PREDICTION OF PHYSICOCHEMICAL PROPERTIES BY ARTIFICIAL NEURAL NETS

PREDICCIÓN DE PROPIEDADES FISICOQUÍMICAS MEDIANTE REDES NEURONALES ARTIFICIALES

Arturo Alvarez[1], Víctor H. Alvarez[2]
(1) School of Mechatronic Engineering, University of Trujillo, La Libertad - Perú
(2) CarbonIP Technologies, Alberta, Canada
(email: victor.alvarez@albertainnovates.ca)

Recibido: 17/05/2023 - Evaluado: 02/06/2023 - Aceptado: 26/06/2023

ABSTRACT

In this study, we employed the Artificial Neural Network-Group COSMO sigma profile method to estimate critical pressure, critical temperature, enthalpy of fusion, and melting temperature of pure chemical compounds. Utilizing a comprehensive database containing 1400 data points of various pure compounds, we developed a robust predictive model. The method demonstrated high accuracy, yielding root mean square errors of 5 bar for critical pressure, 25 K for critical temperature, 2030 kJ/mol for enthalpy of fusion, and 23 K for melting temperature. These results underscore the potential of the Artificial Neural Network-Group COSMO sigma profile method as a reliable tool for predicting critical thermodynamic properties, contributing valuable insights to the field of chemical engineering and material science.

RESUMEN

En este estudio, se aplicó el método de Red Neuronal Artificial-Perfil sigma de Grupo COSMO para estimar la presión crítica, la temperatura crítica, la entalpía de fusión y la temperatura de fusión de compuestos químicos puros. Se investigó una base de datos con 1400 puntos de compuestos puros para proponer el modelo predictivo. Los resultados proporcionaron errores cuadráticos medios de 5 bar, 25 K, 5 kJ/mol y 32 K, para las propiedades estimadas de presión crítica, temperatura crítica, entalpía de fusión y temperatura de fusión, respectivamente.

Keywords: neural net, critical properties, solvation model, thermodynamic models
Palabras clave: red neuronal, propriedades críticas, modelo de solvatación, modelos termodinámicos

INTRODUCTION

Thermodynamic models are essential tools for accurately describing the aqueous solubility of solutes at high pressure and temperature, utilizing critical pressure (Pc), critical temperature (Tc), molar enthalpy of fusion (ΔH_f), and melting point (T_m) as key parameters. However, the experimental data for these properties are often unavailable for numerous chemical compounds, necessitating the development of predictive models.

Group contribution (GC) methods are commonly employed to estimate critical properties and physicochemical properties. These methods assume that the properties of a molecule are derived from the contributions of its constituent elements. The intermolecular forces that determine the constants of interest are largely dependent on the bonds between atoms. Consequently, each molecule is treated as an assemblage of fundamental groups, with each group contributing to the property of interest, which is then calculated by summing the contributions of each group.

Numerous GC methods have been developed to predict critical temperature and pressure, including those by Kudchadker and Zwolinski (1966), Lydersen (1955), Joback and Reid (1987), Marrero and Gani (2001), Thodos (1955), Ambrose and Ghiasse (1987), and Wilson and Jasperson (1999).

Marrero and Gani (2001), advanced several GC methods for predicting ΔHf and Tm. Their third-order GC method, which exhibited the best results for 700 compounds, yielded a standard deviation, average absolute error, and average absolute relative deviation of 3.7, 2.2 kJ/mol, and 15.7%, respectively. The quantitative structure-property relationship (QSPR) method has also been utilized to predict ΔHf, though it is limited to specific chemical families (Dyekjaer & Jonsdottir, 2003; Puri *et al*., 2003; Goodarzi *et al*., 2010). Gharagheizi and Salehi (2011), introduced an artificial neural network group contribution (ANN-GC) method to estimate ΔHf with training and test set deviations of less than 3%.

