Energy 188 (2019) 116091

Contents lists available at ScienceDirect

Energy

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

Machine learning for predicting thermodynamic properties of pure fluids and their mixtures

Yuanbin Liu [a], Weixiang Hong [b], Bingyang Cao [a, *]

[a] Key Laboratory for Thermal Science and Power Engineering of Ministry of Education, Department of Engineering Mechanics, Tsinghua University, Beijing, 100084, China
[b] Institute of Systems Science, National University of Singapore, 119615, Singapore

ARTICLE INFO

Article history:
Received 6 June 2019
Received in revised form
2 September 2019
Accepted 7 September 2019
Available online 9 September 2019

Keywords:
Thermodynamic properties
Machine learning
Support vector regression
Mixtures
Molecular dynamics simulation

ABSTRACT

Establishing a reliable equation of state for largely non-ideal or multi-component liquid systems is challenging because the complex effects of molecular configurations and/or interactions on the thermodynamic properties must generally be taken into account. In this regard, machine learning holds great potential for directly learning the thermodynamic mappings from existing data, thereby bypassing the use of equations of state. The present study outlines a general machine learning framework based on high-efficiency support vector regression for predicting the thermodynamic properties of pure fluids and their mixtures. The proposed framework is adopted in conjunction with training data obtained from a high-fidelity database to successfully predict the thermodynamic properties of three common pure fluids. The predictions demonstrate extremely low mean square errors. Moreover, little loss in the prediction accuracy is obtained for ternary mixtures of the pure fluids at the cost of a modest increase in the volume of training data provided by state-of-the-art molecular dynamics simulations. Our results demonstrate the promising potential of machine learning for building accurate thermodynamic mappings of pure fluids and their mixtures. The proposed methodology may pave the way in the future for the rapid exploration of novel or complex systems with potentially exceptional thermodynamic properties.

1. Introduction

Pressure, volume, and temperature (*PVT*) are the most fundamental thermodynamic properties of matters because they can be relatively easily measured and are intrinsically related to other thermodynamic parameters [1,2]. Nowadays, the *PVT* properties of fluids are of intrinsic interest in many fields, such as energy, physics, chemistry, materials, and environmental sciences, and they are employed in numerous concrete applications such as the design of power generation systems [3,4], the analysis of thermal energy transformations [5,6], and the interpretation of thermochemical and geochemical processes [7,8].

Although experimental measurements are the preferable means of acquiring the *PVT* properties of fluids, conducting experimental measurements can require considerable resources, particularly under extreme environments of high temperature and high pressure and when working with erosive or explosive substances. These issues have been addressed to some extent via the use of molecular dynamics (MD) simulation, which has garnered widespread interest as a state-of-the-art technique for obtaining thermodynamic data related to specific fluids. A large number of studies have demonstrated that the accuracy of the thermodynamic properties obtained by MD simulations is comparable with that of the thermodynamic properties obtained experimentally when the force fields are suitably chosen [9,10]. Nevertheless, MD simulations are extremely time-consuming for large molecular systems. As a result, only limited thermodynamic data can be obtained under discrete *PVT* states experimentally and by MD simulations. This issue can be addressed by developing an equation of state (EOS) to correlate the experimental and simulated data. Remarkable progress has been made in the development of equations of state in recent decades. A series of equations of state have been established and modified, such as the well-known van der Waals, Soave modified Redlich-Kwong (SRK), Peng-Robinson (PR), and Duan-Møller-Weare (DMW) EOS models [11–15]. Unfortunately, no universal EOS has been developed to date due to the complex

* Corresponding author.
E-mail address: caoby@tsinghua.edu.cn (B. Cao).

https://doi.org/10.1016/j.energy.2019.116091
0360-5442/

interactions between molecules in non-ideal fluids. For example, the PR-EOS is one of the most widely applied models, and yet more than 200 modified forms of the model have been developed for pure fluids and over 100 for mixtures to overcome the inherent deficiencies of the original model [14]. Moreover, a majority of these modifications are restricted to the specific properties and substances for which they were developed, and all new modifications and optimizations must be conducted for previously unexplored properties and substances. Additionally, the modified equations of state typically become increasingly complicated to enhance their prediction accuracy, and this significantly reduces the simplicity of the original equations [16–18].

Considerable interest has been generated recently in employing machine learning (ML) methods to emerging applications in the theoretical and computational areas of physics, chemistry, and materials science [19–30]. Presently, these applications mainly involve atomic-level processes, such as substituting the evaluation of atomic-level processes by means of intensive calculations based on density functional theory with rapid predictions based on ML methods [19,24,31–34]. Yet, little work has been devoted toward exploring the thermodynamic properties of pure fluids and their mixtures over a wide range of temperatures by ML methods. Nevertheless, the fact that ML-based approaches are purely data-driven and can potentially generate mappings among the thermodynamic parameters of materials without relying on any concrete expressions or underlying physical insights is extremely attractive. Achieving this goal would create a novel means of rapidly investigating the thermodynamic properties of new or complex compounds.

In this study, we present a general ML framework based on support vector regression (SVR) for predicting the *PVT* properties of pure fluids and their mixtures. Here, SVR is the regression version of the support vector machine. We have selected SVR for the proposed framework because it has demonstrated good performance on multiple non-linear regression problems compared with other ML techniques [35], such as ridge regression [36,37], the least absolute shrinkage and selection operator (LASSO) [37,38], and Gaussian process regression [35,39]. The feasibility and accuracy of our proposed ML framework are demonstrated by quantitatively evaluating the predictions obtained for several practical applications. All training data are derived either from a high-fidelity database or from MD simulations. The goal of the proposed framework is to facilitate the efficient and rapid exploration of new or complex fluids with superior thermodynamic properties.

