Multiple linear regression and thermodynamic fluctuations are equivalent for computing thermodynamic derivatives from molecular simulation

Rahbari, A.; Josephson, T.R.; Sun, Y.; Moultos, O.A.; Dubbeldam, D.; Siepmann, J.I.; Vlugt, T.J.H.

DOI
10.1016/j.fluid.2020.112785

Publication date
2020

Document Version
Final published version

Published in
Fluid Phase Equilibria

Fluid Phase Equilibria 523 (2020) 112785

Contents lists available at ScienceDirect

Fluid Phase Equilibria

journal homepage: www.elsevier.com/locate/fluid

Multiple linear regression and thermodynamic fluctuations are equivalent for computing thermodynamic derivatives from molecular simulation

Ahmadreza Rahbari[a], Tyler R. Josephson[b], Yangzesheng Sun[b], Othonas A. Moultos[a], David Dubbeldam[c], J. Ilja Siepmann[b], Thijs J.H. Vlugt[a,*]

[a] Engineering Thermodynamics, Process & Energy Department, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Leeghwaterstraat 39, Delft, 2628CB, the Netherlands
[b] Department of Chemistry and Chemical Theory Center, University of Minnesota, 207 Pleasant Street SE, Minneapolis, MN 55455-0431, USA
[c] Van't Hoff Institute for Molecular Sciences, University of Amsterdam, Science Park 904, 1098XH Amsterdam, the Netherlands

ARTICLE INFO

Article history:
Received 16 March 2020
Revised 6 July 2020
Accepted 9 August 2020
Available online 11 August 2020

Keywords:
Molecular simulation
Partial molar properties
Open ensembles thermodynamic fluctuations
Linear regression

ABSTRACT

Partial molar properties are of fundamental importance for understanding properties of non-ideal mixtures. Josephson and co-workers (Mol. Phys. 2019, 117, 3589–3602) used least squares multiple linear regression to obtain partial molar properties in open constant-pressure ensembles. Assuming composition-independent partial molar properties for the narrow composition range encountered throughout simulation trajectories, we rigorously prove the equivalence of two approaches for computing thermodynamic derivatives in open ensembles of an n-component system: (1) multiple linear regression, and (2) thermodynamic fluctuations. Multiple linear regression provides a conceptually simple and computationally efficient way of computing thermodynamic derivatives for multicomponent systems. We show that in the reaction ensemble, the reaction enthalpy can be computed directly by simple multiple linear regression of the enthalpy as a function of the number of reactant molecules. Non-linear regression and a Gaussian process model taking into account the compositional dependence of partial molar properties further support that multiple linear regression captures the correct physics.

1. Introduction

Partial molar properties are important quantities for describing multicomponent non-ideal mixtures [1–3]. The partial molar property of component i in a mixture is defined as $\overline{x_i} = (\partial X/\partial n_i)_{T,P,n_{j\neq i}}$, in which X is the corresponding extensive property of the mixture and n_i is the number of moles (or number of molecules) of component i [4]. Alternatively, X is defined as the composition-weighted sum of partial molar properties of the constituent components in the system: $X = \sum_i n_i \overline{x_i}$ [5]. Experimentally, obtaining partial molar properties at extreme conditions is difficult (*i.e.* high temperatures and pressures) [6–9]. Computation of partial molar properties using molecular simulation [5,10,11] is not straightforward, as partial molar properties cannot be determined as a function of atomic positions or momenta of a single configuration in a system [5,10,12–14]. To date, different approaches have been developed based on: (1) direct numerical differentiation [1,10,15], (2) Widom's test particle insertion method [10,12,13], (3) Kirkwood-Buff integrals [16–21], and (4) expanded ensembles [12,22]. Recently, Josephson et al. computed partial molar properties by fitting extensive thermodynamic properties (*e.g.*, enthalpy or volume) as a function of the instantaneous number of molecules of each component [5]. For example, partial molar enthalpies in a binary system, $\overline{h_1}$ and $\overline{h_2}$, can be obtained by fitting the equation $H = n_1\overline{h_1} + n_2\overline{h_2}$ to the instantaneous enthalpy H as a function of the number of molecules of components 1 and 2 (n_1 and n_2, respectively). This requires an ensemble in which the number of molecules of each component fluctuates (*e.g.* the *NPT* version of the Gibbs ensemble [23,24], reaction ensemble [25–27], or grand-canonical ensemble [28,29]). The advantage of this method is that simulation or legacy data (*e.g.*, $H(n_1, n_2)$) can be easily fitted to the linear regression model without additional requirements. The required simulation data files are small compared to trajectory files containing all particle positions [5]. The partial molar properties can be computed from these data files using a single line of code in software such as MATLAB [30].

* Corresponding author.
E-mail address: t.j.h.vlugt@tudelft.nl (T.J.H. Vlugt).

https://doi.org/10.1016/j.fluid.2020.112785

For single-component adsorption in the grand-canonical ensemble, one can fit a linear regression model to simulation data to obtain thermodynamic derivatives [31,32]. In these simulations, partial derivatives of thermodynamic properties can also be obtained from fluctuations [31]. In grand-canonical simulations, chemical potentials, temperature and the volume are imposed, and therefore the number of molecules of each component fluctuates around their average. The relative magnitude of these fluctuations, $\sqrt{\langle n_i^2\rangle - \langle n_i\rangle^2}/\langle n_i\rangle$, scales with $V^{-1/2}$ [11,33]. Similarly, the total numbers of molecules for each species or the total numbers of atoms of each type are imposed in Gibbs ensemble and reaction ensemble simulations, respectively, and the relative magnitude of the fluctuations scales with $n_i^{-1/2}$. For typical simulations, the composition fluctuations are sufficiently small so that within a given simulation, partial properties do not change with composition. The relative composition fluctuations can be large for minority species, and caution would be needed for species that may change aggregation state over these fluctuations. In case of single-component adsorption in the grand-canonical ensemble, it follows from statistical mechanics that $(\partial n/\partial(\beta\mu))_{V,T} = \langle n^2\rangle - \langle n\rangle^2$ and $(\partial U/\partial(\beta\mu))_{V,T} = \langle Un\rangle - \langle U\rangle\langle n\rangle$ in which n is the number of molecules, μ is the imposed chemical potential, V is the volume, U is the potential energy of the adsorbed phase, and $\beta = (k_B T)^{-1}$ [31–35]. The brackets $< \ldots >$ denote ensemble averages. The partial derivative of the total energy of the adsorbed phase with respect to the number of molecules is calculated from [33]:

$$\left(\frac{\partial U}{\partial n}\right)_{V,T} = \left(\frac{(\partial U/\partial(\beta\mu))_{V,T}}{(\partial n/\partial(\beta\mu))_{V,T}}\right) = \frac{\langle Un\rangle - \langle U\rangle\langle n\rangle}{\langle n^2\rangle - \langle n\rangle^2} \tag{1}$$

These fluctuations are encountered in a single simulation of an open ensemble, and do not span the entire range from zero to saturation loading. Fitting the grand-canonical simulation data (*i.e.*, the instantaneous value of U as a function of n) to the simple linear regression model ($U = b_0 + b_1 n + \epsilon$, in which b_0 and b_1 are constants and ϵ represents noise) leads to an identical expression for the slope (b_1) as in Eq. (1) [36], showing the equivalence of thermodynamic fluctuations and linear regression for single-component adsorption in the grand-canonical ensemble. It is important to note that $(\partial U/\partial n)_{V,T} \neq U/n$ [37]. This is due to changes in adsorption behavior as a function of n, as already adsorbed molecules change the adsorption energy of newly adsorbed molecules. For example, this is the case when there is an inflection in the adsorption isotherm [38–40]. At the inflection point, adsorbed molecules start to fill a new adsorption site, or sorbate-sorbate interactions start to play an important role [41], which leads to a change in $(\partial U/\partial n)_{V,T}$. Therefore, when fitting a simple linear regression model to simulation data in the grand-canonical ensemble, the intercept should always be included in the regression model [37].