Although melting temperature can be measured accurately, its prediction has been challenging. While reasonably accurate models have been developed for small subgroups of compounds, relatively few models exist for predicting the Tm of biomolecules or molecules used in pharmaceutical applications (Katritzky *et al*., 2001). Bergström *et al*. (2003), predicted the melting temperatures of a set of 92 drugs, trained on 185 compounds, using electron topology descriptors, reporting a root mean square error (RMSE) of 49.8 K and a squared correlation coefficient (R^2) value of 0.5. Karthikeyan *et al*. (2005), used the Bergström data set as a validation set for their ANN model, trained on a selection of compounds from the Molecular Diversity Preservation International (MDPI) database. For the selection of molecules from MDPI not used in the ANN training, Karthikeyan *et al*. (2005) reported an RMSE of 50.4 K with an R^2 value of 0.64. Hughes et al. (2008), employed a support vector machine model with 2D and 3D descriptors to predict the melting temperature, achieving an RMSE of 53 K for a database of 287 compounds.

An important disadvantage of GC methods is that they cannot be applied to compounds containing groups that are not included in the training set. In addition, these methods lack consideration of the interactions between the different groups present in a molecule and the spatial arrangement of the various groups. Alternative approaches such as the Conductor-like Screening Model-Segment Activity Coefficient (COSMO-SAC), proposed by Lin and Sandler (2002), have gained attention. The COSMO-SAC method uses a sigma profile to represent the charge density distribution within a molecule, providing a more comprehensive understanding of molecular interactions.

Another promising approach is the use of Artificial Neural Networks (ANNs) for predicting thermodynamic properties. ANNs can detect complex relationships between variables and offer robust predictions for a wide range of compounds. However, ANNs are often criticized for their "black-box" nature and the risk of over-fitting, which requires careful training and validation.

In this study, we aim to develop an ANN model to predict Pc, Tc, ΔH_f, and T_m for 1400 chemical compounds. By integrating the sigma profile from the COSMO-SAC model, molecular volume, and molecular weight, our ANN

model seeks to provide accurate and reliable predictions. This approach addresses the limitations of traditional GC methods and leverages the strengths of ANNs in handling complex data.

The research focuses on bridging the gap between traditional GC methods and modern computational techniques, offering a novel solution for predicting properties. The outcomes of this study have significant implications for chemical engineering, providing a valuable tool for researchers and engineers in various industries.

METHODOLOGY

The ANN was meticulously designed and optimized to incorporate the sigma profile of COSMO-SAC, molecular weight, molecular volume, and the number of chemical bonds. A comprehensive database containing 1400 data points for critical temperature (Tc), enthalpy of fusion (ΔH_f), and melting temperature (T_m) of pure compounds from various chemical families was utilized to develop the predictive model. The COSMO-SAC sigma profile for all the chemical compounds was employed in the analysis. The dataset included Tc values ranging from 33 to 1290 K, ΔH_f values ranging from 117 to 99,200 kJ/mol, and T_m values ranging from 13.8 to 870.2 K. The development and application of molecular descriptors based on the COSMO-SAC sigma profile are elaborated in the following sections.

Artificial neural network (ANN)

A computational neural network comprises simple processing units known as neurons. The effectiveness of these neurons is determined by the weights assigned to them. Initially, the inputs are multiplied by their respective connection weights, summed, and then processed through a transfer function to generate the neuron's output.

$$z_j = f\left(\sum_{i=1}^{NI}(wu + b)_i\right) \tag{1}$$

where f is the transfer function, NI is the number of inputs, w represents the connection weights, b is the bias, and ui and z are the i-th input and j-th output of the ANN, respectively. The hyperbolic tangent sigmoid transfer function is the most widely used, as it limits the output value of the neurons between -1 and 1, as shown in equation (2).

$$\tanh(x) = \frac{2}{1 + e^{-2x}} - 1 \tag{2}$$

The multi-layer feedforward (MLF) network has seen extensive use in a variety of applications. MLF networks consist of two or more layers: an input layer, one or more hidden layers, and an output layer. The number of neurons in the input and output layers corresponds to the number of input and output parameters, respectively. The number of neurons in the hidden layers is optimized during the training process. During training, a learning algorithm adjusts the connection weights based on the data for the computational task. Among various learning rules, backpropagation is the most commonly used. In this method, the differences between each ANN output and its desired value are calculated, and the weights are adjusted using error terms. The primary challenge in ANN modeling is to determine a set of optimized weights that minimize the prediction error to an acceptable level.

