PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2021, **23**, 2891

Predicting second virial coefficients of organic and inorganic compounds using Gaussian process regression

Miruna T. Cretu [ab] and Jesús Pérez-Ríos [*b]

Received 21st October 2020,
Accepted 11th January 2021

DOI: 10.1039/d0cp05509c

rsc.li/pccp

We show that by using intuitive and accessible molecular features it is possible to predict the temperature-dependent second virial coefficient of organic and inorganic compounds with Gaussian process regression. In particular, we built a low dimensional representation of features based on intrinsic molecular properties, topology and physical properties relevant for the characterization of molecule-molecule interactions. The featurization was used to predict second virial coefficients in the interpolative regime with a relative error $\lesssim$1% and to extrapolate the prediction to temperatures outside of the training range for each compound in the dataset with a relative error of 2.1%. Additionally, the model's predictive abilities were extended to organic molecules unseen in the training process, yielding a prediction with a relative error of 2.7%. Test molecules must be well-represented in the training set by instances of their families, which are high in variety. The method shows a generally better performance when compared to several semi-empirical procedures employed in the prediction of the quantity. Therefore, apart from being robust, the present Gaussian process regression model is extensible to a variety of organic and inorganic compounds.

1 Introduction

The long-standing goal to establish a relationship between the behaviour of a gas and its microscopic properties has admirably been achieved by the virial equation of state.[1,2] Besides offering a rigorous depiction of pressure, $p(T,\rho)$ as a function of temperature,[3] T, and density, ρ, the virial equation is founded on a solid statistical mechanics framework.[4] The virial equation,

$$\frac{p}{RT\rho} = 1 + \sum_{i=2}^{\mathcal{N}} B_i(T)\rho^{i-1}, \qquad (1)$$

encapsulates the departure from ideality of a gas in an infinite series of temperature-dependent coefficients, $B_i(T)$, which correspond to the molecular interaction in isolated clusters of size i. $B_i(T)$ is the i-th virial coefficient and is related to the role of i-body interactions in a system. In eqn (1) R denotes the ideal gas constant and the series is truncated up to a certain cluster size $\mathcal{N}$.

Two-body interactions are the most relevant interactions to the macroscopic properties of a gas,[5] hence B_2 values have been tabulated for many compounds.[6] Since B_2 can be derived from intermolecular potentials, the latter can be obtained from experimental B_2 through a proper parametrisation of the potential function.[7,8] This is conducive to the calculation of fluid properties such as enthalpy of vaporisation[9] and transport coefficients.[9–11] The knowledge of B_2 also helps to estimate critical points[12] and optimum conditions for crystal growth, which would otherwise require extensive screening experiments.[13]

When it comes to the determination of the second virial coefficient, computational cost and experimental obstacles often come into play. The theoretical approach to estimate B_2 from the interaction potential was developed ever since the 1930s[14,15] and is adapted nowadays to more complex potential functions. However, the process is computationally expensive for all but simple molecules. Furthermore, experimental procedures give accurate results for certain ranges of temperature, however they are faced with the challenge to acquire reliable compressibility data.[16] In the case of empirical approaches, the law of corresponding states[17] leads, in some cases, to very accurate results, whereas in other situations, the accuracy is low.

To provide an alternative to the traditional methods of calculating B_2, we propose tackling the problem within the new paradigm of data-intensive science.[19] The existence of a large and high quality database of temperature-dependent second virial coefficients[6] fulfills the most vital prerequisite for the application of machine learning. The choice of input features for learning is then a matter of physical and computational intuition. Among notable previous works on B_2 estimation is the

[a] *Department of Chemistry, Imperial College London, London SW7 2AZ, UK*
[b] *Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany. E-mail: jperezri@fhi-berlin.mpg.de*

one of Di Nicola *et al.*,[20] which uses thermodynamic input features and artificial neural networks (ANN) to predict B_2 with high accuracy. This method, however, requires the construction of a complex ANN, together with the knowledge of five thermodynamic properties, which are difficult to obtain, as discussed above. As the authors also suggest, this method should only be used when "high accuracy is required",[20] due to its complexity. Furthermore, the prediction of second virial coefficients has been addressed before by Mokshyna *et al.*,[21] by assuming a functional form of the dependency of B_2 on temperature. The methodology involved modelling molecular structures using the Simplex Representation of Molecular Structure (SiRMS),[22] based on which a Random Forest model was trained to predict second virial coefficients.

In this paper, we propose the prediction of second virial coefficients of organic and inorganic compounds in a simple, universal manner. Our approach is based on Gaussian Process Regression (GPR) fed with a low dimensional input featurization scheme (see Fig. 1). To offer perspective to the reader, it is of importance to introduce the potential contexts in which the prediction of second virial coefficients is desired. That is, one might want to predict the quantity outside of an already studied temperature range, as well as predict second virial coefficients for new molecules. Our work addresses both aspects, but we conclude that for accurate results, the new molecule should belong to classes of compounds already studied in the training set.

