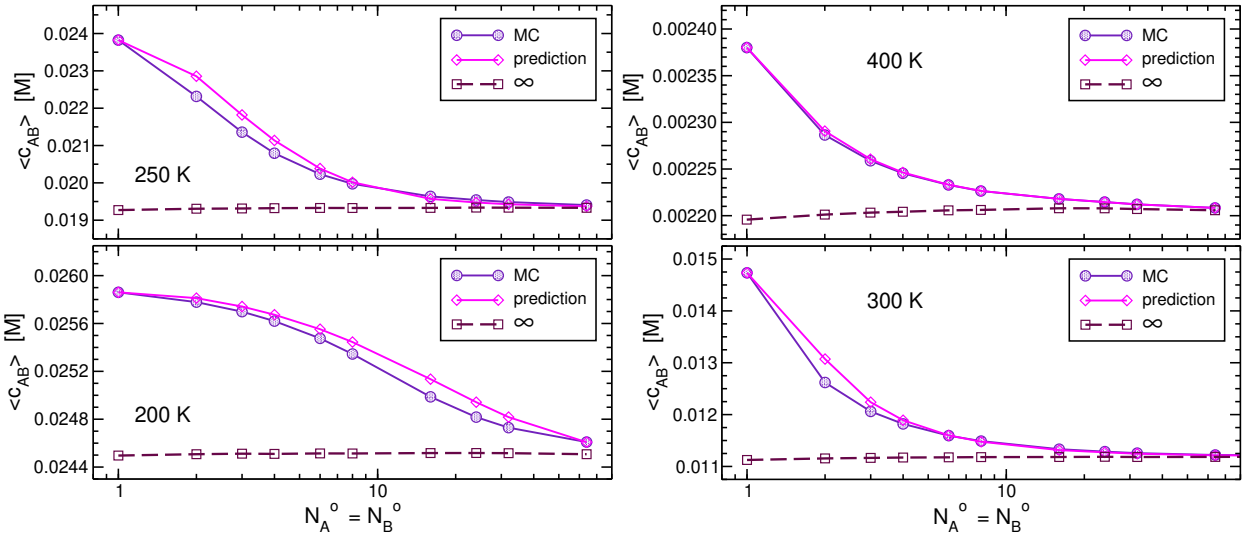

Figure SI-3.5: Concentrations of the bound AB particles calculated by Eq. 30 using approximated predictions for the values of $l(N_{AB}, N_{AB})$ as shown in Fig. SI-3.4, along with the concentrations obtained directly from the MC R1 simulations. The dashed maroon lines are the corresponding values at the thermodynamic limit, $l(N_{AB}, N_{AB}) \rightarrow 0$, calculated by Eq. 31 at each value of $N_A^\circ = N_B^\circ$. For temperatures in the rage $500 - 1200\ K$, the predictions are more accurate than those exhibited at $T = 400K$ (graphs not shown). At $T = 300\ K$, the actual curves end at $N_A^\circ = N_B^\circ = 4096$, however, the last four points are not shown because the predictions obtained are more accurate than that of the last point displayed at $N_A^\circ = N_B^\circ = 64$. At all temperatures, the predictions of the concentrations for $N_A^\circ = N_B^\circ = 1$ are almost identical to those found in the simulations because in this case, the value of the exponent ($\lambda = 0$) given in Eq. 35 makes the expression of $l(N_{AB}, N_{AB})$ in Eq. 34 exact.

