**A symbolic regression based methodology for the construction of interpretable and predictive thermodynamic models**

Sam Kay,a Edgar I. Sanchez Medinab, Kai Sundmacherb,c, Dongda Zhanga\*

*aDepartment of Chemical Engineering, The University of Manchester, Oxford Road, Manchester, M1 3AL, UK.*

*bChair for Process Systems Engineering, Otto-von-Guericke University, Universitätsplatz 2, Magdeburg, 39106, Germany*

*cProcess Systems Engineering, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1, Magdeburg, 39106, Germany*

\*dongda.zhang@manchester.ac.uk.

Abstract

Symbolic regression offers great potential in applied thermodynamics, where traditional modelling efforts have yielded limited improvement. Thermodynamics research revolves around the use of thermodynamic relationships to describe properties such as enthalpy and phase equilibrium amongst others. Yet, these models are often limited to specific temperature and pressure regions, chemical classes and physical states. To address this, we propose a symbolic regression methodology to extend the region of feasibility in which thermodynamic models can be applied reliably. In this study, we generate a residual model, correcting inherent bias within an excess Gibbs energy model by reformulating existing knowledge and introducing new information, namely molecular access systems keys fingerprints (MACCS). To test this approach, a case study was developed, aimed at the prediction of activity coefficients at infinite dilution for a range of temperatures using UNIFAC Dortmund. The results were benchmarked with a Gibbs-Helmholtz graph neural network based method from the literature (Sanchez Medina et al., 2023), in which the advantages and limitations of the proposed symbolic regression based approach were identified, and its interpretibility as well as ease of implementation were well demonstrated. This study, therefore, provides a unique and novel contribution to the field of applied thermodynamics research.

**Keywords**: Symbolic regression, Thermodynamic models, Interpretable machine learning, Molecular fingerprint, Graph neural network

* 1. Introduction

To meet the current net zero targets, there has been a drive to more environmentally focused plant design and operation. In this respect, the integration of novel chemical systems, technologies, and “cleaner” energy sources within process operation are essential. A critical focus for sustainable plant development is the separations process, which account for 10-15% of global energy usage (Mutch, 2022), and 40-90% of capital and operating costs within industry (de Haan et al., 2020). Enhancing the design of separation systems necessitates accurate approximations of phase equilibria which requires thermodynamic data derived from many time-consuming and expensive experimental trials. However, the vast experimental space, encompassing all combinations of current and future synthesisable molecules, leaves extensive gaps within available thermodynamic data (Sanchez Medina et al., 2023). To overcome this, predictive models such as the universal quasichemical functional-group activity coefficients method (UNIFAC) were developed and have become a widespread solution to circumvent experimental limitations (Sanchez Medina et al., 2023).

Activity coefficients describe the deviation from ideal behaviour of a substance in a mixture due to intermolecular interactions and size differences. They are of the upmost importance in describing phase equilibria for non-ideal liquid mixtures, and the successful design of a separations system is naturally largely reliant on their accurate retrieval. In chemical engineering, the activity coefficient at infinite dilution (IDAC) is usually of specific interest for its utility: firstly, they can be used to calculate the activity coefficients at finite dilution (Brouwer et al., 2021; Medina & Sundmacher, 2023); secondly, they provide good initial estimates for solvent performance in solvent selection processes (Cheng et al., 2004).

UNIFAC Dortmund is the most commonly applied variant of the UNIFAC predictive model for identifying IDACs. It is not uncommon however, for the predictions made to be inaccurate, particularly in cases where hydrogen bonding or strong hydrophobic interactions are exhibited (Méndez Sevillano et al., 2014). Furthermore, UNIFAC can only maintain its accuracy for systems under specific temperature and pressure conditions (Muzenda, 2013), and its results worsen for very asymmetric molecules and cannot be directly applied to polymer mixtures. Efforts to improve thermodynamic predictions has led to a recent surge in data driven approaches, comprising techniques from matrix completion methods (Damay et al., 2021), to advanced neural networks and transformers (Sanchez Medina et al., 2023). The use of such black box machine learning methods, although accurate, lacks interpretability, limiting physical understanding one may gain from model construction. In recent literature, symbolic regression has proven effective for knowledge discovery, rediscovery of complex physics and in the identification of descriptors for complex systems. Due to its success across a wide array of fields, it is believed that it may offer a unique solution to overcome the persistent interpretability challenges in traditional data-driven modelling.

