Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning Methods

Q Zang[1], K Mansouri[2], D Allen[1], N Kleinstreuer[1], W Casey[3], R Judson[2]

[1]ILS/NICEATM, RTP, NC, USA; [2]EPA/ORD/NCCT, RTP, NC, USA; [3]NIH/NIEHS/DNTP/NICEATM, RTP, NC, USA

Introduction

- Estimation of physicochemical properties will be key to developing high-throughput approaches to evaluating hazards of environmental chemicals.
- We are developing novel methods for the estimation of six physicochemical properties of environmental chemicals using simple binary molecular fingerprints:
  - Octanol/water partition coefficient (log P)
  - Water solubility (log S)
  - Boiling point (BP) and melting point (MP)
  - Vapor pressure (VP)
  - Bioconcentration factor (BCF)
- This poster presents data on estimation of log P and log S using these methods.

Methods

- The experimentally measured physicochemical properties of a structurally diverse set of 993 environmental chemicals used in this study were obtained from EPI Suite (http://esc.syrres.cm/interkow/EPiSuiteData.htm).
- These organic chemicals cover a wide range of use classes, including industrial compounds, pharmaceuticals, pesticides, and food additives.
- All chemicals were fingerprinted using publicly available SMARTS sets FP3, FP4, PADEL, PubChem, and MACCS from OpenBabel.
- **Figure 1** shows that the experimental values of both log P and log S are normally distributed.
  - Log P spans nearly 13 log units from -4.27 to 8.54 with a median of 2.19.
  - Log S ranges from –9.70 to 1.58 log units and is centered at -2.38.

- **Table 1** lists the summary statistics for log P and log S for the training and test sets.

Figure 1a. Data Distribution of Partition Coefficient (log P)

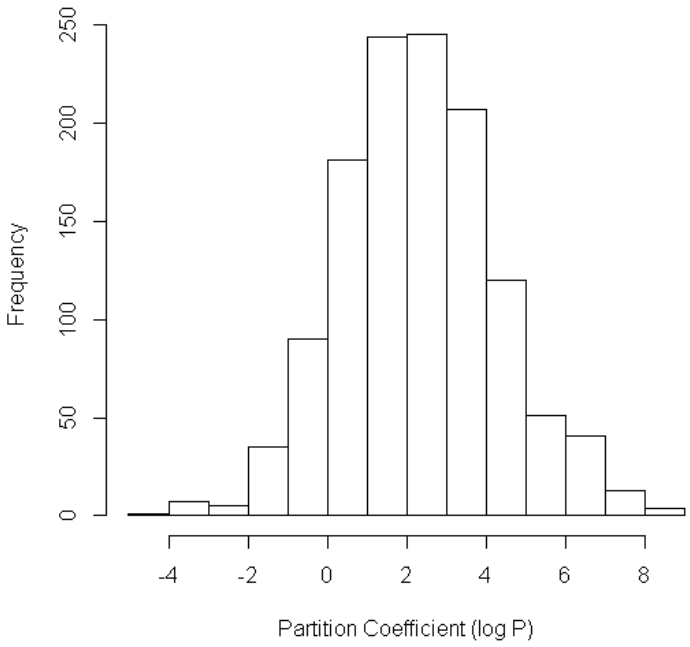
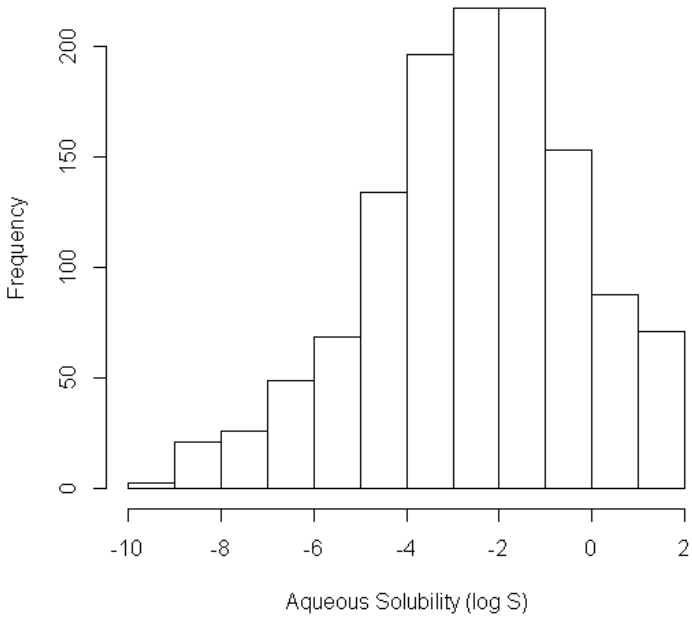


Figure 1b. Data Distribution of Aqueous Solubility (log S)



- Table 1 lists the summary statistics for log P and log S for the training and test sets.