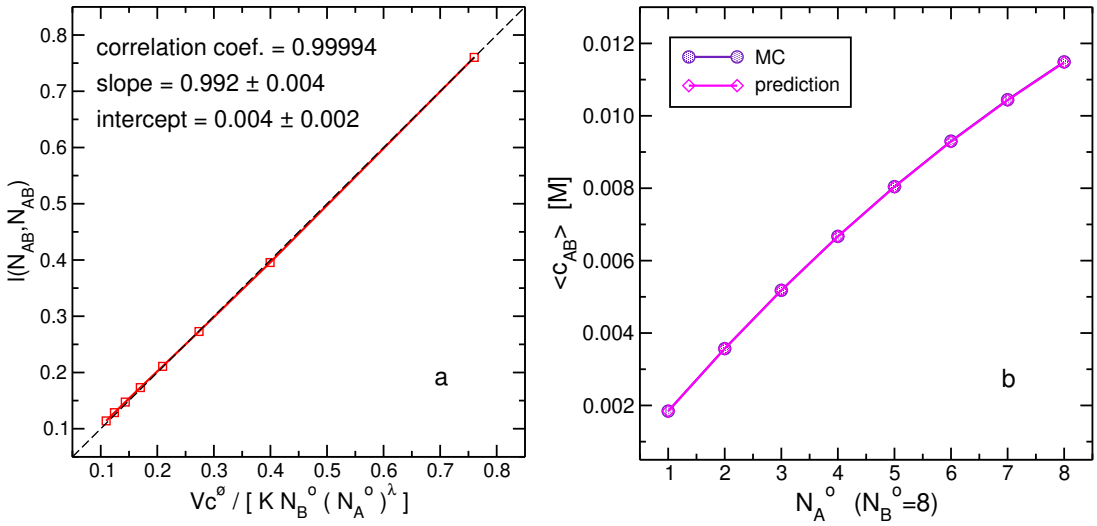

Figure SI-3.6: Approximation results obtained from MC R4 series of simulations. In this series, N_A° and N_B° are not equal and N_A° is not fixed at the value of 1. More specifically, N_A° varied from 1 to 8, whereas $N_B^\circ = 8$, $V = 512\ nm^3$, and $T = 300\ K$ are fixed. (a) The corresponding plot to that of Fig. SI-3.4 and (b) the corresponding plot to Fig. SI-3.5.

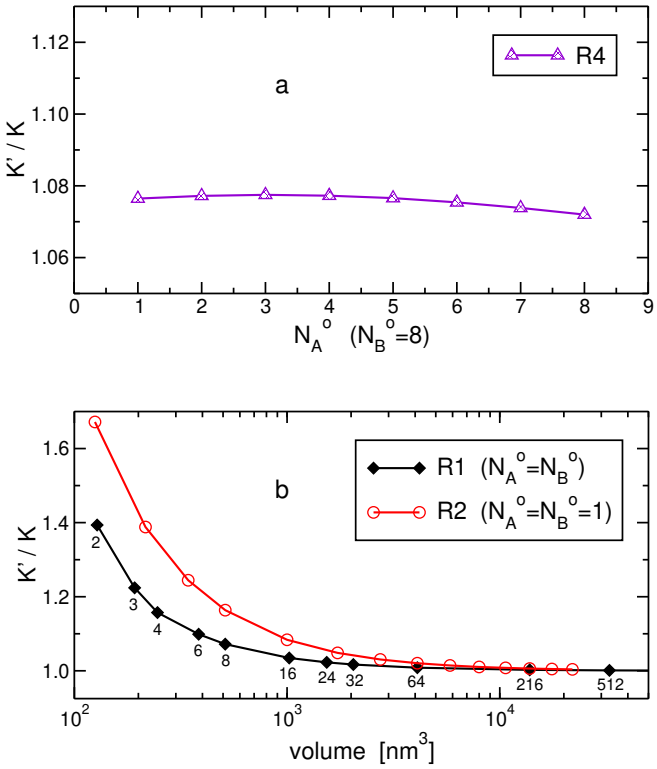

Figure SI-3.7: The ratio between K' and K, which equals $l(N_A, N_B)+1$ and thereby is a measure of correlations between the reactants, from MC simulations at $T = 300\ K$. (a) Results from R4 series where $V = 512\ nm^3$ and $N_B^\circ = 8$ are constants and only N_A° is varied. (b) Resutls from R1 and R2 series as a function of V. In both series $N_A^\circ = N_B^\circ$, however in R2 these numbers equal 1, whereas in R1 their value varies and is indicated below the symbols in the figure.

SI-4 An Alternative Derivation of the Relation between Concentrations and Fluctuations

