Available online at www.sciencedirect.com

SciVerse ScienceDirect

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

Predicting aqueous solubility of environmentally relevant compounds from molecular features: A simple but highly effective four-dimensional model based on Project to Latent Structures

Feng Xiao [a,*], *John S. Gulliver* [a,b], *Matt F. Simcik* [c]

[a] *St. Anthony Falls Laboratory, University of Minnesota, Minneapolis, MN 55414, United States*
[b] *Department of Civil Engineering, University of Minnesota, Minneapolis, MN 55455, United States*
[c] *Division of Environmental Health Sciences, School of Public Health, University of Minnesota, Minneapolis, MN, 55455, United States*

ARTICLE INFO

Article history:
Received 13 February 2013
Received in revised form
30 May 2013
Accepted 7 June 2013
Available online 15 June 2013

Keywords:
QSPRs
Partial least-squares regression
Environmental contaminants
Aqueous solubility
Environmental mobility
Water quality

ABSTRACT

The aqueous solubility (log *S*) of xenobiotic chemicals has been identified as a key characteristic in determining their bioaccessibility/bioavailability and their fate and transport in aquatic environments. We here explore and evaluate the use of a state-of-the-art data analysis technique (Project to Latent Structures, PLS) to estimate log *S* of environmentally relevant chemicals. A large number ($n = 624$) of molecular descriptors was computed for over 1400 organic chemicals, and then refined by a feature selection technique. Candidate predictor descriptors were fitted to data by means of PLS, which was optimized by an internal leave-one-out cross-validation technique and validated by an external data set. The final (best) PLS model with only four variables (*AlogP*, *X1sol*, *Mv*, and *E*) exhibited noteworthy stability and good predictive power. It was able to explain 91% of the data ($n = 1400$) variance with an average absolute error of 0.5 log units through the solubilities span over 12 orders of magnitude. The newly proposed model is transparent, easily portable from one user to another, and robust enough to accurately estimate log *S* of a wide range of emerging contaminants.

1. Introduction

The aqueous solubility (log *S*) of organic compounds is one of the key factors to consider when ranking environmentally significant organic chemicals regarding their mobility in soil and volatility from water. It is also a particular important parameter in studies on the xenobiotic absorption, distribution, metabolisms, and excretion in humans. However, the experimental measurement of log *S* can be difficult because it can either be very time-consuming to reach the solubility equilibrium in the case of apolar compounds or require a large amount of chemicals in the case of highly hydrophilic molecules. In addition, log *S* values of a majority of emerging organic species and high-production-volume substances (Muir and Howard, 2006) remain unknown. Therefore, there is a need for reliable models for estimating log *S* based on the analysis of previously tested compounds. Establishing poly-parameter quantitative structure property relationships (pp-

* *Corresponding author*. Tel.: +1 651 964 5138; fax: +1 612 626 7750.
E-mail addresses: xiaox095@umn.edu, fxiaoee@gmail.com (F. Xiao).
0043-1354/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.watres.2013.06.011

Abbreviations and notations

DBPs	disinfection byproducts
PAHs	polycyclic aromatic hydrocarbons
PBDEs	polybrominated diphenyl ethers
PCBs	polychlorinated biphenyls
PCDDs	polychlorinated dibenzo-*p*-dioxins
PCDFs	polychlorinated dibenzofurans
PCDEs	polychlorinated diphenyl ethers
PFASs	perfluoroalkyl substances
PHCs	petroleum hydrocarbons
PPCPs	pharmaceuticals and personal care products
VOCs	volatile organic compounds
$bCI_{95\%}$	bootstrap 95% confidence intervals
D	Cook's distance
h	leverage
$PRESS_k$	predictive error sum of squares for kth latent component
Q^2_{cum}	the fraction of Y variation explained by all latent components
Q^2_k	the fraction of Y variation explained by a latent component
R^2_{adj}	adjusted coefficient of determination
RSS_{k-1}	residual sum of squares for the $(k-1)$th latent component

QSPRs) would fill this need and also aid in our understanding of the influence of chemical structure on log S.