Parameters for the ANN

The first step in designing a neural network is selecting appropriate inputs. These input parameters must theoretically have a relationship with the outputs. Since Pc, Tc, ΔH_f, and T_m are topology-dependent, molecular weight and molecular volume from COSMO were considered as inputs. Additionally, because the ground state 3D structure of a molecule is influenced by intramolecular interactions, the full COSMO-SAC sigma profile was used

as an input for the ANN. Before using the input parameters for training or testing, they were all normalized to a scale of -1 to 1.

The COSMO-SAC sigma profile is an area-weighted energy profile for each molecule. This profile is a file containing the sigma profile, which represents the probability of a segment having a specific charge density (σ), weighted by the total surface area of the molecule. The sigma profiles were generated from single 3D molecular structures through quantum-mechanical calculations.

First, the equilibrium molecular geometry of each molecule was obtained by minimizing its molecular energy. From this equilibrium geometry, the volume of the cavity (VCOSMO), the area of the cavity (ACOSMO), and the total number of segments (COSMO segments) were estimated using solvation calculations in a perfect conductor. These calculations were performed with the quantum chemistry package developed by Accelrys Materials Studio v4.3. The detailed settings for DMol3 (using the GGA/VWN-BP functional) followed the same parameters used by Mullins *et a*l. (2006). Finally, the COSMO data were utilized to derive the COSMO-SAC sigma profile based on equations previously reported by Mullins *et al*. (2006).

RESULTS AND DISCUSSION

Artificial neural network

A hyperbolic tangent sigmoid transfer function was used in the hidden layer, and a linear transfer function was employed in the output layer for all ANN calculations. The ANN algorithms were implemented in the MATLAB programming language. Optimizing an ANN in MATLAB presents a significant challenge due to the local optimization of parameters. Literature shows that each re-optimization of the ANN can result in different values for parameters such as biases and weights. To address this issue, a global optimization approach was implemented using a genetic algorithm initialized with local optimizations. Experimental data for Tc, ΔH_f, and T_m were sourced from Diadem Public 1.2 (2000).

When using an ANN, it is standard practice to split the collected data into two groups: the training set and the test set. The training set is used to train the network and evaluate its performance. In this study, 80% of the total 1400 data points were randomly selected for training the network, while the remaining 20% were used for validating the developed model (test set).

The accuracy of the developed predictive method was assessed using the root mean square error (RMSE), calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{nANN}\sum_{i=1}^{nANN}\left(lny_i^{exp} - lny_i^{cal}\right)^2} \tag{3}$$

where ln yi is the natural log of the property, exp represents the experimental data, cal denotes the ANN-calculated values, and nANN is the number of data points used for training or testing the ANN model.

The architecture of an ANN is defined by the number of layers, the number of neurons in each layer, the activation function of each layer, and the training algorithm. In this study, various neural network topologies were tested using a trial-and-error procedure, as shown in Table 1.

The final selected structure of the implemented ANN for Pc, Tc, ΔH_f, and T_m consisted of 12 neurons. This architecture was chosen based on the low deviations of the predicted values for the test set, as calculated by RMSE.