2. Methodology

2.1. Support vector regression

We assume a training set with l data points $\chi = \{(\mathbf{y}_1, z_1), (\mathbf{y}_2, z_2), \ldots, (\mathbf{y}_l, z_l)\}$, where $\mathbf{y}_i \in \mathbb{R}^n$ is an input vector, $z_i \in \mathbb{R}^1$ denotes an output or label, and $i = 1, 2, \ldots, l$. The goal of SVR is to find a function $f(\mathbf{y})$ that estimates z with an acceptable small error. The standard SVR algorithm (i.e., ε-SVR) was initially proposed by Vapnik based on the ε-insensitive loss function [40]. Here, ε-SVR was employed to fit a tube of radius ε to data, where absolute deviations between $f(\mathbf{y})$ and labels less than or equal to ε were accepted, and all other deviations greater than ε were rejected. However, specifying an appropriate value of ε beforehand is difficult. Schölkopf et al. [41] addressed this issue by developing the ν-SVR algorithm, which modified ε-SVR by replacing ε with a new parameter ν. The ν-SVR algorithm was demonstrated to be effective for controlling the number of support vectors and training errors by adjusting the value of ν, where $\nu \in (0, 1]$ [41,42]. The regression function in ν-SVR takes the form

$$f(\mathbf{y}) = \sum_{i=1}^{l} (\widehat{\alpha}_i - \alpha_i) \langle \phi(\mathbf{y}), \phi(\mathbf{y}_i) \rangle + b, \tag{1}$$

where $\widehat{\alpha}_i$ and α_i are the Lagrange multipliers, b is a bias parameter, $\phi(\mathbf{y})$ maps $\boldsymbol{y}$ into a higher-dimensional space denoted as the feature space, and $\langle \cdot , \cdot \rangle$ refers to an inner product. The ν-SVR implementation in the present study is based on the LIBSVM package [43].

The role of the mapping $\phi(\mathbf{y})$ is to make the SVR algorithm nonlinear. Even though Eq. (1) maintains a linear form in feature space, it incorporates the nonlinearity in the original input space. However, determining an explicit form for $\phi(\mathbf{y})$ a priori is generally infeasible. Fortunately, determining an explicit form for $\phi(\mathbf{y})$ is unnecessary because determining the inner product $\langle \phi(\mathbf{y}), \phi(\mathbf{y}_i) \rangle$ in feature space is sufficient. This is incorporated in the so-called "kernel trick" that provides a computationally efficient means of determining the inner product via a kernel function given as $\kappa(\mathbf{y}, \mathbf{y}_i) = \langle \phi(\mathbf{y}), \phi(\mathbf{y}_i) \rangle$. The mapping $\phi(\mathbf{y})$ is fully implicit in $\kappa(\mathbf{y}, \mathbf{y}_i)$. In addition, $\kappa(\mathbf{y}, \mathbf{y}_i)$ should be positive semidefinite based on Mercer's theorem [44]. A number of kernels have been commonly adopted, such as linear, polynomial, Gaussian or radial basis function (RBF), Laplacian, and sigmoid kernels. The present work adopts the Gaussian kernel to train our models due to its robust and promising performance in preliminary experiments. The Gaussian kernel is given as

$$\kappa(\mathbf{y}, \mathbf{y}_i) = \exp\left(- \frac{\|\mathbf{y} - \mathbf{y}_i\|^2}{2\sigma^2} \right), \tag{2}$$

where σ denotes the width of the Gaussian, and $\| \cdot \|$ refers to the Euclidean L^2-norm. As a simplification, Eq. (2) can be written as $\kappa(\mathbf{y}, \mathbf{y}_i) = \exp(-\gamma \|\mathbf{y} - \mathbf{y}_i\|^2)$, where $\gamma = (2\sigma^2)^{-1}$.

In ν-SVR, the parameters $\widehat{\alpha}_i$ and α_i are determined by solving the dual optimization problem of convex quadratic programming, which is stated as

$$\max_{\widehat{\alpha}_i,\, \alpha_i} \quad \sum_{i=1}^{l} (\widehat{\alpha}_i - \alpha_i) z_i - \frac{1}{2} \sum_{i,j=1}^{l} (\widehat{\alpha}_i - \alpha_i)(\widehat{\alpha}_j - \alpha_j) \kappa\left(\mathbf{y}_i, \mathbf{y}_j\right), \tag{3}$$

subject to

$$\begin{aligned} &\sum_{i=1}^{l} (\widehat{\alpha}_i - \alpha_i) = 0, \\ &\sum_{i=1}^{l} (\widehat{\alpha}_i + \alpha_i) \le C \cdot \nu, \\ &0 \le \widehat{\alpha}_i, \alpha_i \le C/l, \qquad i = 1, 2, \ldots, l, \end{aligned} \tag{4}$$

where C is a positive regularization parameter determining the trade-off between training error and model complexity. Briefly, the training error typically decreases with increasing C, while overfitting becomes increasingly more likely.

The optimal selections of γ and C may be obtained by a grid search method in conjunction with k-fold cross-validation (CV), where k is a positive integer. This process can be described as follows. First, the training sets are divided into k equally sized subsets. Then, each subset is adopted one time as the testing set, while the remaining $(k-1)$ subsets are utilized for training. Accordingly, the CV error is estimated by the average of the errors of the k generated models. The grid search method employs a grid space consisting of discrete (γ, C) values serving as grid points. The optimal parameters γ and C are then determined as the single grid point within the grid

space yielding the lowest CV error.

2.2. Machine learning framework

We consider mixtures consisting of n distinct molecular components and seek to define the target property P according to the system attributes $\boldsymbol{D} = [V_m\ T\ x_1\ x_2\ \ldots\ x_{n-1}]$, where V_m is the molar volume of the mixture, T is the temperature, and x_j is the molar fraction of the jth molecular component, for $j = 1, 2,\ldots, n-1$. The input vector contains the molar fractions of only $(n-1)$ components because the last component x_n is independent of the others under the constraint $x_n = 1 - (x_1 + x_2 + \ldots + x_{n-1})$. In addition, we note that the attributes of a pure fluid are reduced to $\boldsymbol{D} = [V_m\ T]$.