For multicomponent systems, analytic expressions for $(\partial U/\partial n_i)_{V,T,n_{j\neq i}}$ from fluctuations are complicated [31]. Therefore, comparing the expressions for partial derivatives obtained from fluctuations and multiple linear regression becomes tedious. Here, we show that for an n-component system, identical thermodynamic derivatives are obtained from (1) fitting simulation data to a multiple linear regression model [5,36], and (2) thermodynamic fluctuations [31]. It is important to note that thermodynamic derivatives obtained from regression using an arbitrary function may not necessarily be equal to those obtained from fluctuations [31,35]. For example, for single component adsorption in the grand-canonical ensemble, a polynomial regression of the second order ($U = b_0 + b_1 n + b_2 n^2 + \epsilon$) leads to a different expression for $(\partial U/\partial n_i)_{V,T,n_{j\neq i}}$ [36]:

$$\left(\frac{\partial U}{\partial n}\right)_{V,T} = \times \frac{(\langle Un\rangle - \langle U\rangle\langle n\rangle)\left(\langle n^4\rangle - \langle n^2\rangle^2\right) - \left(\langle Un^2\rangle - \langle U\rangle\langle n^2\rangle\right)\left(\langle n^3\rangle - \langle n\rangle\langle n^2\rangle\right)}{\left(\langle n^2\rangle - \langle n\rangle^2\right)\left(\langle n^4\rangle - \langle n^2\rangle^2\right) - \left(\langle n^3\rangle - \langle n\rangle\langle n^2\rangle\right)^2} \tag{2}$$

which differs from Eq. (1). Derivatives of a fitted function are not necessarily the same as derivatives of the function itself (*e.g.*, fitting $y = x^2$ with linear regression leads to different derivatives). For the general case of k independent variables $x_1, x_2, \ldots, x_k$, the dependent variable y can be fitted with the following multiple linear regression model:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik} + \epsilon \tag{3}$$

in which i indicates an observation from the sample [36]. Solving the regression of Eq. (3) using the least squares method leads to the following set of equations [36]

$$\begin{array}{lllllll}
b_0 N_s & + b_1 \sum_{i=1}^{N_s} x_{1i} & + b_2 \sum_{i=1}^{N_s} x_{2i} & + \ldots + b_k \sum_{i=1}^{N_s} x_{ki} & = & \sum_{i=1}^{N_s} y_i \\
b_0 \sum_{i=1}^{N_s} x_{1i} + b_1 \sum_{i=1}^{N_s} x_{1i}^2 & + b_2 \sum_{i=1}^{N_s} x_{1i}x_{2i} + \ldots + b_k \sum_{i=1}^{N_s} x_{1i}x_{ki} & = & \sum_{i=1}^{N_s} x_{1i}y_i \\
b_0 \sum_{i=1}^{N_s} x_{2i} + b_1 \sum_{i=1}^{N_s} x_{2i}x_{1i} + b_2 \sum_{i=1}^{N_s} x_{2i}^2 & + \ldots + b_k \sum_{i=1}^{N_s} x_{2i}x_{ki} & = & \sum_{i=1}^{N_s} x_{2i}y_i \\
\vdots \quad + \quad \vdots + \quad \vdots \quad + \quad \vdots & \quad \vdots & := & \vdots \\
b_0 \sum_{i=1}^{N_s} x_{ki} + b_1 \sum_{i=1}^{N_s} x_{ki}x_{1i} + b_2 \sum_{i=1}^{N_s} x_{ki}x_{2i} + \ldots + b_k \sum_{i=1}^{N_s} x_{ki}^2 & & = & \sum_{i=1}^{N_s} x_{ki}y_i
\end{array} \tag{4}$$

in which N_s is the sample size. The least squares estimates of the parameters $b_0, b_1, \ldots, b_k$ are obtained by solving Eq. (4) using any appropriate method [36]. It can be shown that the intercept b_0 is obtained from [36]

$$b_0 = \langle y\rangle - b_1\langle x_1\rangle - b_2\langle x_2\rangle \ldots - b_k\langle x_k\rangle \tag{5}$$

in which $< \ldots >$ denotes the sample mean. The numerical results by Josephson et al. [5] strongly suggest that for multicomponent systems, fluctuations and multiple linear regression lead to identical values for thermodynamic derivatives. To the best of our knowledge, there has not been a formal proof showing thermodynamic derivatives obtained from multiple linear regression are thermodynamically consistent with the results obtained from fluctuations in open ensembles. Josephson et al. [5] derived the multiple linear regression expression using the composition-weighted sum definition of the partial molar property based on arguments from continuum thermodynamics. Here, we show this to be equivalent to an alternative derivation from statistical thermodynamics, thereby unifying continuum and statistical thermodynamics. This is analogous to the derivation of the Langmuir adsorption isotherm, which was originally derived using a continuum treatment of adsorption kinetics [42], and was later derived using statistical thermodynamics [43]. We show that multiple linear regression provides a conceptually simple way of computing thermodynamic derivatives in multicomponent systems. We demonstrate this by first deriving this for constant-volume open ensembles, and then extend the result to constant-pressure open ensembles. Depending on the nature of the extensive quantity being fit, multiple linear regression can be used to obtain either residual or total partial molar properties.

This manuscript is organized as follows. In Section 2, we show that performing multiple linear regression and fluctuations lead to

identical expressions for partial properties in the grand-canonical ensemble (constant volume). For two-component mixtures, analytic expressions for $(\partial U/\partial n_i)_{V,T,n_{j\neq i}}$ are provided. In Section 3, the application is extended to systems in which the pressure is constant (instead of constant volume). In Section 4, the compositional dependence of partial molar properties is examined using nonlinear regression and a Gaussian process model [44]. Our conclusions are summarized in Section 5.

2. Fluctuations in the grand-canonical ensemble

For multicomponent adsorption in the grand-canonical ensemble, the derivative $(\partial U/\partial n_i)_{V,T,n_{j\neq i}}$ is obtained from [31]

$$\left(\frac{\partial U}{\partial n_i}\right)_{V,T,n_{j\neq i}} = \sum_{i=1}^{k}\left(\frac{\partial U}{\partial(\beta\mu_k)}\right)_{V,T,\mu_{j\neq k}}\left(\frac{\partial(\beta\mu_k)}{\partial n_i}\right)_{V,T,n_{j\neq i}} \tag{6}$$

where k is the number of components. The second term on the right hand side of Eq. (6) is the change in chemical potential of component i while the number of molecules of all other components are fixed. Using a matrix notation, Eq. (6) can be written as

