Date: December 6, 2018

Technical Report

De : Paul Kowalczyk

À: Jean-Yves Delannoy
Alessio Tamburro

Copie :
Ref :

Pages : 19

In Silico Prediction of Physicochemical Properties

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). All data were retrieved from the publicly available PHYSPROP database. Models were developed using features calculated using the Chemistry Development Kit (CDK) and five machine learning approaches with differing complexity: multiple linear regression (mlr),k-nearest neighbors kNN), and generalized boosted regression models (gbm). Predictions from the various models were tested against a validation set, and all five approaches exhibited satisfactory predictive results, with gbm 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 BCF was the most poorly predicted property with an (R^2) of 0.80. The statistics for other properties were intermediate between BCF and BP with (R^2) equal to 0.92, 0.89, 0.81 and 0.93 for log P, log S, MP, and VP, respectively.

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

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. Consequently, reliable predictions for both physicochemical properties and environmental fate endpoints are needed for risk assessment as well as prioritization for testing. One approach employed is the *in silico* estimation of physicochemical properties.

The most widely used chemical properties in toxicological studies, risk assessment, and exposure studies are associated with bioavailability, permeability, absorption, transport, and persistence of chemicals in the body and in the environment. Measured properties associated with these endpoints include, but are not limited to, the octanol–water partition coefcient, water solubility, melting point, bioconcentration factor, and biodegradability. This study aims to develop robust quantitative structure-property relationships (QSPR) for chemical properties of environmental interest. Specifically, this study presents methods using calculated 1-D and 2-D molecular features 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 QSAR concept is based on the congenericity principle, which hypothesizes that similar structures have similar properties and exhibit similar biological activities. Key to any QSPR study is the calculation of a set of features for each molecule which, in turn, are used to measure the (dis)similarity between molecules.

The Organization for Economic Cooperation and Development (OECD) lists five principles for building robust QSPR models. These principles are:

- a defined endpoint
- an unambiguous algorithm
- a defined applicability domain
- appropriate measures for goodness-of-fit, robustness, and predictivity, and
- a mechanistic interpretation (if possible)

In this study all data was retrieved from the publicly available PHYSPROP database, which has been curated for model creation. For each of the targets modeled each molecule has a single, definite endpoint. The machine learning methds used are open-source and are made available to colleagues, ensuring experimental reproduceability. Principal component analysis was used to test the applicability domains for each target. Five-fold cross-validation and external test set techniques were used to test goodness-of-fit, robustness, and predictivity. Variable (feature) importance calculations are used to address mechanistic interpretations.