Given the setup specified in the manuscript, i.e., a system subject to the process described in Eq. 1 in the canonical ensemble (N_A°, N_B°, V, T) where N_A° and N_B° are the total number of A and B particles, satisfying $N_A^\circ \leq N_B^\circ$. We then express the partition function of the system as,

$$Q = \sum_{i=0}^{N_A^\circ} W^i_{N_A^\circ,N_B^\circ} e^{-\beta\mathcal{H}(i)} = \sum_{i=0}^{N_A^\circ} W^i_{N_A^\circ,N_B^\circ} e^{-\beta[\mathcal{T}+\mathcal{U}(i)]} = \sum_{i=0}^{N_A^\circ} W^i_{N_A^\circ,N_B^\circ} e^{-\beta[\mathcal{T}+i\epsilon_{AB}]} \quad , \qquad \text{(SI-4.1)}$$

where as before, we mapped the sum over energy states onto the sum over $i \equiv N_{AB}$, the number of bound AB particles. The Hamiltonian of the system, $\mathcal{H}(i)$, along with its potential energy component, $\mathcal{U}(i)$, are functions of i, whereas the kinetic energy term, $\mathcal{T}$, is not. In the last equality, $\mathcal{U}(i)$ is expressed explicitly by the energy liberated upon the formation of i bound AB particles, and for simplicity we assume no other intra-molecular potential energy terms. The term $W^i_{N_A^\circ,N_B^\circ}$ which corrects the overcounting due to the indistinguishable character of the particles is defined in Eq. 4.

We start by expressing* $\langle N_{AB}^2 \rangle$,

$$\begin{aligned}
\langle N_{AB}^2 \rangle &= \frac{1}{Q}\sum_{i=0}^{N_A^\circ} i^2\, W^i_{N_A^\circ,N_B^\circ} e^{-\beta\mathcal{H}(i)} = -\frac{1}{Q}\frac{1}{\epsilon_{AB}}\left[\frac{\partial}{\partial\beta}\sum_{i=0}^{N_A^\circ} i\, W^i_{N_A^\circ,N_B^\circ} e^{-\beta\mathcal{H}(i)} + \sum_{i=0}^{N_A^\circ} i\,\mathcal{T} W^i_{N_A^\circ,N_B^\circ} e^{-\beta\mathcal{H}(i)}\right] \\
&= -\frac{1}{\epsilon_{AB}}\left[\frac{1}{Q}\frac{\partial}{\partial\beta}\left(\langle N_{AB}\rangle Q\right) + \langle N_{AB}\mathcal{T}\rangle\right] \\
&= -\frac{1}{\epsilon_{AB}}\left[\frac{\partial\langle N_{AB}\rangle}{\partial\beta} + \frac{\langle N_{AB}\rangle}{Q}\frac{\partial Q}{\partial\beta} + \langle N_{AB}\mathcal{T}\rangle\right] \\
&= -\frac{1}{\epsilon_{AB}}\left[\frac{\partial\langle N_{AB}\rangle}{\partial\beta} + \langle N_{AB}\rangle\frac{\partial \ln Q}{\partial\beta} + \langle N_{AB}\mathcal{T}\rangle\right] \\
&= -\frac{1}{\epsilon_{AB}}\left[\frac{\partial\langle N_{AB}\rangle}{\partial\beta} - \epsilon_{AB}\langle N_{AB}\rangle^2 - \langle N_{AB}\rangle\langle\mathcal{T}\rangle + \langle N_{AB}\mathcal{T}\rangle\right] \\
&= -\frac{1}{\epsilon_{AB}}\frac{\partial\langle N_{AB}\rangle}{\partial\beta} + \langle N_{AB}\rangle^2 \quad , \qquad \text{(SI-4.2)}
\end{aligned}$$

*When writing partial derivatives we will omit the specification of the parameters which are kept constant. That means, in our case of the canonical ensemble, partial derivatives with respect to temperature are taken when N_A°, N_B°, and V are constant.

where the last equality is obtained because, by definition, the value of the kinetic energy in the canonical ensemble is constant. Then, the fluctuations in the number of bound AB particles can be expressed by,

$$L(N_{AB}, N_{AB}) = -\frac{1}{\epsilon_{AB}} \frac{\partial \langle N_{AB} \rangle}{\partial \beta} \quad . \tag{SI-4.3}$$

Using the relation in Eq. 16 we write,

$$\begin{aligned} L(N_{AB}, N_{AB}) &= -\frac{1}{\epsilon_{AB}} \frac{\partial \left[K \langle N_A N_B \rangle \right]}{V c^{\varnothing} \partial \beta} = -\frac{K}{\epsilon_{AB} V c^{\varnothing}} \left[\langle N_A N_B \rangle \frac{1}{K} \frac{\partial K}{\partial \beta} + \frac{\partial \langle N_A N_B \rangle}{\partial \beta} \right] \\ &= -\frac{\langle N_{AB} \rangle}{\epsilon_{AB}} \frac{\partial \ln K}{\partial \beta} - \frac{K}{\epsilon_{AB} V c^{\varnothing}} \frac{\partial \langle N_A N_B \rangle}{\partial \beta} \quad . \end{aligned} \tag{SI-4.4}$$