$$\begin{bmatrix}\left(\frac{\partial U}{\partial n_1}\right)_{n_{j\neq1}}\\ \left(\frac{\partial U}{\partial n_2}\right)_{n_{j\neq2}}\\ \vdots\\ \left(\frac{\partial U}{\partial n_k}\right)_{n_{j\neq k}}\end{bmatrix} = \underbrace{\begin{bmatrix}\left(\frac{\partial(\beta\mu_1)}{\partial n_1}\right)_{n_{j\neq1}} & \left(\frac{\partial(\beta\mu_1)}{\partial n_2}\right)_{n_{j\neq2}} & \cdots & \left(\frac{\partial(\beta\mu_1)}{\partial n_k}\right)_{n_{j\neq k}}\\ \left(\frac{\partial(\beta\mu_2)}{\partial n_1}\right)_{n_{j\neq1}} & \left(\frac{\partial(\beta\mu_2)}{\partial n_2}\right)_{n_{j\neq2}} & \cdots & \vdots\\ \vdots & & \ddots & \vdots\\ \left(\frac{\partial(\beta\mu_k)}{\partial n_1}\right)_{n_{j\neq1}} & \cdots & \cdots & \left(\frac{\partial(\beta\mu_k)}{\partial n_k}\right)_{n_{j\neq k}}\end{bmatrix}}_{\mathbf{M}}\begin{bmatrix}\left(\frac{\partial U}{\partial(\beta\mu_1)}\right)_{\mu_{j\neq1}}\\ \left(\frac{\partial U}{\partial(\beta\mu_2)}\right)_{\mu_{j\neq2}}\\ \vdots\\ \left(\frac{\partial U}{\partial(\beta\mu_k)}\right)_{\mu_{j\neq k}}\end{bmatrix} \tag{7}$$

in which we have defined matrix **M**. The constant temperature and volume notation is dropped to make the equations more compact. In the grand-canonical ensemble, the derivatives $(\partial U/\partial(\beta\mu_i))_{\mu_{j\neq i}}$ in Eq. (7) can expressed as the covariance between U and n_i [31,32,35]:

$$\left(\frac{\partial U}{\partial(\beta\mu_i)}\right)_{\mu_{j\neq i}} = f(U,n_i) \tag{8}$$

where $f(U, n_i)$ equals $\langle Un_i\rangle - \langle U\rangle\langle n_i\rangle$. The terms $f(U, n_i)$ can be calculated directly from simulations in the grand-canonical ensemble. However, the derivatives $(\partial(\beta\mu_i)/\partial n_l)_{n_{j\neq l}}$ in the matrix **M** (Eq. (7)) cannot be directly calculated as ensemble averages, as the chemical potential is imposed. The inverse of the matrix **M** is obtained directly from the reciprocals of the derivatives in **M** [45]. Refs. [31,45] show that

$$\underbrace{\begin{bmatrix}\left(\frac{\partial(\beta\mu_1)}{\partial n_1}\right)_{n_{j\neq1}} & \left(\frac{\partial(\beta\mu_1)}{\partial n_2}\right)_{n_{j\neq2}} & \cdots & \left(\frac{\partial(\beta\mu_1)}{\partial n_k}\right)_{n_{j\neq k}}\\ \left(\frac{\partial(\beta\mu_2)}{\partial n_1}\right)_{n_{j\neq1}} & \left(\frac{\partial(\beta\mu_2)}{\partial n_2}\right)_{n_{j\neq2}} & \cdots & \vdots\\ \vdots & & \ddots & \vdots\\ \left(\frac{\partial(\beta\mu_k)}{\partial n_1}\right)_{n_{j\neq1}} & \cdots & \cdots & \left(\frac{\partial(\beta\mu_k)}{\partial n_k}\right)_{n_{j\neq k}}\end{bmatrix}}_{\mathbf{M}}\underbrace{\begin{bmatrix}\left(\frac{\partial n_1}{\partial(\beta\mu_1)}\right)_{\mu_{j\neq1}} & \left(\frac{\partial n_1}{\partial(\beta\mu_2)}\right)_{\mu_{j\neq2}} & \cdots & \left(\frac{\partial n_1}{\partial(\beta\mu_k)}\right)_{\mu_{j\neq k}}\\ \left(\frac{\partial n_2}{\partial(\beta\mu_1)}\right)_{\mu_{j\neq1}} & \left(\frac{\partial n_2}{\partial(\beta\mu_2)}\right)_{\mu_{j\neq2}} & \cdots & \vdots\\ \vdots & & \ddots & \vdots\\ \left(\frac{\partial n_k}{\partial(\beta\mu_1)}\right)_{\mu_{j\neq1}} & \cdots & \cdots & \left(\frac{\partial n_k}{\partial(\beta\mu_k)}\right)_{\mu_{j\neq k}}\end{bmatrix}}_{\mathbf{M}^{-1}} = \mathbf{I} \tag{9}$$

where **I** is the identity matrix. The advantage of this formulation is that the elements of the matrix $\mathbf{M}^{-1}$ can be calculated directly as a function of ensemble averages in the grand-canonical ensemble [31,45]:

$$\left(\frac{\partial n_l}{\partial(\beta\mu_i)}\right)_{\mu_{j\neq i}} = f(n_i,n_l) \tag{10}$$

where $f(n_i, n_l)$ denotes the covariance $\langle n_in_l\rangle - \langle n_i\rangle\langle n_l\rangle$. By multiplying both sides of Eq. (7) with the inverse matrix $\mathbf{M}^{-1}$ we have

$$\begin{bmatrix}\left(\frac{\partial n_1}{\partial(\beta\mu_1)}\right)_{\mu_{j\neq1}} & \left(\frac{\partial n_1}{\partial(\beta\mu_2)}\right)_{\mu_{j\neq2}} & \cdots & \left(\frac{\partial n_1}{\partial(\beta\mu_k)}\right)_{\mu_{j\neq k}}\\ \left(\frac{\partial n_2}{\partial(\beta\mu_1)}\right)_{\mu_{j\neq1}} & \left(\frac{\partial n_2}{\partial(\beta\mu_2)}\right)_{\mu_{j\neq2}} & \cdots & \vdots\\ \vdots & & \ddots & \vdots\\ \left(\frac{\partial n_k}{\partial(\beta\mu_1)}\right)_{\mu_{j\neq1}} & \cdots & \cdots & \left(\frac{\partial n_k}{\partial(\beta\mu_k)}\right)_{\mu_{j\neq k}}\end{bmatrix}\begin{bmatrix}\left(\frac{\partial U}{\partial n_1}\right)_{n_{j\neq1}}\\ \left(\frac{\partial U}{\partial n_2}\right)_{n_{j\neq2}}\\ \vdots\\ \left(\frac{\partial U}{\partial n_k}\right)_{n_{j\neq k}}\end{bmatrix} = \begin{bmatrix}\left(\frac{\partial U}{\partial(\beta\mu_1)}\right)_{\mu_{j\neq1}}\\ \left(\frac{\partial U}{\partial(\beta\mu_2)}\right)_{\mu_{j\neq2}}\\ \vdots\\ \left(\frac{\partial U}{\partial(\beta\mu_k)}\right)_{\mu_{j\neq k}}\end{bmatrix} \tag{11}$$

Rewriting Eq. (11) using Eqs. (8) and (10) leads to

$$\begin{bmatrix}f(n_1,n_1) & f(n_2,n_1) & \cdots & f(n_k,n_1)\\ f(n_1,n_2) & f(n_2,n_2) & \cdots & f(n_k,n_2)\\ \vdots & & & \vdots\\ f(n_1,n_k) & f(n_2,n_k) & \cdots & f(n_k,n_k)\end{bmatrix}\begin{bmatrix}d_1\\ d_2\\ \vdots\\ d_k\end{bmatrix} = \begin{bmatrix}f(U,n_1)\\ f(U,n_2)\\ \vdots\\ f(U,n_k)\end{bmatrix} \tag{12}$$

