Data in brief 28 (2020) 104821

Contents lists available at ScienceDirect

Data in brief

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

Data Article

Dataset for the quantitative structure-activity relationship (QSAR) modeling of the toxicity equivalency factors (TEFs) of PAHs and transformed PAH products

Gustav Gbeddy [a, *], Prasanna Egodawatta [a], Ashantha Goonetilleke [a], Godwin Ayoko [a], Lan Chen [b]

[a] *Science and Engineering Faculty, Queensland University of Technology (QUT), GPO Box 2434, Brisbane, 4001, Queensland, Australia*
[b] *Institute for Future Environments, Queensland University of Technology (QUT), GPO Box 2434, Brisbane, 4001, Queensland, Australia*

ARTICLE INFO

Article history:
Received 5 September 2019
Received in revised form 8 November 2019
Accepted 11 November 2019
Available online 20 November 2019

Keywords:
Quantitative structure-activity relationship
Predictor variable
Response variable
Toxicity equivalency factor
Factor analysis
PAH
Transformed PAH products
Carcinogenic health risk

ABSTRACT ±

Sixteen significant physicochemical predictor variables for thirty PAHs and transformed PAH products (TPPs) were retrieved individually prior to collation from ChemSpider.com [1] whilst their corresponding toxicity equivalency factor (TEF) end-point was obtained from published articles by Bortey-Sam, Ikenaka [2] and Wei, Bandowe [3]. In order to achieve a 5:1 ratio of the number of observations to predictors which is vital for an effective quantitative structure-activity relationship (QSAR) modelling, factor analysis was used to reduce the data. Four fundamental predictors were obtained whilst the observations were found to cluster into two main groups of nitro-PAHs and other analytes. It is anticipated that the data presented here is highly relevant for future studies on the toxicity and health effects of the analytes in the environment. Secondly, the fate and distribution patterns of PAHs and TPPs are influenced by the parameters in the dataset. In this regard, studies on the behaviour patterns of these environmental pollutants require this information for a comprehensive evaluation and interpretation of results. Researchers across varied fields of environmental science and toxicology will find this dataset very useful. This data currently serves as supplementary information for the

DOI of original article: https://doi.org/10.1016/j.jhazmat.2019.121154.
* Corresponding author.
E-mail address: g.gbeddy@qut.edu.au (G. Gbeddy).

https://doi.org/10.1016/j.dib.2019.104821

research article in the Journal of Hazardous Materials by Gbeddy, Egodawatta [4].

Specifications Table

Subject	Environmental Science; Health, Toxicology and Mutagenesis
Specific subject area	TEFs of PAHs and transformed PAH products (TPPs) are vital for assessing the carcinogenic health risks posed but are found lacking for most species.
Type of data	Table Graph Figure
How data were acquired	The data was acquired from the Royal Society of Chemistry's database/website [1] and published articles by Bortey-Sam, Ikenaka [2] and Wei, Bandowe [3]). According to ChemSpider, the predicted data for some of the descriptor variables were generated using ACD/Labs Percepta Platform PhysChem Module
Data format	Raw and analyzed
Parameters for data collection	As noted by Kunal, Supratik [5]), physicochemical predictor variables such as molecular weight, enthalpy of vapourization, boiling point, vapour pressure, surface tension, octanol water partitioning coefficient and melting point contain essential and significant information in dilating the TEF response variable. As a result, these parameters amongst others were selected.
Description of data collection	A list of thirty PAHs and TPPs with corresponding TEF values was compiled from available information. Secondly, the TEF response and a list of physicochemical predictor variables for the analytes were compiled centred on existing data. Sixteen predictor variables were found to be vastly available for most of the thirty analytes and were subsequently retrieved from the database for each analyte. The thirty TEF response values for each analyte were retrieved from literature.
Data source location	Royal Society of Chemistry, London, UK
Data accessibility	With the article
Related research article	Gustav Gbeddy, Prasanna Egodawatta, Ashantha Goonetilleke, Godwin Ayoko, Lan Chen. Application of quantitative structure-activity relationship (QSAR) model in comprehensive human health risk assessment of PAHs, and alkyl-, nitro-, carbonyl-, and hydroxyl-PAHs laden in urban road dust. J Hazard Mater, 2019.383: p. 121154. https://doi.org/10.1016/j.jhazmat.2019.121154

Value of the Data

- This data is useful in evaluating TEFs via QSAR thereby reducing the time and practical experimentation burden on animals. The data will therefore, help examine the ecotoxicology and health risks posed by these hazardous pollutants. The fate including the transformation, degradation and distribution processes of PAHs and TPPs in the environment can be assessed using this data set.
- Environmentalists, chemists, toxicologists, policy makers, ecologists and health experts can benefit invariably from this data.
- This data can be employed in various modules to prioritize experiments such as in vivo or in vitro toxicological test, biodegradation and photodegradation experiments of these pollutants thereby reducing time and cost. The outcomes of these experiments can help formulate remediation strategies thus protecting human health and the environment.
- Hitherto, the individual parameters in this dataset can be found in different sources. This dataset will therefore serve as a one-stop point for these vital parameters and therefore, likely to attract significant reference.