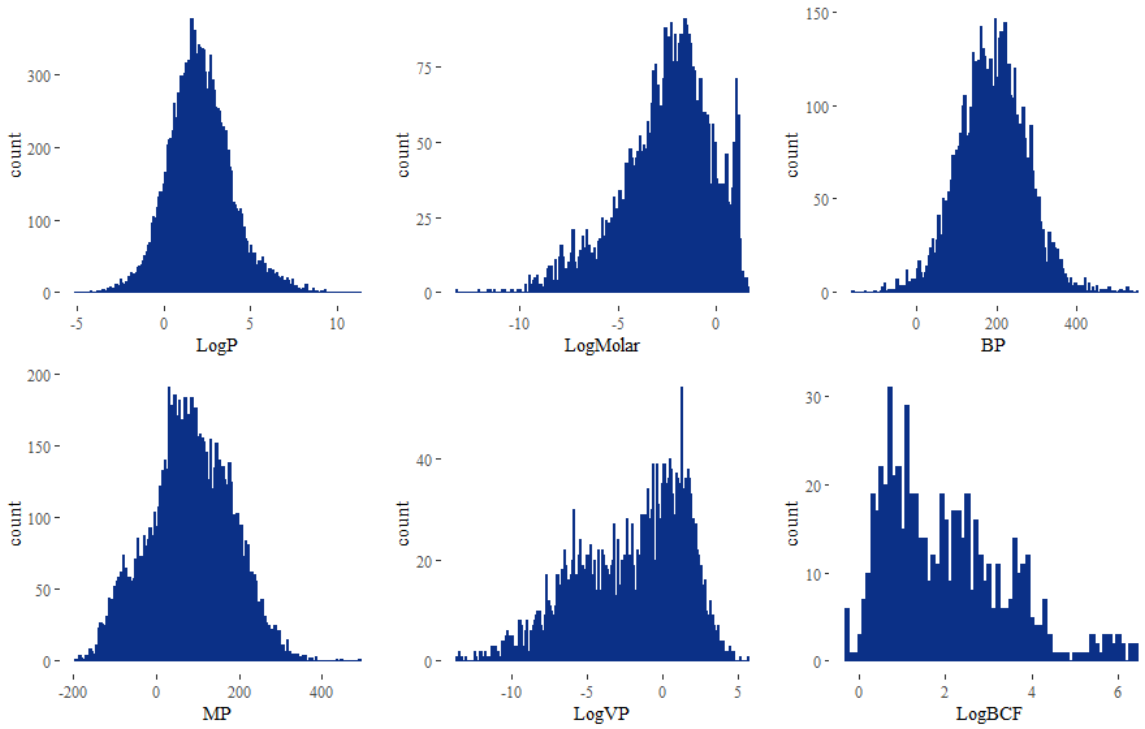
2 Materials & Methods

2.1 Datasets

Experimentally measured physicochemical properties of a structurally diverse set of organic environmental chemicals were obtained from PHYSPROP, a database containing chemical structures, names and physical properties for over 41,000 chemicals (http://www.srcinc.com/what-we-do/environmental/scientific-databases.html#physprop). 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 sets are normally, or nearly normally, distributed.

Figure 1: Distributions of values for each of the physicochemical properties.



Summaries of the values for each physicochemical property are presented in Table 1.

property	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Log P	14026	-5.08	0.89	2.00	2.08	3.16	11.29
Log S	4202	-13.172	-3.877	-2.284	-2.573	-0.988	1.581
BP	5415	-161.5	133.0	189.3	188.9	245.0	548.0
MP	8625	-196.00	16.00	80.00	80.45	151.20	492.50
VP	2701	-13.6778	-4.7696	-1.2573	-2.0395	0.8633	5.6682
BCF	626	-0.350	0.850	1.780	2.002	2.857	6.430

Table 1: Summaries of physicochemical properties. **N**: number of observations; **Min.**: minimum value; **1st Qu.**: first quartile (25th percentile); **Median**: median (50th percentile); **Mean**: mean (average); **3rd Qu.**: third quartile (75th percentile); **Max.**: maximum value.

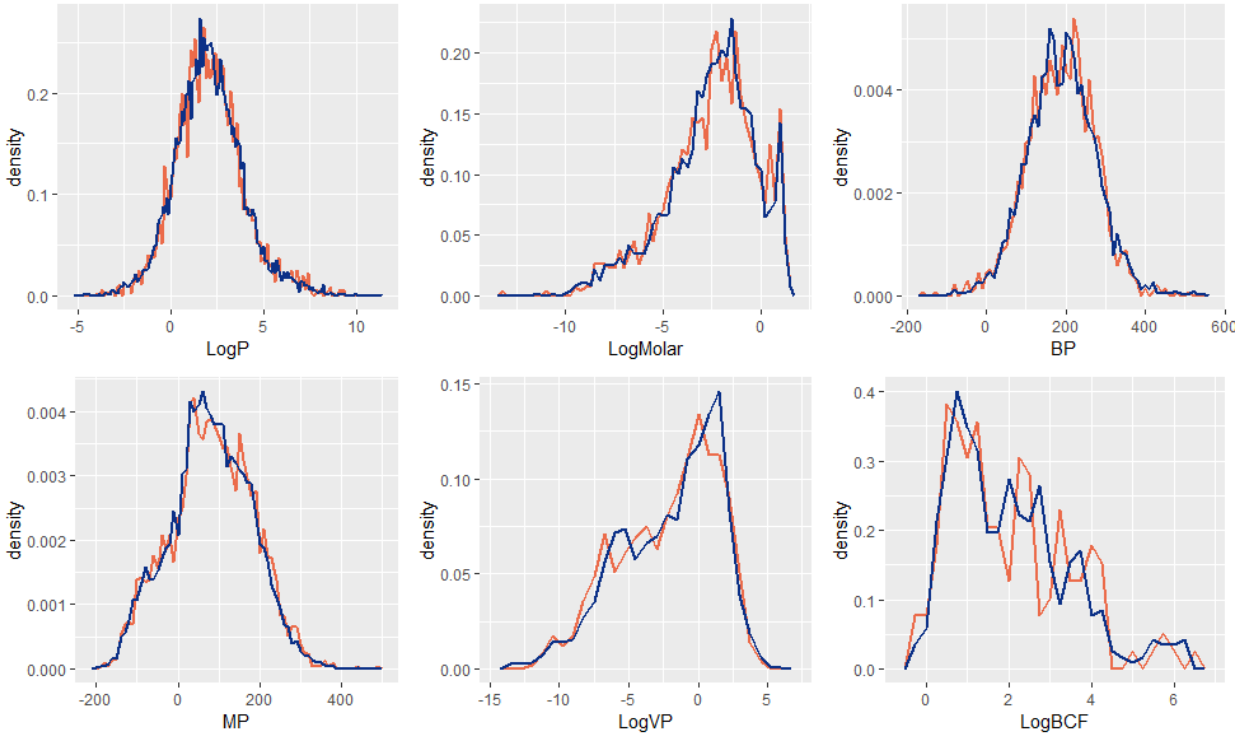
2.2 Selection of Training Sets & Test Sets