* 1. Methodology
		1. UNIFAC

First, we will cover the structure of the UNIFAC Dortmund model. UNIFAC is partitioned into two terms, one being the combinatorial term which holds the entropic contribution due to molecule size and shape differences. The second being the residual term, which represents the enthalpic contribution, and is dependent on the interactions of the constituent subgroups in the mixture. UNIFAC Dortmund was developed in order to correct structural bias in the combinatorial term, to improve predictions on asymmetric systems, and in the residual term, to enhance the temperature dependance of the model (Lohmann et al., 2001). Yet, even so, there exists much room for improvement when compared to recent data driven applications.

* + 1. MACCS Keys

A major characteristic of UNIFAC is its basis in the solution of groups concept, which accounts for the interactions between individual subgroups under the assumption that each subgroup’s properties are independent to the rest of the molecule to which it resides on. To enhance UNIFAC’s accuracy whilst aligning with its solution of groups foundation, it is proposed to introduce structural information about the molecules under study through MACCS fingerprints.

MACCS fingerprints deconstruct each molecule into subgroups, each of which is compared to one of 166 pre-defined keys, obtaining a value of 1 if present and 0 if not; this way, a distinctly representative fingerprint of each molecule can be found. The dimensions of the MACCS fingerprints are impractical for genetic algorithms like those used in symbolic regression so, dimensionality reduction techniques are employed for concision of the solute-solvent binary system MACCS keys. Specifically, an artificial neural network (ANN) was developed and optimised, using Bayesian optimisation, on the residuals between UNIFAC Dortmund’s predictions and the experimentally recorded IDAC values. Then, the final hidden layer was extracted and taken as the reduced input feature space of the MACCS fingerprint in the symbolic regression algorithm. The described methodology is depicted in Figure 1.



Figure 1 - Symbolic regression methodology using the ANN-MACCS input features.

* + 1. Symbolic Regression

We will now cover the concept of genetic algorithms, the cornerstone of symbolic regression. In general, genetic algorithms use a natural selection approach to identify the most fit combination of descriptors. Through successive generations, populations of symbolic equations will become increasingly fit with respect to the objective function, retaining characteristics of the most fit individuals. In this work, individuals within a population are defined via a combination of constants, operators and input variables where the operator space is confined to ($+, -, \*, /, \sqrt{}, exp, ln$), in an attempt to reduce the probability of arriving to unphysical expressions.

The symbolic regression algorithm aims to balance model accuracy with model complexity where the complexity is defined as the sum of process variables, operators and constants in the model. In the symbolic regression algorithm, the performance of the symbolic expressions is optimised using the mean squared error (MSE), defined in Eq. (1.1) and comparison between the best models is completed using the mean absolute error (MAE) metric, shown in Eq. (1.2). Each metric is calculated for a functional representation, $f$, given an input vector $x$and parameters, $θ$,for the $i$th experiment in the output vector $y$**.**

|  |  |  |  |
| --- | --- | --- | --- |
| $$MSE=\frac{1}{N}\sum\_{i=1}^{N}(y\_{i}-f\left(θ\right))^{2}$$ | (1.1) | $$MAE=\frac{1}{N}\sum\_{i=1}^{N}\left|y\_{i}-f\left(θ\right)\right|$$ | (1.2) |

Ideally, it is possible to correct the structure of UNIFAC by making more effective use of the currently provided inputs; hence, statistical features based on said inputs are calculated for use in the symbolic regression model. Explicitly, we target the group interaction parameters, $Ψ$ and the activity of isolated groups in the solution, $Γ$ for which we extract statistical representations such as the mean, standard deviation and summation. In addition to these, the group surface area and volume contribution parameters, $q$ and $r$ are provided to account for potential improvements to the model’s size and shape contribution. Finally, these inputs are appended to the aforementioned ANN-MACCS representations such that model enhancement can proceed via superior use of prior information as well as the introduction of new information.