1. Data

The dataset entails the toxicity equivalency factors (TEFs), physical and chemical (physicochemical) properties of polycyclic aromatic hydrocarbons (PAHs) and their associated transformed products (TPPs) such as carbonyl-, nitro-, and hydroxyl- PAHs as shown in Table S1. Table 1 describes the percentage variance and eigenvalues for the first three factors obtained from the factors analysis (FA). Table 2 describes the VARIMAX rotated factor loadings for the two significant factors. Table 3 indicates the correlation matrix for the variables used in the FA. Fig. 1 shows the pattern recognition via FA biplot among the PAHs and TPPs based on their physicochemical properties and log transformed TEF values.

2. Experimental design, materials, and methods

QSAR model is a mathematical model development for relating biological activities of compounds to their molecular structures. The development of an effective QSAR model entails varied steps but can be classified into four (4) including data preparation, data processing, statistical evaluation of developed model and data interpretation [5].

2.1. Data preparation

The choice of appropriate dataset is a critical step in the development of any QSAR model. Kunal, Supratik [5] noted that any physical or chemical feature having significant information with reference to the response parameter can be employed as predictor variables in a QSAR model. The predictor and response variables can be experimental and theoretically determined parameters. The generic dataset here involves 30 desirable observations (PAHs and TPPs), 16 predictor physicochemical parameters and one response (logTEF) variable as shown in Table S1. The predictor data was obtained from ChemSpider.com [1] whilst the response parameter was retrieved from Bortey-Sam, Ikenaka [6] and Wei, Bandowe [3].

Table 1
Explained variance (Eigenvalues).

Value	Factor 1	Factor 2	Factor 3
Eigenvalue	11.079	3.289	0.850
% of Var.	65.169	19.344	5.001
Cum. %	65.169	84.513	89.514

Table 2
VARIMAX rotated factor loadings.

Variable	Factor 1	Factor 2
Mw	0.90	−0.425
NOR	0.68	−0.680
NOAR	0.62	−0.725
ρ	0.87	−0.057
Hv	0.86	−0.439
Rf	0.76	−0.641
logPl	0.79	−0.602
St	0.84	−0.235
mvol	0.75	−0.581
logKow	0.08	−0.975
MP	0.85	0.064
BP	0.94	0.166
logVp	−0.91	−0.302
Sw	0.18	0.641
logKoc	0.78	−0.520
logBCF	0.07	−0.966
logTEF	0.63	−0.169

Table 3
Factor analysis correlation matrix.

	Mw	NOR	NOAR	ñ	Hv	Rf	logPl	St	mvol	logKow	MP	BP	logVp	Sw	logKoc	logBCF	logTEF
Mw	1.00																
NOR	**0.90**	1.00															
NOAR	**0.86**	**0.92**	1.00														
ρ	**0.81**	0.74	0.58	1.00													
Hv	**0.97**	**0.88**	**0.83**	0.81	1.00												
Rf	**0.96**	**0.96**	**0.95**	0.70	**0.93**	1.00											
logPl	**0.97**	**0.94**	**0.93**	0.72	**0.95**	**0.99**	1.00										
St	**0.84**	**0.85**	0.71	0.96	**0.83**	0.79	0.79	1.00									
mvol	**0.93**	**0.83**	**0.87**	0.56	**0.91**	**0.94**	**0.95**	0.61	1.00								
logKow	0.49	0.73	0.73	0.14	0.52	0.68	0.65	0.29	0.64	1.00							
MP	0.71	0.48	0.49	0.62	0.68	0.59	0.63	0.58	0.64	0.03	1.00						
BP	0.77	0.48	0.45	0.72	0.73	0.61	0.65	0.67	0.64	−0.10	**0.83**	1.00					
logVp	−0.69	−0.37	−0.36	−0.71	−0.63	−0.51	−0.55	−0.63	−0.54	0.24	**−0.84**	**−0.94**	1.00				
Sw	−0.09	−0.24	−0.34	0.16	−0.13	−0.25	−0.23	−0.01	−0.20	−0.51	0.14	0.18	−0.26	1.00			
logKoc	**0.91**	**0.85**	**0.88**	0.61	**0.83**	**0.94**	**0.94**	0.69	**0.92**	0.55	0.68	0.69	−0.62	−0.19	1.00		
logBCF	0.47	0.70	0.73	0.08	0.46	0.68	0.64	0.24	0.64	0.98	0.04	−0.09	0.23	−0.48	0.60	1.00	
logTEF	**0.65**	0.48	0.45	0.45	**0.67**	**0.58**	**0.59**	0.47	**0.64**	0.21	0.41	**0.58**	**−0.49**	−0.01	**0.54**	0.19	1.00

