Chemosphere 127 (2015) 1–9

Contents lists available at ScienceDirect

Chemosphere

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

Prediction of Henry's Law Constants via group-specific quantitative structure property relationships

Darragh R. O'Loughlin, Niall J. English *

School of Chemical and Bioprocess Engineering, University College Dublin, Belfield, Dublin 4, Ireland

HIGHLIGHTS

- Regression and neural-network models for organics' aqueous Henry's Law Constants.
- Class-specific models found to perform better than general ones.
- Neural-network models improve general models' accuracy; not so for class-specific.

ARTICLE INFO

Article history:
Received 26 July 2014
Received in revised form 19 November 2014
Accepted 26 November 2014
Available online 17 January 2015

Handling Editor: I. Cousins

Keywords:
Henry's Law Constant
Neural networks
QSPR
Regression
Group-specific

ABSTRACT

Henry's Law Constants (HLCs) for several hundred organic compounds in water at 25 °C were predicted by Quantitative Structure Property Relationship (QSPR) models, with the division of organic compounds into specific classes to yield more accurate models than generalised ones. Both multiple linear regression (MLR) and artificial neural network (ANN) versions of models were produced for three general cases, encompassing the entire data set; one used the six best descriptors, as determined by maximising the correlation coefficient; another used the twelve best descriptors in a similar manner, whilst the third used the same twelve descriptors as English and Carroll (2001). These achieved, respectively, root-mean square errors (RMSEs) of 0.719, 0.52 and 0.607 log(H_{cc}) units for the MLR version and 0.601, 0.394 and 0.431 for the test set of the ANN models, where H_{cc} is the ratio of the compound's concentration in the vapour phase to that in the liquid phase. These were compared with models for six specific chemical classes: (i) alkanes, (ii) cyclic alkanes, (iii) alkenes, (iv) halogenated compounds, (v) aldehydes, ketones and esters grouped together, and (vi) monoaromatics. These group-specific models had RMSEs of 0.153, 0.141. 0.097, 0.168, 0.122 and 0.104 respectively for the MLR versions and 0.684, 0.719, 0.856, 0.784, 0.875 and 0.861 for the test set of the ANN models. It was found that the class-specific models achieved lower RMSEs than the general models, when using MLR models. The use of ANN was found to improve the predictive accuracy of the general models but failed to improve that for the class-specific models vis-à-vis MLR.

1. Introduction

Henry's Law relates the equilibrium liquid- and vapour-phase concentrations of a solute in the limit of low solute concentrations. For dilute solutions of solute i at moderate pressures, one obtains Henry's law by equating the expressions for the vapour- and liquid-phase fugacities:

$$y_i P = x_i H_{i(px)} \tag{1}$$

where $H_{i(px)}$ is the HLC with dimensions of pressure. In the limit of infinite dilution, $H_{i(px)}$ may be defined as:

$$H_{i(px)} = \gamma^{\infty}_{i(LR)_i} P_i^{SAT} \tag{2}$$

at low to moderate pressures on a Lewis–Randall basis. Despite the importance of HLCs in a wide variety of engineering applications, experimental values are available for no more than a few thousand compounds (Meylan, 1999; Sander, 1999; Yaws, 2003). These span several orders of magnitude, so it is customary to report HLCs on a $\log_{10}(H_{cc})$ basis; the present work considers aqueous solubility at STP (1 bar and 298 K), and most HLC-prediction studies focuses on this.

Quantitative Structure Property Relationship (QSPR) models have become increasingly important and useful in modern research for prediction of physical, biological or chemical characteristics of compounds (Nantasenamat et al., 2009); indeed, the

* Corresponding author.
E-mail address: niall.english@ucd.ie (N.J. English).

http://dx.doi.org/10.1016/j.chemosphere.2014.11.065
0045-6535/

recent European REACH directive has been crafted for the use of QSPR models (Regulation (EC) No 1907/2006, 1907). The prediction of HLCs is very important for a variety of industrial and environmental applications, in terms of determining the fate of compounds released into the atmosphere, and various computer-based prediction methods have become of real value in recent years. Below, we review this progress.

The vapour pressure/aqueous (VP/AS) solubility method determines the HLC from separate solubility and vapour pressure data (Modarresi et al., 2007). This method predicts HLC as a ratio of vapour pressure of the solute to its concentration in the liquid phase (H_{pc}):

$$H_{pc} = \frac{p_i\ MW}{S} \tag{3}$$

where MW is the molecular weight of the solute and its solubility is represented by S.

However, as noted by Mackay et al. (1979), there can be difficulties in getting accurate values for solubility and/or vapour pressure, especially for low-volatility hydrophobic compounds (typically of much environmental interest). It is also difficult to get accurate data for these latter compounds because of high molecular weights, low vapour pressure and often sparing solubility in water; Burkhard et al. noted that, in these cases, the error in measuring vapour pressure is at least 6% (Burkhard et al., 1985). Recorded data are usually listed at higher temperatures than for environmental modelling; extrapolating to the temperature of interest introduces more inaccuracy (Nirmalakhandan and Speece, 1985). Clearly, if VP and AS are not accurate, the error is 'propagated' to the HLC, with the resultant HLC variance being greater than the individual variances of VP and AS. If used with accurate data, the VP/AS approach can be effective (Meylan and Howard, 1991), especially for compounds with low solubility and vapour pressure (Modarresi et al., 2007).