To make the equations more compact, the terms $d_{i\in[1,k]}$ are used to denote the derivatives $(\partial U/\partial n_i)_{n_{j\neq i}}$. We can show that Eq. (12) is identical to the set of equations for multiple linear regression (Eq. (4)) [36]. Writing out the l^{th} row of Eq. (12) leads to

$$d_1f(n_1,n_l) + d_2f(n_2,n_l)\ldots + d_kf(n_k,n_l) = f(U,n_l) \tag{13}$$

Combining Eqs. (13) and (8), (10) leads to

$$\langle U\rangle\langle n_l\rangle - [d_1\langle n_1\rangle\langle n_l\rangle + d_2\langle n_2\rangle\langle n_l\rangle\ldots + d_l\langle n_l\rangle\langle n_l\rangle\ldots + d_k\langle n_k\rangle\langle n_l\rangle] + [d_1\langle n_1n_l\rangle + d_2\langle n_2n_l\rangle\ldots + d_l\langle n_ln_l\rangle\ldots + d_k\langle n_kn_l\rangle] = \langle Un_k\rangle \tag{14}$$

which can be rearranged to

$$\langle n_l\rangle\underbrace{[\langle U\rangle - (d_1\langle n_1\rangle + d_2\langle n_2\rangle\ldots + d_k\langle n_k\rangle)]}_{d_0} + d_1\langle n_1n_l\rangle + d_2\langle n_2n_l\rangle \ldots + d_l\langle n_ln_l\rangle\ldots + d_k\langle n_kn_l\rangle = \langle Un_k\rangle \tag{15}$$

In Eq. (15), we have defined the term d_0. For a sample size of N_s, this leads to the following independent linear equation

$$N_s d_0 + d_1 \sum_{i=1}^{N_s} n_{1i} + d_2 \sum_{i=1}^{N_s} n_{2i} \ldots + d_k \sum_{i=1}^{N_s} n_{ki} = \sum_{i=1}^{N_s} U_i \tag{16}$$

Multiplying Eq. (15) with the sample size N_s and rearranging leads to

$$d_0 \sum_{i=1}^{N_s} n_{li} + d_1 \sum_{i=1}^{N_s} n_{li} n_{1i} + d_2 \sum_{i=1}^{N_s} n_{li} n_{2i} \ldots + d_l \sum_{i=1}^{N_s} n_{li}^2 \ldots + d_k \sum_{i=1}^{N_s} n_{li} n_{ki} = \sum_{i=1}^{N_s} U_i n_{li} \tag{17}$$

Eqs. (16) and (17) are identical to the equations for least squares multiple linear regression as in Eq. (4). Therefore, it is shown that identical expressions are obtained for (1) thermodynamic derivatives using least square multiple linear regression, and (2) thermodynamic fluctuations in the grand-canonical ensemble. While we demonstrate this equivalence for calculation of $(\partial U/\partial n_i)_{V,T,n_{j\neq i}}$, any extensive property X may be substituted for U.

For a two component system, it is still possible to solve Eq. (6) without resorting to matrix algebra. Using multiple linear regression ($U = b_0 + b_1 n_1 + b_2 n_2 + \epsilon$), the partial derivatives in the grand-canonical ensemble can be written as

$$\left(\frac{\partial U}{\partial n_1}\right)_{V,T,n_2} = \frac{(\langle U n_1\rangle - \langle U\rangle\langle n_1\rangle)\left(\langle n_2^2\rangle - \langle n_2\rangle^2\right) - (\langle U n_2\rangle - \langle U\rangle\langle n_2\rangle)(\langle n_1 n_2\rangle - \langle n_1\rangle\langle n_2\rangle)}{\left(\langle n_1^2\rangle - \langle n_1\rangle^2\right)\left(\langle n_2^2\rangle - \langle n_2\rangle^2\right) - (\langle n_1 n_2\rangle - \langle n_1\rangle\langle n_2\rangle)^2} \tag{18}$$

and

$$\left(\frac{\partial U}{\partial n_2}\right)_{V,T,n_1} = \frac{(\langle U n_2\rangle - \langle U\rangle\langle n_2\rangle)\left(\langle n_1^2\rangle - \langle n_1\rangle^2\right) - (\langle U n_1\rangle - \langle U\rangle\langle n_1\rangle)(\langle n_1 n_2\rangle - \langle n_1\rangle\langle n_2\rangle)}{\left(\langle n_1^2\rangle - \langle n_1\rangle^2\right)\left(\langle n_2^2\rangle - \langle n_2\rangle^2\right) - (\langle n_1 n_2\rangle - \langle n_1\rangle\langle n_2\rangle)^2} \tag{19}$$

Similar to Eq. (5), the intercept is obtained from

$$b_0 = \langle U\rangle - b_1\langle n_1\rangle - b_2\langle n_2\rangle \tag{20}$$

It is important to note that computing fluctuations in the number of molecules in the grand-canonical ensemble is not always computationally efficient [11]. This is for example the case at high molecule densities (where insertion and removal of molecules is difficult [11]), or near inflection points in the adsorption isotherm (where $(\partial n/\partial\mu) \approx 0$) [38–40]. In those cases, one may consider linear regression on a collection of *NVT* ensembles with varying number of molecules, which is similar to the numerical differentiation approach [1,10,15]. We will not consider this further here.

3. Application to constant pressure ensembles

In the previous sections, we have considered systems at constant volume. Partial molar properties are defined as partial derivatives of thermodynamic properties with respect to the number of molecules of component i at constant pressure, while keeping the number of molecules of all other components constant [1,10]. To calculate partial molar properties, one can either perform simulations directly at constant pressure, or perform a translation from a constant volume ensemble to a constant pressure ensemble as explained in Ref. [16]. For convenience, we choose to perform our simulations in constant pressure ensembles [46,47], such as the reactive isothermal-isobaric ensemble [25–27], or the isothermal-isobaric version of the Gibbs ensemble [24]. It is not practical to perform simulations in which only the pressure, temperature and the chemical potential are fixed (μPT ensemble) [11], as the pressure, temperature, and chemical potential are all intensive variables. In this ensemble, the size of the system in the simulation may decrease or increase without bound, making simulations generally unstable [48]. To impose an upper bound on the system size, at least one extensive variable should be fixed (*e.g.* in the grand-canonical ensemble, the volume is fixed). To constrain the system size at constant pressure, there are several possibilities: (1) Performing simulation in the *NPT* version of the Gibbs Ensemble (where the total number of molecules is fixed) [11,24]; (2) Performing simulations in the *NPT* version of the reaction ensemble (total number of atoms is fixed) [25,27,49–51]. The reaction ensemble can also be considered as a grand-canonical ensemble in which the chemical potentials of reactants and reaction products are imposed in such a way that chemical equilibrium is obtained [25–27]. To calculate the partial derivatives of thermodynamic properties at constant pressure, one can fit the multiple linear regression model (Eq. (4)) to the simulation data [5]. Since the partial derivatives in Eqs. (8) and (10) do not change when the pressure is kept constant instead of the volume, the resulting expressions from fluctuations (Eq. (6)) do not change either. At constant pressure, it is possible to show that the intercept from the linear regression model will be zero [5]. As an example, one can write the instantaneous total volume as the sum of the composition-weighted partial molar volumes (k components) [5]

$$V = n_1\overline{\upsilon_1} + n_2\overline{\upsilon_2} \ldots + n_k\overline{\upsilon_k} \tag{21}$$

in which $\overline{\upsilon_i}$ is the partial molar volume of component i. By fitting the least squares multiple linear regression model, ($V = n_0 + n_1\overline{\upsilon_1} + n_2\overline{\upsilon_2} + \ldots + n_k\overline{\upsilon_k} + \epsilon$, in which ϵ represents noise), the expression for the intercept follows from 5:

$$n_0 = \langle V\rangle - \overline{\upsilon_1}\langle n_1\rangle - \overline{\upsilon_2}\langle n_2\rangle - \ldots - \overline{\upsilon_k}\langle n_k\rangle \tag{22}$$

From Eqs. (21) and (22) it follows directly that the intercept n_0 should be zero, and this can be used as a test for the numerical calculations. We have carried out such a test for the systems in this study at constant pressure and observed that the values obtained for the intercept are statistically indistinguishable from zero. Nonetheless, we recommend constraining the intercept to be zero when capturing partial molar properties.