We show that the chosen features succeed to incorporate the most relevant characteristics of the second virial coefficient, in that the model succeeds to predict B_2 of an unseen molecule with a relative error of 2.7%, over whole ranges of temperatures. Moreover, the power of our model is reinforced by the successful extrapolation (relative error 2.1%) to temperatures outside of the training range for any molecule in the dataset. Our method's universality stems from its applicability to compounds belonging to a wide range of families and from the availability and accessibility of input features for any compound. The simplicity stems from the facile practice to generate input features and from the ease of applying computationally inexpensive GPR (for the number of data points considered in this work). Different featurization combinations were tested to yield the best, lowest dimensional scheme finally. All the input data were generated using RDKit,[18] an open-source toolkit for cheminformatics implemented in Python. While most of the features we used are basic molecular properties of compounds, the Morgan fingerprint is a representation of the connectivity of atoms in a molecule.[23] This mostly caters for molecular characterisation and for identifying common fragments within different molecules.

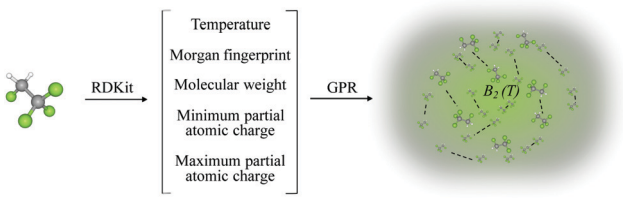

Fig. 1 Schematic representation of the method designed for the prediction of $B_2(T)$ using Gaussian process regression. A chosen molecule is characterized by a set of input features obtained using RDKit.[18] From the input data, the trained GPR model is used to predict $B_2(T)$ for the desired molecule.

2 The dataset

A comprehensive database of second virial coefficients for pure organic and inorganic substances is made available through the compilations of Dymond *et al.* and Gmehling *et al.*,[24] totalling over 9300 values for a temperature range from 0.63 to 1473.15 K. As the experimental errors in the determination of second virial coefficients have not been reported in the database, a handful of data points for different molecules and temperatures were inspected. This showed that experimental errors generally vary between 0.5 and 12%. It is worth emphasizing that each compound has data for the second virial coefficient in a particular range of temperatures compatible to the inherent thermo-physico-chemical properties of the compound under consideration. Subsequent to filtering, our dataset comprises 1720 data points for inorganic and 5213 data points for organic compounds, which are divided in diverse types of classes (see Fig. 2). While for some compounds, experimental values of $B_2(T)$ were reported for more than 200 temperatures, for other substances there existed only one data point in the set. When different B_2 values were registered for the same compound at the same temperature, an average of the B_2 values was taken. Further filtering of the data was performed by leaving out compounds with less than 3 data points and by eliminating the values which were off the temperature-dependent trend.

The diversity of data is notable with regard to the physical and chemical properties of molecules. For instance, the inorganic compounds cover a broad spectrum of molecules and atoms starting from noble gas atoms to polyatomic molecules such as boranes. Whereas within the organic compounds, one finds ketones, which have important industrial applications,[25] carbonyl compounds that appear as a natural product of pollution[26] or siloxanes: an incredibly versatile class of molecules that has been proposed as a candidate for Bethe–Zeldovich–Thompson fluids,[27] or that shows exciting properties as a surfactant.[28]

3 Machine learning model

3.1 Gaussian process regression

In the context of solving non-linear regression problems, Gaussian process regression (GPR) can be viewed as a non-parametric approach. In other words, GPR does not assume any functional form to find the fitting to a given data set. Rather, GPR employs a Gaussian distribution of functions to match the observed variables. Next, Bayesian inference, *i.e.*, the estimation of the probability of an event given the occurrence of a previous one, allows a prior distribution of data to develop into a posterior one. In the case of GPR, the prior distribution over