Bold numbers represent (i) correlation among predictors contributing ≥ 0.7 to Factor 1 in Table 1, and (ii) correlation between predictors in (i) and logTEF with a coefficient > 0.50 or ≤ -0.49.

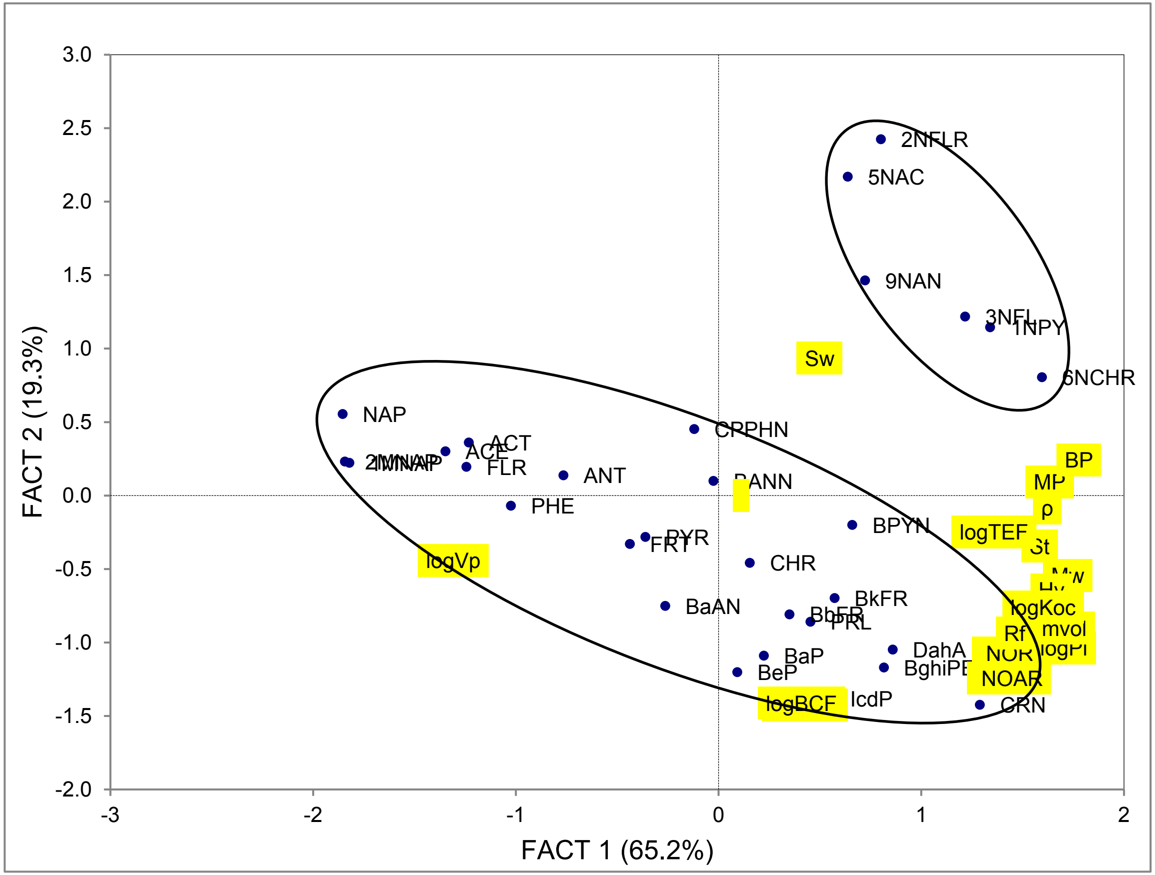

Fig. 1. Factor Analysis biplot.

