Predicting second virial coefficients of inorganic and organic compounds using Gaussian Process Regression

Miruna T. Cretu\textsuperscript{1,2} and Jesús Pérez-Ríos\textsuperscript{2}\textsuperscript{,}\textsuperscript{17}

\textsuperscript{1)Department of Chemistry, Imperial College London, London SW7 2AZ, United Kingdom}
\textsuperscript{2)Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany}

(Dated: 8 September 2020)

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 using Gaussian process regression. In particular, we find that a low dimensional representation of features based on intrinsic molecular properties, topology and physical properties relevant for the characterization of molecule-molecule interactions, succeeds to predict the second virial coefficient of any molecule with a relative error \(\lesssim 1\%\).

I. 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\textsuperscript{1,2}. Besides offering a rigorous depiction of the pressure, \(p(T, \rho)\) over a wide range of temperature\textsuperscript{3,4} \(T\), and density, \(\rho\), the virial equation is founded on a solid statistical mechanics framework\textsuperscript{4}. The virial equation,

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

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 \(i\). Therefore, \(B_i(T)\) is the \(i\)-th virial coefficient and is related to the role of \(i\)-body interactions in a system. In Eq. (1) \(R\) denotes the ideal gas constant and the series is truncated up to a certain cluster size \(N\).

Two-body interactions are the most relevant to the macroscopic properties of a gas\textsuperscript{5}, hence \(B_2\) values have been tabulated for many gases\textsuperscript{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\textsuperscript{7,8}. This is conducive to the calculation of fluid properties such as enthalpy of vaporisation\textsuperscript{9} and transport coefficients\textsuperscript{2,10}. The knowledge of \(B_2\) also helps to estimate critical points\textsuperscript{11} and optimum conditions for crystal growth, which would otherwise require extensive screening experiments\textsuperscript{12}.

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\textsuperscript{13} 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\textsuperscript{14}. In the case of empirical approaches, the law of corresponding states\textsuperscript{15} 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\textsuperscript{16}. The existence of a large and high quality database of temperature-dependent second virial coefficient\textsuperscript{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 one of Di Nicola et al.\textsuperscript{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”\textsuperscript{20} due to its complexity.

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), yielding a prediction of the second virial coefficient with a relative error \(\lesssim 1\%\). Our method’s universality stems from its applicability to compounds belonging to a wide range of families and from the

\textsuperscript{a}Electronic mail: jerezri@fhi-berlin.mpg.de

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\textsuperscript{18}. From the input data, the trained GPR model is used to predict \(B_2(T)\) for the desired molecule.
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, respectively, 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.

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\textsuperscript{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\textsuperscript{21}. This mostly caters for molecular characterisation and for identifying common fragments within different molecules.

II. 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\textsuperscript{22}, totalling over 9300 values for a temperature range from 0.63 to 1473.15 K. Subsequent to filtering, our dataset comprises 1720 inorganic and 5213 organic compounds, 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 and were therefore of unphysical meaning.

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\textsuperscript{23}, carbonyl compounds that appear as a natural product of pollution\textsuperscript{24} or siloxanes: an incredibly versatile class of molecules that has been proposed as a candidate for Bethe-Zel’dovich-Thompson fluids\textsuperscript{25}, or that shows exciting properties as a surfactant\textsuperscript{26}.

III. GAUSSIAN PROCESSES

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, $p(f|x)$ is a multivariate Gaussian distribution, usu-
ally with a zero mean function \( m(x) \) and with a covariance matrix defined by a kernel designated by the user, \( K(x, x') \), which stores information about the correlation between the input points.\(^{[27]}\) The distribution of functions is therefore defined as:

\[
f(x) \sim GP(m(x), K(x, x')).
\]  

(2)

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

IV. FEATURIZATION METHODS

The choice of input features for our model was primarily guided by chemical and physical intuition, as well as by domain expertise. The features 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\(^{[13]}\). A further explanation for the choice of physical and topology features is outlined below.

- The minimum and maximum partial charges of a molecule are correlated with the molecule’s dipole moment, which in turn has an influence on the interaction potential. The presence of a dipole moment in a molecule leads to a dipole-dipole interaction instead of 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.\(^{[28]}\) This shows a direct relationship between \( B_2 \) and the magnitudes of the minimum and maximum partial charges.

- 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 Morgan’s extended connectivity algorithm.\(^{[22]}\) 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.\(^{[21]}\) Furthermore, Morgan fingerprints can be used to generate a similarity score to a reference molecule. This can be obtained through commands implemented in RDKit\(^{[13]}\). 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 is attributed to each molecule in the database, as an input feature.

- 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.\(^{[21]}\) 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.