When applied to the reaction ensemble [26,27,49,50,52–55], the regression method proposed by Josephson et al. [5] can be used to calculate the reaction enthalpy without computing partial molar enthalpies of each species in the mixture. As an example, we consider the ammonia synthesis reaction ($N_2 + 3H_2 \rightleftharpoons 2NH_3$) from our earlier work [12]. To facilitate reaction trial moves, the Continuous Fractional Component (CFC) version of the reaction ensemble was used [49]. The reaction enthalpy of the ammonia synthesis reaction is obtained directly by fitting the multiple linear regression model to instantaneous data obtained from a single simulation. Here, enthalpy of the system (H) and the number of nitrogen molecules (n_{N_2}) are related by

$$H = b_0 + b_1 n_{N_2} + \epsilon \tag{23}$$

in which b_1 and b_0 are the slope and intercept of the fitted regression line, respectively, and ϵ is the error term. Just as in the grand-canonical ensemble, the intercept b_0 will not be zero here. From simple linear regression (Eq. (1)) it follows that

$$b_1 = \frac{\langle H n_{N_2}\rangle_{RxMC} - \langle H\rangle_{RxMC}\langle n_{N_2}\rangle_{RxMC}}{\langle n_{N_2}^2\rangle_{RxMC} - \langle n_{N_2}\rangle_{RxMC}^2} \tag{24}$$

Table 1
Residual reaction enthalpy ($\Delta\bar{h}$) of the ammonia synthesis reaction, per mole of N_2, at $T = 573$ K and $P = 400$ bar and $P = 800$ bar, computed both from simulations in the reaction ensemble (RxMC) [49] using Eq. (24), and the CFC-NPT ensemble [12,22]. The reported residual reaction enthalpies are relative to the reaction enthalpy at ideal gas conditions, as the contributions from the enthalpies of formation of N_2, H_2, and NH_3 are not included here in $\Delta\bar{h}$. The magnitude of the fluctuations for the number of ammonia molecules, $\xi_{NH_3} = \sqrt{\langle n_{NH_3}^2\rangle - \langle n_{NH_3}\rangle^2}/\langle n_{NH_3}\rangle$, in the reaction ensemble is provided. Numbers in brackets are standard deviations of average volumes $\langle V\rangle$, from 5 independent simulations.

	RxMC [49]			CFC-*NPT* [12]
P/[bar]	$\Delta\bar{h}$/[kJ mol^{-1}]	ξ_{NH_3}	$\langle V\rangle$/Å^3	$\Delta\bar{h}$/[kJ mol^{-1}]
400	−45.6(6)	0.0120(3)	81,800(200)	−45(2)
800	−75(1)	0.0070(1)	40,000(100)	−75(2)

in which the brackets $\langle\ldots\rangle_{RxMC}$ denote an ensemble average in the reaction ensemble. The slope in Eq. (24) is the reaction enthalpy per mole of nitrogen ($b_1 = \Delta\bar{h}$). Similarly, the reaction enthalpy can be calculated by fitting the total enthalpy of the system as a function of the number of ammonia or hydrogen molecules. Since the fluctuations of both reactants and reaction products depend on the extent of the reaction, the final results will be identical. To verify Eq. (24), simulations of the ammonia synthesis reaction at $P = 400$ bar and $P = 800$ bar are performed in the CFC version of the reaction ensemble at $T = 573$ K in a similar manner as described in Ref. [49]. The results are compared to the reaction enthalpies obtained from multiple simulations in the Continuous Fractional Component *NPT* (CFC-*NPT*) ensemble as in Ref. [12]. In these simulations, the reaction enthalpy follows from the partial molar enthalpies of all reactants and reaction products in the system, which are computed separately. For simulation details, the reader is referred to Refs. [12,49]. The resulting residual reaction enthalpies are shown in Table 1. The residual $\Delta\bar{h}$ reported in Table 1 is with respect to the ideal gas contribution (102.07 kJ per mole of N_2 [12]). At $P = 400$ bar and $P = 800$ bar, the residual $\Delta\bar{h}$ values are non-zero which shows non-ideal behavior of the system at high pressures. Excellent agreement is observed between the results from the CFC-*NPT* ensemble simulations and direct computation of the reaction enthalpy using least squares linear regression. As shown in Table 1, the relative magnitude of the fluctuations in the reaction ensemble, encountered throughout a single simulation is small. Therefore, partial molar properties do not change in these simulations. It is important to note that for simulations both in the reaction and CFC-*NPT* ensembles, the enthalpy of the system includes contributions from so-called fractional molecules [35]. When the number of fractional molecules is less than 1% of the total number of molecules (as is the case here), the presence of fractional molecules does not significantly (given current computational resources) affect the computed ensemble averages [56].

4. Compositional dependence of partial molar properties

The simplistic linear regression model Eq. (3) assumes constant partial molar properties. However, this assumption can be removed to allow for more flexibility in the choice of the model [5]. Non-linear models can detect and characterize non-linearities in the system. A non-linear model that still yields approximately constant partial molar properties would further strengthen the justification that the linear model is correct for the subset of compositions encountered in simulations of a system with more than 100 molecules for the minority component. Of course, for the global composition space, compositional dependence of partial molar properties cannot be neglected.

More generally, a function $Y = F(\mathbf{n})$ predicting an extensive thermodynamic property Y from the numbers of molecules in the system can be regressed using the trajectory of a simulation, starting from an equilibrated state. When $F(\mathbf{n})$ is a non-linear function, the partial molar properties obtained from regressing the trajectory will be dependent on the composition of the system. Therefore, a more robust alternative compared to linear regression, for example, Gaussian processes [44,57], may be employed as the functional form of $F(\mathbf{n})$ in a purely data-driven perspective. It is important to note that $F(\mathbf{n})$ cannot be arbitrary as Y is an extensive property. When $F(\mathbf{n})$ is a Gaussian process, the predicted extensive property may not double when the number of molecules in the system is doubled, making the model thermodynamically inconsistent. Therefore, the fundamental mathematical feature of an extensive property should be considered when constructing the functional form of $F(\mathbf{n})$, that is, it needs to be homogeneous of $\mathbf{n}$ with degree 1, that is,

$$\alpha F(\mathbf{n}) = F(\alpha\mathbf{n}) \tag{25}$$

Although $F(\mathbf{n})$ is thermodynamically constrained, an arbitrary differentiable function operating on the mole fractions of the system $f(\mathbf{x})$ can be used to satisfy this constraint, that is,

$$Y = F(\mathbf{n}) = Nf\left(\frac{\mathbf{n}}{N}\right) = Nf(\mathrm{x}) \tag{26}$$

where $N = \sum_{j=1}^{k} n_k$, and N is the total number of molecules in the system for each frame in the simulation trajectory. The non-linearity in $f(\mathbf{x})$ allows for composition dependence of partial molar properties, and $f(\mathbf{x})$ can be any type of model including polynomials, neural networks [58], or Gaussian processes [44,57,59]. Since $f(\mathbf{x}) = Y/N$, it is fitted by regressing the molar property as a function of the mole fractions.