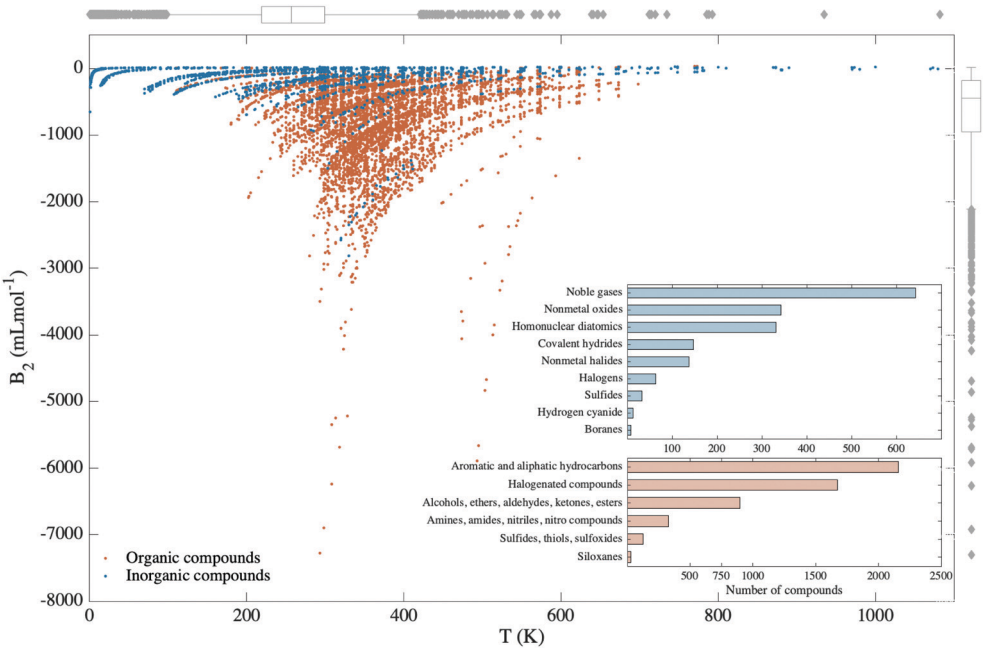

Fig. 2 Experimental values of second virial coefficients from the filtered database as a function of temperature. The two bar charts in the inset show classifications of inorganic and organic compounds in the dataset, as well as the number of data points for each class of compounds. The associated box plots for temperature and second virial coefficients values are also shown, with a 1.5 maximum whisker length.

the space of functions, $p(f|\mathbf{x})$ shows a joint multivariate Gaussian distribution, usually with a zero mean function $m(\mathbf{x})$ and with a covariance matrix defined by a kernel designated by the user, $K(\mathbf{x},\mathbf{x}')$, which stores information about the correlation between the input points.[29] The Gaussian Process is therefore defined as:

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}),K(\mathbf{x},\mathbf{x}')). \quad (2)$$

The posterior distribution, $p(f|\mathbf{x},y)$, which is also normal multivariate, is obtained by conditioning the joint Gaussian prior distribution on the observations (y). This allows to make predictions (y_*) for new, unobserved data.

The models were developed and analysed using MATLAB's already implemented tools.

3.2 Featurization methods

The choice of input features for our model was primarily guided by chemical and physical intuition, as well as by domain knowledge. Recent efforts have shown that universal descriptors for machine learning, which capture targeted information about systems (*e.g.* features related to the fitting of potential energy surfaces, such as Coulomb matrices) give high accuracy in many tasks of predicting molecular properties. This has tremendously helped obtain valuable information about systems without *a priori* information, in an "automated" manner. In this work, we show that good accuracy can also be obtained with the use of physico-chemical properties and molecular fingerprints solely selected using domain intuition and judgment, rather than canonical feature selection algorithms. This was proven previously in the work of Liu *et al.* in predicting dipole moments of diatomic molecules, explaining the good performance of intuitively chosen predictors over abstract, general purpose ones, when using small datasets.[30] To subsequently validate the chosen featurization scheme, an embedded type feature selection mechanism was implemented (see Section 4), which learned feature importance as part of the model learning process. A comparison of the performances of various combinations of predictors is also provided.

That being said, we devised what properties would be most relevant to describing the second virial coefficient, from a physical perspective. The features used in this work belong to three categories: physical properties that can describe molecular interactions (partial atomic charges and valence electrons), topology features that characterize similarity and complexity of compounds, deduced from cheminformatics (Morgan and E-state fingerprints) and intrinsic properties of molecules (molecular weight), to account for their different sizes. All of the input features are available and easy to compute, and in our case, they were generated using RDKit.[18] A further explanation for the choice of physical and topological features is outlined below.

• The minimum and maximum partial charges of a molecule are correlated with the molecule's dipole moment. This is supported by the recent work of Veit *et al.*, which implements a partial-charge model to predict dipole moments of molecules.[31] The presence of a dipole moment in a molecule leads to a dipole–dipole interaction apart from the van der Waals interaction of non-polar molecules. The dipole moment of a molecule is proven to increase the attractive forces between molecules and therefore to lower B_2 for a given temperature.[32] This shows a direct relationship between B_2 and the magnitudes of the minimum and maximum partial charges. The partial charges used in this work were computed using RDKit, which employs the procedure described by Gasteiger.[33] This computes

charge distribution in molecules based on the identities of individual atoms and their connectivities.

• Morgan fingerprints represent a well-known method for molecular characterization in terms of topology and connectivity within a molecule. In particular, a molecule is characterized by a fingerprint that contains 1024 bits, and each of these bits represents a fragment, *i.e.*, a possible scenario of individual atoms and their environment (meaning all neighbouring atoms within a diameter of four chemical bonds) within the molecule. The "extended connectivity" of atoms is computed using Morgans extended connectivity algorithm.[23] Therefore, the complexity of a molecule can be assessed by counting how many bits out of 1024 are needed to describe connectivities in a molecule, as well as element types, charges and atomic masses.[23] Furthermore, Morgan fingerprints can be used to generate a similarity score to a reference molecule. This can be obtained through commands implemented in RDKit.[18] In our study, the reference molecule was chosen to be the one with the highest number of nonzero bits in the Morgan fingerprint, *i.e.*, the most complex molecule from this point of view: 2-ethylthiophene. In this way, a similarity score to the fingerprint of the reference molecule was attributed to each molecule in the database, as an input feature. Intuitively, this is a measure of comparison between environments, connectivities and chemical features within different molecules, which can further be relevant to comparing 2-body interactions for different molecules.

• The E-state fingerprint has also been used to characterize the molecules in the data set. This fingerprint is based on the electrotopological state indices of atoms within a molecule.[23] These encode information related to the valence state, electronegativity of atoms and the molecule's topology. In particular, we translate the information for each molecule into a numerical descriptor through the ratio between the total summation of E-state indices for all atoms and the summation of the number of times each possible atom type appears in the molecule.