We now evaluate the first partial derivative after the last equality by using the definition of K in Eq. 5,

$$\begin{aligned} \left(\frac{\partial \ln K}{\partial \beta} \right)_V &= -\frac{1}{R} \frac{\partial (\Delta G^{\varnothing}/T)}{\partial \beta} = -\frac{1}{R} \frac{\partial (\Delta F^{\varnothing}/T + V \Delta P^{\varnothing}/T)}{\partial \beta} = -\frac{1}{R} \frac{\partial (\Delta F^{\varnothing}/T)}{\partial \beta} \\ &= \frac{T^2}{N_{\text{Avogadro}}} \frac{\partial (\Delta F^{\varnothing}/T)}{\partial T} = -\frac{\Delta U^{\varnothing}}{N_{\text{Avogadro}}} = -\epsilon_{AB} \quad . \end{aligned} \tag{SI-4.5}$$

The third equality in Eq. SI-4.5 holds for ideal gases, $V \Delta P^{\varnothing}/T = R \Delta n^{\varnothing}$ (for reactions described by Eq. 1, the change in the number of moles of gas particles under standard conditions, $\Delta n^{\varnothing}$, equals 1) and for reactions in solution where the change in pressure is negligible, $V \Delta P^{\varnothing} \simeq 0$. It is worth pointing that Eq. SI-4.5 is the equivalent of the van't Hoff relation, which is applicable at constant pressure, to processes at constant volume.

Next, we evaluate the second partial derivative after the last equality in Eq. SI-4.4,

$$\begin{aligned} &\frac{\partial \langle N_A N_B \rangle}{\partial \beta} = \frac{\partial}{\partial \beta} \left[\frac{1}{Q} \sum_{i=0}^{N_A^\circ} (N_A^\circ - i)(N_B^\circ - i) \; W^i_{N_A^\circ, N_B^\circ} e^{-\beta \mathcal{H}(i)} \right] \\ &= -\frac{1}{Q} \sum_{i=0}^{N_A^\circ} \mathcal{H}(i) N_A(i) N_B(i) \; W^i_{N_A^\circ, N_B^\circ} e^{-\beta \mathcal{H}(i)} - \frac{1}{Q^2} \frac{\partial Q}{\partial \beta} \sum_{i=0}^{N_A^\circ} N_A(i) N_B(i) \; W^i_{N_A^\circ, N_B^\circ} e^{-\beta \mathcal{H}(i)} \quad , \end{aligned} \tag{SI-4.6}$$

where $N_A(i) = (N_A^\circ - i)$ designates the number of unbound A particles, and a corresponding

notation, $N_B(i)$, designates the unbound B particles. We continue,

$$\begin{aligned}\frac{\partial\langle N_A N_B\rangle}{\partial\beta} &= -\langle\mathcal{H}N_A N_B\rangle - \frac{\partial \ln Q}{\partial\beta}\langle N_A N_B\rangle = -\langle(\mathcal{T}+\mathcal{U})N_A N_B\rangle + \langle\mathcal{H}\rangle\langle N_A N_B\rangle \\ &= -\langle\mathcal{T}N_A N_B\rangle - \langle\epsilon_{AB}N_{AB}N_A N_B\rangle + \langle\mathcal{T}+\epsilon_{AB}N_{AB}\rangle\langle N_A N_B\rangle \\ &= -\epsilon_{AB}\langle N_{AB}N_A N_B\rangle + \epsilon_{AB}\langle N_{AB}\rangle\langle N_A N_B\rangle = -\epsilon_{AB}L(N_{AB}, N_A N_B) \quad . \quad \text{(SI-4.7)}\end{aligned}$$

Again, because the kinetic energy in the canonical ensemble is constant, the two terms containing the value of $\mathcal{T}$ cancel each other. Now we take the results obtained in Eq. SI-4.5 and Eq. SI-4.7 and insert them into Eq. SI-4.4 to calculate $L(N_{AB}, N_{AB})$,

$$L(N_{AB}, N_{AB}) = \langle N_{AB}\rangle + \frac{K}{Vc^{\varnothing}}L(N_{AB}, N_A N_B) = \langle N_{AB}\rangle + \frac{\langle N_{AB}\rangle}{\langle N_A N_B\rangle}L(N_{AB}, N_A N_B) \quad . \quad \text{(SI-4.8)}$$

If we divide both sides of Eq. SI-4.8 by $\langle N_{AB}\rangle^2$ we can express a relation between two relative deviations as,

$$l(N_{AB}, N_{AB}) = \frac{1}{\langle N_{AB}\rangle} + l(N_{AB}, N_A N_B) \quad , \quad \text{(SI-4.9)}$$

which is identical to Eq. 27.

