
Article:
Cresswell, A.J., Wheatley, R.J., Wilkinson, R.D. et al. (2016) FDCCS16 Molecular simulation of the thermophysical properties and phase behaviour of impure CO_2 relevant to CCS. Faraday Discussions, 192. pp. 415-436. ISSN: 1359-6640

https://doi.org/10.1039/C6FD00026F

CREATED USING THE RSC LaTeX PCCP ARTICLE TEMPLATE – SEE www.rsc.org/electronicfiles FOR DETAILS
PAPER www. rsc.org/faraday_d | Faraday Discussions

Molecular simulation of the thermophysical properties and phase behaviour of impure CO_2 relevant to CCS†

Alexander J. Cresswell[a], Richard J. Wheatley[b], Richard D. Wilkinson[c] and Richard S. Graham[a]*

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/c000000x

Impurities from the CCS chain can greatly influence the physical properties of CO_2. This has important design, safety and cost implications for the compression, transport and storage of CO_2. There is an urgent need to understand and predict the properties of impure CO_2 to assist with CCS implementation. However, CCS presents demanding modelling requirements. A suitable model must both accurately and robustly predict CO_2 phase behaviour over a wide range of temperature and pressure, and maintain that predictive power for CO_2 mixtures with numerous, mutually interacting chemical species. A promising technique to address this task is molecular simulation. It offers a molecular approach, with foundations in firmly established physical principals, along with the potential to predict the wide range of physical properties required for CCS. The quality of predictions from molecular simulation depends on accurate force-fields to describe the interactions between CO_2. Unfortunately, there is currently no universally applicable method to obtain force-fields suitable for molecular simulation.

In this paper we present two methods of obtaining force-fields: the first being semi-empirical and the second using ab-initio quantum-chemical calculations. In the first approach we optimise the impurity force-field against measurements of the phase and pressure-volume behaviour of CO_2 binary mixtures with N_2, O_2, Ar and H_2. A gradient-free optimiser allows us to use the simulation itself as the underlying model. This leads to accurate and robust predictions under conditions relevant to CCS. In the second approach we use quantum-chemical calculations to produce ab-initio evaluations of the interactions between CO_2 and relevant impurities, taking N_2 as an exemplar. We use a modest number of these calculations to train a machine-learning algorithm, known as a Gaussian Process, to describe these data. The resulting model is then able to accurately predict a much broader set of ab-initio force-field calculations at low comparatively numerical cost. This approach has the potential to lead to first-principles simulation of the thermody-

† Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/c000000x/

[a] *School of Mathematical Sciences, University of Nottingham, Nottingham NG7 2RD, UK. Tel: +44 115 951 3850; E-mail: Richard.Graham@nottingham.ac.uk*
[b] *School of Chemistry, University of Nottingham, Nottingham NG7 2RD, UK.*
[c] School of Mathematics and Statistics, University of Sheffield, Western Bank Sheffield, S10 2TN UK.

namic properties of impure CO_2, without fitting to experimental data.

1 Introduction

1.1 CCS Problem and Pipeline Operating Window

The carbon capture and storage depends upon safe and economical transport of CO_2. Pipelines already widely used hydrocarbon transport and are a strong possibility for this transport. Pipeline transport of CO_2 allows a large volume of material to moved directly from the source to the storage[1]. However, CO_2 output from large facilities is likely to contain a varying number and percentage of components, depending upon the capturing process. These impurities can significantly shift the thermophysical properties of the mixture, relative to that of pure CO_2. Impurities can change the pressure-density behaviour of a CO_2 fluid, with important consequences for pipeline efficiency and CO_2 metering. Furthermore, impurities can modify the vapour-liquid equilibrium (VLE) of CO_2, and, in particular, can extend the two-phase region. There are importance cost and safety issues associated with two-phase flow in pipelines so defining the minimum pressure to ensure homogeneous phase flow a key task. Therefore accurate modelling of the homogeneous phase density and the VLE of impure CO_2 is an important focus for CCS modelling. Recent work has defined the expected operating conditions for CCS pipelines[1,2] and the most efficient way of transporting CO_2 is in the homogeneous phase, at pressures close to the critical point. The upper transport temperature will be set by the compressor discharge temperature and the temperature limits of the pipeline and the lower temperature will correspond to the winter ground temperature of the surrounding soil[3]. Expected impurity levels are about $\lesssim 4\%$, with N_2 , O_2, Ar and H_2 being key impurities[1,2,4,5]. This range of pressure, temperature and impurity level define pipeline operating conditions and provide a target window for CCS-oriented modelling. However, CCS-relevant models should aspire to model a wider range of impurity fraction. This will ensure the robustness of the model, capture the impurity-rich vapour phase in two-phase flow and make the models applicable to abnormal pipeline operation.

1.2 Relevant Experiments

Literature reviews of data relevant to the CCS transport problem have been recently produced by the IMPACTS project[6] and from the thesis of Demetriades[7]. There is an extensive range of data for CO_2+N_2 mixtures, several data sets for CO_2+O_2 and CO_2+Ar mixtures and some very recently published data[5,8] on CO_2+H_2.

1.3 Equations of state

Equations of state (EoS) are a widely used modelling tool in CCS. EoS postulate expressions for the volume and temperature-dependence of either the fluid pressure[9,17] or the Gibbs free energy[11,12]. These models contain empirical terms that

described the deviation from ideal gas behaviour. To optimise their accuracy, EoS need to be calibrated by fitting their parameters to experimental measurements on CO_2 mixtures.

There is considerable uncertainly over which EoS is most appropriate for CCS modelling. Options range from simple cubic EoS, such as the Peng-Robinson model[9] , which are mathematically simple and numerically cheap, but which often fail to accurately predict the liquid properties. For pure CO_2, the Span-Wagner EoS[11] covers from the triple-point temperature up to very high pressures and temperatures with very high accuracy. Furthermore an EoS by Yokozeki[13] captures solid-liquid coexistence of pure CO_2. There are complex EoS for CO_2 mixtures, including the SAFT[14], PC-SAFT[15], GERG[12] , EOS-CG[6,16] and Demetriades models[17]. Some comparisons to CCS-relevant measurements have been made[6,17,18].

The majority of EoS are fundamentally based on empirical parametric expressions, which leads to several limitations. The quality of the agreement is very sensitive to the postulated parametric expression and the effectiveness of the parameter fitting; and there are no systematic and universally reliable approaches to either of these. Furthermore, the predictions can only be as good as the available measurements. Thus effective data fitting requires careful and laborious construction and fitting of the EoS to comprehensive experimental data. Even then extrapolation of the model is dubious and modest extrapolations in pressure, temperature or impurity concentration can lead to wayward model predictions, for even the most well-established EoS. For example, a quantitative failing of the GERG EoS[12] when predicting CO_2-H_2 phase behaviour has recently been demonstrated[5,17]. The SAFT family of EoS[14,15] has a basis in physical modelling, meaning they have the potential to address some of these issues. However, they still require fitting to experiments and it is not yet clear whether these outperform empirical EoS in terms of accuracy and robustness[6,17,18]. There is a clear need for modelling based on robust physical and chemical principles to complement EoS approaches.

1.4 Molecular simulation

CCS presents demanding modelling requirements. A suitable model must both accurately and robustly predict CO_2 phase behaviour over a wide range of temperature and pressure, and maintain that predictive power for CO_2 mixtures with numerous, mutually interacting chemical species. A promising technique to address this task is molecular simulation. It offers a molecular approach, with foundations in firmly established physical principals, along with the potential to predict the wide range of physical properties required for CCS.