The proposed ML framework for predicting the thermodynamic properties of pure fluids or their mixtures is illustrated in Fig. 1 based on the above discussion. This process is also compared in Fig. 1 with that employed when adopting equations of state for generating mappings among thermodynamic parameters. For the ML framework, the sample space of the thermodynamic parameters $\boldsymbol{D}$ must first be obtained from some high-fidelity databases, experiments, or MD simulations. Then, the data within the sample space must be rescaled to decrease the differences in their magnitudes by various means, such as normalization, regularization, or logarithmic transformation. This is a significant step prior to submitting the data for training to avoid the impacts of data with extremely imbalanced magnitudes on the training process. The rescaling process, however, does not apply to the molar fractions because they are naturally normalized. We note from preliminary experiments that logarithmic transformation tends to maintain the relative magnitudes of PVT data better than normalization and regularization and therefore provides better results for our tasks. The logarithmic transformation equations are given as follows:

$$P' = \log_{10}(P), \tag{5}$$

$$V'_m = \log_{10}(1000 \times V_m) \tag{6}$$

$$T' = \log_{10}(T). \tag{7}$$

Here, P is given in MPa, V_m in L/mol, and T in K. Accordingly, the input vector $\boldsymbol{y}$ and the label z employed by SVR are determined in training as $[V'_m\ T'\ x_1\ x_2\ \ldots\ x_{n-1}]$ and P', respectively. Afterward, the rescaled sample space is divided into a training set and a testing set. Then, the grid search method is adopted with fivefold cross-validation (CV) to determine the optimal values of γ and C in ν-SVR. The entire training set with the optimal values of γ and C is then employed to optimize the objective function of our ν-SVR implementation with respect to the model parameters using the Gaussian kernel. As a result, a predictive model is constructed.

After constructing the predictive model, its feasibility can be evaluated by comparing its ML mappings with the original training data, and the testing set can be employed to validate the accuracy and generalization ability of the predictive model. Two evaluation criteria are commonly adopted for conducting quantitative assessments of the feasibility, accuracy, and generalization ability of a predictive model. The first is the mean square error (MSE)

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{y}_i) - z_i)^2, \tag{8}$$

which is employed to represent the training error of the predictive model with respect to the training set and the generalization error of the predictive model with respect to the testing set. The other is the squared correlation coefficient r^2, given as

$$r^2 = \frac{\left(N\sum_{i=1}^{N} f\left(\mathbf{y}_i\right) z_i - \sum_{i=1}^{N} f\left(\mathbf{y}_i\right)\sum_{i=1}^{N} z_i\right)^2}{\left[N\sum_{i=1}^{N} f(\mathbf{y}_i)^2 - \left(\sum_{i=1}^{N} f(\mathbf{y}_i)\right)^2\right]\left[N\sum_{i=1}^{N} z_i^2 - \left(\sum_{i=1}^{N} z_i\right)^2\right]}, \tag{9}$$

which is used to quantify the strength of the linear correlations between the labels and the training results with respect to the training set and the prediction results with respect to the testing set. It worth noting that there are several significant factors that will bring in errors with v-SVR: (a) regions of low data density, this could be avoided by adopting the relatively uniform data distribution to train the model; (b) high variance of magnitude of inputs, this could be solved by the scaling of inputs; (c) underfitting, it might be overcome by appropriately decreasing the regularization strength; (d) overfitting, it might be settled by providing sufficient training data and appropriately selecting the regularization strength and the kernel width with the fivefold cross-validation.

2.3. Details on molecular dynamics simulations

The parameters of the potential models employed in our MD simulations have been carefully chosen. Accordingly, the TIP4P [45], EMP2 [46], and two-site models [47] were employed for simulating H_2O, CO_2, and H_2 molecules, respectively. A series of MD simulations are performed by using the LAMMPS package [48]. Simulated

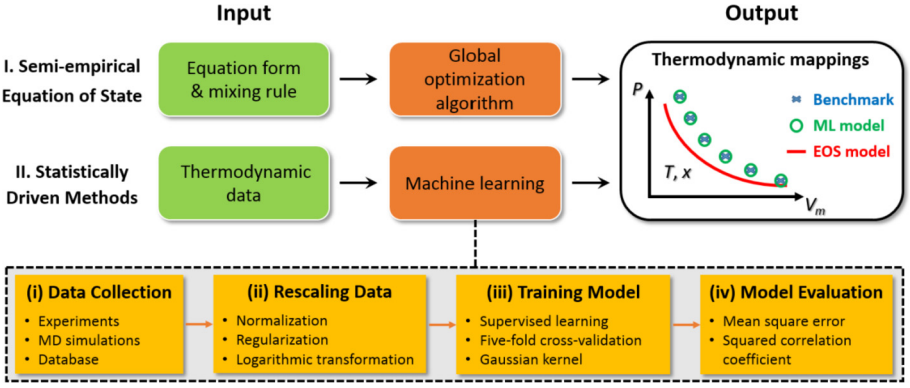

Fig. 1. Outline of the two methods for generating mappings among thermodynamic parameters. One method involves establishing an empirical equation of state by fitting experimental or simulated data, as illustrated in the upper half of the figure. The other method is based on the proposed machine learning framework, as illustrated in the lower half of the figure.

Table 1
Details regarding the sample spaces. Here, N_1 and N_2 denote the sizes of the training sets and the testing sets, respectively.

System	x_{H_2O}	x_{CO_2}	x_{H_2}	T (K)	V_m (L/mol)	P (MPa)	N_1	N_2
H_2O	1	–	–	400–2000	0.05–0.95	0.25–426	646	132
CO_2	–	1	–	300–1800	0.05–0.95	2–770	591	120
H_2	–	–	1	300–1000	0.05–0.95	3–236	285	56
H_2O-CO_2-H_2	0.1–0.8	0.1–0.8	0.1–0.8	650–1150	0.05–1.50	3–339	504	156

systems contain 2500 molecules for the ternary mixtures. All molecules are located in a cubic box with periodic boundary conditions. The long-range Coulombic interactions are handled by the particle-particle/particle-mesh (PPPM) whereas the short-range interaction potential is cut off beyond 12 Å. A Nosé-Hoover thermostat [49] is coupled to systems to control temperatures for the canonical (NVT) ensembles along which the pressure of systems is calculated. The time step is of 1 fs in all simulations. Initial equilibration periods are 600 ps, followed by simulation runs of 600 ps to record meaningful data.

3. Results and discussion

In the following sections, we apply the proposed ML framework to predict the thermodynamic properties of three pure fluids and ternary mixtures of the three. The pure fluids include water (H_2O), carbon dioxide (CO_2), and hydrogen (H_2) because these are the most common substances and also include polar and non-polar molecules. Moreover, these pure fluids play important roles in recently emerging technologies such as the extraction of supercritical fluids [50] and coal gasification in supercritical water [51]. Predictions for all cases are conducted over a wide range of T, which include near-critical and supercritical regions of their phase space.

