Fluid Phase Equilibria 384 (2014) 1–13

Contents lists available at ScienceDirect

Fluid Phase Equilibria

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

A chemical structure based model for the estimation of refractive indices of organic compounds

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

[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 14 August 2014
Accepted 4 October 2014
Available online 21 October 2014

Keywords:
QSPR
Refractive index
Databank
Sequential search
Genetic function approximation

ABSTRACT

In this communication, the quantitative structure–property relationship (QSPR) strategy is applied to estimate the refractive indices of pure organic chemical compounds. In order to propose a comprehensive, reliable, and predictive model, a large dataset of 11,918 pure organic compounds was exploited in the development of the model. The sequential search mathematical strategy coupled with the genetic function approximation method has been observed to be the only viable technique capable of selection of the proper model parameters (molecular descriptors) which are then used in the correlation of the refractive indices. In order to allocate data to the training, validation, and test sets, the K-means clustering technique was applied. The leverage approach is used to check whether the newly developed model is statistically correct and valid. In the leverage approach, the statistical hat matrix, Williams plot, and the residuals of the model results assist in the identification of the probable data outliers. Finally, an analysis was performed to determine the validity and accuracy of the model for various atomic elements contained in the molecules, i.e., an elemental analysis with regard to the model performance. Using the dedicated strategy, satisfactory results were obtained and are quantified by the following statistical parameters: average absolute relative deviation of the predicted properties from existing literature values: 0.9%, and squared correlation coefficient: 0.892.

1. Introduction

The refractive index (n) in the visible range is an important optical property and is frequently used to assess the purity of liquid compounds in material science and thus evaluate the applicability of materials for various purposes [1]. Moreover, it is an important physical property in the identification and characterization of pure organic compounds. Precise values of this optical property are required at different temperatures and wavelengths [2]. Refractive index is related to other physicochemical properties such as density, surface tension, critical temperature, polarizability, and boiling point [1].

Fatty acids are compounds of major interest as a result of their industrial usefulness as food precursors, lubricants, and reagents. Also, many alkanoic acids play an important role in the manufacture of many synthetic materials [3]. However, few systematic studies of optical and thermo-optical properties of these apparently very common organic compounds are reported. It is for these purposes that a reliable model for the prediction of refractive indices would be very useful.

All properties of organic molecules-physical, biological, chemical, and technological-depend on their chemical structure and vary with it in a systematic way. The establishment of quantitative correlations between chemical structure and diverse molecular properties is now of great and significant importance to society in estimating and improving technological, medicinal, and environmental aspects of life [1]. These are defined as quantitative structure–property relationships (QSPR) that relate chemical, physical, or physicochemical properties of compounds to their structures. Finding a mathematical relationship between the property of interest and one or more descriptive parameters

* Corresponding author.
** Corresponding author.
E-mail addresses: a.h.m@irgcp.fr (A.H. Mohammadi), ramjuger@ukzn.ac.za (D. Ramjugernath).

http://dx.doi.org/10.1016/j.fluid.2014.10.005
0378-3812/

(descriptors) derived from the structure of the molecule can be one of the major goals of QSPR studies. Unlike the molar refraction, the refractive index was not used in many QSPR studies before 1990.

Several techniques are available for estimating the refractive indices of liquids in addition to QSPR methods. Khodier [4] determined the refractive index of eight standard oils from Physikalisch Technische Bundesanstalt, Germany with an accuracy of $\pm 1 \times 10^{-4}$ by using a Abbe refractometer. The measurements were performed at a temperature 293 K in the spectral range 0.4–0.7 μm. Measurements of the refractive index from 293 to 321 K at four fixed wavelengths, from 587.6 to 404.7 nm, reported for several acids by Rubio et al. [2]. Furthermore, they also reported the wavelength and temperature dependencies of the refractive indices obtained from a least-squares routine. The agreement between the measured and calculated refractive indices lay within the experimental uncertainty. Xu et al. [5] developed linear and nonlinear QSPR models for the estimation of refractive indices of polymers based on a diverse dataset of 120 polymers by using multi-linear regression analysis and feed-forward artificial neural networks. They calculated descriptors of the polymers from their corresponding cyclic dimer structures. Their results showed that nonlinear model performed better in comparison with linear QSPR model.

