Date: November 26, 2018

Technical Report

De : Paul Kowalczyk

À: Jean-Yves Delannoy
Alessio Tamburro

Copie :
Ref :

Pages : 16

In Silico Prediction of Physicochemical Properties

Final Report.

Quantitative structure-property relationship (QSPR) models were developed for the prediction of six physicochemical properties of environmental chemicals: octanol–water partition coefficient (log P), water solubility (log S), boiling point (BP), melting point (MP), vapor pressure (VP) and bioconcentration factor (BCF). Models were developed using simple binary molecular fingerprints and four approaches with differing complexity: multiple linear regression, random forest regression, partial least squares regression, and support vector regression (SVR). To obtain reliable and robust regression models with high prediction performance, genetic algorithms (GA) were employed to select the most information-rich subset of fingerprint bits. Predictions from the various models were tested against a validation set, and all four approaches exhibited satisfactory predictive results, with SVR outperforming the others. BP was the best-predicted property, with a correlation coefficient (R^2) of 0.95 between the estimated values and experimental data on the validation set while MP was the most poorly predicted property with an (R^2) of 0.84. The statistics for other properties were intermediate between MP and BP with (R^2) equal to 0.94, 0.93, 0.92 and 0.86 for log S, log P, VP and BCF, respectively. The prediction results for all properties were superior to those from Estimation Program Interface (EPI) Suite (R^2 values ranged from 0.63 to 0.94), a widely used tool for property prediction. This study demonstrates that (1) molecular fingerprints are useful descriptors, (2) GA is an efficient feature selection tool from which selected descriptors can effectively model these properties, and (3) simple methods give comparable results to more complicated methods.

It's hoped that in describing how one might build predictive models for chemical datasets, colleagues will be encouraged to discuss how these same workflows may be applied to Solvay project data.

Paul KOWALCZYK Jean-Yves DELANNOY

Contents

Chapter 1

Introduction

- Current tools for testing the biological activity and toxicity of chemicals are time-consuming and costly. Thus, only a fraction of these chemicals have been fully characterized for their potential hazard and risks to both human health and the environment.
- In vitro and in silico approaches are being developed as more efficient tools for chemical hazard characterization and prioritization. One of these approaches is in silico estimation of physicochemical properties.
- This study presents novel methods using simple binary molecular fingerprints for the estimation of six physicochemical properties of environmental chemicals:
  - Octanol–water partition coefficient (log P)
  - Water solubility (log S)
  - Boiling point (BP)
  - Melting point (MP)
  - Vapor pressure (VP)
  - Bioconcentration factor (BCF)
- The goal of this project is to produce models that can be easily used by Solvay colleagues and that adhere to internationally accepted validation principles defined by the Organisation for Economic Co-operation and Development (OECD 2004).

Chapter 2

Characteristics of the Chemical Sets

- Experimentally measured physicochemical properties of a structurally diverse set of organic environmental chemicals were obtained from EPI Suite (EPA 2012 and EPI Suite Data). These chemicals represent a wide range of use classes, including industrial compounds, pharmaceuticals, pesticides, and food additives.

- Figure 1 shows that values for the physicochemical properties of the chemical set are normally or nearly normally distributed.

  - Log P (Figure 1a) ranges from -4.27 to 8.54 log units with a median of 2.19.
  - Log S (Figure 1b) ranges from -9.70 to 1.58 log units (mol/L) with a median of -2.38.
  - BP (Figure 1c) ranges from -88.60 to 548.00 C with a median of 189.20 C.
  - MP (Figure 1d) ranges from -199.00 to 385.00 C with a median of 85.00 C.
  - VP (Figure 1e) ranges from -13.68 to 5.89 log units (mmHg) with a median of -2.11.
  - BCF (Figure 1f) ranges from -0.35 to 5.97 log units with a median of 1.73.

Chapter 3

Definition of Training and Test Sets

The chemicals were randomly partitioned into training sets (80% of the chemicals) to build the models and test sets (20% of the chemicals) to validate the predictive power of each model. Table 1 lists the summary statistics for physicochemical properties of the training and test sets.