3. Data processing

A dataset consisting of 30 observations and 16 predictors implies the potential availability of redundant and inter-correlated data. It is reported that for satisfactory robustness and reliability of the developed QSAR model, the ratio of observations to descriptors should be 5:1 [5]. In this regard, the generic QSAR data (Table S1) was subjected to factor analysis (FA) pre-treatment procedure using Microsoft Office Professional Plus Excel 2010 StatistiXL plug-in version 1.8 software. FA simplifies the dimensionality of the data by reducing the number and multicollinearity of variables with minimum loss of information thereby serving as a data reduction and variable selection method. As a result of the differences in the units of the variables, the data matrix was auto-scaled (standardized) and a reduced correlation matrix formed. The initial variables are transformed into new sets of variables called factors. Factors are derived from the calculation of eigenvalues and the corresponding eigenvectors. Further information on the explanation of eigenvalues and eigenvectors can be found in Kunal, Supratik [5]). The factors are composed of factor loadings and orthogonal combinations of the initial variables. Factor loading depicts how much a variable contributes to the factor. Thus a high factor loading signifies that the dimensions of the factor are highly accounted for by the variable. The factors are extracted using the principal component (PC) method and VARIMAX rotation in order to prevent collinearity in the new data. The first, second, third etc. Factors exhibit decreasing trend of residual variance in the data [5,7,8].

The results of analysis as indicated in Table 1 show that two factors are significant (i.e. eigenvalue >1) accounting for 84.5% variance in the data. The factor loadings in Table 2 indicate that eleven (11) variables including Mw, ρ, Hv, Rf, logPl, St, mvol, MP, BP, logVp and logKoc have high factor loadings. From Table 3 and Fig. 1, BP, MP and ρ predictors are highly correlated with each other. However, amongst these three variables, BP has the highest correlation with the response variable logTEF. Secondly, Mw, St, Hv, logKoc, Rf, mvol, logPl, NOR and NOAR predictor variables are significantly correlated. However, amongst these parameters Mw exhibited the greatest correlation with other predictors. Mw, Hv and mvol have the highest correlation with logTEF. The only predictor variable with appreciable negative correlation with logTEF is logVp as shown in Table 3 and Fig. 1. The 16 predictors can therefore be reduced to four (4) relevant parameters. The outcomes of this pre-treatment process was used in formulating the QSAR table for predicting the TEF values of untested PAH and TPPs as stipulated in the associated research article by Gbeddy, Egodawatta [4].

The PAH and TPPs clustered into two major groups indicated in Fig. 1. All the nitro-PAHs (N-PAHs) entailing 1NPY, 2NFLR, 3NFL, 5NAC, 6NCHR and 9NAN form one cluster whilst all the other species form another large cluster. This may be an indication of similar distribution and behaviour patterns, and potential sources of these classes of analytes in the environment. PAHs and oxygenated PAHs (O-PAHs) potentially emanate from combustion sources while N-PAHs originate from photochemical processes [3,9].

3.1. Statistical evaluation of developed QSAR model

In order to promote the robustness, reliability and predictability of developed QSAR model, the model output must be subjected to detailed statistical tests. These tests can be categorized as statistical quality test and validation tests. The validation test results points to the predictive potential of the developed model [5].

1) Statistical quality tests involves the following:

a. Mean average error (MAE)

MAE can be estimated using Equation (1). A low MAE value implies that the developed model is good.

$$MAE = \frac{\sum |Y_{obs} - Y_{calc}|}{n} \tag{1}$$

where Y_{obs}, Y_{calc} and n refer to individual original response, calculated response and number of observations respectively.

b. Determination coefficient (R^2)

$$R^2 = 1 - \frac{\sum (Y_{obs} - Y_{calc})^2}{\sum (Y_{obs} - \overline{Y}_{obs})^2} \tag{2}$$

where $\overline{Y}_{obs}$ refers to the average of original response values. A deviation of R^2 from 1 implies a reduction in the fitting quality of the developed model. The square root of R^2 ($\sqrt{R^2}$) is the regression correlation coefficient (R). A good model will potentially have an R value closer to 1.

c. Adjusted determination coefficient (R_a^2)

In order to highlight the fraction of the data variance explained by the developed model, R_a^2 is used.

$$R^2{}_a = \frac{\left((n-1)^*R^2\right) - p}{n-p-1} \tag{3}$$

where p refers to the number of predictor variables.

d. Variance ratio (F)

The overall significance of the regression coefficients (R) can be estimated using F. F is related to R by the equation as follows. A high F value indicates that R is significant.

$$F = \frac{(n-p-1)^*R^2}{\left(1-R^2\right)^*p} \tag{4}$$

e. Standard error of estimate (s)

$$s = \sqrt{\frac{\sum (Y_{obs} - Y_{calc})^2}{n-p-1}} \tag{5}$$

A good model should have a low s value.

f. Root mean square error of calibration (RMSEC)

$$RMSEC = \sqrt{\frac{\sum \left(Y_{obs(training)} - Y_{calc(training)}\right)^2}{n}} \tag{6}$$

where $Y_{obs(training)}$ and $Y_{calc(training)}$ refer to the response values in the training set and the corresponding calculated response. A low RMSEC denotes a good developed model.