Table 1: Root means square error for the data test set using various numbers of neurons in one hidden layer.

in-layer	Pc	Tc	ΔH_f	T_m
5	6.6	29.4	5.0	41.3
6	6.9	28.5	5.3	38.1
7	5.4	30.0	4.8	38.0
8	7.3	26.5	4.8	37.6
9	7.5	28.6	4.6	37.1
10	5.2	25.0	5.1	35.1
12	6.0	27.3	5.1	31.8
15	7.3	27.9	4.5	41.6

For the Pc testing data, 280 compounds were considered with the highest and lowest values of 103 and 10.7 bar for the Bromine and 1,2-benzene dicarboxylic acid, dinonyl ester, respectively. For the Tc testing data, the highest and lowest values of 1290 and 191 K for the phosphorus sulfide and methane, respectively. For the testing data of the ΔH_f, the highest and lowest values of 97.9 and 0.5 kJ/mol for the pentaerythritol tetranitrate and pentane, 2,2,3,4-tetramethyl-, respectively. For the Tm test data, the highest and lowest values of 625.15 and 13.95 K for melamine and hydrogen, respectively.

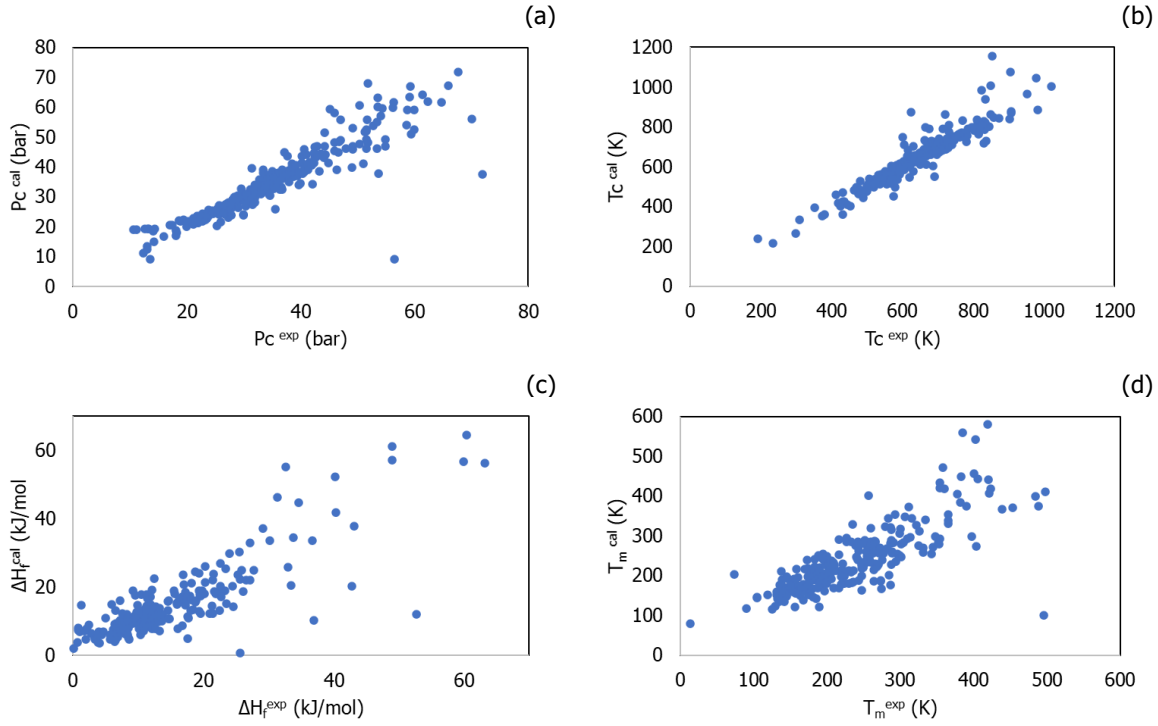

Fig. 1: Property predicted by ANN versus experimental value for test data for the critical pressure (a). critical temperature (b), enthalpy of fusion (c), and melting temperature (d).

Furthermore, Fig. 1 shows that the ANN-Sigma profile model can reproduce the trend of the values for the experimental properties. But there were high deviations at high values of enthalpy of fusion suggesting that ΔH_f is the most challenging property to predict.

CONCLUSIONS

This study presents predictions for four challenging properties of pure chemical compounds: critical pressure, critical temperature, enthalpy of fusion, and melting temperature. An artificial neural network model was developed to predict these properties using inputs from the COSMO-SAC model, molecular weight, and the number of chemical bonds. Among these properties, the enthalpy of fusion proved to be the most difficult to predict. The study confirms that the sigma profile serves as a useful descriptor in this context.