The chemicals were randomly partitioned into training sets (75% of the chemicals) to build the models and test sets (25% of the chemicals) to validate the predictive power of each model. The property value distributions for these training sets and test sets are presented in Figure 2. That these distributions are coincident supports the assertion that training sets and test sets represent equivalent sample populations from the respective full datasets. Numerical summaries for property values in the training sets and test sets are presented in Table 2. The complementarity of these data summaries further speaks to the equivalence of the training sets and the test sets.

property	set	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Log P	train	10520	-5.080	0.890	2.000	2.075	3.160	11.290
	test	3506	-3.950	0.910	2.000	2.096	3.170	9.300
Log S	train	3149	-12.060	-3.825	-2.284	-2.580	-1.007	1.581
	test	1063	-13.172	-3.929	-2.284	-2.553	-0.945	1.541
BP	train	4062	-103.7	134.0	189.0	189.3	245.0	548.0
	test	1353	-161.5	131.0	190.5	187.5	244.5	512.0
MP	train	6463	-196.00	15.95	79.00	79.61	150.00	437.30
	test	2162	-187.60	16.62	83.00	82.97	156.88	492.50
VP	train	2024	-13.678	-4.770	-1.222	-2.005	0.919	5.668
	test	677	-11.796	-4.737	-1.396	-2.142	0.732	4.717
BCF	train	469	-0.300	0.860	1.800	2.006	2.820	6.360
	test	157	-0.350	0.780	1.700	1.990	2.960	6.430

Table 2: Summaries of physicochemical properties for training and test sets. **N**: number of observations; **Min.**: minimum value; **1st Qu.**: first quartile (25th percentile); **Median**: median (50th percentile); **Mean**: mean (average); **3rd Qu.**: third quartile (75th percentile); **Max.**: maximum value.

Figure 2: Distributions of values for training sets and test sets.



2.3 Descriptor Calculation

A key requirement for the predictive modeling of molecular properties and activities are molecular descriptors - numerical charaterizations of the molecular structure. The CDK implements a variety of molecular descriptors, categorized into topological, constitutional, geometric, electronic and hybrid (`https://cdk.github.io/cdk/1.5/docs/api/org/openscience/cdk/qsar/descriptors/molecular/package-summary.html`). In total, 115 1-dimensional and 2-dimensional descriptors are calculated for each molecule.

For each of the datasets, any descriptor having one unique value (*i.e.*, zero variance descriptors) is removed. These descriptors have no information, and are discarded without consequence.

Further, highly correlated descriptors are removed. Redundant descriptors often add more complexity to a model than information they provide to the model. Using highly correlated descriptors – in techniques like linear regression – can result in highly unstable models, numerical errors, and degraded predictive performance. In these studies we've chosen a cutoff = 0.85; a minimum nimber of descriptors is removed to ensure that the absolute value of all pairwise correlations is below 0.85.

Table 3 reports the number of zero variance and highly correlated descriptors removed from each dataset, prior to modeling.

property	# Zero Variance	# Highly Correlated	# Descriptors Remaining
Log P	9	32	74
Log S	9	34	72
BP	10	39	66
MP	4	34	77
VP	11	37	67
BCF	11	34	70

Table 3: Summary of the number of zero variance and highly correlated descriptors removed from each dataset, prior to modeling

2.4 Model Development

Each dataset was modeled using five machine learning algorithms: multiple linear regression (MLR), partial least squares regression (PLS), support vector machines (SVM), k-nearest neighbors (k-NN), and gradient boosted machines (gbm). Each of these algorithms is briefly described in the Supplement to this report.