However, if separate accurate data for VP and AS are not available, it is then generally more accurate and pragmatic to attempt to predict HLC directly, in view of the accumulation of variances mentioned above. This is within the ambit of QSPRs. However, it is important to be aware of a number of important points, before discussing and comparing different QSPR models. Many models are trained on different data sets; when comparing, Modarresi et al. remarked that, strictly, the same data set should be used to judge their performance (Modarresi et al., 2007); many within the community would tend to agree. Indeed, Gharagheizi et al. have 'reinforced' this by recommending that they should not only be compared with the same data set, but developed on similar data, with the same level of uncertainties; when quoting results, researchers should use same definitions for deviations from true values (Gharagheizi et al., 2012). Nevertheless, these desiderata cannot always be fulfilled, or it may not always be practical to do so.

Hine and Mookerjee (1975) correlated $\log(\gamma)$ values using both bond- and group-contribution methods. By least-squares regression, they correlated 34 bond contributions and found that the difference between predicted and experimental values had a standard deviation of 0.41 $\log(H_{cc})$ units; henceforth, all references to HLC or (predictive) errors thereof shall be in $\log(H_{cc})$ units. Hine and Mookerjee speculated that deviations were mainly due to interactions between polar bonds. Their group-contribution model yielded a more accurate result, with the standard deviation of the difference as 0.12; they noted that this may not have been much larger than the experimental errors. Although the authors remarked that the group-contribution method is more accurate, bond contribution is more widely applicable: it is not always possible to determine values of all group contributions. Indeed, Lin and Sandler (2002) commented that there are two major limitations with the group contribution: no accounting for compounds' different isomers, with larger errors tending to occur for non-alkyl functional groups, although these can be countered somewhat by empirically-determined correction factors. Naturally, the group-contribution approach is not effective for compounds with functional groups not in the training set (Modarresi et al., 2007).

Cabani et al. (1981) introduced correction factors for compounds containing more than one functional group. They reported a standard deviation of 0.5 for 209 compounds, slightly improved over Hine and Mookerjee (1975). Interestingly, Dearden and Schüüurmann (2003) conducted a review of available QSPR models, each applied to a diverse set of 700 compounds, to predict the Ostwald solubility coefficient (the ratio of molar concentration of the compound in water vis-à-vis air). They found Hine-Mookerjee group contribution to yield more accurate results with a SE of 0.92 (applied to 263 compounds), compared to 2.38 (for 302) for Cabani et al.

Meylan and Howard (1991) increased the number of bond definitions from 34 to 59 bonds using least-squares analysis of 345 organic compounds, obtaining a SE of 0.34 after applying 15 correction factors. They updated this with 64 bond definitions and 57 correction factors for the bond-contribution model, while a group-contribution version contains 93 definitions, codified as HENYWIN (Meylan and Howard, 2012). In the Dearden–Schüürmann comparison (Dearden and Schüüurmann, 2003), HENRYWIN's group-contribution model obtained the most accurate value for Ostwald solubility ratio (SE of 0.88), but, however, this was only applicable to 392 of the 700 compounds. HENRYWIN's bond-contribution method was applicable to all, with SE of 1.03.

Lin and Sandler (2002) took into account electrostatic charges on nearby functional groups in the same molecule using multipole corrections. They related liquid fugacity and activity coefficient at infinite dilution to solvation free energy, formulating HLC in terms of these parameters and other related ones, *e.g.*, charge and dipole moment. The overall RMSE was 0.34 $\log(H_{cc})$ units for 395 organic compounds, comparing favourably to Meylan–Howard's models (Meylan and Howard, 1991, 2012), which had RMSEs on the same data of 0.43 and 0.52 for bond- and group-based models respectively. However, the Meylan–Howard models were developed on completely different data, containing not only organic compounds.

English and Carroll (2001) produced an early example of artificial neural networks (ANNs) for multivariate regression, producing ten- and twelve-descriptor models with both ANN and linear regression, based on 303 diverse organic compounds. A new, less localised descriptor was developed as an alternative to Kier–Hall connectivity indices (Kier and Hall, 1986; Kier and Hall, 1987). The SE and r^2 values of the ten- and twelve-descriptor ANN models were 0.202 and 0.999, and 0.224 and 0.987 respectively. The SEs were lower for ANN than linear regression. It was noted one model was better for certain types of compounds and the other for different classes. The models' SEs compared well to other studies; the authors demonstrated the increased accuracy of ANNs over linear regression for their study.

Yaffe et al. (2003) developed QSPRs based on 'fuzzy' ARTMAP and back-propagation neural networks using a heterogeneous set of 495 organic compounds. Quantum chemical, PM3-level molecular-orbital theory descriptors included polarisability, dipoles, ionisation potential and heat of formation. Average absolute errors of 0.03 and 0.13 log units were obtained for the overall data and the test set, respectively. The optimal back-propagation model was less accurate and exhibited larger average absolute errors of 0.28 and 0.27 for the validation and test sets, respectively. The fuzzy ARTMAP-based QSPR was superior to back-propagation and linear-regression models.

Modarresi et al. (2005), like Lin and Sandler (2002), also used the solvation free energy to predict the HLC, but introduced "cavity ovality", which is a 'sphericity' factor. For a set of 189 hydrocarbons, an ANN model gave a RMSE of 0.22 compared to 0.4 when cavity