Table 1. Summary Statistics for Training and Test Sets

	Minimum	Maximum	Mean	Median	Standard Deviation
Log P: Training	-4.27	8.54	2.29	2.18	1.98
Log P: Test	-3.89	8.39	2.39	2.29	2.03
Log S: Training	-9.70	1.58	-2.54	-2.18	2.24
Log S: Test	-9.21	1.57	-2.58	-2.39	2.28

- Genetic algorithms (GA) and RF methods were employed to select the most information-rich subset of descriptors for obtaining reliable and robust regression models.
- Mathematical processing for data standardization, multivariate regression analysis, and statistical model building were performed using the statistical software package *R* (version 3.0.2, GNU Public License v3) (R Development Core Team 2008).
- Quantitative structure-property relationship (QSPR) models were developed using four approaches with differing complexity: multiple linear regression (MLR), random forest (RF) regression, partial least squares regression (PLSR), and support vector regression (SVR).

- These were implemented by the packages *subselect*, *randomForest*, *stats*, *pls* and *e1071*, respectively.

- QSPR model performance was evaluated by establishing a correlation between the experimental and calculated values with a set of parameters: R^2 (correlation coefficient) and RMSE (root mean squared error) in log units:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(p_i - \hat{p}_i)^2}{\sum_{i=1}^{n}(p_i - \bar{p})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{p}_i)^2} \quad (2)$$

where p_i and $\hat{p}_i$ are the measured and predicted values for chemical *i*, respectively; and $\bar{p}$ is the mean of all chemicals (*n*) in the data set.

Results

- The property of a chemical calculated from a set of molecular fingerprints can be described by a general equation:

$$\log \text{Property} = \sum_{j=1}^{m} c_j f_j \quad (3)$$

where

- *log Property* is the logarithm of the physicochemical property
- c_j is the contribution coefficient which is determined by regression analysis
- f_j is the binary bit of the *jth* fingerprint, with presence or absence denoted by the numeric value 1 or 0

- The validation results show a significant correlation between the estimated and measured values for the training and test sets (**Figure 2)**.

- For log P, $R^2 = 0.936$, corresponding to a minimum RMSE of 0.492 log units for the test set when using 200 fingerprint bits selected by GA, compared to $R^2 = 0.961$ for the training set (**Figure 2a**).
- For log S, $R^2 = 0.927$ corresponded to a minimum RMSE of 0.588 log units for test set when using 250 fingerprint bits selected by GA, compared to $R^2 = 0.945$ for training set (**Figure 2b).**

Figure 2a. Estimated Values Versus Experimental Values for Training and Test Sets of log P

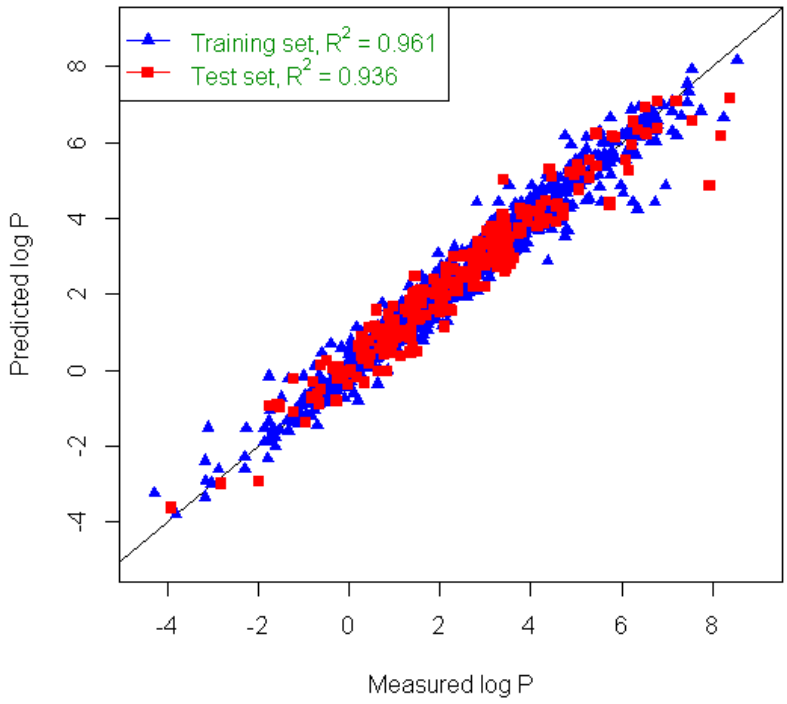
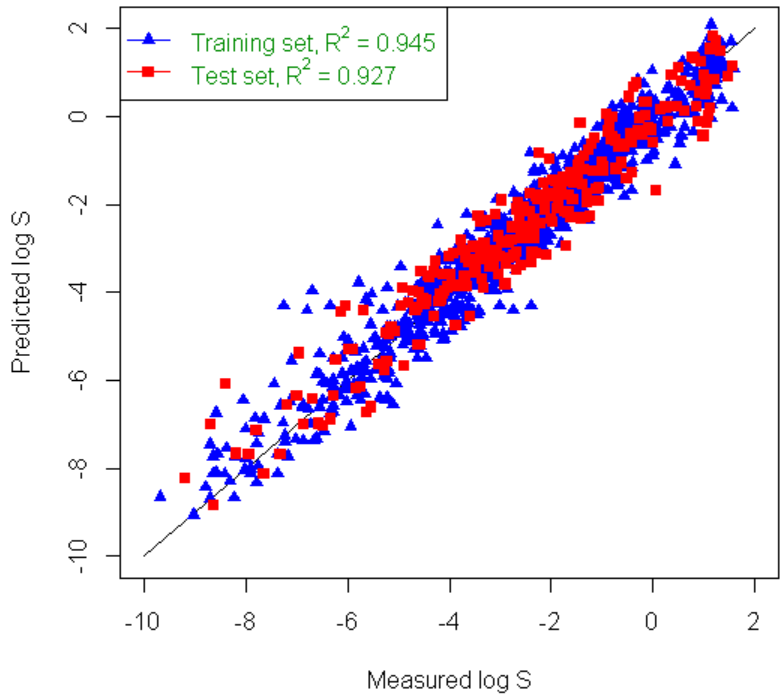