The thermodynamic properties of H_2O, CO_2, and H_2 have been extensively studied via experiments and simulations, and this has generated an abundance of high-fidelity datasets. Accordingly, the sample spaces for these pure fluids have been obtained directly from the database provided by the prestigious National Institute of Standards and Technology (NIST). However, experimental and simulation data regarding the thermodynamic properties of H_2O-CO_2-H_2 ternary mixtures are presently scarce. Therefore, sufficient training data pertaining to the thermodynamic properties of ternary H_2O-CO_2-H_2 mixtures required by the present study have been obtained by conducting MD simulations ranging from the near-critical region to the supercritical region, which yielded 660 datasets. Details regarding the adopted sample spaces of the aforementioned systems are provided in Table 1.

The ML framework was first applied to the pure fluids, where $\boldsymbol{y} = [V'_m\ T']$ and $z = P'$. The results for H_2O are described in detail, whereas only the main results for CO_2 and H_2 are presented owing to space limitations. The coarse- and fine-grained grid searches for the optimal parameters γ and C for H_2O are illustrated in Figs. 2(a) and (b), respectively, as a function of $\log_2(\gamma)$ and $\log_2(C)$. Choosing such exponentially increasing grid sizes is a typical strategy employed to increase the efficiency of the grid search method. In addition, this strategy ensures that the grid space can span a sufficiently wide range of both γ and C. The total calculation times involved in the grid search method can be further reduced by first conducting a coarse grid search to determine an approximate region for which the optimal parameters may be located prior to conducting searches over finer-grained grids. This approximate region for H_2O is enclosed by the dashed black line in Fig. 2(a). Then, smaller grid sizes are employed, as shown in Fig. 2(b), to pinpoint the optimal parameters. The results indicate that the two optimal parameters lie within the grid space, which confirms that our original grid spans are sufficiently large to include the optimal parameters. The contour lines in Fig. 2(b) suggest that multiple parameter values correspond with an equivalent CV accuracy. Under this condition, the optimal parameters may be determined as those lying with the minima of C to reduce the possibility of overfitting. The above-discussed procedure yields optimal γ and C values of 16 and 8, respectively, for the H_2O model. The optimization results for other models are listed in Table 2.

The optimal values of γ and C are then utilized with the training

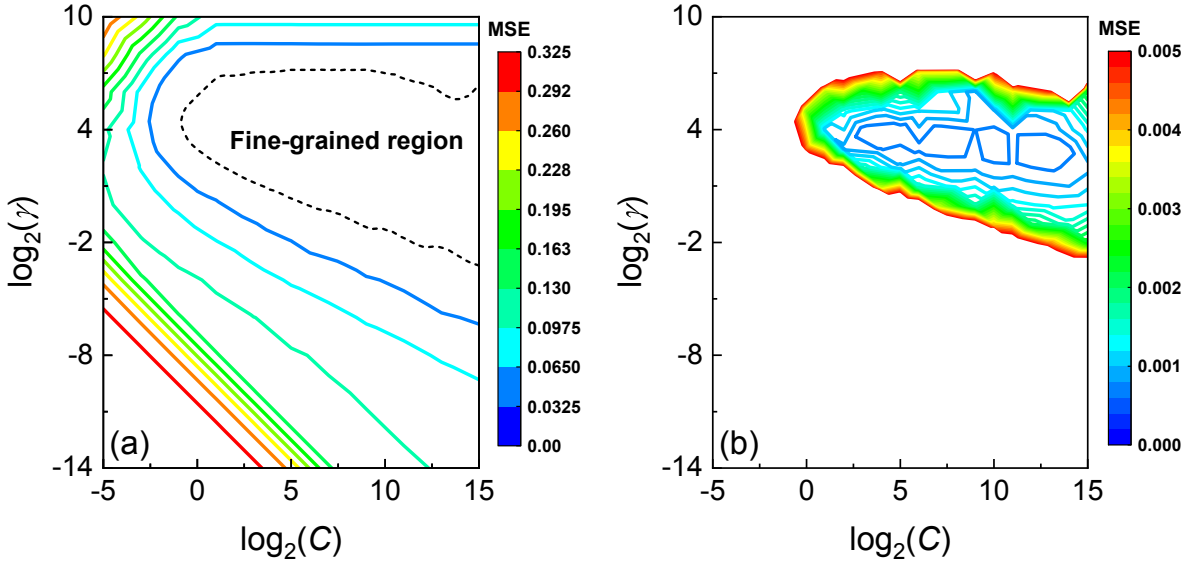

Fig. 2. Contour lines of the CV error as a function of $\log_2(\gamma)$ and $\log_2(C)$ for pure H_2O. The coarse- and fine-grained grid searches are illustrated in (a) and (b), respectively.

set derived from the NIST basis sets for training the ν-SVR, and the final ML mappings are generated according to the ν-SVR with the Gaussian kernel. Figs. 3(a) and (b) present comparisons between the NIST basis sets with the data generated from the ML mappings for H_2O based on a training set and a testing set, respectively. The NIST training set data are marked by the "+" symbols in Fig. 3(a) whose colors represent the values of T based on the color scale given to the right of the figure, while the corresponding ML mappings are described by the smooth grey curves. Similarly, the NIST testing set data in Fig. 3(b) are marked by various symbols according to the values of V_m pertaining to the data, while the corresponding ML mappings are described by the curves. The accuracy of the ML mappings based on the training set serves as a necessary factor for validating the feasibility of the data-driven ML framework, whereas the prediction accuracy of the ML mappings based on the testing set serves the role of estimating the generalization performance of the framework for data not included in the original dataset. It is observed that all ML outputs for H_2O agree remarkably well with the training and testing set data over a wide range of T. Furthermore, we quantify and summarize the precision of the ML mappings for reproducing the pressure of H_2O, CO_2, and H_2 in Table 3. It is noted that the training and prediction errors of ML mappings are rather small. In addition, all values of r^2 are quite close to 1, indicating that a strongly linear relationship exists between the basis sets and the mapping and prediction data. These results verify the applicability of our ML framework for various pure fluids.