3.3 Model performance evaluation

GPR, as a general fitting approach, needs a method to characterize its performance. In other words, an error estimation is needed for the proper evaluation of GPR models and the posterior identification of outliers of the model.

One of the most common error estimators is the mean absolute error (MAE), defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i^*|, \tag{3}$$

where N is the total number of values in the data set, y_i are the true values of second virial coefficients, and y_i^* are the predictions.

In GPR, the predictions are being made after examining correlations between input features and observations in the training set. This is done without prior knowledge of the test set and, implicitly, no weighting on it, making the root mean squared error (RMSE), which is defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^*)^2}, \tag{4}$$

a practical evaluation tool. The RMSE of predictions on the test data will be used along this work. However, when predicting physical or chemical quantities, it may be better to have a dimensionless error estimator. The normalized error (r_E), given as

$$r_\mathrm{E} = \frac{\mathrm{RMSE}}{y_\mathrm{max} - y_\mathrm{min}}, \tag{5}$$

does not have units since it is defined as the ratio between the RMSE and the extension of the data. Therefore, the normalised error is an important error estimator regarding GPR, and it will be used throughout this work.

4 Results

Second virial coefficients are learned at a given temperature through a GPR model from molecular and cheminformatics-based properties of compounds. A filtered dataset of 6933 second virial coefficients for different ranges of temperatures was used to train and test the GPR model. All data was randomly divided into train and test sets to find the best featurization scheme for our predictions and to investigate the performance of the chosen model, together with its covariance function and parameters. For model selection and to avoid overfitting, 5-fold cross validation was implemented. The model's extrapolation capabilities were evaluated by testing on temperatures outside of the training range for each compound. Finally, applicability and transferability were analysed by testing the model on 42 different organic molecules, left out of the training process.

4.1 Featurization performance analysis

To assess the performances of the proposed features and to decide on the best featurization scheme, a GPR model based on a rational quadratic kernel function was implemented:

$$K(\mathbf{x},\mathbf{x}') = \sigma^2\left[1 + \frac{(\mathbf{x}-\mathbf{x}')^2}{2\alpha l^2}\right]^{-\alpha}, \tag{6}$$

where σ^2 is the signal variance, l is the characteristic length scale of the function and α determines the weighting between different length scales. l, $\alpha > 0$.

The results of using temperature, molecular weight, minimum and maximum partial atomic charges, and similarity of the Morgan fingerprint to that of a reference molecule, as input features, are shown in Fig. 3. The training was done on 5547 randomly selected data points, using 5-fold cross-validation, and 1386 predictions were made on the remaining "out-of-sample" points. It is easily noticed from the figure that most of the predicted values for the second virial coefficient agree with the true experimental values, translating into an excellent performance and predictive capability of the model at hand. This performance is characterized by an RMSE of 58 mL mol^{-1} and a normalized error of 0.8%, as it is shown in Table 1. For reference, a multiple linear regression (MLR) model with the same descriptors was implemented and registered an RMSE of $\sim$525 mL mol^{-1}. This indicated the impact of non-linearity on the relationship between second virial coefficients and the

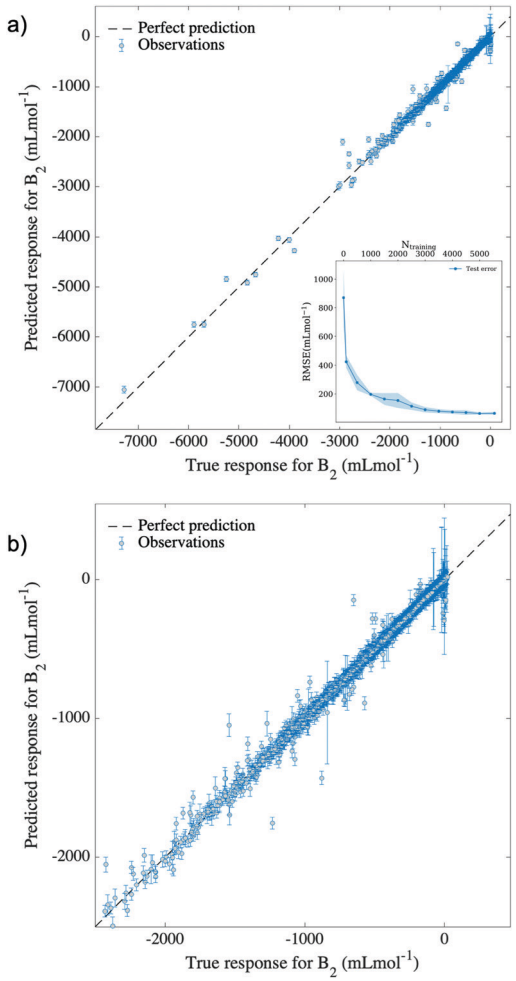

Fig. 3 (a) GPR 5-fold cross-validated predictions of temperature-dependent second virial coefficients, $B_2(T)$, on "out-of-sample" randomly chosen test data, representing 20% of the entire dataset. Error bars represent the uncertainty in the GPR prediction. A 5 dimensional representation of input features is used (temperature, molecular weight, minimum and maximum partial atomic charges and similarity of Morgan fingerprint to that of a reference molecule). The inset shows the corresponding learning curve for this model, in which 1386 "out-of-sample" randomly selected test points were used. The shaded area stands for the error bars after 5 different iterations. (b) A zoom into the plot from panel (a).

proposed descriptors, and implicitly the requirement of a non-linear method such as GPR in this matter.