Molecular simulations have a number of advantages over EoS. Macroscopic physical properties are predicted from models of the interactions between constituent molecules. Thus simulations have a more rigorous treatment of molecular mixing and the effect of temperature than EoS. Therefore we might expect improved robustness with respect to changes due to temperature and new impurities. Simulations can readily compute pressure-volume behaviour, phase separation and other properties such as specific heat, viscosity and speed of sound[19–21]. Simulations have some disadvantages, primarily their computational cost, partic-

ularly if results with very low statistical uncertainty are required. However, the ongoing growth in computing power is mitigating this. A further difficulty is that the accuracy of prediction depends on there being an effective and fast method to compute the interactions between the constituent molecules of the fluid of interest. Unfortunately, there is currently no universally applicable methods to compute intermolecular potentials that are suitable for molecular simulation.

In this paper we present two methods of obtaining force-fields: the first being semi-empirical and the second being from *ab-initio* quantum-chemical calculations. In the first approach we optimise the impurity force-field against measurements of the phase and pressure-volume behaviour of CO_2 mixtures, containing impurities that are common in CCS. We use a gradient-free optimiser, which enables us to use the simulation itself as the underlying model. Our approach leads to accurate and robust predictions under conditions relevant to CCS. In the second approach we use quantum-chemical calculations to produce *ab-initio* evaluations of the force-field between CO_2 and relevant impurities. We use a modest number of these calculations to train a machine-learning algorithm, known as a Gaussian Process, to describe these data. The resulting model is then able to accurately predict a much broader set of *ab-initio* force-field calculations at low comparatively numerical cost. This second approach has the potential to lead to first-principles simulation of the thermodynamic properties of impure CO_2, without fitting to experimental data.

2 Simulations from semi-empirical molecular interactions

All molecular simulations require a method to evaluate the interactions between molecules. Semi-empirical force-fields are most commonly used. Here a simple mathematical form is chosen for the potential between pairs of molecules, with a small number of parameters characterising the interaction between alike molecules, while further parameters define the cross-species interaction. The total interaction is the sum over all binary pairs. This summation neglects the influence of the surrounding molecules on the pair interactions. However, these "non-additive" interactions are known to be important for VLE and other properties[22]. In practice, the neglect of these higher-order interactions is compensated for by fitting binary force-field parameters to experiments. However, this fundamentally semi-empirical approach will, unavoidably, lead to a finite range of temperature and pressure applicability for a given parameterisation.

Currently available molecular force-fields for common CCS impurities are fitted to experiments on the pure fluid around its critical point[23]. This often involves temperatures that are much lower than the CCS operating window. Therefore they are generally unsuited to simulate mixtures at these higher temperatures. Some binary mixture simulations have been performed that optimise the mixing parameters but these are also for lower temperatures[24]. In this section we run molecular simulations for CO_2 mixtures at CCS-relevant temperatures, using literature force-fields. We then perform a re-optimisation of the impurity parameters, fitting directly to mixture measurements at CCS-relevant temperatures.

2.1 Background

The significant increase in available computing power in the last few decades has made computer simulation a very valuable and powerful tool. The Monte Carlo simulation method has become widely used to generate equilibrium properties. For example, a direct NPT simulation, where the number of molecules, pressure and temperature are specified and the density is predicted, leads to the pressure-density behaviour in the homogeneous phase. Furthermore very effective methods exist to model VLE behaviour.

2.1.1 Grand Equilibrium Monte Carlo. Grand Equilibrium Monte Carlo (GEMC) is a molecular simulation method that predicts two-phase coexistence behaviour[21]. The GE algorithm involves running Monte-Carlo (MC) simulations on two separate boxes, corresponding to the liquid gas phases, which sample the properties of these phases in coexistence. These simulations involve imposing the temperature and mol fraction on the liquid phase and then predicting the coexistence pressure, the gas mol fraction and the densities of the two coexisting phases. Full details of the method are described elsewhere[21,25].

2.1.2 Molecular Potentials. In semi-empirical force-fields, the Van der Waals forces are often represented by the Lennard-Jones (LJ) potential which details the repulsive and attractive interactions between atoms

$$U(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma^6}{r}\right)\right], \tag{1}$$

where r is the separation between two particles and ε and σ are the energy and size parameters respectively. For interactions between the particles of the same species, the LJ potential remains as above but for interactions between different particles, ε and σ are found using the Lorenz-Bertholet combining rules

$$\sigma_{ij} = \eta\frac{\sigma_i + \sigma_j}{2} \tag{2}$$

$$\varepsilon_{ij} = \xi\sqrt{\varepsilon_i\varepsilon_j}. \tag{3}$$

$$\tag{4}$$

These two new parameters η and ξ scale all of the LJ interactions between a given pair of particle types i and j and effectively allow arbitrary modification to the combination rules between two molecules while leaving the self interaction unchanged.

Electrostatic interactions can be problematic in molecular simulation because of the slow decay of the interaction with molecular separation. However, an efficient way to model electrostatic interactions in neutrally charged molecules is to use a point quadrupole potential, the formula for which is given in equation 18 of Deublein *et al.*[25]. This quadruple potential requires a single parameter per molecular species, the quadrupole moment, Q. As this interaction is electrostatic the expression is valid for the interaction between alike and distinct molecules.

2.1.3 Force-fields for CO_2 and CCS impurities. Force-Fields for more complex diatomic, triatomic and larger molecules can be constructed by combining multiple LJ and quadrupole sites together, with each individual contribution

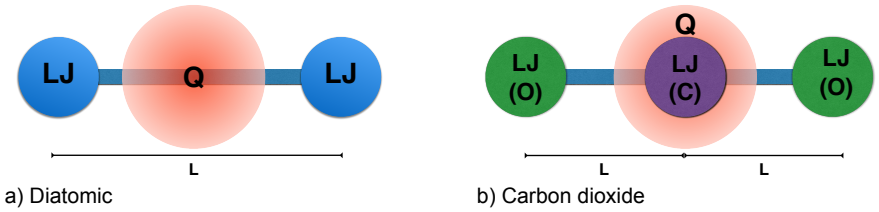


Fig. 1 Diagrams of two force-field types: a two centre Lennard-Jones with quadrupole (a) used for diatomic molecules and a three centre Lennard-Jones with quadrupole (b) used for CO_2.

summed to give the total potential. For example, a simple diatomic molecule can be created by fixing two identical LJ sites a specific bond length from one another and placing a quadrupole site at the centre (see figure 1a), an arrangement known as the two-centre Lennard-Jones plus quadrupole (2CLJQ) potential[26]. This force-field requires 4 parameters, ε and σ for the LJ terms, the bond length, L and the quadrupole moment, Q. 2CLJQ force-fields are available for diatomic molecules, including N_2, O_2[23] and H_2[5]. Similarly, a CO_2 force-field has been obtained using a three-centre Lennard-Jones plus point quadrupole model, to represent this triatomic molecule[27] (see figure 1b and table 1). As Ar is monatomic and non-polar, we use a single Lennard-Jones site with no quadrupole[23]. These literature force-fields were obtained by fitting self interactions to experimental data for the corresponding pure fluid around its critical point. For CO_2 this is close to the working temperature of CCS transport pipelines. Indeed, simulation results from this force-field are within $\sim 2\%$ of experiments on pure CO_2 for the density and vapour pressure in the CCS transport window[23]. However, the temperature ranges used to fit the impurity force-fields for N_2 and O_2 are less well matched to the CCS transport problem. The critical points for both of these materials are closer to 100K which is significantly lower than that of the 273-310K CCS region. These impurity force-fields perform well for the pure material around its critical point but, because of the empirical nature of the original fitting, they are unlikely to be perform as well at this much higher temperature. To compound this we also require that they predict properties of a mixture of which the impurity is only a small percentage. This is in contrast to the original fitting, which only used self interactions to optimise for the pure material properties[23].