In contrast to the pure liquids, the input vector $\boldsymbol{y}$ of the H_2O-CO_2-H_2 mixtures would be rendered as $[V'_m\ T'\ x_{H_2O}\ x_{CO_2}]$, $[V'_m\ T'\ x_{H_2O}\ x_{H_2}]$, or $[V'_m\ T'\ x_{CO_2}\ x_{H_2}]$, which are mutually equivalent because the last molar fraction component is independent of the others under the above-discussed constraint. Despite the increased number of dimensions, the ML procedures implemented here are identical to those employed for the pure liquids. Specific composition for the ternary mixtures and corresponding dataset size are presented in Fig. 4. Again, the optimized values of γ and C are obtained, as listed in Table 2. From Fig. 5(a), we can observe that the ML mappings for the ternary mixtures achieve good thermodynamic mapping and high prediction accuracies from the near-critical region to the supercritical region. This is also indicated by the results listed in Table 3 for the ternary mixtures, where the mapping and prediction errors are extremely low, and the values of r^2 are all nearly equal to 1. We also compare the results of Fig. 5(a) with the results obtained from the three most typical equations of state for the H_2O-CO_2-H_2 mixtures in Fig. 5(b), namely PR-EOS, SRK-EOS, and DMW EOS. Here, extending an EOS to mixtures requires the use of a mixing rule. We adopted the commonly used van der Waals mixing rule for the PR-EOS and SRK-EOS, and the corresponding equation parameters were obtained from the available literature [52–54]. We note from the figure that the prediction

Table 2
Optimal values of parameters γ and C obtained for ν-SVR by fivefold CV.

System	γ	C
H_2O	16	8
CO_2	8	16
H_2	2	4
H_2O-CO_2-H_2	2	32

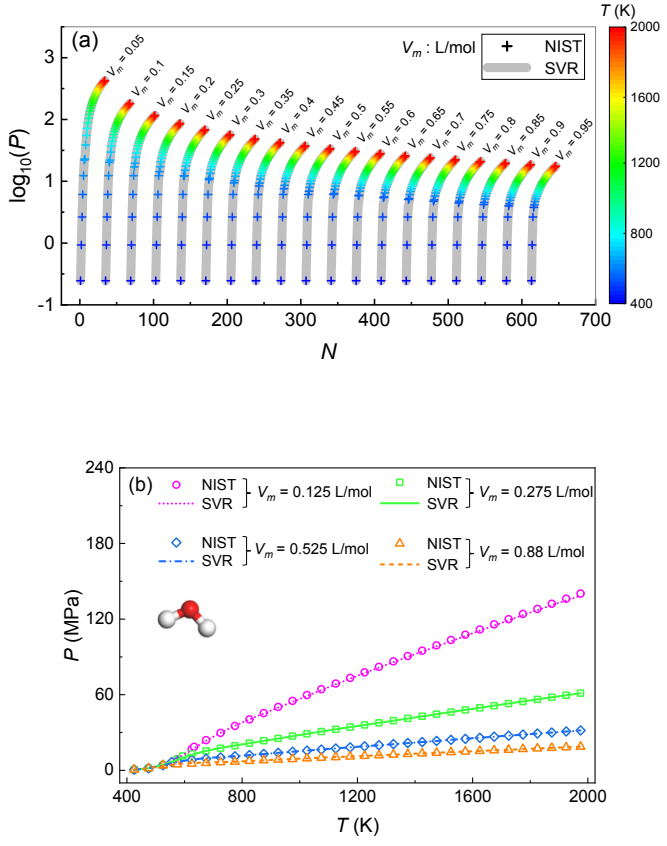

Fig. 3. Comparisons of the NIST basis sets with the data generated from the ML mappings for H_2O: (a) ML mappings based on the training set as a function of the training set size N; and (b) ML mappings based on the testing set as a function of T. Both datasets are derived from the NIST database.

Table 3
Accuracies and correlations for H_2O, CO_2, H_2, and their ternary mixtures.

System	Training set		Testing set	
	MSE	r^2	MSE	r^2
H_2O	2.52×10^{-4}	0.999223	2.41×10^{-4}	0.999115
CO_2	9.65×10^{-6}	0.999958	5.18×10^{-6}	0.999978
H_2	4.44×10^{-7}	0.999997	2.47×10^{-7}	0.999999
H_2O-CO_2-H_2	1.70×10^{-6}	0.999992	2.52×10^{-5}	0.999894

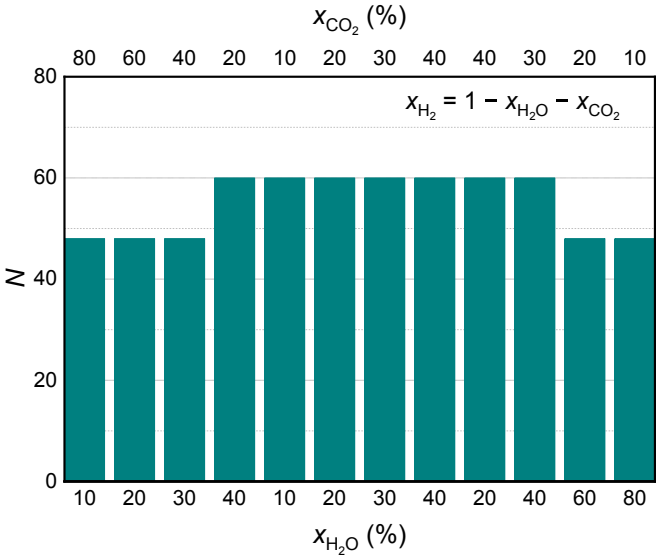

Fig. 4. Chosen specific composition and corresponding dataset size for the H_2O-CO_2-H_2 ternary mixtures.