To further analyze the performance of our GPR model we have calculated its learning curve, *i.e.*, the model performance as a function of the number of points in the training set while keeping the number of data points in the test set constant, which is shown in the inset of Fig. 3. As a result, it is observed that the model's learning capabilities are converged around 4000 data points of the training set. Therefore, the interpolation performance of our model cannot benefit from having a larger number of training points for the families of compounds already present in the training set.

The combination and the number of input features for our model were selected after the implementation of different featurization schemes and the comparison of their performances. The results of this procedure are shown in Table 1. Here, it is noticed that when the number of valence electrons is used as a predictor instead of the partial atomic charges, a much poorer performance is obtained, at the same dimensionality (5D). This is suggestive of the importance of minimum and maximum partial atomic charges as predictors in our model, presumably succeeding to account for the strength of interactions between molecules, more than just for their internal electronic structure. In addition, we notice that although the E-state fingerprint contains additional information concerning the valence state of atoms, it does not show an improved performance to that of the Morgan fingerprint in a 5-dimensional representation. Indeed, this correlates with our previous statement about the major role of partial charges in comparison with the number of valence electrons regarding molecular interactions.

To get a measure of the importance of individual predictors relative to each other, the automatic relevance determination (ARD)[34] rational quadratic kernel function was used in GPR:

$$K(\mathbf{x},\mathbf{x}'|\theta) = \sigma^2\left[1+\frac{1}{2\alpha}\sum_{m=1}^{d}\frac{(x_m-x'_m)^2}{{\sigma_m}^2}\right]^{-\alpha}, \qquad (7)$$

where $\theta_m = \log(\sigma_m)$ for $m = 1,2,\ldots d$ with $\theta_{d+1} = \log(\sigma)$, and d is the total number of predictors. ARD allows the assignment of separate length scales for each predictor, instead of the same one for all of them. If an input's length scale is large, the distance one needs to move in the input space so that the function values become uncorrelated is also large, so that the covariance will become almost independent of that input.

Table 1 Predictors ranking by the test RMSE score of the 5-fold cross-validated GPR model. The symbols used in the table are assigned as follows: T is the temperature, M_W stands for the molecular weight, δ_{min}, δ_{max} are minimum and maximum partial atomic charges, respectively, $MF_{nonzeros}$ is the number of nonzero bits in the Morgan fingerprint, $MF_{similarity}$ is the similarity of the compound's fingerprint to that of the reference compound, E-state encodes information on the E-state fingerprint and VE is the number of valence electrons. The results are obtained using GPR trained on 5547 randomly selected training points with 5-fold CV, being tested on 1386 "out-of-sample" points. Errors were reported after doing 10 different iterations

Dimension	Features	Test RMSE (mL mol^{-1})	Test MAE (mL mol^{-1})	Test r_E (%)
4D	(T,M_W,δ_{min},δ_{max})	80 ± 6	27 ± 0.6	1.1 ± 0.1
5D	(T,M_W,$\delta_{min}\delta_{max}$, $MF_{similarity}$)	58 ± 2	22 ± 0.6	0.8 ± 0.03
5D	(T,M_W,δ_{min},δ_{max}, $MF_{nonzeros}$)	68 ± 7	23 ± 0.5	0.9 ± 0.1
5D	(T,M_W,δ_{min},δ_{max}, E-state)	67 ± 4	24 ± 0.4	0.9 ± 0.1
5D	(T,M_W,VE, $MF_{nonzeros}$, $MF_{similarity}$)	155 ± 10	56 ± 3	2.1 ± 0.1
6D	(T,M_W,VE,δ_{min},δ_{max}, $MF_{nonzeros}$)	60 ± 2	22 ± 0.4	0.8 ± 0.03
6D	(T,M_W,VE,δ_{min},δ_{max}, $MF_{similarity}$)	57 ± 3	22 ± 0.4	0.8 ± 0.04

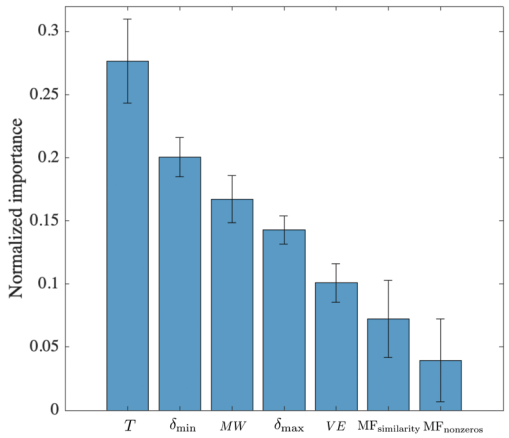

Fig. 4 Ranking of predictors based on the characteristic length scale of each predictor, obtained with an ARD rational quadratic kernel. The symbols used in the figure were defined in the caption of Table 1. The errors associated with each weight are the result of performing 5 iterations.

