Fluid Phase Equilibria 382 (2014) 21–30

Contents lists available at ScienceDirect

Fluid Phase Equilibria

journal homepage: www.elsevier.com/locate/fluid

# A group contribution model for the prediction of the freezing point of organic compounds

Farhad Gharagheizi[a,b], Poorandokht Ilani-Kashkouli[a,b], Arash Kamari[a], Amir H. Mohammadi[a,c,*], Deresh Ramjugernath[a,*]

[a] Thermodynamics Research Unit, School of Engineering, University of KwaZulu-Natal, Howard College Campus, King George V Avenue, Durban 4041, South Africa
[b] Department of Chemical Engineering, Buinzahra Branch, Islamic Azad University, Buinzahra, Iran
[c] Institut de Recherche en Génie Chimique et Pétrolier (IRGCP), Paris Cedex, France

ARTICLE INFO

*Article history:*
Received 3 March 2014
Received in revised form 19 August 2014
Accepted 20 August 2014
Available online 28 August 2014

*Keywords:*
Group Contribution
Freezing Point
Sequential Search
Model
Database

ABSTRACT

The freezing point is a fundamental thermo-physical property which is important in describing the transition between the liquid and solid phases. As this property is required for describing phase behavior and the design of separation unit operations, an efficient, applicable and reliable method which can predict it is of great importance, especially for compounds where there are no experimental data available. In this article, an efficient and reliable group contribution (GC) model is developed for the determination of the freezing point of organic compounds. The sequential search mathematical approach is used in this study to select an optimal collection of functional groups (112 functional groups) and subsequently to develop the model. A large dataset of freezing point data for about 17,000 pure mostly organic compounds was used to develop and validate the model. A comparison between the model results and the database shows a squared correlation coefficient of 0.735 ($R^2$). Moreover, the proposed group contribution model is able to predict the freezing point of organic compounds to within an average absolute relative deviation of 10.76%, which is of adequate accuracy for many practical applications. Furthermore, the leverage approach (Williams plot) is used to determine the applicability domain of the model and to detect probable erroneous data points.



## 1. Introduction

At the freezing point of a solution, solid solvent is in equilibrium with the solvent in solution. As with the melting point, an increased pressure normally raises the freezing point. The freezing point is lower than the melting point, in the case of mixtures and for certain organic compounds such as fats [1]. As a mixture freezes, the solid that forms first normally has a composition different from that of the liquid, and formation of the solid changes the composition of the remaining liquid, normally in a way that steadily lowers the freezing point. This principle is utilized in purifying mixtures, successive melting and freezing gradually separating the components [1]. Consequently, the fusion heat (heat required to melt a solid) must be removed from the liquid to freeze it. Some liquids can be supercooled (cooled below the freezing point) without solid crystals forming. The addition of a seed crystal into a supercooled liquid triggers freezing, whereupon the release of the heat of fusion raises the temperature rapidly to the freezing point [1].

Freezing point and/or melting point (depending on some considerations in their descriptions) are fundamental physical property specifying the transition temperature between liquid and solid phases [2]. Furthermore, they have been used for the prediction of other physical properties such as aqueous solubility [3–5]. Hence, accurate prediction of this fundamental thermo-physical property seems an essential necessity. To date, there have been a few quantitative structure-property relationships (QSPR) methods, such as the property–property relationships (PPR) [6], and group contribution methods [7–9] applied in attempt to estimate freezing/melting point. There are some successful estimations of melting points, e.g. for 24 normal alkanes ($R^2$ = 0.998) using topological indices like the carbon number, Wiener index, and the Balaban distance sum connectivity index [10]. Nevertheless, some models such as the QSPR models proposed by Needham et al. [11] indicate poor predictability ($R^2$ = 0.570) for their use of 56 normal and branched alkanes. A

* Corresponding authors at: Institut de Recherche en Génie Chimique et Pétrolier (IRGCP), Paris, France.
*E-mail addresses:* a.h.m@irgcp.fr, amir_h_mohammadi@yahoo.com (A.H. Mohammadi), ramjuger@ukzn.ac.za (D. Ramjugernath).

http://dx.doi.org/10.1016/j.fluid.2014.08.025
0378-3812/

QSPR model [12] for melting point using a dataset containing 443 mono- and di-substituted benzenes which was correlated with a set of structural parameters, and a nine-parameter model showed a $R^2$ of 0.837. Descriptors related to hydrogen bonding ability, molecular packing in crystals, and other intermolecular interactions such as charge transfer and dipole–dipole interactions contributed to the prediction of melting point.

Burch et al. [13] recently proposed multi parameters models to estimate melting points of alkanes having 10–20 carbon atoms and only one methyl group, which are of special interest to petroleum engineers manufacturing synthetic diesel fuel. A nonlinear regression model with satisfactory predictability was acquired based on the Wiener path numbers, the number of carbon atoms, the methyl locant index, and the mean Wiener index.

A comprehensive review on the previous methods developed for the prediction of the freezing point of chemical compounds demonstrates that most of them have been developed for small chemical groups/families of compounds using small databanks. Hence, in this study, a very large database is used to develop a general group contribution relationship for the prediction of the freezing point of organic compounds.