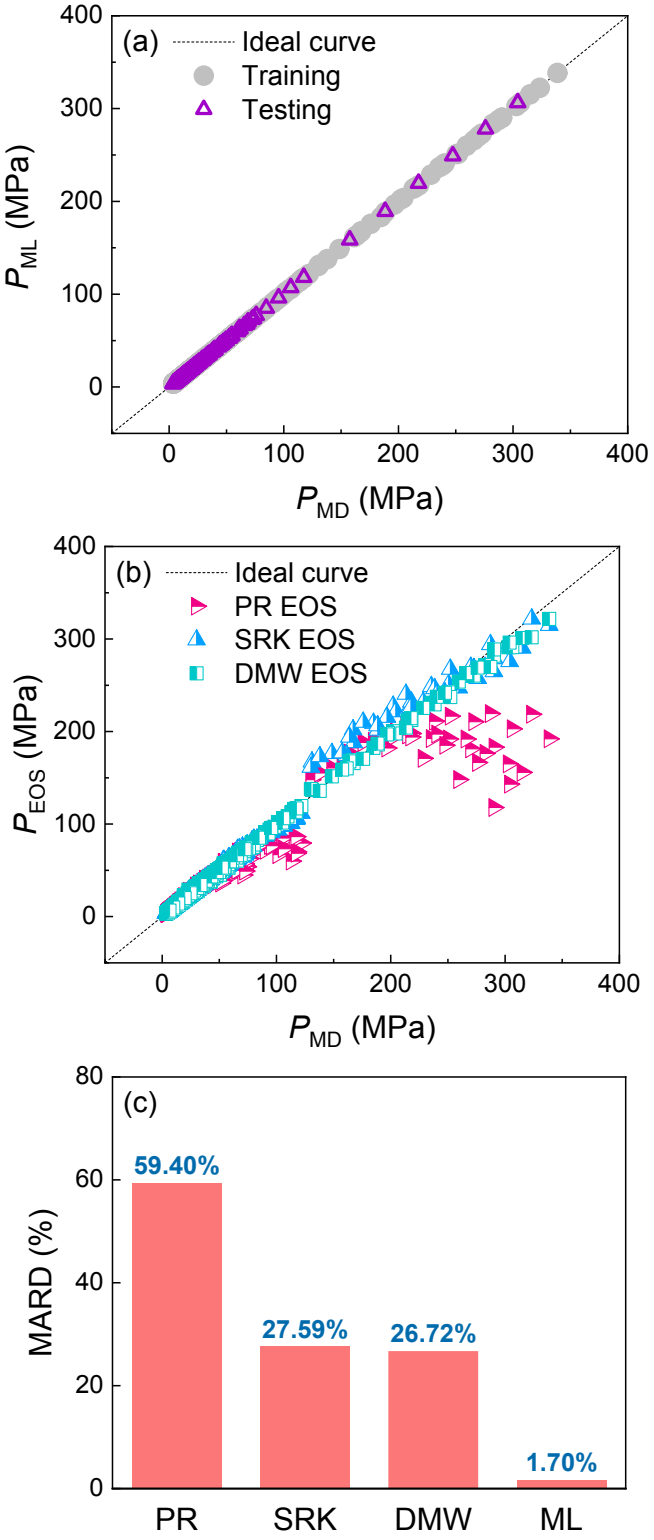

Fig. 5. Accuracy of the ML mappings and the EOS models. Comparison of the datasets from the MD simulations with the outputs from (a) the ML mappings for the training and testing sets and (b) the three typical equations of state for H_2O-CO_2-H_2 mixtures. (c) Comparison of the maximum absolute relative deviations between the ML mappings and the EOS models.

performance of the PR-EOS is satisfactory within the relatively low pressure regime. However, the prediction results of the PR-EOS deviate from the MD simulations in the high pressure and temperature regimes, and the deviation increases with increasing pressure. The mixing rules employed may be the primary factor detracting from the performance of the PR-EOS. It is shown that determining an appropriate mixing rule for ternary mixtures over varying molar fractions is quite challenging due to the complexity and uncertainty associated with interactions between unlike molecules in non-ideal mixtures. It is no surprise that some of equations of state can provide a good prediction for the ternary mixtures, because the coefficients in those equations and mixing rules have been frequently modified and verified by the fitting to the experimental data of pure members and binary mixtures of H_2O, CO_2, and H_2. To further evaluate the accuracy of the ML mappings and the EOS models, the maximum absolute relative deviations (MARD) between the MD simulations and the above models are calculated as

$$\mathrm{MARD} = \left|\frac{P_{\mathrm{model}} - P_{\mathrm{MD}}}{P_{\mathrm{MD}}}\right| \times 100\%, \tag{10}$$

where the subscripts 'MD' and 'model' represent the results from the MD simulations and the ML mappings or the EOS models, respectively. We can be observed from Fig. 5(c) that the DMW EOS exhibits a minimum MARD among those three EOS models. It is because unlike the cubic EOS truncated at the third virial coefficient, the DMW EOS has a rather complex form with more coefficients, generally presenting better predictability for supercritical fluid mixtures. However, the ML mappings can give a prediction with extremely low MARD of 1.7%, an order of magnitude improvement over those three equations of state. At this stage in the development of equations of state, the proposed data-driven ML method, which requires no knowledge of the motions and intrinsic interactions of molecules, may hold greater promise for generating accurate thermodynamic mappings for multi-component mixtures.

4. Conclusions

In summary, we introduced a general ML framework for predicting the thermodynamic properties of pure fluids and their mixtures. The feasibility, accuracy, and generalization ability of the proposed ML approach for constructing the thermodynamic mappings of fluids and predicting their thermodynamic properties were evaluated via practical applications using pure H_2O, CO_2, and H_2, and their ternary mixtures. The ML mappings of the predictive model were thereby demonstrated to yield extremely satisfactory mapping and prediction results. In contrast to the prediction of thermodynamic properties based on the development of equations of state, which is generally a very challenging process involving ever increasingly complex analyses of the effects of molecular configurations and/or interactions on the forms of the EOS and various mixing rules, our approach is advantageous for directly learning thermodynamic mappings from existing data with no prior knowledge of the underlying physical mechanisms. Even so, the further development of equations of state through unremitting efforts is invaluable and should not be replaced by ML approaches. We envision that the generation of thermodynamic data by ML mappings may be extremely conductive toward investigating physical correlations among thermodynamic parameters and may well promote the development of equations of state.

Acknowledgements

This work was supported by the National Key R&D Program of China (Grant No. 2016YFB0600100).

Appendix

Soave modified Redlich-Kwong *equation of state*. The SRK-EOS for pure fluids is given by Ref. [11].

$$P=\frac{RT}{V_m-B}-\frac{A}{V_m(V_m+B)}, \tag{11}$$

with

$$A=A_c\beta, \tag{12}$$

$$A_c=0.42747\frac{R^2T_c^2}{P_c}, \tag{13}$$

$$\beta=\left[1+\kappa\left(1-\sqrt{T_r}\right)\right]^2, \tag{14}$$

$$\kappa=0.48508+1.55171\omega-0.15613\omega^2, \tag{15}$$

$$B=0.08664\frac{RT_c}{P_c}, \tag{16}$$

where R is the ideal gas constant; the subscript 'c' denotes the properties at the critical point; T_r represents a reduced temperature; ω is the acentric factor of molecules.