Pp-QSPRs have attracted increasing attention among environmental scientists (Chen et al., 2010; de Ridder et al., 2010; Delgado et al., 2012; Gramatica et al., 2007; Katritzky et al., 2010; Lee and von Gunten, 2012; Mauffret et al., 2010; Nguyen et al., 2005; Redding et al., 2009; Yangali-Quintanilla et al., 2010). Compared to one-parameter QSPRs, pp-QSPRs can take into account of various factors potentially affecting the chemical fate or property of interest. Several studies have developed pp-QSPRs in order to predict log S of organic chemicals, especially drugs (see references in Jorgensen and Duffy, 2002), from physicochemical properties (i.e., melting points (Jain and Yalkowsky, 2001)) and/or fragments (i.e., the $-CH_2-$ fragment (Klopman and Zhu, 2001)). Most of these previous regression exercises are based on multiple linear regression (MLR). However, MLR or ordinary least-squares regression has a set of assumptions (e.g., homoscedasticity and linearity) to be imposed before data mining (Mundry and Nunn, 2009). It is important to examine whether the basic assumptions of MLR analysis have been violated too greatly while using MLR (Cronin and Schultz, 2003; Dearden et al., 2009). In addition, computer-controlled MLR practices can add predictor variables having no significant correlation with the response variable (Mundry and Nunn, 2009). There is always a temptation to add many predictor variables in a pp-QSPR just to increase R^2 by small amounts. An example of this case is a log S pp-QSPR containing 55 predictor variables (Eros et al., 2004). In addition to wasting degrees of freedom and generating a cumbersome model, including many but insignificant variables into a pp-QSPR may also overfit the data, leading to unsatisfactory performance of the model on an external data set.

Artificial neural networks and other machine learning techniques such as support vector machines have also been used in QSPR studies (Engkvist and Wrede, 2002; McElroy and Jurs, 2001; Vapnik, 1995; Wegner and Zell, 2003). They are highly efficient for data of high-order interactions and missing values, and have been successfully applied in environmental pattern recognition (Ha and Stenstrom, 2003). However, a major disadvantage of machine learning techniques is that they cannot generate an exploratory model, and thus the resulting QSPRs are not easily portable from one user to another.

Recently, Project to Latent Structures (PLS) (Abdi, 2010; Eriksson et al., 2003; Wold et al., 2001) has been recognized as a powerful data analysis technique in various disciplines including environmental science, computational chemistry, toxicology, and pharmaceutical science (Carroll et al., 2009; Morel et al., 2004; Platikanov et al., 2007; Qin et al., 2012; Reed et al., 2011; Weiss et al., 2009; Wold et al., 2001; Yangali-Quintanilla et al., 2010). In PLS, predictor variables are reduced to latent components, followed by a regression step where the latent components but not all predictor variables are used to predict the dependent variable. PLS shares many similarities with the principal component analysis. However, unlike the principal component analysis decomposing X to obtain components which best explain X, PLS identifies components from X (i.e., molecular descriptors) that best predict Y (i.e., log S) (Abdi, 2010). (The reader wishing to obtain a more detailed description of the PLS multivariate methodology is referred to the comprehensive reviews and relevant texts on this subject (Abdi, 2010; Barker, 2010; Barker and Rayens, 2003; Eriksson et al., 2001; Esposito Vinzi et al., 2010; Wold et al., 2001)). The purpose of this paper is to explore and evaluate the use of PLS in building a log S pp-QSPR that is exploratory and explanatory in the sense of providing the insights to the structural influence on log S. The new model derived from a set of 1400 measured solubilities is generated using PLS against a group of pertinent molecular descriptors. Molecular descriptors are state-of-the-art mathematical expressions of the structured information contained in a molecule. They are believed to be "the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment" (Todeschini and Consonni, 2000).

2. Methods

2.1. Data set

The data set (see Table S1 of the Supplementary data) contains 1400 organic compounds in more than 22 chemical classes, gleaned from the database of the Syracuse Research Corporation (http://www.syrres.com). Most of the solubilities were determined at 20–25 °C under the ambient pressure