ACKNOWLEDGEMENTS

The authors are grateful to CarbonIP Technologies for funding this project.

REFERENCES

Ambrose, D. & Ghiasse, N.B. (1987) Vapor pressures and critical temperatures and critical pressures of some alkanoic acids: C1 to C10. *J. Chem. Thermodyn., 19*, 505–519.

Bergstrom, C.A., Norinder, U., Luthman, K. & Artursson, P. (2003). Molecular descriptors influencing melting point and their role in the classification of solid drugs. *J. Chem. Inf. Comput. Sci.*, *43*, 1177-1185.

Diadem Public 1.2. (2000). The DIPPR Information and Data Evaluation Manager.

Dyekjaer, J.D. & Jonsdottir, S.O. (2003). QSPR models based on molecular mechanics and quantum chemical calculations. 2. Thermodynamic properties of alkanes, alcohols, polyols, and ethers. *Ind. Eng. Chem. Res.*, *42*, 4241-4259.

Gharagheizi, F. & Salehi, G.R. (2011). Prediction of enthalpy of fusion of pure compounds using an Artificial Neural Network-Group Contribution method. *Thermochim. Acta*, *521*, 37-40.

Goodarzi, M., Chen, T. & Freitas, M.P. (2010). QSPR predictions of heat of fusion of organic compounds using Bayesian regularized artificial neural networks. *Chemometr. Intell. Lab., 104*, 2010, 260-264.

Hughes, L.D., Palmer, D.S., Nigsch, F. & Mitchell, J.B.O. (2008). Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and Log P. *J. Chem. Inf. Model., 48*, 220-232.

Joback, K.G. & Reid, R.C. (1987). Estimation of pure component properties from group contributions. *Chem. Eng. Commun., 57*, 233–243.

Karthikeyan, M., Glen, R.C. & Bender, A. (2005). General melting point prediction based on a diverse compound data set and artificial neural networks. *J. Chem. Inf. Model., 45*, 581-590.

Katritzky, A.R., Jain, R., Lomaka, A., Petrukhin, R., Maran, U. & Karelson, M. (2001). Perspective on the relationship between melting points and chemical structure. *Cryst. Growth Des.*, *1*, 261-265.

Kudchadker, A.P. & Zwolinski, B.J. (1966). Vapor pressures and boiling points of normal alkanes C21 to C100. *J. Chem. Eng. Data, 11*, 253–255.

Lin, S.T. & Sandler, S.I. (2002). A priori phase equilibrium prediction from a segment contribution solvation model. *Ind. Eng. Chem. Res., 41*, 899-913.

Lydersen, A.L. (1955). *Estimation of Critical Properties of Organic Compounds*, Eng. Exp. Stn. Rep. 3; University of Wisconsin College Engineering: Madison, WI.

Marrero, J. & Gani, R. (2001). Group-contribution based estimation of pure component properties. *Fluid Phase Equilibr.*, *183–184*, 183-208.

Mullins, E., Oldland, R., Liu, Y.A., Wang, S., Sandler, S.I., Chen, C.C., *et al*. (2006). Sigma-profile database for using COSMO-based thermodynamic methods. *Ind. Eng. Chem. Res. 45*, 4389-4415.

Puri, S, Chickos, J.S. & Welsh, W.J. (2003). Three-dimensional quantitative structure-property relationship (3D-QSPR) models for prediction of thermodynamic properties of polychlorinated biphenyls (PCBs): enthalpy of vaporization*. J. Chem. Inf. Comp. Sci., 43*, 55-62.

Wilson, G.M. & Jasperson, L. V. (1996) *Critical constants Tc and Pc, estimation based on zero, first and second order methods*, AIChE Spring National Meeting, 25-29 February, New Orleans, LA-USA.

Thodos, G. (1955). Critical constants of saturated aliphatic hydrocarbons. *AIChE J., 1*, 168–173.