Peng-Robinson equation of state. The PR-EOS takes the form for pure fluids [12].

$$P=\frac{RT}{V_m-B}-\frac{A}{V_m(V_m+B)+B(V_m-B)}, \tag{17}$$

with

$$A=A_c\beta, \tag{18}$$

$$A_c=0.45724\frac{R^2T_c^2}{P_c}, \tag{19}$$

$$\beta=\left[1+\kappa\left(1-\sqrt{T_r}\right)\right]^2, \tag{20}$$

$$\kappa=0.37464+1.54226\omega-0.26992\omega^2, \tag{21}$$

$$B=0.07780\frac{RT_c}{P_c}. \tag{22}$$

van der Waals mixing rule. Mixing rules extend the applications of equations of state from pure fluids to mixtures. In this work, van der Waals mixing rule is used as follows:

$$A=\sum_{j=1}^{n}\sum_{q=1}^{n}x_jx_qA_{jq} \tag{23}$$

$$A_{jq}=(1-\delta_{jq})\sqrt{A_jA_q} \tag{24}$$

$$B=\sum_{j=1}^{n}x_jB_j \tag{25}$$

where δ_{jq} is the binary interaction parameter for the components j and q. All parameters in the above mixing rule for the H_2O-CO_2-H_2 mixtures can be found in other literature [52–54].

Duan-Møller-Weare equation of state. Compared with cubic equations of state, the DMW-EOS has a more complicated form with fourteen coefficients [15].

$$Z=\frac{P_zV_z}{RT_z}=1+\frac{a_1+a_2/T_z^2+a_3/T_z^3}{V_z}+\frac{a_4+a_5/T_z^2+a_6/T_z^3}{V_z^2}+\frac{a_7+a_8/T_z^2+a_9/T_z^3}{V_z^4}+\frac{a_{10}+a_{11}/T_z^2+a_{12}/T_z^3}{V_z^5}+\frac{a_{13}}{T_z^3V_z^2}\left(1+\frac{a_{14}}{V_z^2}\right)\exp\left(-\frac{a_{14}}{V_z^2}\right), \tag{26}$$

with

$$P_z=\frac{3.0626\tau^3P}{\mu}, \tag{27}$$

$$T_z=\frac{154T}{\mu}, \tag{28}$$

$$V_z=V_m\left(\frac{\tau}{3.691}\right)^{-3}, \tag{29}$$

where Z is the compression factor; τ and μ are the Lennard-Jones parameters; V_m is in L/mol. For mixture applications of the DMW-EOS, the Lorentz-Berthelot rules is used to mix the parameters μ and τ

$$\mu=\sum_{j=1}^{n}\sum_{q=1}^{n}x_jx_qC_{1,jq}\sqrt{\mu_j\mu_q}, \tag{30}$$

$$\tau=\sum_{j=1}^{n}\sum_{q=1}^{n}x_jx_qC_{2,jq}(\tau_j+\tau_q)/2, \tag{31}$$

where $C_{1,jq}$ and $C_{2,jq}$ are the mixing parameters describing the binary interaction between components j and q.

References

[1] Zhang ZG, Duan ZH. An optimized molecular potential for carbon dioxide. J Chem Phys 2005;122(21).
[2] Feng X-J, Liu Q, Zhou M-X, Duan Y-Y. Gaseous PVTx properties of mixtures of carbon dioxide and propane with the burnett isochoric method. J Chem Eng Data 2010;55(9):3400–9.
[3] Zhang XR, Yamaguchi H, Fujima K, Enomoto M, Sawada N. Theoretical analysis of a thermodynamic cycle for power and heat production using supercritical carbon dioxide. Energy 2007;32(4):591–9.
[4] Fan J, Hong H, Jin H. Power generation based on chemical looping combustion: will it qualify to reduce greenhouse gas emissions from life-cycle assessment? ACS Sustainable Chem Eng 2018;6(5):6730–7.
[5] Chakraborty A, Saha BB, Koyama S, Ng KC. Thermodynamic modelling of a solid state thermoelectric cooling device: temperature–entropy analysis. Int J Heat Mass Transf 2006;49(19):3547–54.
[6] Bang-Møller C, Rokni M. Thermodynamic performance study of biomass gasification, solid oxide fuel cell and micro gas turbine hybrid systems. Energy Convers Manag 2010;51(11):2330–9.
[7] Gilfillan SMV, Lollar BS, Holland G, Blagburn D, Stevens S, Schoell M, et al. Solubility trapping in formation water as dominant CO_2 sink in natural gas