The partial molar properties can be analytically evaluated given the partial derivatives (gradient) of $f(\mathbf{x})$. It is important to note that since mole fractions of n components sum to 1, the partial molar volumes cannot be simply evaluated as differentiating $f(\mathbf{x})$ with respect to molar fractions [16]. In Ref. [1], it is shown that the partial molar property of Y for the ith component (assuming constant T and P) can be expressed as

$$y_i = \left(\frac{\partial Y}{\partial n_i}\right)_{P,T,n_{j\neq i}} = \left(\frac{\partial (Nf(\mathbf{x}))}{\partial n_i}\right)_{P,T,n_{j\neq i}} \tag{27}$$
$$= f(\mathbf{x}) + \left(\frac{\partial f(\mathbf{x})}{\partial x_i}\right)_{P,T,x_{j\neq i}} - \sum_{j=1}^{k} x_j \left(\frac{\partial f(\mathbf{x})}{\partial x_j}\right)_{P,T,x_{j\neq i}}$$

For a linear function of $f(\mathbf{x})$, y_i is a constant number because the second term and the third term always cancel. When a linear combination of mole fractions is used for $f(\mathbf{x})$, the well-known approach of calculating partial molar properties from Eqs. (27) and (26) is equivalent to multiple linear regression (Eq. (21)). Although the regression data for Eq. (26) are the mole fractions instead of the numbers of molecules, the effect of the fluctuating total number of molecules is negligible, and Eqs. (26) and (21) yield statistically indistinguishable partial molar properties.

As a comparison with linear regression, the partial molar properties for a natural gas condensate system containing methane, *n*-butane, and *n*-decane [5] were also calculated using a quadratic function using

$$f(\mathbf{x}) = \mathbf{x}\mathbf{C_2}\mathbf{x}^T + \mathbf{x}\mathbf{C_1} + C_0 \tag{28}$$

and a Gaussian process [44,57] using

$$f(\mathbf{x}) = \mathrm{GP}(\mathbf{x}) \tag{29}$$

Both models take the vector of mole fractions as the input and outputs the extensive property Y divided by the total number of

Table 2
Partial molar properties (including intramolecular potential energy contributions but excluding kinetic energy terms) from regressing simulation trajectories of natural gas condensates. Since the conformational distributions of large, flexible molecules may differ between liquid and vapor phases (*i.e.*, good and poor solvents) [60], the internal potential energy and enthalpy need to include the intramolecular potential energy contributions. The thermodynamic constraints for the Gibbs ensemble of the ternary mixture were $N_{C1} = 1276$, $N_{C4} = 425$, and $N_{C10} = 125$, $P = 16.22$ MPa, and $T = 333$ K. In the two-box Gibbs ensemble, the number fluctuations in the two boxes are identical, and the relative fluctuations are: $\xi_{C1} = 0.109_4$, $\xi_{C4} = 0.065_2$, and $\xi_{C10} = 0.034_1$ in the liquid phase and $\xi_{C1} = 0.092_4$, $\xi_{C4} = 0.176_5$, and $\xi_{C10} = 0.50_1$ in the gas phase. The mean values and uncertainties were obtained from 64 independent simulations. The input data for regression contains instantaneous values of the relevant properties from for each independent simulation at 100-cycle intervals. Parameters for quadratic regression were obtained using the analytical form of least-squares regression and hyperparameters for Gaussian process were selected using 5-fold cross-validation. For comparison, molar properties of the pure species are also provided for the stable phase at the same temperature and pressure as the ternary VLE simulations: $N_{C1} = 1000$, $N_{C4} = 800$, or $N_{C10} = 320$, and $P = 16.22$ MPa, and $T = 333$ K in 16 independent *NPT* simulations (with 50,000 MC cycles each). Uncertainties are reported as the standard error of the mean from 64 or 16 independent simulations.

		Linear regression Sections 2 and 3	Quadratic regression Eq. (28)	Gaussian process Eq. (29)	Pure component
Liquid $\overline{V_i}$	C1	0.0892_4	0.0891_4	0.0889_4	–
$[m^3/mol] \times 10^{-3}$	C4	0.0927_9	0.0927_9	0.0929_9	0.102970_{10}
	C10	0.140_2	0.135_3	0.139_2	0.196578_{14}
MAD $[nm^3]$		1.26_1	1.26_1	1.26_2	
Vapor $\overline{V_i}$	C1	0.1535_2	0.1520_5	0.1535_2	0.15221_3
$[m^3/mol] \times 10^{-3}$	C4	-0.016_1	-0.015_1	-0.016_1	–
	C10	-0.253_6	-0.253_6	-0.265_7	–
MAD $[nm^3]$		3.23_1	3.22_1	3.24_4	
Liquid $\overline{U_i}$	C1	-0.63_2	-0.63_2	-0.63_2	–
[kJ/mol]	C4	-11.38_7	-11.39_8	-11.40_8	-11.084_2
	C10	-17.8_1	-18.1_3	-17.8_1	-16.316
MAD [K]		$1.712_8 \times 10^2$	$1.710_8 \times 10^2$	$1.72_1 \times 10^2$	
Vapor $\overline{U_i}$	C1	-1.09_1	-1.12_1	-1.08_1	-1.5990_4
[kJ/mol]	C4	-8.26_5	-8.24_6	-8.26_6	–
	C10	-12.9_2	-12.9_2	-12.0_2	–
MAD [K]		94_1	94_1	94_1	
Liquid $\overline{H_i}$	C1	0.79_2	0.78_3	0.78_3	–
[kJ/mol]	C4	-9.7_1	-9.8_1	-9.8_1	-9.420_4
	C10	-15.7_2	-16.0_4	-15.6_2	-13.134_{14}
MAD [K]		$2.67_2 \times 10^2$	$2.67_2 \times 10^2$	$2.67_3 \times 10^2$	
Vapor $\overline{H_i}$	C1	1.42_1	1.38_3	1.42_1	0.8716_{13}
[kJ/mol]	C4	-8.6_1	-8.6_1	-8.6_1	–
	C10	-16.8_3	-16.8_3	-16.2_4	–
MAD [K]		$2.33_2 \times 10^2$	$2.33_2 \times 10^2$	$2.33_3 \times 10^2$	

molecules in the system. The parameters in Eq. (28) (quadratic model) were fitted using the analytic solution of polynomial least squares regression, and the functional form of Eq. (29) (Gaussian process) was calculated from the set of training data points in each cross-validation split [44]. The partial molar properties were calculated using Eq. (27), treating the mole fractions as independent variables. The code used for computing the partial molar properties can be downloaded from https://github.com/SiepmannGroup/PartialMolarProperties.

Table 2 lists the partial molar properties and mean absolute deviations (MAD) obtained by linear regression, quadratic regression, and a Gaussian process. In Table 2, MAD is the measure of how accurately the linear (or nonlinear) model captures the extensive property and it corresponds to the magnitude of the fluctuations in the system. For all cases except for the partial molar enthalpy in the vapor box where the Gaussian Process overfits each trajectory, all three models achieved almost perfect agreement in regressing the partial molar properties of the system. These results demonstrate that the quadratic model and Gaussian process have learned the same pattern as the linear model from the simulation trajectories. This shows the physical correctness of the linear regression method for the subset of composition space sampled by the simulations. It should be noted that the partial molar volumes for butane and decane in the gas phase are negative. Negative molar volumes are impossible for pure compounds at any state point, but can be observed for non-ideal mixtures. For comparison, we also carried out single-phase simulations for pure methane, butane, and decane in the isobaric-isothermal ensemble. The chosen state point is above the critical point for pure methane and this phase is considered as vapor here, but butane and decane are found in stable liquid phases. Comparing the molar properties for pure systems to the partial molar properties in the corresponding phase (see Table 2) shows statistically different values for all molecules.