ε_C/k_B (K)	σ_C (Å)	ε_O/k_B (K)	σ_O (Å)	Q (DÅ)	L (Å)
12.3724	2.8137	100.493	2.9755	4.0739	1.2869

Table 1 3CLJQ Force-field for pure CO_2[27].

There are some CO_2 binary mixtures that have been simulated with the available force-fields and good agreement with VLE data was obtained with the optimisation of the Lorenz-Bertholet mixing parameters[24]. However, this comparison does not use the detailed CO_2 force-field[27] and was performed for temperatures below the CCS transport window.

2.2 Simulations with literature force-fields

	Source	ε/k_B (K)	σ (Å)	Q (DÅ)	L (Å)	η	ξ
N_2	Vrabec[23]	34.897	3.3211	1.4397	1.0464	1.0000	1.0000
	Optimised	35.272	3.4267	1.3836	1.0594	1.0646	0.9970
O_2	Vrabec[23]	43.183	3.1062	0.8081	0.9699	1.0000	1.0000
	Optimised	43.649	3.1315	0.8036	0.9741	1.0010	1.0216
Ar	Vrabec[23]	116.79	3.3952	N/A	N/A	1.0000	1.0000
	Optimised	114.87	3.5462	N/A	N/A	1.0497	0.9839
H_2	Tenorio[5]	12.500	2.5900	-0.4233	0.7400	1.0000	1.1000
	Optimised	12.745	2.6742	-0.4243	0.7019	0.9929	1.1428

Table 2 2CLJQ Force-fields for nitrogen, oxygen, argon and hydrogen from the literature and optimised force-fields from this work.

We computed the homogeneous density and VLE behaviour of CO_2 binary mixtures using the GE method[21], implemented in the ms2 software[25], for each impurity in table 2. These simulations used the 3CLJQ force-field for CO_2 and 2CLJQ force-fields for the impurity, as detailed in table 1 and 2 respectively. In both the liquid and gas simulations we used 20,000 MC steps for equilibration and 100,000 production steps, from which the ensemble-average properties were extracted. The liquid simulations involved 800 particles and the gas runs had an average of 500 particles. The chemical potential in the liquid simulation was computed using Widom's insertion method, with 2,000 test insertions per MC step. The Lennard-Jones cutoff radius was chosen to be the maximum value allowed by the box size, which was typically larger than 17Å. These simulation settings have previously been shown to be adequate for quantitative simulations of CO_2 mixtures with comparable diatomic molecules[24].

The GEMC method imposes a temperature T and a liquid mole fraction x from which the coexistence properties are found. The method requires an initial estimate of the coexistence pressure P_0, although it does not have to be exact as the true value will be found from the vapour simulation. Therefore, we took the initial pressure estimate from experimental data for the mixture under investigation, interpolating to the desired mol fraction where necessary. If the simulated coexistence pressure differed significantly from the initial estimate P_0, we performed a second set of simulations where P_0 was set to the previously simulated value, thus ensuring the system had converged with respect to the choice of P_0. Using the simulation parameters above, a full coexistence simulation, comprising an NPT run followed by a GE run, takes around 20 hours on a single core of a 3GHz processor. A homogeneous density simulation takes between 3 and 4 hours.

Using the simulation parameters and literature force-fields in table 2, we performed a set of simulations for mixtures of CO_2 with the CCS-relevant impurities, N_2 O_2 Ar and H_2 and compared to coexistence and homogeneous measurements in the CCS temperature and pressure. We use CO_2+N_2, here, as an exemplar and the results are shown in figure 2. The agreement for the vapour mole

fraction is moderately poor everywhere, except for lower pressures at $T = 293K$. The coexistence pressure agreement for both temperatures worsens as the simulation pressure increases and, while the pressure performs better than that of the vapour mole fraction, the disagreement is sometimes outside the statistical error of the simulation. For the homogeneous, supercritical phase (figure 2b) the general agreement for the density is good with the exception of a single point in $T = 300K$. The results for O_2, Ar and H_2, shown in the Supplementary Information are entirely comparable to N_2. Generally, the homogeneous phase density is predicted well, but the predictions are unacceptable for the coexistence pressure and are more erroneous still for the gas mol fraction.

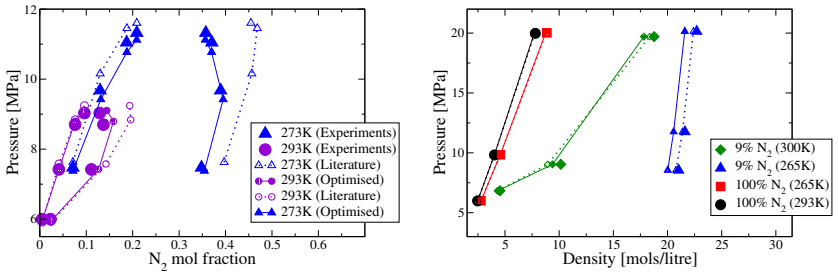


Fig. 2 Comparison between CO_2 + N_2 measurements (shaded symbol) and simulations using the literature force-field (open symbols) and the optimised force-field (shaded symbols), from table 2: (a) coexisting mol fraction (experimental data at 273K [28] and 293K [29]); and (b) homogeneous phase pressure-density behaviour (experimental data for mixtures [30] and pure N_2 [31]).

2.3 Optimisation of impurity force-fields

2.3.1 The simplex method. For each impurity in this study, we optimised the force-field to CCS-relevant data to improve the quantitative accuracy of simulations of CO_2 mixtures in the CCS region of temperature and pressure. Using binary mixture measurements as a basis for comparison, the parameters of the impurity force-fields were changed using an iterative optimisation method. The experimental data for CO_2 mixtures used for this optimisation is a combination of coexisting mol fraction, coexisting densities (where available) and homogeneous phase density data, along with density data for the pure impurity, all at several temperatures. We used a reduced set of representative data, to control the number of concurrent simulations needed. The experimental data used for optimisation correspond exactly to those data presented in the figures in the Supplementary Information.

During this optimisation, we held the CO_2 force-field fixed and made small adjustments to the values of the six impurity force-field parameters (ε, σ, Q, L, η and ξ, where the Lorenz-Bertholet mixing parameter refer to the interaction between CO_2 and the impurity), using the literature force-fields as a starting point. To quantify the disagreement between the simulation and experiments, we con-

structed an error function that sums over the square fractional deviation between experiments and simulations,

$$E(\varepsilon,\sigma,Q,L,\eta,\xi) = \sum_{i=1}^{N_c}\left(\frac{x_{\mathrm{v}i}^E - x_{\mathrm{v}i}^S}{x_{\mathrm{v}i}^E}\right)^2 + \sum_{i=1}^{N_c}\left(\frac{P_{\mathrm{v}i}^E - P_{\mathrm{v}i}^S}{P_{\mathrm{v}i}^E}\right)^2 + \sum_{i=1}^{N_{c\rho}}\left(\frac{\rho_{\mathrm{c}i}^E - \rho_{\mathrm{c}i}^S}{\rho_{\mathrm{c}i}^E}\right)^2 + \sum_{i=1}^{N_{hm}}\left(\frac{\rho_{\mathrm{m}i}^E - \rho_{\mathrm{m}i}^S}{\rho_{\mathrm{m}i}^E}\right)^2 + \sum_{i=1}^{N_p}\left(\frac{\rho_{\mathrm{p}i}^E - \rho_{\mathrm{p}i}^S}{\rho_{\mathrm{p}i}^E}\right)^2, \tag{5}$$

where E and S denote experiment and simulation, respectively, x_{v} is the coexisting vapour mol fraction, P_{v} is the coexisting vapour pressure, $\rho_{\mathrm{c}i}$ is the coexisting density (liquid or vapour), ρ_{m} is the homogeneous phase density for the CO_2 mixture, $\rho_{\mathrm{p}}$is the homogeneous phase density for the pure impurity, and N_c, $N_{c\rho}$, N_{hm} and N_p are the number of data point for the coexisting mol fraction, coexisting density, mixture density and pure additive density, respectively.