fields. Nature 2009;458:614.
[8] Gaillard F, Scaillet B, Arndt NT. Atmospheric oxygenation caused by a change in volcanic degassing pressure. Nature 2011;478:229.
[9] Duan Z, Zhang Z. Equation of state of the H_2O, CO_2, and H_2O–CO_2 systems up to 10 GPa and 2573.15K: molecular dynamics simulations with ab initio potential surface. Geochem Cosmochim Acta 2006;70(9):2311–24.
[10] Guàrdia E, Martí J. Density and temperature effects on the orientational and dielectric properties of supercritical water. Phys Rev E 2004;69(1):011502.
[11] Soave G. Equilibrium constants from a modified Redlich-Kwong equation of state. Chem Eng Sci 1972;27(6):1197–203.
[12] Peng D-Y, Robinson DB. A new two-constant equation of state. Ind Eng Chem Fundam 1976;15(1):59–64.
[13] Wilczek-Vera G, Vera JH. Understanding cubic equations of state: a search for the hidden clues of their success. AIChE J 2015;61(9):2824–31.
[14] Lopez-Echeverry JS, Reif-Acherman S, Araujo-Lopez E. Peng-Robinson equation of state: 40 years through cubics. Fluid Phase Equilib 2017;447:39–71.
[15] Duan Z, Møller N, Weare JH. A general equation of state for supercritical fluid mixtures and molecular dynamics simulation of mixture PVTx properties. Geochem Cosmochim Acta 1996;60(7):1209–16.
[16] Ma J, Li J, He C, Peng C, Liu H, Hu Y. Thermodynamic properties and vapor–liquid equilibria of associating fluids, Peng–Robinson equation of state coupled with shield-sticky model. Fluid Phase Equilib 2012;330:1–11.
[17] Arvelos S, Rade LL, Watanabe EO, Hori CE, Romanielo LL. Evaluation of different contribution methods over the performance of Peng–Robinson and CPA equation of state in the correlation of VLE of triglycerides, fatty esters and glycerol+CO2 and alcohol. Fluid Phase Equilib 2014;362:136–46.
[18] Zhao H, Lvov SN. Phase behavior of the CO2–H2O system at temperatures of 273–623 K and pressures of 0.1–200 MPa using Peng-Robinson-Stryjek-Vera equation of state with a modified Wong-Sandler mixing rule: an extension to the CO2–CH4–H2O system. Fluid Phase Equilib 2016;417:96–108.
[19] Rupp M, Tkatchenko A, Müller K-R, von Lilienfeld OA. Fast and Accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett 2012;108(5):058301.
[20] Rossouw D, Burdet P, de la Peña F, Ducati C, Knappett BR, Wheatley AEH, et al. Multicomponent signal unmixing from nanoheterostructures: overcoming the traditional challenges of nanoscale x-ray analysis via machine learning. Nano Lett 2015;15(4):2716–20.
[21] Raccuglia P, Elbert KC, Adler PDF, Falk C, Wenny MB, Mollo A, et al. Machine-learning-assisted materials discovery using failed experiments. Nature 2016;533:73.
[22] Huan TD, Batra R, Chapman J, Krishnan S, Chen L, Ramprasad R. A universal strategy for the creation of machine learning-based atomistic force fields. npj Comput Mater 2017;3(1):37.
[23] Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: recent applications and prospects. npj Comput Mater 2017;3(1):54.
[24] Brockherde F, Vogt L, Li L, Tuckerman ME, Burke K, Müller K-R. Bypassing the Kohn-Sham equations with machine learning. Nat Commun 2017;8(1):872.
[25] Choi S, Shin JH, Lee J, Sheridan P, Lu WD. Experimental demonstration of feature extraction and dimensionality reduction using memristor networks. Nano Lett 2017;17(5):3113–8.
[26] Chmiela S, Tkatchenko A, Sauceda HE, Poltavsky I, Schütt KT, Müller K-R. Machine learning of accurate energy-conserving molecular force fields. Sci Adv 2017;3(5).
[27] Liu Z, Zhu D, Rodrigues SP, Lee K-T, Cai W. Generative model for the inverse design of metasurfaces. Nano Lett 2018;18(10):6570–6.
[28] Hang Z, Hippalgaonkar K, Buonassisi T, Løvvik O M, Sagvolden E, Ding D. Machine learning for novel thermal-materials discovery: early successes, opportunities, and challenges. ES Energy Environ 2018;2:1–8.
[29] Geng Z, Yang X, Han Y, Zhu Q. Energy optimization and analysis modeling based on extreme learning machine integrated index decomposition analysis: application to complex chemical processes. Energy 2017;120:67–78.
[30] Yuan X, Tan Q, Lei X, Yuan Y, Wu X. Wind power prediction using hybrid autoregressive fractionally integrated moving average and least square support vector machine. Energy 2017;129:122–37.
[31] Hansen K, Montavon G, Biegler F, Fazli S, Rupp M, Scheffler M, et al. Assessment and validation of machine learning methods for predicting molecular atomization energies. J Chem Theory Comput 2013;9(8):3404–19.
[32] Schutt KT, Glawe H, Brockherde F, Sanna A, Muller KR, Gross EKU. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. Phys Rev B 2014;89(20).
[33] Hansen K, Biegler F, Ramakrishnan R, Pronobis W, von Lilienfeld OA, Müller K-R, et al. Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. J Phys Chem Lett 2015;6(12):2326–31.
[34] Deringer VL, Csanyi G. Machine learning based interatomic potential for amorphous carbon. Phys Rev B 2017;95(9).
[35] Seko A, Maekawa T, Tsuda K, Tanaka I. Machine learning with systematic density-functional theory calculations: application to melting temperatures of single- and binary-component solids. Phys Rev B 2014;89(5):054303.
[36] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970;12(1):55–67.
[37] Faber F, Lindmaa A, von Lilienfeld OA, Armiento R. Crystal structure representations for machine learning models of formation energies. Int J Quantum Chem 2015;115(16):1094–101.
[38] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 1996;58(1):267–88.
[39] Rasmussen CKIW CE. Gaussian processes for machine learning. Cambridge: MIT Press; 2006.
[40] Vapnik VN. The nature of Statistical learning theory. Springer; 1995.
[41] Schölkopf B, Smola AJ, Williamson RC, Bartlett PL. New support vector algorithms. Neural Comput 2000;12(5):1207–45.
[42] Chang C-C, Lin C-J. Training *v*-support vector regression: theory and algorithms. Neural Comput 2002;14(8):1959–77.
[43] Libsvm Chang C-C, Lin C-J. ACM Trans Intell Syst Technol 2011;2(3):1–27.
[44] Schölkopf B, Smola AJ, Bach F. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press; 2002.
[45] Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys 1983;79(2):926–35.
[46] Harris JG, Yung KH. Carbon dioxide's liquid-vapor coexistence curve and critical properties as predicted by a simple molecular model. J Phys Chem 1995;99(31):12021–4.
[47] Yang Q, Zhong C. Molecular simulation of adsorption and diffusion of hydrogen in metal–organic frameworks. J Phys Chem B 2005;109(24):11862–4.
[48] Plimpton S. Fast parallel algorithms for short-range molecular dynamics. J Comput Phys 1995;117(1):1–19.
[49] Hoover WG. Canonical dynamics: equilibrium phase-space distributions. Phys Rev A 1985;31(3):1695–7.
[50] Mouahid A, Crampon C, Toudji S-AA, Badens E. Supercritical CO_2 extraction of neutral lipids from microalgae: experiments and modelling. J Supercrit Fluids 2013;77:7–16.
[51] Li Y, Guo L, Zhang X, Jin H, Lu Y. Hydrogen production from coal gasification in supercritical water with a continuous flowing system. Int J Hydrogen Energy 2010;35(7):3036–45.
[52] Yang X, Xu J, Wu S, Yu M, Hu B, Cao B, et al. A molecular dynamics simulation study of PVT properties for $H_2O/H_2/CO_2$ mixtures in near-critical and supercritical regions of water. Int J Hydrogen Energy 2018;43(24):10980–90.
[53] Meng L, Duan Y-Y, Wang X-D. Binary interaction parameter kij for calculating the second cross-virial coefficients of mixtures. Fluid Phase Equilib 2007;260(2):354–8.
[54] Nishiumi H, Gotoh H. Generalization of binary interaction parameters of Peng-Robinson equation of state for systems containing hydrogen. Fluid Phase Equilib 1990;56:81–8.