2) Validation tests; internal validation method was employed and this includes the following metrics.

a. Predicted residual sum of squares (PRESS)

By employing leave-one-out (LOO) cross-validation approach, the predictivity of a developed model can be ascertained via PRESS especially for small QSAR data matrix.

$$PRESS = \sum \left(Y_{obs} - Y_{pred}\right)^2 \tag{7}$$

where Y_{pred} refers to the corresponding predicted response value for an observed response Y_{obs} during the LOO process. A good model should have a very low PRESS value [10].

b. Cross-validated determination coefficient (Q^2)

$$Q^2 = 1 - \frac{PRESS}{\sum \left(Y_{obs(training)} - \overline{Y}_{training}\right)^2} \tag{8}$$

where $\overline{Y}_{training}$ represents the mean of the observed response of the training set. An acceptable model should have a Q^2 value great than a predetermined value of 0.5 (i.e. $Q^2 > 0.5$).

c. Standard deviation of error of prediction (SDEP)

$$SDEP = \sqrt{\frac{PRESS}{n}} \tag{9}$$

A very small SDEP suggests a good developed model.

These statistical metrics were employed by Gbeddy, Egodawatta [4] in verifying the quality, robustness and reliability of the developed QSAR model.

Acknowledgments

The authors will like to acknowledge the Queensland University of Technology (QUT) for providing the postgraduate research award to Gustav Gbeddy to undertake this study. The Central Analytical Research Facility (CARF) under the Institute of Future Environments, QUT where the data employed in this paper were acquired is highly appreciated. Access to CARF was facilitated by generous funding from the Science and Engineering Faculty, QUT. The significant role of the Ghana Atomic Energy Commission (GAEC) is highly appreciated for granting study leave to Gustav Gbeddy in order to embark upon this study.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2019.104821.

References

[1] ChemSpider.com, ChemSpider search and share chemistry, 2019. (Accessed 8 March 2019).

[2] N. Bortey-Sam, et al., Levels, potential sources and human health risk of polycyclic aromatic hydrocarbons (PAHs) in particulate matter (PM(10)) in Kumasi, Ghana, Environ. Sci. Pollut. Res. Int. 22 (13) (2015) 9658–9667.

[3] C. Wei, et al., Polycyclic aromatic hydrocarbons (PAHs) and their derivatives (alkyl-PAHs, oxygenated-PAHs, nitrated-PAHs and azaarenes) in urban road dusts from Xi'an, Central China, Chemosphere 134 (2015) 512–520.

[4] G. Gbeddy, et al., Application of quantitative structure-activity relationship (QSAR) model in comprehensive human health risk assessment of PAHs, and alkyl-, nitro-, carbonyl-, and hydroxyl-PAHs laden in urban road dust, J. Hazard Mater. 383 (2019) 121154.

[5] R. Kunal, K. Supratik, N.D. Rudra, Understanding The Basics Of QSAR for Applications In Pharmaceutical Sciences And Risk Assessment, Academic Press, an imprint of Elsevier, Amsterdam; Boston, 2015.

[6] N. Bortey-Sam, et al., Occurrence, distribution, sources and toxic potential of polycyclic aromatic hydrocarbons (PAHs) in surface soils from the Kumasi Metropolis, Ghana, Sci. Total Environ. 496 (2014) 471–478.

[7] X. Wan, et al., Source apportionment of PAHs in atmospheric particulates of Dalian: factor analysis with nonnegative constraints and emission inventory analysis, Atmos. Environ. 40 (34) (2006) 6666–6675.

[8] A.G. Yong, S. Pearce, A beginner's guide to factor Analysis: focusing on exploratory factor Analysis, Tutorials in Quantitative Methods for Psychology 9 (2) (2013) 79–94.

[9] G. Gbeddy, et al., Transformation and degradation of polycyclic aromatic hydrocarbons (PAHs) in urban road surfaces: influential factors, implications and recommendations, Environ. Pollut. (2019).

[10] J.N. Miller, J.C. Miller, Statistics and Chemometrics for Analytical Chemistry, sixth ed., Pearson Education Limited, England, 2010.