SI-5 Transformations between g(r) of Systems with Different Sizes

As demonstrated in Fig. 4c, the radial distribution function of the product, $g_{ab}(r)$, depends on the system size even if the total concentrations of the A and B particles (c_A° and c_B°) are not altered. Obviously, this is because the equilibrium concentrations do depend on the size of the system. However, because we can predict $\langle c_{_{AB}} \rangle$ for a macroscopic system from a finite-system (Eq. 31), we can perform the corresponding transformation for $g_{ab}(r)$. If the formation of trimers can be ignored, as we actively prevented in our model, the transformation of $g_{ab}(r)$ for the bound state, i.e. for distances around the first-minumum and lower, $r < r_{_{fm}}$, can be performed by using the ratio of the concentrations as a scaling-factor,

$$g_{ab}(r)_\infty = g_{ab}(r)_{finite} \cdot \frac{\langle c_{_{AB}} \rangle_\infty}{\langle c_{_{AB}} \rangle_{finite}} \qquad \text{for} \quad r < r_{_{fm}} \qquad . \tag{SI-5.1}$$

On the other hand, the scaling-factor for larger distances, $r \geq r_{_{fm}}$, is different. To obtain it, we calculate the probability of finding a and b sites at distances $r \geq r_{_{fm}}$ apart,

$$P_{ab}(r \geq r_{_{fm}}) = \frac{N_A^\circ N_B^\circ - \langle N_{AB} \rangle}{N_A^\circ N_B^\circ} = 1 - \frac{\langle N_{AB} \rangle}{N_A^\circ N_B^\circ} \qquad , \tag{SI-5.2}$$

where we subtracted in the numerator the average number of bound particles from the overall possible number of pairs. We consider this probability, for both, finite and macroscopic systems. For the latter case, given $-\beta\epsilon_{_{AB}}$ is not too large, we have $P_{ab}(r \geq r_{_{fm}})_\infty \to 1$, thus $g_{ab}(r)_\infty$ for distances larger than the first-minimum, $r \geq r_{_{fm}}$, can be obtained by,

$$g_{ab}(r)_\infty = g_{ab}(r)_{finite} \cdot \frac{P_{ab}(r \geq r_{_{fm}})_\infty}{P_{ab}(r \geq r_{_{fm}})_{finite}} = g_{ab}(r)_{finite} \cdot \frac{1}{1 - \frac{\langle N_{AB} \rangle_{finite}}{(N_A^\circ N_B^\circ)_{finite}}} \qquad \text{for} \quad r \geq r_{_{fm}} \qquad . \tag{SI-5.3}$$

This conversion of $g_{ab}(r)_{finite}$ obtained at a finite system to that of a macroscopic system is demonstrated in Fig. SI-5.1 utilizing Eq. 31 to calculate $\langle c_{_{AB}} \rangle_\infty$. Although the region describing the bound state and the unbound state are very well reproduced, the transition region, not surprisingly, is not. In addition in this transition region, the conversion from the system of $N_A^\circ = N_B^\circ = 1$ exhibits larger deviations compare to those from any other finite systems. Plausibly because this system, $N_A^\circ = N_B^\circ = 1$, is the only one that does not contain the pure repulsion between the like-type sites (thus between $a \cdots a$ or between $b \cdots b$ sites).