We optimised the model parameters by using the simplex method[32] to improve the error function. The simplex method is a simple downhill optimiser that will locate a local minimum in the error, which is desirable for this system as straying too far from the literature force-field would be unphysical. The simplex method is also gradient free, which is important for our system as generating accurate gradients of the error function by simulation is very expensive. For the initial simplex, we used the literature force-field and formed the simplex nodes by individually scaling-up all parameters by 5%, apart from η and ξ, which were scaled-up by 10%. This approach leads to a seven point simplex. As approximately 20 data points were used for each impurity, the initial simplex requires 140 simulations (the longest of which is $\sim$ 20 hours). However, as these simulations are independent they can be run in parallel on separate cores. Subsequent iterations require only a single simplex point, reducing the number of simulations to $\sim$ 20 per iteration. The simplex optimisation is ended when the difference between error terms at different points on the simplex becomes of the same order as the statistical uncertainties of the simulations. At this point simplex had shrunk sufficiently that no further meaningful improvement in the agreement could be achieved. This typically required 10-20 iterations.

2.3.2 CO_2 + N_2. Figure 2 shows for the optimised force-field, there is a clear improvement in the coexistence predictions for both temperatures in figure 2a) when predicting vapour pressure and, particularly, vapour mole fraction. Figure 2b) shows that, in the homogeneous single phase, predictions of the pure N_2 properties remain consistent with the experimental data as we intended and, while the mixture density does suffer slightly at 265K, this is more than compensated for by the large increase in accuracy in the coexistence phase. For CO_2+N_2 we held back the coexisting density data, to test whether the resulting optimised force-field could predict these data. Predictive results from the optimised force-field for these measurements are shown in figure 3b). The results are reasonable, but there is a noticeable deficiency in the predictions for high densities in both the homogeneous and coexisting phase (figure 2b and c). To address this for the optimisations of O_2, Ar and H_2) we include coexisting density data when

available. These optimisations proceeded in very similar manner to CO_2+N_2 and are detailed in the Supplementary Information.

2.4 Predictive simulations

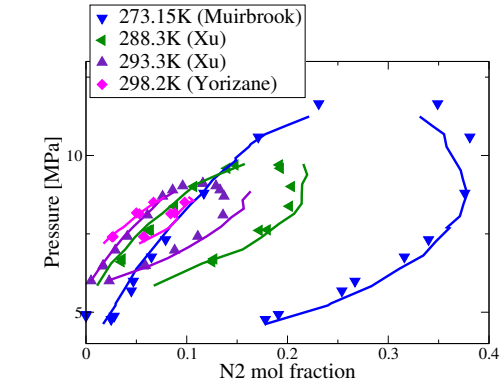
To test the robustness of the predictions from the optimised force-field we ran a series of predictive simulations to compare with CCS-relevant measurements. In all cases these included coexisting mol fractions and densities, along with homogeneous phase densities. The range of temperature, pressure and impurity fraction encompassed by these comparisons is much more extensive that the data used for fitting. For the coexistence simulations in the vicinity of the critical point pressure, we found noticeable noise in the gas mol fraction. To address this we used larger simulations, with 150,000 run steps and 1200 and 750 particles in the liquid and gas runs, respectively. We also averaged the predictions over 10 independent simulation runs of the same conditions. Comparison of the simulation predictions with a wide range of measurements are shown in figures 3- 7.

2.4.1 CO_2+N_2. In figure 3a) the liquid mol fraction is predicted very accurately, at all temperatures. The gas mol fraction is also predicted well but at high pressures minor noise and systematic deviation become evident. The coexisting density (figure 3b) shows good agreement for the gas density, apart from minor deviations approaching the critical pressure. This deviation is probably due to inaccuracies in the coexisting gas mol fraction, rather than an inability to predict the gas density at the correct mol fraction. The coexisting liquid density shows reasonable agreement but has clear underprediction in the approach to the critical point. This slight underprediction of the density in the high density region is also evident in the homogeneous phase predictions in figure 3c), where low and moderate densities are predicted well.

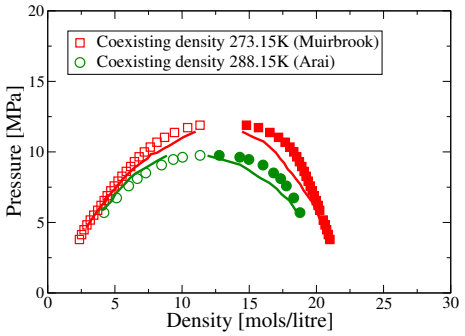
2.4.2 CO_2+O_2. The mol fraction predictions for CO_2+O_2 mixtures are shown in figure 4a). Here, as with CO_2+N_2, the liquid mol fraction is predicted accurately at all temperatures and the gas mol fraction is well captured everywhere except close to the critical point, where there is some systematic deviation, along with noise in the simulation results. In an improvement over the CO_2+N_2 performance, the coexisting and homogeneous phase densities are predicted accurately everywhere, including at high densities (see figure 4 b and c).

2.4.3 CO_2+Ar. Figure 5 shows that the CO_2+Ar results are comparable to the CO_2+N_2 results, in that the measurements are generally predicted accurately, except for the gas mol fraction close to the critical point and the pressure-density behaviour at high densities.

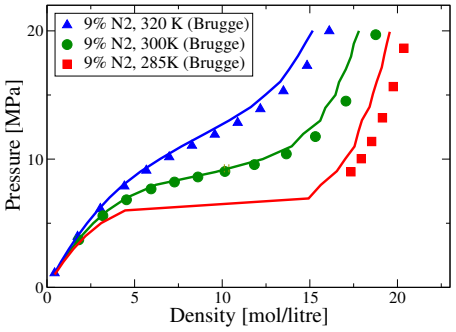
2.4.4 CO_2+H_2. In figure 6a) the simulations predict accurately the liquid mol fraction. The gas mol fraction is predicted reasonably accurately, although the predictions are systematically a little too high. At 295.K the measured mol fractions approach rapidly from about 12MPa, to give a critical point at abut 15MPa, a feature that the simulations do not capture even though the liquid mol fraction is correctly predicted at all pressures. It is noteworthy that the widely-used GERG equation of state drastically overpredicts these data [17], giving a critical pressure of 25MPa. The simulations predict accurately the density in the coexisting and homogeneous phases everywhere (see figure 6 b and c).



a)

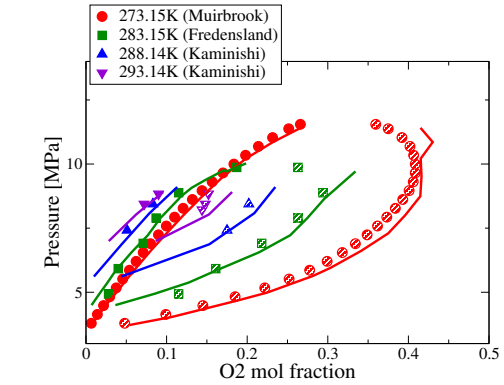


b)