5. Conclusions

We have shown that in the grand-canonical ensemble, expressions for thermodynamic derivatives obtained from least squares multiple linear regression are identical to the expressions obtained from thermodynamic fluctuations. This provides a conceptually simple and computationally efficient approach to obtain thermodynamic properties from fluctuations in multicomponent systems. Multiple linear regression is thermodynamically consistent with fluctuations both in constant-volume and constant-pressure ensembles. We also show in the reaction ensemble that the reaction enthalpy can be obtained directly from a single simulation by fitting the enthalpy as a function of the number of reactant molecules with simple linear regression. In this work, we have assumed that composition range due to fluctuations encountered in a given simulation is small, so that partial molar properties are locally constant within a given simulation, noting that partial molar properties in non-ideal mixtures are not constant but vary over the global composition space. Nonlinear regression models capable of capturing compositional dependence of partial molar properties do not perform better than the linear model when applied to simu-

lations at a single state point, thus providing strong support that multiple linear regression captures the essential physics.

Declaration of Competing Interest

None.

Acknowledgments

This work was supported by NWO Exacte Wetenschappen (Physical Sciences) for the use of supercomputer facilities, with financial support from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organization for Scientific Research, NWO). TJHV acknowledges NWO-CW for a VICI grant. This work was also supported by the Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences, under Award DE-FG02-17ER16362 (TJR, YS, and JIS). Computational resources from the Minnesota Supercomputing Institute are also gratefully acknowledged.