A. 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

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - y_i^*|,
\]  

(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:

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - y_i^*)^2},
\]  

(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

\[
r_E = \frac{\text{RMSE}}{y_{\text{max}} - y_{\text{min}}},
\]  

(5)

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

Second virial coefficients are learned at a given temperature through a GPR model and from molecular and cheminformatics-based properties of compounds. The results of using GPR based on a rational quadratic kernel (see appendix) and of using temperature, molecular weight, minimum and maximum partial atomic charges, and similarity of the Morgan fingerprint to that of a reference molecule, are shown in Fig. 3. 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 astonishing performance is characterized by an RMSE of 59.58 ± 1.43 mLmol⁻¹ and a normalized error of 0.81 % as it is shown in Table I.

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 performance of our model cannot benefit from having a larger dataset.

The combination and the number of input features for our model are selected after the implementation of different featurization schemes and the comparison of their performances. The results of this procedure are shown in Table I. 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.

To get a measure of the importance of individual predictors relative to each other, the automatic relevance determination (ARD) rational quadratic kernel function (see appendix) is used in GPR. 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. The predictor data is standardized to allow for consistency. In this way, a weight is assigned to each input feature, as shown in Fig. 4. The ranking is consistent with our previous evaluation of the featurization schemes’ performances (see Table I): temperature is the most important, followed by partial atomic charges and/or molecular weight. Morgan fingerprint similarity is expected to
It is noticed that the output data for a specific compound, when not many points are available for it. This is a characteristic behaviour of fitting operations over a small number of points, that our GPR model appears to comply with.

The inorganic molecules that are not well represented by our model are ammonia-D₃, oxygen at temperatures within the interquartile range and helium at high temperatures (see Fig. 2), meaning that it is not the case that every temperature is represented for each 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 on a total of 6933 training and test points, from 5 iterations.

The applicability of a model can be evaluated by performing a residuals analysis. The best performing model was chosen for analysis, with a 5 dimensional featureization scheme and 5-fold cross validation GPR based on a rational quadratic kernel. For regression models, a normal distribution of residuals usually suggests that they are random and independent of the true B₂ values, meaning that the predictive information is well captured. While our residuals do not perfectly follow a normal distribution, the box plot in Figure 5 shows that the residuals are almost symmetrically distributed and that the data is very compact, gathering around the mean. This figure also shows the structures and corresponding temperatures of the main outliers of our model. It is noticed that the outliers’ temperatures and true B₂ values fall within the interquartile ranges (see Fig. 2), meaning that it is not the case that the model fails to extrapolate for extreme temperatures or B₂ values. Most of the outliers are organic compounds, with 6 of them having cyclic structures. It is interesting to notice this, since the database comprises more organic than inorganic compounds, so that one could have expected a more accurate prediction for organic compounds, rather than for inorganic. The inorganic molecules that are not well represented by our model are ammonia-D₃, oxygen at temperatures within the interquartile range and helium at high temperatures (∼1300K), whose residuals cluster at low values of B₂ (see Fig. 3). Most of the outliers that can be seen in Figure 5 occur due to a tendency in overestimating B₂ for compounds at temperatures which are either a minimum or a maximum in the set of input data for a specific compound, when not many points are provided for it. This is a characteristic behaviour of fitting operations over a small number of points, that our GPR model appears to comply with.

FIG. 5. Residuals for the prediction of B₂ with GPR and their associated box plot. (T, MW, δ_min, δ_max, MF_similarity) were used as input features, with 5-fold cross validation. The most notable residuals are labeled with the compound’s structure and the corresponding temperature.

| Dimension | Features | CV RMSE (mLmol⁻¹) | CV MAE(mLmol⁻¹) | CV rₑ (%) |
|-----------|----------|-------------------|-----------------|-----------|
| 4D        | (T, MW, δ_min, δ_max) | 79.18±9.37       | 26.59±0.79      | 1.08±0.13 |
| 5D        | (T, MW, δ_min, δ_max, MF_similarity) | 59.58±1.43       | 22.85±0.24      | 0.81±0.02 |
| 5D        | (T, MW, δ_min, δ_max, MF_nonzeros) | 72.10±8.21       | 22.53±0.44      | 0.99±0.11 |
| 5D        | (T, MW, δ_min, δ_max, E-state) | 66.88±8.01       | 23.68±0.58      | 0.91±0.11 |
| 5D        | (T, MW, VE, MF_nonzeros, MF_similarity) | 158.12±12.20     | 54.94±2.91      | 2.16±0.17 |
| 6D        | (T, MW, VE, δ_min, δ_max, MF_nonzeros) | 59.60±1.58       | 22.72±0.34      | 0.82±0.02 |
| 6D        | (T, MW, VE, δ_min, δ_max, MF_similarity) | 60.25±4.70       | 22.59±0.32      | 0.82±0.06 |
| 7D        | (T, MW, VE, δ_min, δ_max, MF_nonzeros, MF_similarity) | 55.80±2.65       | 20.85±0.35      | 0.76±0.04 |