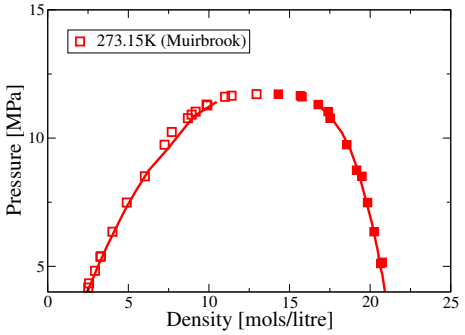


c)

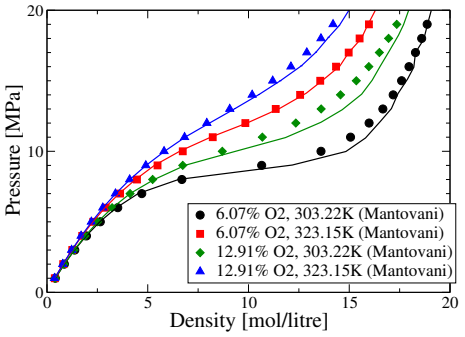
Fig. 3 Simulations (lines) for CO_2 + N_2 mixtures using our optimised force-field, compared with measurements (points) of coexisting mol fraction[28,29,33] (a); coexisting density[28,34] (b); and the pressure-density behaviour of the homogeneous phase[30] (c).



a)

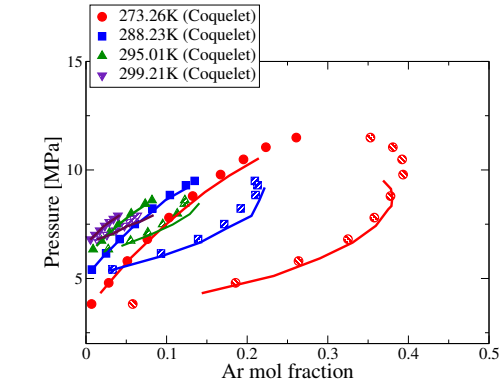


b)

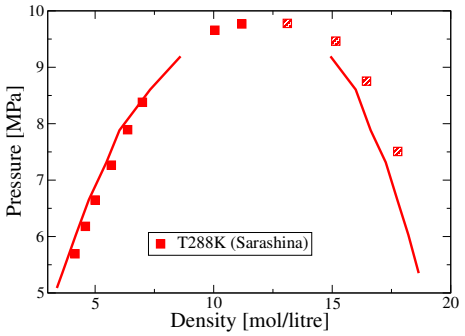


c)

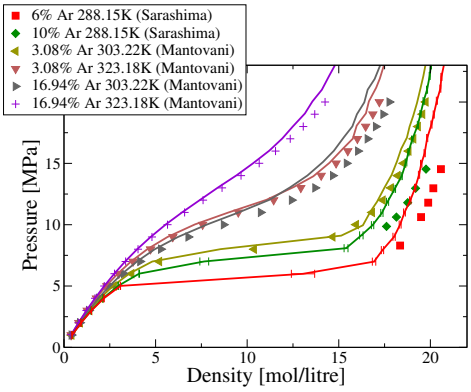
Fig. 4 Simulations (lines) for CO_2 + O_2 mixtures using our optimised force-field, compared with measurements (points) of coexisting mol fraction[28,35,36] (a); coexisting density[28] (b); and the pressure-density behaviour of the homogeneous phase[37](c).



a)

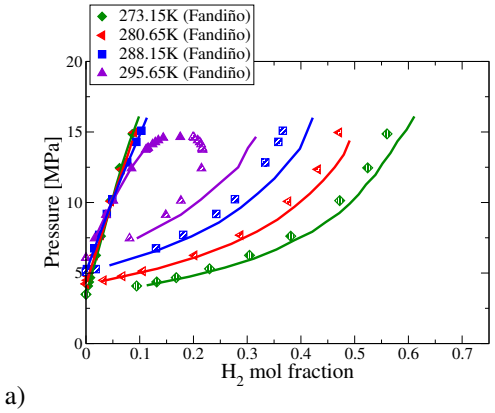


b)

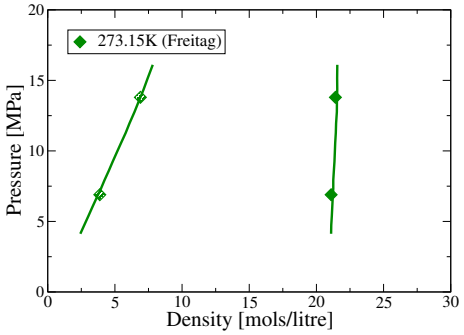


c)

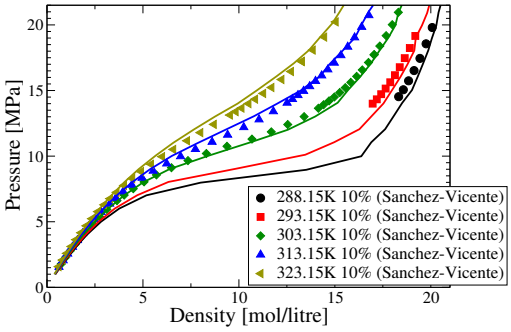
Fig. 5 Simulations (lines) for CO_2 + Ar mixtures using our optimised force-field, compared with measurements (points) of coexisting mol fraction[38] (a); coexisting density[39] (b); and the pressure-density behaviour of the homogeneous phase[37,39](c).



a)



b)



c)

Fig. 6 Simulations (lines) for CO_2 + H_2 mixtures using our optimised force-fields, compared with measurements (points) of coexisting mol fraction[8] (a); coexisting density[41] (b); and the pressure-density behaviour of the homogeneous phase[42](c).

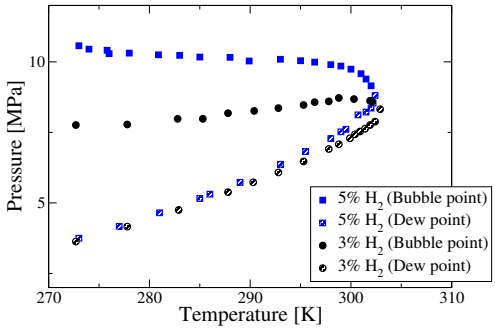


Fig. 7 Simulations (lines) for the CO_2 + H_2 bubble point computed from our optimised force-fields, compared with measurements of the bubble point[5] (closed symbols). Also shown are the measured dew points (patterned symbols).

Need to add predictions to figure 7: Figure 7 shows the results for the CO_2+H_2 phase envelope for H_2 concentrations of 3% and 5%, compared to measurements. The bubble point predictions were computed by direction simulation. However, the GE method is unsuitable for direction simulation of the dew point[5] so we obtained dew point predictions by interpolating the simulation data in figure 6a). The simulations predict accurately the bubble and dew points for all measurements. Also included is the prediction of the GERG EoS for the bubble point[40], which at 5%H_2, significantly underpredicts the bubble point. This has implications for pipeline design as the bubble-point defines the limit of the safe regime of operating pressures. The stronger molecular basis of our simulations leads to a more successful description of these data.