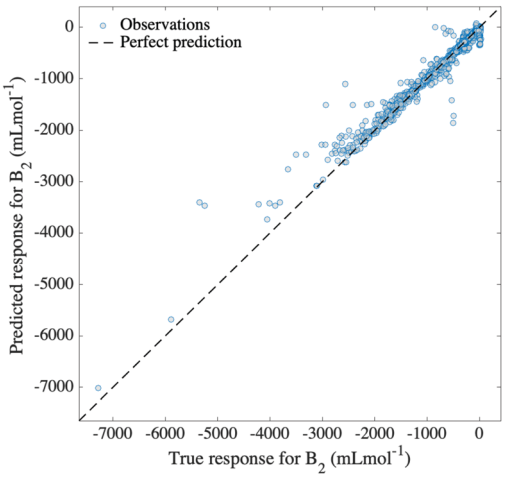

Fig. 5 GPR 5-fold cross-validated predictions of $B_2(T)$ for marginal temperatures of each compound (representing 20% of the entire data set). A 5 dimensional representation of input features is used (temperature, molecular weight, minimum and maximum partial atomic charges and similarity of Morgan fingerprint to that of a reference molecule).

This is known as an embedded method for feature selection, as the selection is done during model training. The predictor data was standardized to allow for consistency. In this way, a weight was assigned to each input feature and was normalized, as shown in Fig. 4. The ranking is consistent with our previous evaluation of the featurization schemes' performances (see Table 1): temperature is the most important, followed by partial atomic charges and/or molecular weight. Morgan fingerprint similarity is expected to perform better than the number of nonzero bits in the fingerprint. It is worth noticing that partial charges are better ranked than the number of valence electrons. This confirms our intuition on the relevance of partial charges over number of valence electrons, which is also supported by the results shown in Table 1.

4.2 Extrapolation to marginal temperatures

To evaluate our model's extrapolation capability, the data was divided as follows: for each molecule, data points corresponding to marginal temperatures (meaning the lowest 10% and the highest 10% temperatures) were used for testing, whereas the rest were used for training. This selection naturally yielded a training set comprising 80% of the total data, on which 5-fold cross-validation was applied. A GPR model based on a rational quadratic kernel was implemented, using the best featurization scheme obtained in the previous subsection (temperature, molecular weight, minimum and maximum partial atomic charges and similarity of Morgan fingerprint to that of a reference molecule). This allowed a prediction of second virial coefficients characterized by an RMSE of 157 mL mol^{-1} and by a relative error of 2.1%, which is portrayed in Fig. 5. By this means, it can be noticed that the model achieves successful extrapolation for low, as well as for high temperatures corresponding to compounds in our dataset, with a few exceptions. Among these, the most distinguishable outliers are generated by ethanenitrile, toluene, methyl ethanoate and ethanol. In some cases, these molecules are poorly represented in the training set, either by the absence of other molecules of their class, or by large spacing between the temperatures at which B_2 values are measured. In other cases, the descriptors fail to describe the molecule.

While Gaussian process regression serves well as an approach to smooth interpolation,[29] it usually performs worse in extrapolation tasks. Our model illustrates well the former statement, exhibiting excellent interpolation in the previous subsection. In addition to this, the representation of input features, as well as the chosen kernel function prove to lead to reasonable extrapolation capability, with only a few molecules experiencing the limitations of the model. The rational quadratic kernel differs from the squared exponential one (a popular choice of kernel function for GPR) in that it contains an additional parameter (α) which determines the relative weighting between large and small-scale variations.[29] This eventually leads to a better generalization of the long term trend, compared to a squared exponential kernel, and therefore better extrapolation. The fact that the same model achieves both interpolation and extrapolation translates into considerate choice of both input features and kernel function.

4.3 Applicability and transferability

Finally, to estimate the applicability of our model, 42 different organic molecules were left out of the training process and were tested on. This generated a training set comprising 6258 data points ($\sim$90%) and a test set which registered 675 data points ($\sim$10%) for organic molecules only. The molecules in the test set were selected so that they cover the widest range of organic families possible (see Fig. 2), having, at the same time, corresponding examples of their families in the training set. The choice to only include organic molecules in the test set is rooted in the general poorer performance of inorganic compounds when compared to organic ones. While the model succeeds in predicting B_2 values for inorganics in interpolative regimes and

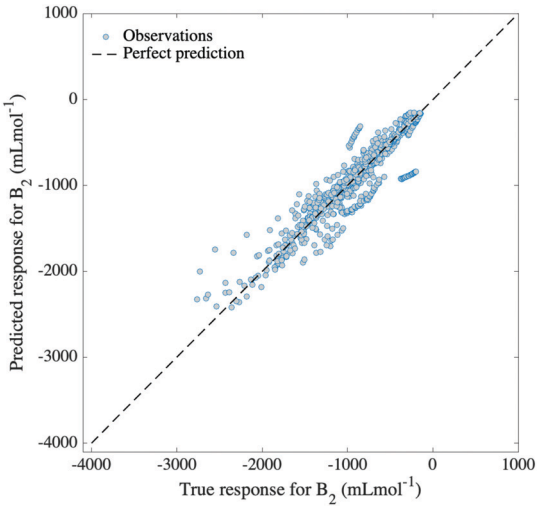

Fig. 6 GPR 5-fold cross-validated predictions of $B_2(T)$ for organic molecules absent from the training set. A 5 dimensional representation of input features is used (temperature, molecular weight, minimum and maximum partial atomic charges and similarity of Morgan fingerprint to that of a reference molecule).

when extrapolating to marginal temperatures, it is not ideally applicable to inorganic molecules which are unseen in the training set.