Figure 2b. Estimated Values Versus Experimental Values for Training and Test Sets of log S



- The number of significant principal components (PCs) for the partial least squares algorithm was determined using a 10-fold cross-validation (CV) procedure on the training set. The relation of the standard error of prediction (SEP) versus the number of PCs is displayed in **Figure 3**.
    - The gray lines were produced by repeating this procedure 100 times. The black line represents the lowest SEP value from a single 10-fold CV, while the dashed vertical lines represent the optimal number of PCs.
    - For the all-descriptor model, initially SEP decreases with PCs, and then starts to rebound after a certain point when the model begins to simulate the noise as the complexity of the model increases.

Figure 3. Relationship Between Number of Principal Components and Standard Error of Prediction for Log P Models.

Figure 3a. All Fingerprints

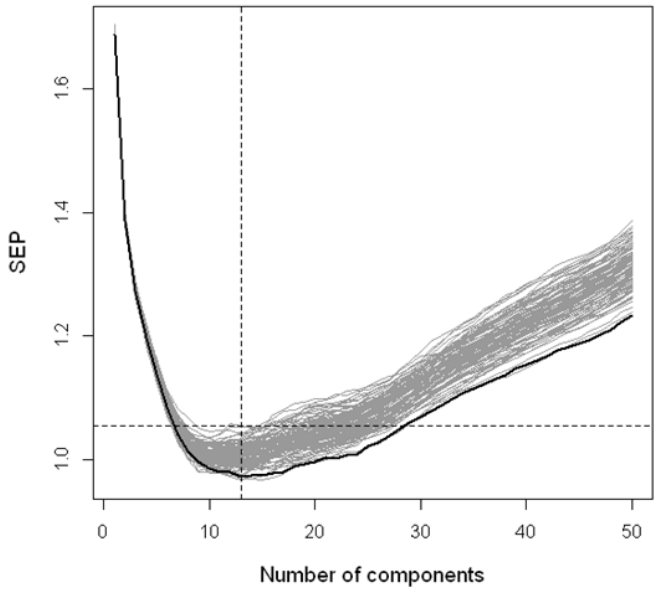
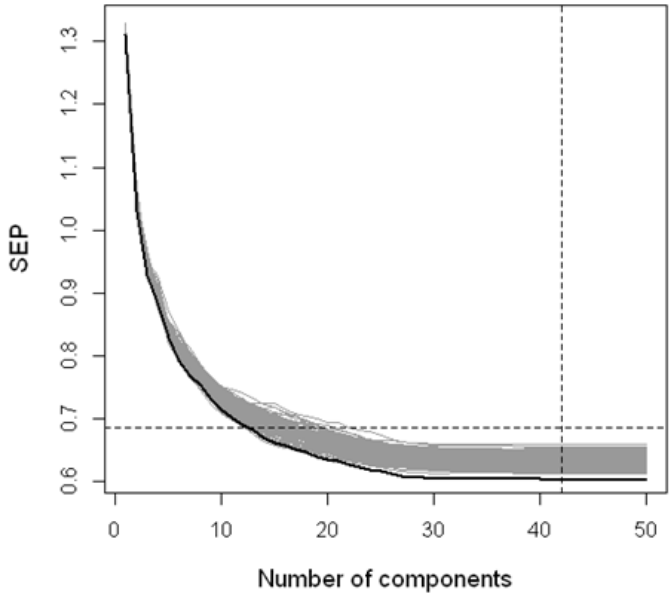


Figure 3b. 250 Fingerprints Selected by GA.