2.5 Application to EoS

We note that CO_2 thermophysical properties are usually modelled using EoS. EoS are significantly cheaper numerically than simulations and, with sufficiently effective fitting to high-quality measurements, give more accurate correlations of the data. However, the ability of EoS to extrapolate to regions where measurements are unavailable, is questionable, even for small changes in temperature or impurity fraction (for example, see the GERG predictions in figure 7). This is due to either their lack of physical basis or mathematical approximations. In contrast to an EoS, molecular simulation provides a physical model, arising from the interactions between CO_2 and the relevant impurity. Thus, if a suitable force-field is available, simulations can provide more robust predictions in regions where no optimisation or fitting has been done. Thus there is a key complementary role for molecular simulation in CCS modelling. We propose, here, two methods by which simulations can improve EoS fitting. Firstly, where experimental data are sparse, simulation predictions can be used as surrogate experimental data and included in the fitting, to impose improved extrapolation onto the EoS. This will

be particularly useful in regions where experiments are prohibitively difficult, expensive or unsafe. Secondly, often EoS are fitted to coexistence data via constrained optimisation[10,17]. The constraints are more readily applied if the full complement of coexistence data is available for each pressure and temperatures. This requires measurements of the density at coexistence for both the liquid and the vapour phases. For many experimental techniques the co-existing density is not measured. Indeed we recently developed an *ad hoc* method of interpolating to these density data when fitting an EoS[17]. However, replacing this method with simulation data is an attractive alternative. Using simulation data in EoS fitting will provide a more complete picture of coexistence and numerically more straightforward fitting. Both of these exploit the superior robustness of simulation predictions, with respect to changes in temperature, pressure and impurity fraction.

2.6 Summary and future extensions

We performed simulations of binary mixtures of CO_2 with N_2, O_2, Ar and H_2 using two-centre Lennard-Jones plus quadrupole force-fields and compared the results against mixture data in the CCS region of temperature and pressure. Although the literature force-field for pure CO_2 was suitable for CCS transport modelling, force-fields for the impurities did not given sufficiently accurate predictions at temperatures relevant to CCS. Therefore we re-optimised the impurity force-fields by fitting to selected VLE and homogeneous phase measurements for binary CO_2 mixtures, under conditions relevant to CCS transport. We then used these new force-fields to predict a much broader set of CCS-relevant measurements. For all impurities, the simulations generally gave very good predictions for all liquid mol fraction. The gas mol fraction predictions were good, except close to the critical point where they showed some minor noise and systematic deviation from the measurements. The density predictions at high density were somewhat erroneous for CO_2 with N_2 and Ar but were very good throughout for O_2 and H_2. We have focused on binary mixtures herein, however the impurity force-fields developed herein can be used to compute the properties of CO_2 mixtures with many impurity species. The force-fields will also provide useful predictions for quantities such as specific heat, viscosity and speed of sound. Our optimisation method will allow force-fields for other impurities to be similarly optimised where experimental data exists. In the longer term, the optimisation method in this work can be applied to lower temperatures than we studied here, down to the triple point of CO_2. This will aid in modelling burst CO_2 transport pipes, as the mixture will cool when it escapes the high pressure environment within the pipe. This may necessitate impurity force-fields that include a weak temperature dependence if they are to be effective over a larger temperature range.

3 *Ab-initio* molecular interactions

We demonstrated in section 2 the potential for molecular simulation to improve upon and complement modelling by EoS. Above we used semi-empirical molecular force-fields, which were optimised to measurements in the relevant regime

of pressure, temperature and impurity concentration. Despite it successes, this approach has several limitations. Even with fitting, there is not close agreement with experiments everywhere. Furthermore, fitting simulations is numerically expensive. Finally, the semi-empirical method requires experimental data for fitting, although it needs less extensive data than EoS fitting. These limitation arise from the semi-empirical nature of the fitted force-field. In contrast, fully ab initio approaches[22] are also possible, in principle, and have the potential to address these issues.

In this section we illustrate a nascent technique, with the potential, ultimately, to lead to completely ab initio predicts of thermophysical properties, such as those studied above. Computational chemistry has advanced to the stage where calculations of intermolecular potential energies can be accurately computed, for small molecules, from *ab initio* quantum calculations. However, the computational cost of evaluating the energy at a single point is too large (often minutes or hours of time per pair of molecules) to be practical within a molecular simulations. Thus it is necessary to fit or interpolate calculated energy data to produce a potential energy surface. Approaching this problem with traditional parametric fitting techniques is often laborious and its effectiveness is contingent on a good initial choice of parametric function. Good choices for a particular molecular interaction may not translate to other chemistries.

We follow a procedure, proposed by Uteva *et al*[43] to produce intermolecular potential energy surfaces efficiently from a relatively small number of training data. This non-parametric approach uses a machine learning technique to directly learn the mathematical structure from the data; no selection of parametric function is necessary. The use of machine learning suggests that this algorithm may be more readily generalised to new interactions. Indeed the technique has been readily applied to several distinct chemical systems, without modification[43]. Here we apply the algorithm to interpolation of the CO_2-N_2 interaction. We use quantum-chemical calculations to produce ab-initio evaluations of the intermolecular potential between the CO_2-N_2 binary pair. Using a modest number of these calculations, we to train a machine-learning algorithm, known as a Gaussian Process, to describe these data. The resulting model is then able to accurately predict a much broader set of ab-initio force-field calculations at low comparatively numerical cost.

3.1 Intermolecular potential data for CO_2-N_2

Data sets of the intermolecular interaction energy of the bimolecular complex CO_2+N_2 are calculated as a function of their configurational geometry. All molecules are approximated as linear rigid rotors in their vibrational ground state, with fixed bond lengths of 1.1632Å and 1.0975Å for the C-O and N-N bonds, respectively. Energy calculations are carried out in Molpro[44] using second-order Möller-Plesset perturbation theory (MP2) and augmented correlation-consistent triple-zeta (aug-cc-pVTZ) basis sets. Basis set superposition errors are corrected using the full counterpoise correction procedure. An energy cutoff of $E_{\mathrm{cut}} = 0.005\ E_{\mathrm{h}}$ is imposed and configurations with intermolecular potentials above this cutoff are excluded from the training and test data sets. Configurations are also excluded if any interatomic distance is below 1.5 Å or if all interatomic distances

are above 8.5 Å.

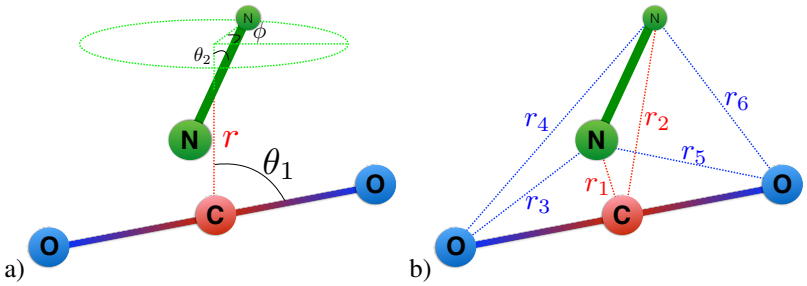


Fig. 8 The geometry of a CO_2+N_2 pair, described by angles and centre of mass distance (a) and interatomic distances (b).

As with many molecular systems, the CO_2-N_2 interaction contains several symmetries. For example, the intermolecular potential is unchanged by the interchange of both N atoms or both O atoms (see ESI for full details). Because of these symmetries we generate data over the region $1.5\text{Å} < r < 10\text{Å}$, $0 < \theta_1 < \pi/2$, $0 < \theta_2 < \pi/2$ and $0 < \phi < \pi$, where r is the distance between the molecular centres, θ_1 is the angle between r and the CO_2 axis, θ_2 is the angle between r and the N_2 axis, and ϕ is the torsional angle of the N_2 axis (see figure 8a). This defines a symmetry-distinct sub-region, which is the smallest region of space from which the behaviour for all space can be inferred via the symmetries.