Model Validation. The performance of each QSPR model was evaluated by examining the correlation between the experimental and predicted values using the following parameters: 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.

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

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{p}_i)^2} \tag{3}$$

$$RMSE_{cv} = \sqrt{\frac{1}{\nu}\sum_{i=1}^{\nu}(p_i - \hat{p}_i)^2} \tag{4}$$

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.

$$SEP = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(p_i - \hat{p} - bias)} \tag{5}$$

$$bias = \frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{p}) \tag{6}$$

2.5 Model Validation

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).

2.6 Applicability Domain

Applicability Domain. The applicability domain of a QSPR model is the use case or query molecule space a model can reliably predict. A diverse set of approaches have been developed to determine if a query molecule is in the AD of a model. Generally, AD approaches can be grouped into two categories: 1) assessing reliability based on a defined chemical descriptor space overlap between the training set and query molecule (i.e. outlier detection) and 2) estimating model prediction uncertainty of a query molecule. Outlier detection, when comparing multi-variate distributions, becomes problematic when nonlinear relationships exists between descriptors and when assessing prediction reliability at distribution boundaries. Rather, we use an estimate of uncertainty approach, specific to random forest, which does not need a predefined descriptor space overlap and takes into account nonlinear relationships between variables. Briefly, the prediction is more reliable and considered in the AD of a model if the variance of the predicted property among the trees of the global model is less than the variance of the cross-validated models.

3 Results & Discussion

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.

property	mlr	pls	svm	kNN	gbm
Log P	0.8125	0.8098	0.9052	0.8546	0.9234
Log S	0.8317	0.8315	0.8770	0.8113	0.8924
BP	0.8978	0.8875	0.9487	0.8829	0.9535
MP	0.6849	0.6656	0.7841	0.7253	0.8106
VP	0.8834	0.8823	0.9385	0.8662	0.9327
BCF	0.0668	0.0398	0.7690	0.6916	0.8085

Table 4: Adjusted R^2.

4 Conclusions

5 Supplemental Material

5.1 Machine Learning Algorithms

5.1.a Multiple Linear Regression.

Multiple linear regression (MLR) is widely used in the modeling of property data. We used MLR to produce a linear model to describe the relationship between a physicochemical property and the calculated molecular descriptors:

$$property = \sum_{j=1}^{m} c_j f_j \tag{7}$$

In eq 7, *property* is one of the six physicochemical properties (logP, logS, logBCF, BP, MP or logVP); c_j is the contribution coefficient, which is determined by regression analysis; and f_j is the value of the jth descriptor.

5.1.b Partial Least Squares Regression.

Partial least-squares regression (PLSR) is a widely used multivariate analytical technique in QSPR studies. 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 calculated descriptors 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 descriptors, 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).

5.1.c Support Vector Machines

The basic concept of support vector regression is mapping the original data **X** nonlinearly into a higher dimensional feature space and solve a linear regression problem in this feature space.

5.1.d k-Nearest Neighbors

The k-nearest neighbor algorithm (k-NN) is a method to classify objects based on closest examples in the feature space. k-NN uses feature similarity to predict the value of any new data point. The new data point is assigned a vlue based on how closely that data point resembles points in the training set.