Abbreviation: SEP = standard error of prediction.
Black = single of 10-fold CV; Gray =100 repetitions of the 10-fold CV.

- A significant correlation between log P and log S experimental values was observed (R^2 = 0.761), and molecular weight (MW) is moderately correlated to log S (R^2 = 0.463) (**Figures 4a** and **4b**). These data suggest that log S is more closely related to log P than to MW.

Figure 4a. Aqueous Solubility (log S) Versus Partition Coefficient (log P)

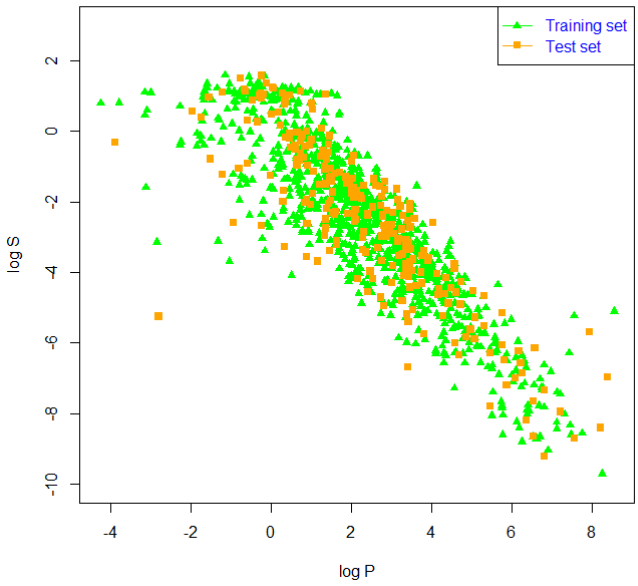
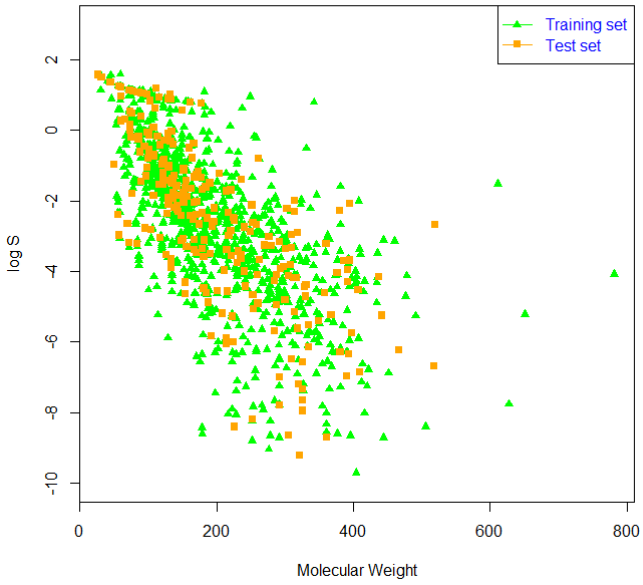


Figure 4b. Aqueous Solubility (log S) Versus Molecular Weights (MW)



- MLR, PLSR, and SVM exhibited satisfactory predictive results with low prediction errors, and all substantially outperformed RF **(Table 2).**

Table 2. Comparison of the Best Models from the Four Methods for the Test Set

	MLR	PLSR	SVM	RF
Log P: R^2	0.915	0.916	0.936	0.835
Log P: RMSE	0.535	0.529	0.492	0.666
Log S: R^2	0.917	0.916	0.927	0.880
Log S: RMSE	0.594	0.599	0.588	0.696

Abbreviations: log P = partition coefficient; log S = aqueous solubility; MLR = multiple linear regression; PLSR = partial least squares regression; R^2 = correlation coefficient; FR = random forest; RMSE = root mean squared error; SVR = support vector regression.

Conclusions

- This study demonstrates that
  - Molecular fingerprints are useful descriptors.

- GA is an efficient feature selection tool from which selected descriptors can effectively model these properties.
- Simple methods such as MLR give results comparable to more complicated methods under optimal conditions.
- There are multiple ways for deriving regression models with similar statistics

Acknowledgements

- The Intramural Research Program of the National Institute of Environmental Health Sciences (NIEHS) supported this poster. Technical support was provided by ILS under NIEHS contracts N01-ES 35504 and HHSN27320140003C.
- The views expressed above do not necessarily represent EPA and NIH policy or the official positions of any Federal agency. Since the poster was written as part of the official duties of the authors, it can be freely copied.
- A summary of NICEATM activities at the Ninth World Congress is available on the National Toxicology Program website at http://ntp.niehs.nih.gov/go/41583.