Chapter 4

Development of QSPR Models

- Molecular fingerprints, a series of binary bits that represent the presence (1) or absence (0) of particular substructures in a molecule, were used as independent variables.
- Genetic algorithm (GA; Wegner et al. 2003) was employed to select the most information-rich subset of variables for obtaining reliable and robust regression models.
- Quantitative structure–property relationship (QSPR) models were developed using four approaches with differing complexity in ascending order: multiple linear regression (MLR), partial least squares regression (PLSR), random forest regression (RFR), and support vector regression (SVR).
- Mathematical processing for data standardization, multivariate regression analysis, and statistical model building were performed using the statistical software package R (version 3.0.2)(R Development Core Team 2008). GA, MLR, RFR, PLSR and SVR were implemented by the packages subselect, stats, randomForest, pls and e1071, respectively.
- The performance of each QSPR model is evaluated by establishing a correlation between the experimental and calculated values with a set of parameters:
  - R^2 and RMSE are the coefficient of determination and root mean squared error for training or test set with n chemicals.
  - Q^2 and RMSEcv are the coefficient of determination and root mean squared error for 10-fold cross validation (CV) with v chemicals not included in the CV

Molecular Fingerprints. The chemicals were represented by fingerprints derived from their molecular structures. Fingerprints were calculated using a wide variety of publicly available SMARTS systems implemented in PaDEL:51,52 Estate (79 bits), Extended (1024 bits), Substructure (307 bits), Klekota Roth (4860 bits), PubChem (881 bits), Atom Pairs 2D (780 bits), and MACCS (166 bits). A total of 8097 binary bits were generated, with 1 and 0 denoting the presence or absence, respectively, of a specific structural fragment. Fingerprint bits with zero variance (i.e., uniform observations across the set) were removed. To obtain reliable models, sufficient occurrences of the fingerprint bits throughout the entire data sets are necessary and thus, bits with low occurrences (<2eliminated. Following removal of highly correlated and infrequently occurring bits, the resulting numbers of bits retained and employed to build the regression models were: 1681 for logP; 1061 for logS; 450 for logBCF; 1050 for BP; 1424 for MP; and 1145 for logVP. A genetic algorithm (GA)53,54 was used to reduce the feature space by assigning an initial population of chromosomes to two times the number of variables (fingerprint bits). The crossover probability on each chromosome in a population and mutation rate on each gene in a chromosome were set to 50% and 1%, respectively. There were no improvements in the fitness score after 1000 generations.

Multiple Linear Regression. Multiple linear regression (MLR) is widely used in the modeling of property data.40,55 We used MLR to produce a linear model to describe the relationship between a physicochemical property and the molecular fingerprint bits:

$$property = \sum_{j=1}^{m} c_j f_j \tag{4.1}$$

In eq 4.1, property is one of the six physicochemical properties (logP, logS, logBCF, BP, MP or logVP); cj is the contribution coefficient, which is determined by regression analysis; and f j is the binary bit of the jth fingerprint, with its presence or absence represented by the numeric values 1 or 0 respectively. Any fragment that occurred in a molecule was counted only once for that molecule, no matter how many times it occurred in the molecule.

Partial Least Squares Regression. Partial least-squares regression (PLSR) is a widely used multivariate analytical technique in QSPR studies.56,57 The advantage of PLSR over MLR lies in its ability to build a regression model based on highly correlated descriptors, extract the relevant information, and reduce data dimensions. We employed PLSR to generate linear statistical models based on the fingerprint bits and the physicochemical property being predicted. A set of orthogonal latent variables or principal components (PCs) were first generated through a linear combination of the original molecular fingerprint bits, which served as new variables for regression with the response variables (i.e., the physicochemical properties) to build QSPR models. The optimal number of PCs was determined by 10-fold cross-validation (CV).

Random Forest Regression. Random forest (RF) is a nonlinear consensus method based upon an ensemble of decision trees which are grown from separate bootstrap samples of the training data.58 Bootstrap sampling is conducted via random selection with replacement from the training chemicals during tree growth. The chemicals that are not selected in the construction of the forest are called out-of-bag (OOB) samples, which are used to evaluate the prediction accuracy as trees are added to the forest. Each tree gives a prediction for its OOB chemicals, and the average of these results over all trees provides an overall unbiased external validation. There are three possible model parameters for RF regression: ntree - the number of trees in the forest; mtry - the number of variables randomly sampled at each tree node; and nodesize - the minimum node size below which nodes are not further subdivided. In the present study, the RF model was trained based upon a parameter combination of ntree = 500, nodesize = 5, and mtry = 1/3 the number of fingerprint bits.