3.2 Gaussian Processes

Gaussian processes (GPs)[45] are used extensively in machine learning and statistics as regression models. They are 'non-parametric' models of functions, which generalise the Gaussian distribution. The prior specification of a GP consists of a mean function (often taken as zero) and a covariance function $k(\mathbf{x},\mathbf{x}')$, expressing the covariance between $f(\mathbf{x})$ and $f(\mathbf{x}')$, where f is the function being interpolated. Training data, consisting of observations of the value of f at various locations, update the mean and covariance functions to give a posterior model which predicts the function at any location.

Properties of the GP model are inherited from the covariance function, for example, symmetry, differentiability and stationarity. Stationarity is a common assumption when using GPs. It assumes that the covariance function depends only on the distance $|\mathbf{x}-\mathbf{x}'|$ and not the individual positions $\mathbf{x}$ and $\mathbf{x}'$. However, the intermolecular energy is a non-stationary function of distance, as it varies rapidly at small interatomic separations, but more gently at larger separation. To deal with this non-stationarity we use the inverse interatomic distances as covariates in the GP, to achieve approximate stationarity. Thus the GP coordinates are $\mathbf{x} = (1/r_1, \ldots, 1/r_{N_D})$ where r_i is the interatomic distance, running over all pairs of nuclei on different molecules. Thus for $CO_2{}^+N_2$ this results in an over-specified system, with $N_D = 6$ dimensions, of which 4 are independent.

Our GP has a zero mean function and a squared-exponential covariance func-

tion

$$\kappa(\mathbf{x},\mathbf{x}') = \prod_{i=1}^{N_D} \exp\left[-\frac{(x_i - x_i')^2}{2l_i^2}\right] \tag{6}$$

where l_i is the correlation length for each dimension. We generalise this function to respect the symmetries of the system. The symmetries mean that the intermolecular potential is invariant under several permutations of the inverse interatomic distances x (see the ESI for full details). Let G represent the permutation group containing permutations of elements of $\mathbf{x}$ under which the energy surface is unchanged. If $l_i = l_j$ for all coordinates x_i and x_j that swap for some permutation in G, then a covariance function of the form

$$k_{\mathrm{sym}}(\mathbf{x},\mathbf{x}') = \sigma_f^2 \sum_{g\in G} \kappa(g\mathbf{x},\mathbf{x}'), \tag{7}$$

where σ_f^2 is the signal variance, results in a GP that shares the symmetries of the energy surface (see the Supplementary Material). The 'symmetric model' based on this covariance function gives predictions that respect the relevant symmetries, and usually significantly improves the performance[43], even within the symmetry-invariant region covered by the test data, as shown below.

3.2.1 Latin hypercube data The training and test data should ideally cover evenly a single symmetry-distinct sub-region of $\mathbf{x}$ space, and respect the geometric constraint. We generate candidate co-ordinate sets of the desired size from Latin hypercube (LHC) sampling of $1/r$, $\cos\theta_1$, $\cos\theta_2$ and ϕ, on the ranges specified above. We then delete points that violate the geometric constraints, reject the entire LHC if it does not contain at least the target number of points and compute the minimum separation of the remaining points in $\mathbf{x}$ space. We repeat this process over a large number of iterations and the candidate data set with the largest minimum separation is then used in Molpro energy calculations. This 'maximin' approach aims to cover evenly the symmetric distinct sub-region of $\mathbf{x}$ space.

3.3 Results

Results are obtained using the GPy package[46], modified to include symmetric covariance functions. Zero-mean Gaussian observation error[45] is assumed on the function outputs (referred to as *nugget* in geostatistics), with standard deviation σ_n. Thus the model's hyperparameters are σ_f, σ_n and $\{l_i\}$. These hyperparameters are estimated by optimising the log-likelihood over ≈ 30 random restarts, which typically is sufficient to find the optimal values multiple times. We repeated this process for a range of different sizes of training data, all generated via the LHC algorithm described above. The accuracy of the interpolation was then tested by computing the root mean square error (RMSE) for the GP interpolations against a much larger test data set. The test set comprised of a very large LHC of size $\sim 15,000$ points (also generated via the same algorithm). Figure 9 shows the RMSE for increasing number of training points. When compared a fixed number of training points, the symmetric covariance function typically gives an RMSE that is 3-7 time more accurate than the non-symmetric version. An RMSE of 5×10^{-6} E_h corresponds to a mean error of 0.5% of the high energy cut off, and this is achieved by the symmetric model with ~ 200 training points.

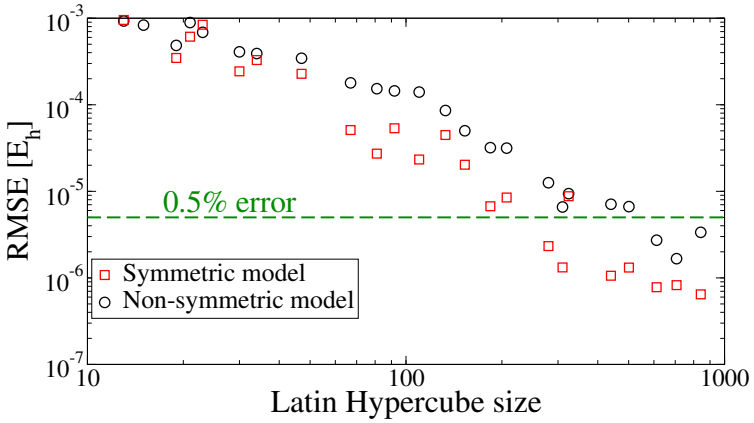


Fig. 9 RMSE against number of training points (LHC size) for CO_2+N_2. The lowest energy in the test data is -1.38×10^{-3} E_h.

3.4 Summary

We used quantum-chemical calculations to produce ab-initio evaluations of the interactions between CO_2 and N_2. Via a modest number of these calculations, we trained a machine-learning algorithm, known as a Gaussian Process, to describe these data. The resulting model accurately predicts a much broader set of ab-initio force-field calculations at low comparatively numerical cost. Several pieces of future work are necessary to be able to implement GP force-fields into molecular simulations such as those from our semi-empirical force-fields in section 2. These are extension of the technique to non-additive (three-body) interactions; optimisation of the number and placement of training data and streamlining of the computation of the predicted potential using the symmetric covariance function. This work leads directly from the algorithm of Uteva[43], used here, and has the potential to lead to first-principles simulation of the thermodynamic properties of impure CO_2, without fitting to experimental data.

4 Conclusions

The goal of this work was to develop methods by to predict the thermophysical properties of impure CO_2, to aid in the design and safe operation of CCS technology, particularly the pipeline transport. Robust modelling methods are needed because available experimental data do not comprehensively cover the relevant temperature, pressure and impurity regime. Simulation is a promising molecular approach, with foundations in firmly established physical principals. However, molecular simulation requires suitable force-fields to describe the interaction between CO_2 molecules and the impurities and we explored two methods to obtain such force-fields. We demonstrated that, even with semi-empirical force-fields this leads to VLE and pressure-density predictions that, when compared to EoS, are more robust to changes to temperature, pressure and impurity fraction and re-

quire far less comprehensive fitting data. We also described a method to produce ab initio potential energy surfaces from quantum-chemical calculations, applying the method to the CO_2+N_2 binary pair. This method has the potential to deliver first-principles simulation of the thermodynamic properties, provided the extensions to the method outlined above can be achieved.

Future work for the semi-empirical forcefields will involve simulating a wider range of CCS-relevant properties, such as specific heat, viscosity and speed of sound, comparing to measurements on ternary and higher order mixtures, and extending the range of pressure and temperature. To obtain first-principles the method for ab initio potential energy surfaces needs to be generalised to three-body interactions and implemented within a molecular simulation.