Nonetheless, the model succeeds to predict second virial coefficients for organic molecules with an RMSE of 195 mL mol^{-1} (see Fig. 6), which is indicative of the greater capability of the input features to describe organic compounds, rather than inorganic ones. This is naturally expected, as Morgan fingerprints incorporate valuable information for organic molecules concerning

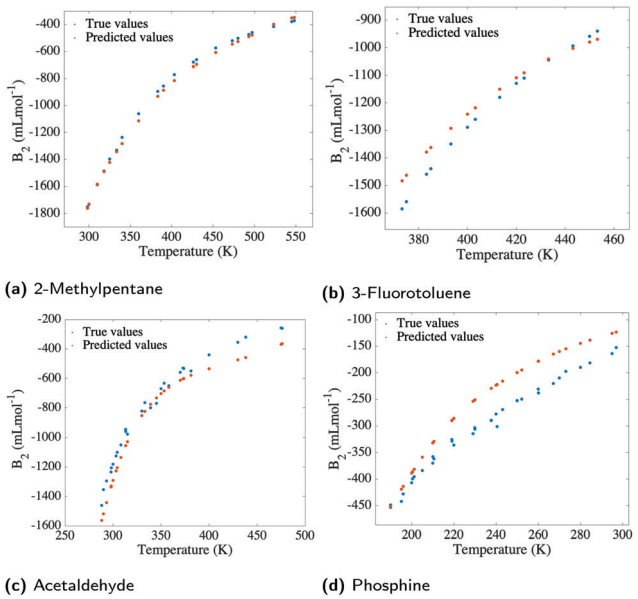

Fig. 7 Examples of plots showing predicted and true values for second virial coefficients as a function of temperature. Training was done on all but the presented molecules, using a GPR model with 5-fold cross validation and a 5 dimensional representation of input features.

Table 2 Percentage deviations of second virial coefficient predictions from experimental values by various methods. The deviations are normalised and expressed as $\mathrm{RMS} = 100\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{|y_i - y_i^*|}{y_i}\right)^2}$, where y_i and y_i* are experimental and calculated values of B_2, respectively. Abbreviations for B_2 calculation methods are as follows: PC: Pitzer and Curl (1957),[35] TD: Tarakad and Danner (1977),[36] ECS: Xiang (2002),[37] GPR: Gaussian process regression method described in this work

Compound	PC RMS	TD RMS	ECS RMS	GPR RMS
1,3-Dimethylbenzene	8	8	6	4
Phenol	10	6	6	2
1,2-Dichloroethane	10	9	9	11
1-Propanol	13	13	16	12
Bromoethane	9	17	9	13
Water	10	34	20	26
1-Butanol	11	10	23	4
2-Pentanone	21	7	6	7

topology, element types and atomic charges. Fig. 7 reveals some examples of predicted, as well as true values for B_2 plotted against temperature. Instances from three different organic families are presented, for which the predicted curve is smooth. At the same time, it can be seen that the inorganic molecule phosphine performs poorly compared to the other instances.

The model was trained using 5-fold cross-validation, having a training RMSE of 62 mL mol^{-1} and a test RMSE of 195 mL mol^{-1} (relative error 2.7%). Generally, the model predicts with great accuracy second virial coefficients for molecules well-represented in the training set, such as hydrocarbons (see Fig. 7a). However, the model is not transferable to molecules which have no resemblance to the training set, being limited, in this case, to "out-of-sample" compounds that interpolate. Nonetheless, the wide range of families present in our database offer an optimistic view towards the applicability of our model, as various commonly encountered classes of organic compounds are encompassed.

Finally, the performance of our model was analysed in contrast to the performances of established methods for the calculation of B_2 for polar substances through empirical equations and through the corresponding states principle.[37] Table 2 contains deviations from experimental values of second virial coefficients calculated through four different methods. The data reinforces the points made previously about our model. That is, the model is characterised by a poor performance in predicting second virial coefficients for inorganic compounds, but proves a generally better performance relative to other methods in the prediction for organic compounds.

5 Conclusions

We have developed a method for estimating second virial coefficients using Gaussian process regression with a relative error $\lesssim$1% in the interpolative regime. The same model was used to predict second virial coefficients for marginal temperatures of all compounds (relative error 2.1%), as well as for "out-of-sample" organic molecules resembling the training set (relative error 2.7%). This has been possible through the use of a low-dimensional representation of predictors based on accessible,

intuitive, and reproducible molecular features, conveniently obtained through RDKit. The applicability of our model is characterized by great performance for molecules well-represented in the training set by instances of their families, which are high in variety. When compared to traditional techniques used to calculate second virial coefficients, our method stands out in particular through its simplicity and through its efficiency, avoiding the difficulties posed by computational cost or by experimental obstacles. The input features are readily obtained through RDKit and the time required to train our best model is approximately 74 seconds on a 2 GHz Intel Quad-Core i5 machine. Besides, our method shows a generally better performance than that of several established semi-empirical procedures employed for the prediction of the quantity. Finally, it is worth emphasising the important role the existence of a comprehensive and high quality database has played in this work.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank Prof. Gerard Meijer and Dr Stefan Truppe for reading the manuscript and for useful suggestions to improve it, as swell as Xiangyue Liu for fruitful discussions and for great recommendations regarding the error estimation.