## 2. Data collection

The key step in developing thorough predictive models is the selection of an informative, inclusive and representative dataset [14–16]. The essential criteria for a satisfactory predictive model are the availability of a set of data of adequate size, diversity and measured under the same (or similar) conditions with satisfactory reproducibility and accuracy [2]. Consequently, there are relatively few previous studies in literature which report the handling of large datasets to derive a group contribution method. Hence, a dataset of freezing point values for 16,941 diverse mostly organic compounds extracted from Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds [17] was used in this study.

An analysis of the compounds within the dataset indicates that the freezing points range between 54.26 and 914.05 K. The compounds are composed of hydrogen (1 to 200 atoms per compound), carbon (1 to 99 atoms per compound), nitrogen (1 to 8 atoms per compound), oxygen (from 1 to 18 atoms per compound), phosphorus (1 to 4 atoms per compound), sulfur (1 to 8 atoms per compound), fluorine (1 to 45 atoms per compound), chlorine (1 to 10 atoms per compound), bromine (1 to 10 atoms per compound), iodine (1 to 5 atoms per compound) and boron (1 to 3 atoms per compounds). Other elements in the database are aluminum, silicon, iron, germanium, arsenic, selenium, cadmium, tin, antimony, tellurium, mercury, lead and bismuth. The maximum atom numbers of these elements are reported as, 3, 1, 8, 1, 1, 2, 4, 1, 2, 2, 2, 1, 1 and 1, respectively.

There are 2460 hydrocarbons in the dataset whose freezing points range from 85.47 to 710.55 K. The dataset includes 5900 nitrogen compounds whose freezing points range from 90.35 to 636.85 K. An elemental composition analysis of the dataset further indicates that there are 9777 oxygen compounds whose freezing points range from 36.45 to 1131.15 K. There are 1339 sulfur compounds in the dataset having freezing points that range from 104.2 to 755.15 K. The dataset includes 268 phosphorous compounds whose freezing points range from 175.95 to 618.65 K. There are a significant number of halogen compounds within the dataset: 1066 fluorine-containing compounds with freezing points between 74 and 723.15 K; 2046 chlorine containing compounds with freezing point between 74 and 604.15 K; 1059 bromine-containing compounds having freezing points that range from 105.15 to 644.66 K; and 469 iodine-containing compounds whose freezing points range from 151.15 to 641.18 K.

## 3. Development of the group-contribution model

For the prediction of pure component properties, group-contribution models such as those developed by Lyman et al. [18], Lydersen et al. [19], Joback and Reid [20], Horvath [21], Ambrose [22], and Klincewicz and Reid [23] have been widely utilized. In these models, the property of a compound is a function of structurally-dependent parameters, which are determined by summing the frequency of each group occurring in the molecule and multiplying by its contribution. These techniques provide the advantage of quick prediction without requiring substantial computational resources [24]. In proposing an efficient and reliable group-contribution model for the prediction of freezing point, the chemical structures of all the compounds were tested thoroughly to find out the most efficient sub-structures. Hence, having defined the compounds present in our database, the chemical structures of all of the studied compounds were analyzed to recognize the chemical substructures. These functional groups are normally selected from a series involving approximately 500 varying chemical groups [25].

In the next step, the frequency of appearance of each of the chemical substructures was counted in each compound. The pair correlation between each pair of the chemical substructures was then evaluated to avoid entering irrelevant parameters into the final model. Next, if the pair correlation of a pair of chemical substructures was more than the threshold value of 0.9, one of them was removed while the other was kept for the next step. Conducting these steps, the collection of the chemical substructures was decreased to nearly 300 chemical substructures.

In order to select the optimal subset of chemical substructures which affect the freezing point and finally proposing the final group-contribution model, the sequential search strategy was implemented [25]. The method is suitable for the subset variable selection in terms of its capability of handling the large number of data, as well as for an acceptable computational run-time. As a matter of fact, the major target of a sequential search is to find an optimal subset of chemical substructures for a specified model size [26]. The basic idea of the method is to replace each chemical substructure, one at a time, with all the remaining ones and see whether a better model is obtained.

Normally in the group contribution modeling, the selected literature dataset is divided into three subsets which are the training, validation and test sets. The "Training" set is applied to generate the model structure, while the "Validation" set as well as the "Test" set are employed to investigate its prediction validity and capability. In other words, the first set is for developing the model, the second set is for evaluation of the internal validity of the group-contribution model, and the final set is for assessing the predictive capability. In splitting the dataset into sub-data sets, several distributions have been used to avoid local minima and accumulation of the data in the feasible region of the problem. Consequently, an adequate distribution is the one with homogeneous accumulations of the data in the domain of the three sub-data sets [27]. In this study, the *K*-means clustering technique is implemented to partition the main dataset into the training, the validation, and the test sets. The *K*-means clustering method is a means of cluster analysis which aims to split $n$ observations into $k$ clusters in which each observation belongs to the cluster with the nearest mean. In other words, it would be of great interest if we could divide the main dataset so that all the subsets are uniform and have almost the same ranges and means. This procedure resolves the issue of inappropriate allocation of datasets. Another point is the quota of each sub-dataset from the main dataset. As a consequence, we assigned 80% (13,533 points), 10% (1694 points), and 10% (1694 points) of the main databank to each of the training, the validation, and the test sets, respectively.