Model Validation. The performance of each QSPR model was evaluated by examining the correlation between the experimental and predicted values using the following parameters:61,62 R2 (coefficient of determination) and RMSE (root mean squared error) for training or test sets with n chemicals; Q2 (coefficient of determination) and RMSEcv for 10-fold CV with v chemicals not included in the CV model building set. The 10-fold CV procedure was completed using only the training set.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(p_i - \hat{p}_i)^2}{\sum_{i=1}^{n}(p_i - \bar{p})^2} \tag{4.2}$$

$$Q^2 = 1 - \frac{\sum_{i=1}^{\nu}(p_i - \hat{p}_i)^2}{\sum_{i=1}^{\nu}(p_i - \bar{p})^2} \tag{4.3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{p}_i)^2} \tag{4.4}$$

$$RMSE_{cv} = \sqrt{\frac{1}{\nu}\sum_{i=1}^{\nu}(p_i - \hat{p}_i)^2} \tag{4.5}$$

In eqs 4.2 – 4.5, p_i and $\hat{p}_i$ are the measured and predicted property values for chemical i, respectively, and $\hat{p}$ is the mean of all chemicals in the data set. In addition, standard error of prediction (SEP) was employed as a criterion to select the optimal principal components in the PLSR analysis.

$$SEP = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(p_i - \hat{p} - bias)} \tag{4.6}$$

$$bias = \frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{p}) \tag{4.7}$$

Applicability Domain. Three distance-based measures, i.e., leverage, distance from centroid and k-nearest neighbors (kNN), were applied to assess the applicability domain (AD) of each regression model. The distance of a test chemical from a defined point in the descriptor space of the training set was calculated and compared to a predefined threshold. The test chemical is considered to be within AD if its distance is less than or equal to the threshold. Leverage is defined as the diagonal element of the covariance matrix for a given data set, and the leverage of a test chemical is proportional to Hotellings T2 statistic and its Mahalanobis distance. The threshold was set to three times the average of the leverage ($3\frac{m}{n}$, with m being the number of variables and n the number of training chemicals). For the measure of distance from centroid, the distance of a test chemical from the training set centroid is compared with a threshold, which is determined as follows: (1) calculate the distances of training chemicals from their centroid; (2) sort the vector of distances in ascending order; (3) set the distance value corresponding to 95th percentile as the threshold. The kNN measure defines the model's AD based on the similarity between a test chemical and the training chemicals. The average distance of the test chemical from its five nearest neighbors in the training set is compared with a threshold, which is the 95th percentile of average distance of training chemicals from their five nearest neighbors.

Statistical Analysis. Mathematical processing for data standardization, multivariate regression analysis, and statistical model building were performed using the R statistical computing environment for Windows (version 3.2.1).66 Genetic algorithm, multiple linear regression, partial leastsquares regression, random forest regression, support vector regression and distance of k-nearest neighbors were implemented by the R packages subselect, stats, pls, randomForest, e1071, and FNN, respectively. The R code for feature selection and regression analysis is provided in the Supporting Information.

Pearson correlation coefficients (r)

$$r = \frac{n\sum p_k p_l - \sum p_k \sum p_l}{\sqrt{n\sum p_k^2 - (\sum p_k)^2}\sqrt{n\sum p_l^2 - (\sum p_l)^2}} \tag{4.8}$$

Chapter 5

Correlation Between Estimated and Measured Values

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

$$property = \sum_{j=1}^{m} c_j f_j \tag{5.1}$$

In equation (5.1):

- textitproperty is the value 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 quality of the model depends heavily on the number of selected fingerprint bits, and the predictive performance of the model is enhanced remarkably when an appropriate number of fingerprint bits were selected from GA (Figure 2). Results show that the prediction for the training set is improved continuously with increasing feature number. In contrast, the test set followed a different pattern, *i.e.*, the RMSE value initially decreased, attained a minimum at a medium number of bits, and then gradually increased afterwards.

- For log P, the modeling statistics are not sensitive to the bit number, and the model performance does not vary considerably with different subsets of fingerprint bits for the test set (Figure 2a).
- For log S, the lowest prediction errors occurred on the models with moderate complexity around 250 and 300 bits

The validation results show a significant correlation between the estimated and measured values in the test set.

- For log P, R^2 of 0.925 corresponded to a minimum RMSE of 0.516 log units for test set when using 600 fingerprint bits selected by GA, compared to R2 of 0.980 for training set (Figure 3a).
- For log S, R^2 of 0.935 corresponded to a minimum RMSE of 0.559 log units for test set when using 250 fingerprint bits selected by GA, compared to R2 of 0.955 for training set (Figure 3b).