* + 1. Data Source

The data used in this work consists of a large dataset of binary systems collected as a subset of the DECHEMA Chemistry Data Series Vol. IX, where only datapoints extrapolated from finite dilution phase equilibria measurements were removed due to their tendency to be inaccurate. The data collected consists of experimental IDAC values for a wide range of binary systems.

* 1. Results and Discussion

The data is partitioned into two sets, that is training and testing, each of which contain 25,075 and 6,457 experiments respectively. A validation set of size 10,075 is then removed from the training set to be used in model construction. To reduce the dimensionality of the input space further, gradient boosting was applied in which feature importance charts were generated to identify the most important variables to be used in the subsequent symbolic regression model.

In this figure, features 0 to 11 represent the ANN-MACCS inputs, and any feature above f11 relates to a statistical feature based on UNIFAC inputs. The 12 most crucial features were selected for use in the symbolic regression procedure; 9 of which constitute the ANN-MACCS features, whilst variables 30, 23 and 24 denote the solute surface area contribution parameter $q\_{solute}$, and the summation and mean statistics of the activities of isolated groups in the solution respectively. Construction of symbolic expressions using the selected input features prioritised the use the ANN-MACCS features, neglecting the remaining statistics, the three best resulting expressions of which are shown in Table 1. The finalised expression resulting from symbolic expression was a linear combination of the ANN-MACCS inputs; the average training, validation and testing results are 0.143, 0.170 and 0.171 respectively. In comparison, UNIFAC Dortmund yields 0.265, 0.266 and 0.267 for the training, validation and testing MAEs. Parity plots of the testing results are shown in figure 3.

Figure 2 - Feature importance chart depicting importance of the provided input features.

Table 1 – Three best performing expressions derived from symbolic regression.

|  |  |
| --- | --- |
| **Symbolic Expression** | **Testing MAE** |
| $$\sum\_{}^{}a\_{i}∙X\_{i}, i=\{2,7,8\}$$ | 0.21 |
| $$\sum\_{}^{}a\_{i}∙X\_{i}+X\_{3}^{3} i=\{2,7,8\}$$ | 0.19 |
| $$\sum\_{}^{}(a\_{i}∙X\_{i})+b, i=\{0,2,5,6,7,8\}$$ | 0.17 |



Figure 3 - Parity plot for SR corrected UNIFAC (left) and UNIFAC Dortmund (right).

From Figure 3, in the regions where UNIFAC provides relatively high errors, the generated symbolic expression is consistently able to improve the predictions; this suggests that the symbolic regression (SR) model has an increased spread of binary systems for which it can accurately predict compared to UNIFAC. Secondly, the SR model offers an improvement of 36.0 % with regards to the MAE over UNIFAC Dortmund, clearly indicating an improved capturing of underlying physical phenomena.

To assess the SR models performance against literature data driven approaches, we use the work of Sanchez Medina et al., (2023) as a benchmark. Sanchez Medina et al., (2023) combined a graph neural network with the Gibbs-Helmholtz derived expression (GH-GNN) to improve the predictions on IDAC; the results they reported are compared to UNIFAC Dortmund, and the generated SR corrected UNIFAC model in Table 2.