References

1 KNAW, Proceedings, 1901–1902, 4, 125–147.
2 J. Lennard-Jones, *Physica*, 1937, **4**, 941–956.
3 D. T. Haar, *Proc. Phys. Soc., London, Sect. A*, 1953, **66**, 847–848.
4 D. McQuarrie, *Statistical Mechanics*, University Science Books, 2000.
5 G. Marcelli and R. J. Sadus, *J. Chem. Phys.*, 1999, **111**, 1533–1540.
6 K. M. M. Frenkel, *Virial Coefficients of Pure Gases*, Springer-Verlag Berlin Heidelberg, 2002.
7 R. B. B. Joseph, O. Hirschfelder and C. F. Curtiss, *Molecular Theory of Gases and Liquids*, Wiley, New York, 1964.
8 R. J. Sadus, *J. Chem. Phys.*, 2019, **150**, 024503.
9 D. McQuarrie and J. D. Simon, *Molecular Thermodynamics*, University Science Books, 1999.
10 S. Montero and J. Pérez-Ríos, *J. Chem. Phys.*, 2014, **141**, 114301.
11 V. M. Zhdanov, *Transport Processes in Multicomponent Plasma*, Taylor & Francis, London, 2002.
12 G. A. Vliegenthart and H. N. W. Lekkerkerker, *J. Chem. Phys.*, 2000, **112**, 5364–5369.
13 B. L. Neal, D. Asthagiri, O. D. Velev, A. M. Lenhoff and K. W. Kaler, *J. Cryst. Growth*, 1999, **196**, 377–387.
14 J. O. Hirschfelder, R. B. Ewell and J. R. Roebuck, *J. Chem. Phys.*, 1938, **6**, 205–218.
15 R. H. Fowler, *Math. Proc. Cambridge Philos. Soc.*, 1925, **22**, 861–885.
16 E. S. Barkan, *J. Eng. Phys.*, 1983, **44**, 651–657.
17 T. W. Leland and P. S. Chappelear, *Ind. Eng. Chem.*, 1968, **60**, 15–43.
18 G. Landrum, RDKit: Open-source cheminformatics, http://www.rdkit.org, http://www.rdkit.org.
19 G. Bell, T. Hey and A. Szalay, *Science*, 2009, **323**, 1297–1298.
20 G. D. Nicola, G. Coccia, M. Pierantozzi, S. Tomassetti and R. C. Grifoni, *Chem. Eng. Commun.*, 2018, **205**, 1077–1095.
21 E. Mokshyna, P. G. Polishchuk, V. I. Nedostup and V. E. Kuzmin, *Mol. Inf.*, 2015, **34**, 53–59.
22 V. E. Kuz'min, A. G. Artemenko and E. N. Muratov, *J. Comput.-Aided Mol. Des.*, 2008, **22**, 403–421.
23 R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley-VCH Verlag GmbH & Co. KGaA, 2009.
24 *Springer Materials*, https://materials.springer.com/interactive/overview?propertyId=PhysProp-25r9sjgdc7k3l7s7nf3-k7m1k2vmblff6, accessed August 2020.
25 O. D. Sparkman, Z. E. Penton and F. G. Kitson, *Gas Chromatography and Mass Spectrometry (Second Edition)*, Academic Press, Amsterdam, 2nd edn, 2011, pp. 345–349.
26 G. Hanrahan, *Key Concepts in Environmental Chemistry*, Academic Press, Boston, 2012, pp. 215–242.
27 P. Colonna, A. Guardone and N. R. Nannan, *Phys. Fluids*, 2007, **19**, 086102.
28 R. M. Hill, in *Siloxane surfactants*, ed. I. D. Robb, Springer, Netherlands, Dordrecht, 1997, pp. 143–168.
29 C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006.
30 X. Liu, G. Meijer and J. Pérez-Ríos, *Phys. Chem. Chem. Phys.*, 2020, **22**, 24191–24200.
31 M. Veit, D. M. Wilkins, Y. Yang, R. A. DiStasio and M. Ceriotti, *J. Chem. Phys.*, 2020, **153**, 024113.
32 C. Vega, C. McBride and C. Menduiña, *Phys. Chem. Chem. Phys.*, 2002, **4**, 3000–3007.
33 J. Gasteiger and M. Marsili, *Tetrahedron*, 1980, **36**, 3219–3228.
34 F. R. Burden, M. G. Ford, D. C. Whitley and D. A. Winkler, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1423–1430.
35 K. S. Pitzer and R. F. Curl, *J. Am. Chem. Soc.*, 1957, **79**, 2369–2370.
36 R. Tarakad and R. Danner, *AIChE J.*, 1977, **23**, 685–695.
37 H. W. Xiang, *Chem. Eng. Sci.*, 2002, **57**, 1439–1449.