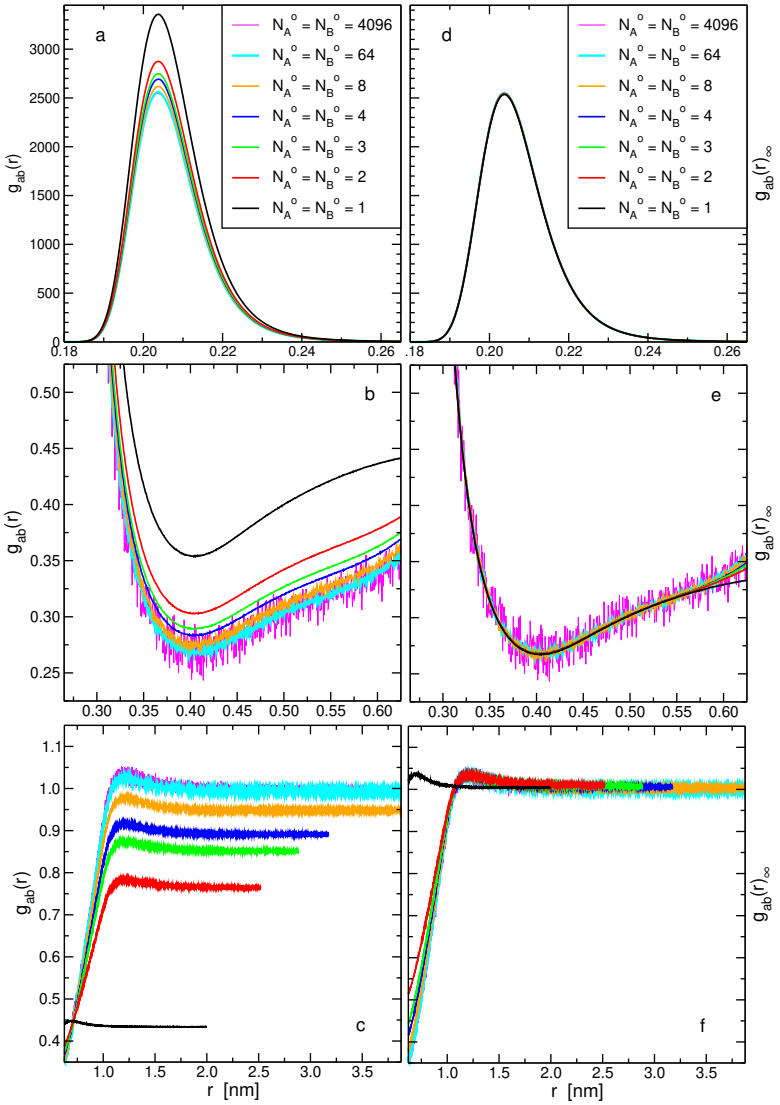

Figure SI-5.1: Transformations of the radial distribution functions, $g_{ab}(r)$ (of R1 MC simulations, also displayed in Fig. 4c), obtained at different system sizes and shown on the left panel (a)-(c), to a corresponding distribution of a system with an infinite-size, $g_{ab}(r)_\infty$, shown on the right panel (d)-(f). The segment of the distribution up to around the minimum defining the bound state, $r < 0.625\ nm$ thus (a) and (b), is converted by applying the ratio of the bound-state concentrations in the two systems as the scaling factor (Eq. SI-5.1). The segment of the distribution with larger values of r, (c), is converted according to Eq. SI-5.3. These transformations break-down in the range, $0.57\ nm < r < 1.00\ nm$, whereas for $N_A^\circ = 1$ it is not valid for a wider range, up to $r \sim 1.5\ nm$. The x-axes in (a) and (d) end at a point where the x-axes of (b) and (e) start, and the latter end at a point where the x-axes of (c) and (f) start.

References

[1] W. G. McMillan and J. E. Mayer, *J. Chem. Phys.*, 1945, **13**, 276–305.

[2] D. A. McQuarrie, *Statistical Thermodynamics*, University Science Books, Mill Valley, CA, 1973.

[3] M. P. Allen and D. J. Tildesley, *Computer Simulations of Liquids*, Oxford Science Publications, Oxford, 1987.

[4] D. Frenkel and B. Smit, *Understanding Molecular Simulations: From Algorithms to Applications*, Academic Press, London, 2002.

[5] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *J. Chem. Phys.*, 1953, **21**, 1087–1092.

[6] B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, *J. Chem. Theory Comput.*, 2008, **4**, 435–447.

[7] S. Nosé, *J. Chem. Phys.*, 1984, **81**, 511–519.

[8] W. G. Hoove, *Phys. Rev. A*, 1985, **31**, 1695–1697.

[9] G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, 014101.

[10] G. J. Martyna, M. L. Klein and M. Tuckerman, *J. Chem. Phys.*, 1992, **97**, 2635–2643.