Proposing predictive models for refractive indices based simply on molecular structures is of great significance since synthesis and experimental determination of the refractive indices of new compounds is costly, laborious, and even dangerous to the researcher or the environment if the compound concerned is radioactive or has other hazardous properties. This study presents a new method for determination of refractive indices of pure organic compounds based on a QSPR modeling approach using an available dataset collected from previously published literature [6]. Moreover, the sequential search mathematical strategy has been observed to be the only viable search technique capable for feature (molecular descriptor) reduction for a dataset as large as the one which is utilized in this study [7–9]. In addition, the genetic function approximation method is used to select the most efficient subset of molecular descriptors and develop the final model. In order to partition the dataset into the training, the validation, and the test sets, the K-means clustering technique is applied. To check whether the newly developed model is statistically correct and valid, the leverage approach is applied, in which the statistical hat matrix, Williams plot, and the residuals of the model results lead to identification of the probable data outliers. Finally, an elemental analysis was performed to determine the validity and accuracy of the model for various atomic elements contained in the molecules.

2. Experimental Databank

The refractive index (RI) of a substance is associated to the speed of light in that substance [10]. Values of refractive index can be measured experimentally and are normally applied to correlate density and/or other physical properties of chemicals [11]. Therefore, information obtained from the RI measurements is used in various chemical engineering calculations. The enormous compilation of refractive index data for 11,918 diverse organic compounds drawn from the Yaws' handbook of thermodynamic and physical properties of chemical compounds [6] was used to develop the QSPR model. A universal model should exploit all information, as much as possible, regarding the studied property. From this point of view, this study includes the most comprehensive dataset collected for refractive index modeling to date. On the other hand, the applicability, reliability and accuracy of the models for estimation of physical properties depend on the comprehensiveness of the dataset employed in their development [12–18].

3. Mathematical Methodology

3.1. Determination of molecular descriptors

In developing a QSPR model, molecular descriptors are one of the most important ingredients. Furthermore, the molecular descriptors are the final result of a logical and mathematical procedure in technical terms, which transform chemical structure information, encoded within a symbolic representation of a molecule, into a useful number or the result of some standardized experiment [19]. The optimized molecular structures are a necessity to calculate molecular descriptors. The molecular structures are optimized with accurate Dreiding force fields as defined by Chemaxon's JChem [20]. In order to calculate the molecular descriptors, the optimized molecular structures must be loaded into Dragon software [21]. It is capable of calculating over 3000 descriptors from several diverse classes. These classes consist of topological indices, Burden eigen values, constitutional descriptors, connectivity indices, information indices, 2d autocorrelations, walk and path counts, functional group counts, atom-centered fragments, molecular properties, edge-adjacency indices, topological charge indices, eigenvalue-based indices, geometrical descriptors, randic molecular profiles, 3D-MORSE descriptors, RDF descriptors, WHIM descriptors, GETAWAY descriptors, charge descriptors, 2D binary fingerprint, and 2D frequency finger print. The descriptors obtained were analyzed carefully and those which were not able to be calculated for some compounds were neglected completely.

3.2. Model development

After calculation of the descriptors, the next step is to gather the subset from the descriptor pool which can correlate the refractive index well. One of the crucial issues in this study is the handling of a large number of compounds as their associated descriptors in model development. From experience gained in previous studies [8,22], it can be concluded that the sequential search strategy [23] is the correct choice for the subset variable selection/reduction, in terms of its capability of handling the large number of data, as well as acceptable computational run-times. However, the algorithm uses a simple scheme to select the best features and develop the final model. As a result, it is better to implement a two-step strategy as follows;

1. Reduction of the number of molecular descriptors to several tens so that the application of a more sophisticated method such as genetic function approximation (GFA) is feasible.
2. Selection of the best subset of molecular descriptors from the output of the previous step to develop the final model.

In the first step, a sequential search mathematical strategy is applied to reduce the number of molecular descriptors to several tens of descriptors. The basic idea is to replace each variable one at a time with all the remaining ones and determine whether a better model is obtained. It should be mentioned that the sequential forward selection (SFS) with a percentage of average absolute relative deviation as an objective function is successfully implemented for selection of variables. In this approach, features (descriptors) are added sequentially to the empty nominee until the addition of more descriptors reduces the percentage of average absolute relative deviation by a threshold value less than 0.01.

In the next step, the GFA is applied for selection of the most efficient subset of variables from the small subset of variables selected by the sequential search algorithm in the previous step. The GFA – as a genetically based variable selection approach – includes the combination of multivariate adaptive regression