TABLE I. Predictors ranking by the RMSE score from 5-fold cross validation (CV). The symbols used in the table are assigned as follows: T is the temperature, MW 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 on a total of 6933 training and test points, from 5 iterations.
VI. CONCLUSIONS

We have developed a method for estimating second virial coefficients using Gaussian process regression with a relative error $\leq 1\%$. 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. There is no requirement to make a distinction between organic and inorganic or small and large molecules, as the input features succeed to describe well a generous number of types of compounds. The analysis of residuals generated by our model reinforces the applicability of the model to molecules belonging to a large variety of classes. When compared to traditional techniques used to calculate second virial coefficients, our method stands out to make a distinction between organic and inorganic or small and large molecules, as the input features succeed to describe well a generous number of types of compounds. The analysis of residuals generated by our model reinforces the applicability of the model to molecules belonging to a large variety of classes. When compared to traditional techniques used to calculate second virial coefficients, our method stands out

ACKNOWLEDGMENTS

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

VII. DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Appendix A: Details on the GPR model

In this work, the Gaussian process we used was defined by a rational quadratic kernel function:

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

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

The ARD rational quadratic kernel function was also used:

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

where $\theta_m = \log(\sigma_m)$ for $m = 1, 2, \ldots d$ and $\theta_{d+1} = \log(\sigma)$. $d$ is the total number of predictors.

1^1Expression of the equation of state of gases and liquids by means of series,” KNAW, Proceedings 4, 125–147 (1901-1902).

1^2L. J. Lennard-Jones, “The equation of state of gases and critical phenomena,” Physica 4, 941 – 956 (1937).

1^3T. Haar, “The second virial coefficient near absolute zero,” Proceedings of the Physical Society. Section A 66, 847–848 (1953).

1^4D. McQuarrie, Statistical Mechanics (University Science Books, 2000).

1^5G. Marcelli and R. J. Sadus, “Molecular simulation of the phase behavior of noble gases using accurate two-body and three-body intermolecular potentials,” The Journal of Chemical Physics 111, 1533–1540 (1999).

1^6K. M. M. Frenkel, Virial Coefficients of Pure Gases (Springer-Verlag Berlin Heidelberg, 2002).

1^7R. B. B. Joseph O. Hirschfelder, Charles F. Curtiss, Molecular Theory of Gases and Liquids (Wiley, New York, 1964).

1^8R. J. Sadus, “Two-body intermolecular potentials from second virial coefficient properties,” The Journal of Chemical Physics 150, 024503 (2019).

1^9J. D. S. D. A. McQuarrie, Molecular Thermodynamics (University Science Books, 1999).

1^10S. Montero and J. Pérez-Ríos, “Rotational relaxation in molecular hydrogen and deuterium: Theory versus acoustic experiments,” The Journal of Chemical Physics 141, 114301 (2014).

1^11V. M. Zhidanov, Transport Processes in Multicomponent Plasma (Taylor & Francis, London, 2002).

1^12G. A. Vliegenthart and H. N. W. Lekkerkerker, “Predicting the gas–liquid critical point from the second virial coefficient,” The Journal of Chemical Physics 112, 5364–5369 (2000).

1^13Why is the osmotic second virial coefficient related to protein crystallization?” Journal of Crystal Growth 196, 377 – 387 (1999).

1^14J. O. Hirschfelder, R. B. Ewell, and J. R. Roebuck, “Determination of intermolecular forces from the joule-thomson coefficients,” The Journal of Chemical Physics 6, 205–218 (1938).

1^15R. H. Fowler, “Assemblies of imperfect gases by the method of partition functions,” Mathematical Proceedings of the Cambridge Philosophical Society 22, 861–885 (1925).

1^16E. S. Barkan, “Determination of compressibility and virial co-efficients by the burnett method,” Journal of Engineering Physics 44, 651–657 (1983).

1^17T. W. Leland and P. S. Chappalear, “The corresponding states principle—a review of current theory and practice,” Industrial & Engineering Chemistry 60, 15–43 (1968).

1^18G. D. Nicola, G. Coccia, M. Pierantozzi, S. Tomassetti, and R. C. Grifoni, “Artificial neural network for the second virial coefficient of organic and inorganic compounds: An ann for b of organic and inorganic compounds,” Chemical Engineering Communications 205, 1077–1095 (2018).

1^19R. M. Hill, “Siloxane surfactants,” in Specialist Surfactants edited by I. D. Robb (Springer Netherlands, Dordrecht, 1997) pp. 143–168.

1^20C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning (The MIT Press, 2006).

1^21C. Vega, C. McBride, and C. Menduhiu, “The second virial coefficient of the dipolar two center lennard-jones model,” Phys. Chem. Chem. Phys. 4, 3000–3007 (2002).