References

[1] S.I. Sandler, Chemical, Biochemical, and Engineering Thermodynamics, fourth ed., John Wiley & Sons, Hoboken, N.J., USA, 2006.
[2] M.J. Moran, H.N. Shapiro, Fundamentals of Engineering Thermodynamics, fifth ed., John Wiley & Sons, West Sussex, England, 2006.
[3] S.K. Schnell, P. Englebienne, J.-M. Simon, P. Krüger, S.P. Balaji, S. Kjelstrup, D. Bedeaux, A. Bardow, T.J.H. Vlugt, How to apply the Kirkwood–Buff theory to individual species in salt solutions, Chem. Phys. Lett 582 (2013) 154–157.
[4] J.M. Smith, H.C. Van Ness, M.M. Abbott, Introduction to Chemical Engineering Thermodynamics, seventh ed., McGraw-Hill, New York, USA, 2005.
[5] T.R. Josephson, R. Singh, M.S. Minkara, E.O. Fetisov, J.I. Siepmann, Partial molar properties from molecular simulation using multiple linear regression, Mol. Phys. 117 (2019) 3589–3602.
[6] I.M. Abdulagatov, A.R. Bazaev, E.A. Bazaev, M.B. Saidakhmedova, A.E. Ramazanova, Volumetric properties of near-critical and supercritical water + pentane mixtures: molar, excess, partial, and apparent volumes, J. Chem. Eng. Data 43 (1998) 451–458.
[7] I.M. Abdulagatov, A.R. Bazaev, E.A. Bazaev, S.P. Khokhlachev, M.B. Saidakhmedova, A.E. Ramazanova, Excess, partial, and molar volumes of n-alkanes in near-critical and supercritical water, J. Solution Chem. 27 (1998) 731–753.
[8] H. Liu, J.P. O'Connell, On the measurement of solute partial molar volumes in near-critical fluids with supercritical fluid chromatography, Ind. Eng. Chem. Res. 37 (1998) 3323–3330.
[9] T. Chang, R.W. Rousseau, P.K. Kilpatrick, Methanol synthesis reactions: calculations of equilibrium conversions using equations of state, Ind. Eng. Chem. Process Des. Dev. 25 (1986) 477–481.
[10] P. Sindzingre, G. Ciccotti, C. Massobrio, D. Frenkel, Partial enthalpies and related quantities in mixtures from computer simulation, Chem. Phys. Lett. 136 (1987) 35–41.
[11] D. Frenkel, B. Smit, Understanding Molecular Simulation: From Algorithms to Applications, second ed., Academic Press, San Diego, California, 2002.
[12] A. Rahbari, R. Hens, I.K. Nikolaidis, A. Poursaeidesfahani, M. Ramdin, O.A. Moultos, D. Dubbeldam, T.J.H. Vlugt, I.G. Economou, Computation of partial molar properties using continuous fractional component Monte Carlo, Mol. Phys. 116 (2018) 3331–3344.
[13] P. Sindzingre, C. Massobrio, G. Ciccotti, D. Frenkel, Calculation of partial enthalpies of an argon-krypton mixture by NPT molecular dynamics, Chem. Phys. 129 (1989) 213–224.
[14] B. Smit, D. Frenkel, Calculation of the chemical potential in the Gibbs ensemble, Mol. Phys. 68 (1989) 951–958.
[15] S.M. Walas, Phase Equilibria in Chemical Engineering, Butterworth-Heinemann, USA, 1985.
[16] S.K. Schnell, R. Skorpa, D. Bedeaux, S. Kjelstrup, T.J.H. Vlugt, J.M. Simon, Partial molar enthalpies and reaction enthalpies from equilibrium molecular dynamics simulation, J. Chem. Phys. 141 (2014) 144501.
[17] J.G. Kirkwood, F.P. Buff, The statistical mechanical theory of solutions. I, J. Chem. Phys. 19 (1951) 774–777.
[18] P. Krüger, S.K. Schnell, D. Bedeaux, S. Kjelstrup, T.J.H. Vlugt, J.-M. simon, Kirkwood–Buff integrals for finite volumes, J. Phys. Chem. Lett. 4 (2012) 235–238.
[19] N. Dawass, P. Krüger, J.-M. Simon, T.J.H. Vlugt, Kirkwood–Buff integrals of finite systems: shape effects, Mol. Phys. 116 (2018) 1573–1580.
[20] N. Dawass, P. Krüger, S.K. Schnell, T.J.H. Vlugt, J.-M. Simon, Kirkwood-Buff integrals from molecular simulation, Fluid Phase Equilib. 486 (2019) 21–36.
[21] P. Krüger, T.J.H. Vlugt, Size and shape dependence of finite-volume Kirkwood-Buff integrals, Phys. Rev. E 97 (2018) 051301–051305.
[22] A. Rahbari, R. Hens, O.A. Moultos, D. Dubbeldam, T.J.H. Vlugt, Multiple free energy calculations from single state point continuous fractional component Monte Carlo simulation using umbrella sampling, J. Chem. Theory Comput. 16 (2020) 1757–1767.
[23] A.Z. Panagiotopoulos, Direct determination of phase coexistence properties of fluids by Monte Carlo simulation in a new ensemble, Mol. Phys. 61 (1987) 813–826.
[24] A.Z. Panagiotopoulos, N. Quirke, M. Stapleton, D. Tildesley, Phase equilibria by simulation in the Gibbs ensemble: alternative derivation, generalization and application to mixture and membrane equilibria, Mol. Phys. 63 (1988) 527–545.
[25] J.K. Johnson, A.Z. Panagiotopoulos, K.E. Gubbins, Reactive canonical Monte Carlo: a new simulation technique for reacting or associating fluids, Mol. Phys. 81 (1994) 717–733.
[26] W.R. Smith, B. Triska, The reaction ensemble method for the computer simulation of chemical and phase equilibria. I. theory and basic examples, J. Comput. Phys. 100 (1994) 3019–3027.
[27] H.C. Turner, J.K. Brennan, M. Lisal, W.R. Smith, K.J. Johnson, K.E. Gubbins, Simulation of chemical reaction equilibria by the reaction ensemble Monte Carlo method: a review, Mol. Simul. 34 (2008) 119–146.
[28] D.J. Adams, Chemical potential of hard-sphere fluids by Monte Carlo methods, Mol. Phys. 28 (1974) 1241–1252.
[29] G.E. Norman, V.S. Filinov, Investigations of phase transitions by a Monte-Carlo method, High Temp. Res. USSR 7 (1969) 216–222.
[30] Optimization, Toolbox User's Guide, MathWorks, Inc, The MathWorks, USA, 2016.
[31] F. Karavias, A.L. Myers, Isosteric heats of multicomponent adsorption: thermodynamics and computer simulations, Langmuir 7 (1991) 3118–3126.
[32] S. Ban, T.J.H. Vlugt, F. Kapteijn, J.v den Bergh, Modeling the loading dependency of diffusion in zeolites: the relevant site model extended to mixtures in DDR-type zeolite, J. Phys. Chem. C 113 (2009) 21856–21865.
[33] M.P. Allen, D.J. Tildesley, Computer Simulation of Liquids, second ed., Oxford University Press, Oxford, United Kingdom, 2017.
[34] T.J.H. Vlugt, D. Dubbeldam, S. Ban, S. Calero, E. García-Pérez, Computing the heat of adsorption using molecular simulations: the effect of strong coulombic interactions, J. Chem. Theory Comput. 4 (2008) 1107–1118.
[35] A. Torres-Knoop, A. Poursaeidesfahani, T.J.H. Vlugt, D. Dubbeldam, Behavior of the enthalpy of adsorption in nanoporous materials close to saturation conditions, J. Chem. Theory Comput. 13 (2017) 3326–3339.
[36] R.E. Walpole, R.H. Myers, S.L. Myers, K. Ye, Probability & Statistics for Engineers & Scientists, ninth ed., Prentice Hall, Boston, USA, 2012.
[37] A. Poursaeidesfahani, A. Torres-Knoop, M. Rigutto, N. Nair, D. Dubbeldam, T.J.H. Vlugt, Computation of the heat and entropy of adsorption in proximity of inflection points, J. Phys. Chem. C 120 (2016) 1727–1738.
[38] T.J.H. Vlugt, R. Krishna, B. Smit, Molecular simulations of adsorption isotherms for linear and branched alkanes and their mixtures in silicalite, J. Phys. Chem. B 103 (1999) 1102–1118.
[39] T.J.H. Vlugt, W. Zhu, F. Kapteijn, J.A. Moulijn, B. Smit, R. Krishna, Adsorption of linear and branched alkanes in the zeolite silicalite-1, J. Amer. Chem. Soc. 120 (1998) 5599–5600.
[40] R. Krishna, B. Smit, T.J.H. Vlugt, Sorption-induced diffusion-selective separation of hydrocarbon isomers using silicalite, J. Phys. Chem. A 102 (1998) 7727–7730.
[41] R.F. DeJaco, B. Elyassi, M.D. de, N. Mittal, M.M Tsapatsis, J.I. Siepmann, Understanding the unique sorption of alkane-α, ω-diols in silicalite-1, J. Comput. Phys. 149 (2018) 072331.
[42] I. Langmuir, The adsorption of gases on plane surfaces of glass, mica and platinum, J. Amer. Chem. Soc. 40 (9) (1918) 1361–1403.
[43] M. Volmer, P. Mahnert, Über die auflösung fester körper in flüssigkeitsoberflächen und die eigenschaften der dabei entstehenden schichten, Z. Phys. Chem. 115 (1) (1925) 239–252.
[44] M. Seeger, Gaussian processes for machine learning, Int. J. Neural Syst. 14 (2004) 69–106.
[45] A.Z. Panagiotopoulos, R.C. Reid, On the relationship between pairwise fluctuations and thermodynamic derivatives, J. Chem. Phys. 85 (1986) 4650–4653.
[46] W.W. Wood, Monte Carlo calculations for hard disks in the isothermal-isobaric ensemble, J. Chem. Phys. 48 (1968) 415–434.
[47] W.W. Wood, NpT-ensemble Monte Carlo calculations for the hard-disk fluid, J. Chem. Phys. 52 (1970) 729–741, doi:10.1063/1.1673047.
[48] J.R. Ray, Ensembles and computer simulation calculation of response functions, Handb. Mater. Model. (2005) 729–743.
[49] A. Poursaeidesfahani, R. Hens, A. Rahbari, M. Ramdin, D. Dubbeldam, T.J.H. vlugt, Efficient application of continuous fractional component Monte Carlo in the reaction ensemble, J. Chem. Theory Comput. 13 (2017) 4452–4466.
[50] T.W. Rosch, E.J. Maginn, Reaction ensemble Monte Carlo simulation of complex molecular systems, J. Chem. Theory Comput. 7 (2011) 269–279.
[51] E.O. Fetisov, I.-F.W. Kuo, C. Knight, J. VandeVondele, T.V. Voorhis, J.I. Siepmann, First-principles Monte Carlo simulations of reaction equilibria in compressed vapors, ACS. Cent. Sci. 2 (2016) 409–415.
[52] C.H. Turner, J.K. Johnson, K.E. Gubbins, Effect of confinement on chemical reaction equilibria: the reactions $2NO \leftrightarrow (NO)_2$ and $N_2 + 3H_2 \leftrightarrow 2NH_3$ in carbon micropores, J. Comput. Phys. 114 (2001) 1851–1859.
[53] S.P. Balaji, S. Gangarapu, M. Ramdin, A. Torres-Knoop, H. Zuilhof, E.L. Goetheer, D. Dubbeldam, T.J.H. Vlugt, Simulating the reactions of CO_2 in aqueous monoethanolamine solution by reaction ensemble Monte Carlo using the continuous fractional component method, J. Chem. Theory Comput. 11 (2015) 2661–2669.
[54] R.G. Mullen, S.A. Corcelli, E.J. Maginn, Reaction ensemble Monte Carlo simulations of CO_2 absorption in the reactive ionic liquid triethyl(octyl)phosphonium 2-cyanopyrrolide, J. Phys. Chem. Lett. 9 (2018) 5213–5218.

[55] M. Lísal, M. Bendová, W.R. Smith, Monte Carlo adiabatic simulation of equilibrium reacting systems: the ammonia synthesis reaction, Fluid Phase Equilib. 235 (2005) 50–57.
[56] A. Rahbari, R. Hens, D. Dubbeldam, T.J.H. Vlugt, Improving the accuracy of computing chemical potentials in CFCMC simulations, Mol. Phys. 117 (2019) 3493–3508.
[57] J.Q. Shi, T. Choi, Gaussian Process Regression Analysis for Functional Data, CRC Press, Boca Raton, USA, 2011.
[58] S. Haykin, Neural Networks: a Comprehensive Foundation, Prentice Hall PTR, USA, 1994.
[59] D.F. Specht, A general regression neural network, IEEE Trans. Neural Netw. 2 (1991) 568–576.
[60] N.D. Zhuravlev, M.G. Martin, J.I. Siepmann, Vapor-liquid phase equilibria of triacontane isomers: deviations from the principle of corresponding states, Fluid Phase Equilib. 202 (2002) 307–324.