Chapter 6

Relationship Between Number of Principal Components and Standard Error of Prediction

- The number of significant principal components (PCs) for the PLS algorithm was determined using 10-fold cross-validation (CV) procedure on the training set (Zang et al. 2011). The relation of the standard error of prediction (SEP) versus the number of PCs is displayed in Figure 4.
- 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. The dashed vertical lines represent the optimal number of PCs and the dashed horizontal lines indicate the SEP value for the test set when the optimal PCs are applied.
- 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 4a). For the 600-bit model, the SEP decreases monotonically and gradually approaches a stable value, and the model with 42 PCs gave a minimum RMSE (Figure 4b).

Chapter 7

Applicability Domain

- An applicability domain (AD) is a chemical, structural, or physicochemical space of the training set.
- The AD of the models was assessed using a leverage-based approach that compares a predefined threshold to the distance of query compounds from a defined point within the descriptor space. The approach is based on the covariance matrix derived from center-scaled variables. The threshold is three times the average of the leverage that corresponds to m/n, the ratio of m, the number of model variables, to n, the number of training compounds.
- Figure 5 displays the relationship between leverage and standardized residuals (William plot [Sahigara et al. 2012]).
  - For log P, 39 out of 2998 (1.30%) test chemicals are located outside the AD (Figure 5a).
  - For log S, 18 out of 457 (3.94%) test chemicals are located outside the AD (Figure 5b).

Chapter 8

Comparison of the Models

SVR substantially outperformed the other three approaches in predicting log P, log BCF, BP and MP with a low error rate (Table 3). However, performance of SVR was similar to the other three approaches for predicting log S and log VP.

Chapter 9

Conclusions

This study demonstrates that:

- Molecular fingerprints are useful descriptors for modeling the six properties.
- GA is an efficient feature selection tool from which selected descriptors can effectively model these properties.
- Simple methods such as MLR give similar results to more complicated methods under optimal conditions for modeling log S and log VP.
- There are multiple ways for deriving regression models with similar statistics.
- When compared to other procedures currently in use, these methods present better accuracy for a wider range of chemicals of interest, are highly stable and reliable, and are in line with the validation principles put forth by the OECD. They thus have broad applicability for property estimation of many classes of compounds.

FICHE RESUME/BIBLIOGRAPHICAL FORM DOCUMENTUM

SERVICE	Laboratory Group	SM@RT
Type de document	Document type	Rapport de Projet
Date	Application Date	Decembre 2014
TITRE en Anglais	English Title	COMPNANOCOMP FP7 Project : Final Report
TITRE en Français	French Title	Projet FP7 COMPNANOCOMP : Rapport Final
Entreprise	Enterprises/Sponsors	R&I/AIO/Advanced Materials Platform
Auteur(s)	Authors	Jean Yves DELANNOY
PROJET	Projects	COMPNANOCOMP
Collaborateurs	Collaborators	Cédric Feral-Martin Aurélie Papon Magali Fontana Olivier Sanseau
N^0 d'affaire	Business code	
RESUME Anglais		The objective of this document is to offer a "digest" of the main results obtained in this project that aims at the development of multiscale simulation methodology and software for predicting the morphology (spatial distribution and state of aggregation of nanoparticles), thermal (glass temperature), mechanical (viscoelastic storage and loss moduli, plasticity, fracture toughness and compression strength), electrical and optical properties of soft and hard polymer matrix nanocomposites from the atomic-level characteristics of their constituent nanoparticles and macromolecules and from the processing conditions used in their preparation. The document gives a short summary of the results obtained within this project by Solvay and its partners. It also emphasizes the interest for Solvay R&I of the developments obtained.

RESUME Français		L'objectif de ce rapport est de fournir un résumé des points essentiels développés dans le rapport final du projet européen FP7 COMPNANOCOMP dont le but est de développer une méthodologie de simulation multi-échelle permettant de prédire la morphologie (distribution spatiale et état d'agrégation des nanoparticules), et les propriétés thermiques, mécaniques optiques et électriques de nanocomposites de polymères. Ce travail s'effectue sur la base des caractéristiques atomiques des constituants du matériaux et doit prendre en compte les conditions du procédé ayant permis sa réalisation. Ce document donne une vision des résultats obtenus par Solvay et ses collaborateurs et met en avant l'intérêt pour le groupe des développements effectués.
Mots Clés Anglais	English Keywords	RUBBER, REINFORCEMENT, SILICA, COUPLING, MODELING, MORPHOLOGY, FP7
Mots Clés Français		CAOUTCHOUC ; RENFORT ; SILICE ; COUPLAGE ; MODELISATION ; MORPHOLOGIE, FP7
RNCAS	RNCAS	
Destinataires	Addressees	
CONFIDENTIEL	Confidential	YES