Table 2 - Table comparing the testing MAE for the literature derived GH-GNN, the proposed symbolic regression model and UNIFAC Dortmund

|  |  |
| --- | --- |
| **Model** | **MAE (test set)** |
| UNIFAC Dortmund | 0.27 |
| SR corrected UNIFAC Dortmund | 0.17 |
| GH-GNN | 0.12 |

From Table 2, UNIFAC Dortmund performs the worst by a large margin, however, there also exists a notable difference in the performance of the proposed SR model and the GH-GNN model. This is to be expected since the architecture of the GNN is much more intricate than of the symbolic regression model, allowing for more system complexity to be captured. Even with this, it is clear that the difference in performance between the literature GH-GNN model and the SR method is much less than the difference between the SR method and UNIFAC Dortmund, this grants justification of the lessened accuracy by potential increases in interpretability.

* 1. Conclusion

In conclusion, the use of symbolic regression offers a unique solution to the lack of interpretability offered by current “black box” data driven approaches, whilst maintaining a high level of accuracy. The introduction of new structural information into the UNIFAC Dortmund method permitted the establishment of a robust framework in which accurate estimation of complex thermodynamic properties can be made. Furthermore, the application of symbolic regression enables one to identify the statistically most important input features rather easily when compared to other data driven alternatives. This methodology facilitates the extension of UNIFAC Dortmund to a greater range of temperature studies for a larger range of molecules which traditionally would lead to poor performance using UNIFAC Dortmund. Overall, this work demonstrates the combining innovative feature engineering of new structural inputs with dimensionality reduction techniques and symbolic regression based modelling, for the improvement of traditional thermodynamic predictive models.

**Acknowledgements**

E.I.S.M. is affiliated with the International Max Planck Research School for Advanced Methods in Process and Systems Engineering–IMPRS ProEng at the Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg

References

Brouwer, T., Kersten, S. R. A., Bargeman, G., & Schuur, B. (2021). Solvent pre-selection for extractive distillation using infinite dilution activity coefficients and the three-component Margules equation. *Separation and Purification Technology*, *276*, 119230.

Cheng, J.-S., Tang, M., & Chen, Y.-P. (2004). Correlation and comparison of the infinite dilution activity coefficients in aqueous and organic mixtures from a modified excess Gibbs energy model. *Fluid Phase Equilibria*, *217*, 205–216.

Damay, J., Jirasek, F., Kloft, M., Bortz, M., & Hasse, H. (2021). Predicting Activity Coefficients at Infinite Dilution for Varying Temperatures by Matrix Completion. *Industrial and Engineering Chemistry Research*, *60*(40), 14564–14578.

de Haan, A. B., Eral, H. B., & Schuur, B. (2020). *Industrial Separation Processes - Fundamentals*. De Gruyter.

Ivan, E., Medina, S., Linke, S., Stoll, M., & Sundmacher, K. (2023). Gibbs–Helmholtz graph neural network: capturing the temperature dependency of activity coefficients at infinite dilution. *Digital Discovery*, *2*(3), 781–798.

Lohmann, J., Joh, R., & Gmehling, J. (2001). From UNIFAC to modified UNIFAC (Dortmund). *Industrial and Engineering Chemistry Research*, *40*(3), 957–964.

Medina, E. I. S., Linke, S., Stoll, M., & Sundmacher, K. (2023). Gibbs–Helmholtz graph neural network: capturing the temperature dependency of activity coefficients at infinite dilution. *Digital Discovery*, *2*(3), 781–798.

Medina, E. I. S., & Sundmacher, K. (2023). Solvent pre-selection for extractive distillation using Gibbs-Helmholtz Graph Neural Networks. *Computer Aided Chemical Engineering*, *52*, 2037–2042.

Méndez Sevillano, D., van der Wielen, L. A. M., Hooshyar, N., & Ottens, M. (2014). MPP-UNIFAC, a predictive activity coefficient model for polyphenols. *Fluid Phase Equilibria*, *384*, 82–88.

Mutch, G. A. (2022). Electrochemical separation processes for future societal challenges. *Cell Reports Physical Science*, *3*(4), 100844.

Muzenda, E. (2013). *From UNIQUAC to modified UNIFAC Dortmund: a discussion*.