Acknowledgements

This work was supported by the Materials for Next Generation CO_2 Transport Systems (MATTRAN) project, funded by the EPSRC and E.ON (EP/G061955/1). We thank Mike George, Jie Ke and the other MATTRAN project partners for very useful discussions concerning this work. We thank Prof. Roland Span for making available to us the literature experimental data compiled during the IMPACTS project. We are grateful for access to the University of Nottingham High Performance Computing Facility used in this work.

References

1 P. N. Seevam, D. J. M. Race and P. M. J. Downie, *Journal of Pipeline Engineering*, 2007, 1–14.
2 E. de Visser, C. Hendriks, M. Barrio, M. J. M. lnvik, G. de Koeijer, S. Liljemark and Y. L. Gallo, *International Journal of Greenhouse Gas Control*, 2008, **2**, pp. 478—484.
3 M. Mohitpour, P. Seevam, K. K. Botros, B. Rothwell and C. Ennis, *Pipeline Transportation of Carbon Dioxide Containing Impurities*, ASME Press, New York, NY, USA, 2012.
4 R. T. Porter, M. Fairweather, M. Pourkashaniana and R. M. Woolley, *International Journal of Greenhouse Gas Control*, 2015, **36**, 161174.
5 M.-J. Tenorio, A. J. Parrott, J. A. Calladine, Y. Sanchez-Vicente, A. J. Cresswell, R. S. Graham, T. C. Drage, M. Poliakoff, J. Ke and M. W. George, *International Journal of Greenhouse Gas Control*, 2015, **41**, 68–81.
6 G. J. Gernert, *PhD thesis*, Faculty of Mechanical Engineering, University of Bochum, 2013.
7 T. A. Demetriades, *PhD thesis*, University of Nottingham, 2014.
8 O. Fandiño, J. M. Trusler and D. Vega-Maza, *International Journal of Greenhouse Gas Control*, 2015, **36**, 78–92.
9 D. Peng and D. B. Robinson, *Ind Eng Chem Fund*, 1976, **15**, 59–64.
10 T. A. Demetriades, T. C. Drage and R. S. Graham, *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering*, 2013, **227**, 117–124.
11 R. Span and W. Wagner, *J. Phys. Chem. Ref. Data*, 1996, **25**, pp. 1509—1596.
12 O. Kunz, R. Klimeck, W. Wagner and M. Jaeschke, *The GERG-2004 wide-range reference equation of state for natural gases*, Fortschr.-Ber. VDI, VDI-Verlag, Düsseldorf, 2007.
13 A. Yokozeki, *Fluid Phase Equilibria*, 2004, **222-223**, 55–66.
14 W. G. Chapman, K. E. Gubbins, G. Jackson and M. Radosz, *Fluid Phase Equilibria*, 1989, **52**, pp. 31—38.
15 J. Gross and G. Sadowski, *Ind. Eng. Chem. Res.*, 2001, **40**, 1244–1260.
16 R. Span, J. Gernert and A. Jäger, *Energy Procedia*, 2013, **37**, 2914–2922.
17 T. A. Demetriades and R. S. Graham, *The Journal of Chemical Thermodynamics*, 2016, **93**, 294–304.
18 N. I. Diamantonis, G. C. Boulougouris, E. Mansoor, D. M. Tsangaris and I. G. Economou, *Ind. Eng. Chem. Res.*, 2013, 130227083947003.

19 M. Allen and D. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, 1989.
20 D. C. Rapaport, *The Art of Molecular Dynamics Simulation*, Cambridge University Press, 2004.
21 J. Vrabec and H. Hasse, *Molecular Physics*, 2002, **100**, 3375–3383.
22 M. T. Oakley and R. J. Wheatley, *The Journal of Chemical Physics*, 2009, **130**, 034110.
23 J. Vrabec, J. Stoll and H. Hasse, *The Journal of Physical Chemistry B*, 2001, **105**, 12126–12133.
24 J. Stoll, J. Vrabec and H. Hasse, *Aiche J*, 2003, **49**, 2187–2198.
25 S. Deublein, B. Eckl, J. Stoll, S. V. Lishchuk, G. Guevara-Carrion, C. W. Glass, T. Merker, M. Bernreuther, H. Hasse and J. Vrabec, *Computer Physics Communications*, 2011, **182**, 2350–2367.
26 J. Stoll, J. Vrabec, H. Hasse and J. Fischer, *Fluid Phase Equilibria*, 2001, **179**, 339–362.
27 T. Merker, C. Engin, J. Vrabec and H. Hasse, *The Journal of chemical physics*, 2010, **132**, 234512.
28 N. Muirbrook and J. Prausnitz, *AIChE Journal*, 1965, **11**, 1092–1096.
29 N. Xu, J. Dong, Y. Wang and J. Shi, *Fluid phase equilibria*, 1992, **81**, 175–186.
30 H. B. Brugge, J. C. Holste, K. R. Hall, B. E. Gammon and K. N. Marsh, *Journal of Chemical & Engineering Data*, 1997, **42**, 903–907.
31 E. Lemmon, M. McLinden and D. Friend, *Thermophysical Properties of Fluid Systems in NIST Chemistry WebBook, NIST Standard Reference Database Number 69,Eds. P.J. Linstrom and W.G. Mallard, National Institute of Standards and Technology, Gaithersburg MD, 20899*, 2011, `http://webbook.nist.gov`.
32 W. H. Press, S. A. Teukolsky, B. P. Flannery and W. T. Vetterling, *Numerical Recipes in C:The Art of Scientific Computing Second Edition*, Cambridge University Press, 1988.
33 M. Yorizane, S. Yoshimura and H. Masuoka, *Chemical Engineering*, 1970, **34**, 953–957,a1.
34 Y. Arai, G. Kaminishi and S. Saito, *J. Chem. Eng. Japan*, 1971, **4**, 113–122.
35 A. Fredenslund and G. A. Sather, *Journal of Chemical and Engineering Data*, 1970, **15**, pp. 17—22.
36 G. Kaminishi and T. Toriumi, *Kogyo Kagaku Zasshi*, 1966, **69**, 175–178.
37 M. Mantovani, P. Chiesa, G. Valenti, M. Gatti and S. Consonni, *The Journal of Supercritical Fluids*, 2012, **61**, 34–43.
38 C. Coquelet, A. Valtz, F. Dieu, D. Richon, P. Arpentinier and F. Lockwood, *Fluid Phase Equilibria*, 2008, **273**, 38–43.
39 E. Sarashina, Y. Arai and S. Sasto, *Journal of Chemical Engineering of Japan*, 1971, **4**, 379–381.
40 O. Kunz and G. E. de Recherches Gazières, *The GERG-2004 wide-range equation of state for natural gases and other mixtures*, VDI Verlag, 2007.
41 N. P. Freitag and D. B. Robinson, *Fluid Phase Equilibria*, 1986, **31**, pp. 183—201.
42 Y. Sanchez-Vicente, T. C. Drage, M. Poliakoff, J. Ke and M. W. George, *International Journal of Greenhouse Gas Control*, 2013, **13**, 78–86.
43 E. Uteva, R. S. Graham, R. D. Wilkinson and R. J. Wheatley, *Chemical Communications*, Under review.
44 H. J. Werner *et al.*, MOLPRO *version 2012.1: A package of ab initio programs*, http://www.molpro.net, 2012.
45 C. Rasmussen and C. K. I. William, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
46 *GPy: A Gaussian process framework in python*, http://github.com/SheffieldML/GPy, 2012–2015.