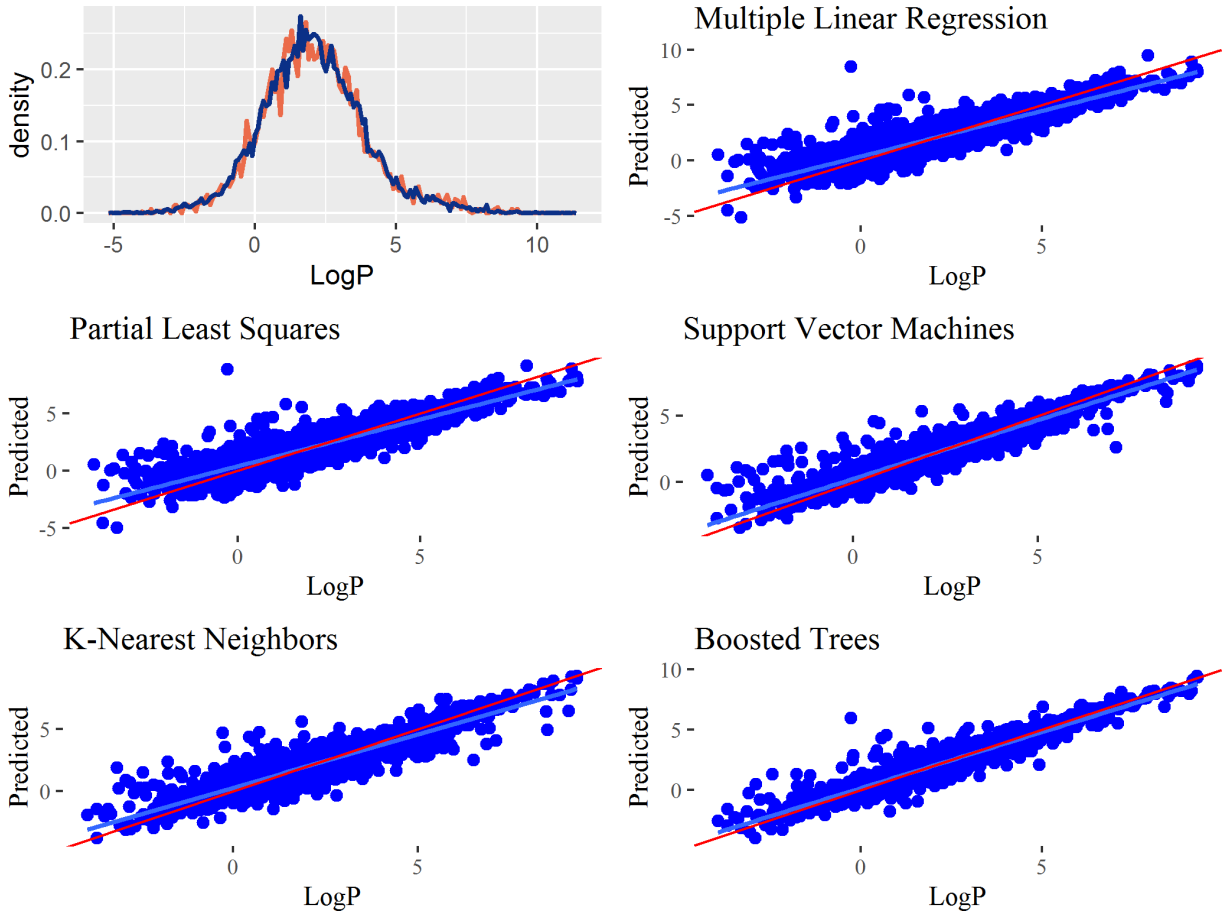
5.1.e Gradient Boosted Machines

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

5.2 Physicochemical Property Modeling Summaries

5.2.a Log P

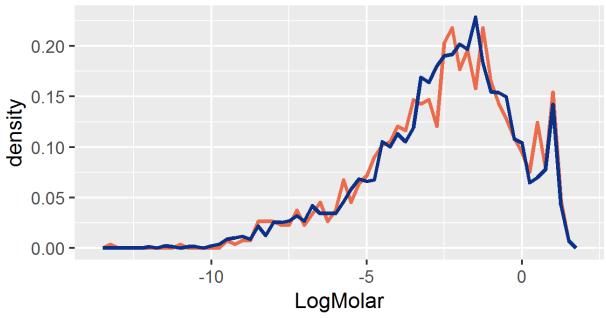
property	set	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Log P	train	10520	-5.080	0.890	2.000	2.075	3.160	11.290
	test	3506	-3.950	0.910	2.000	2.096	3.170	9.300

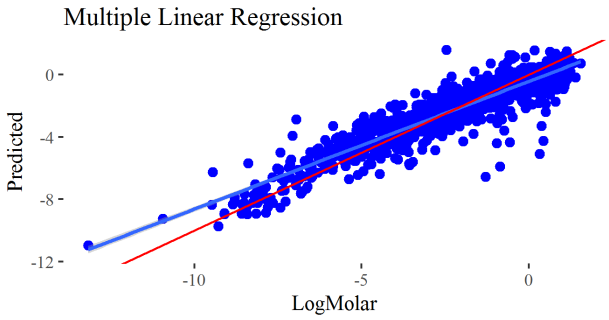


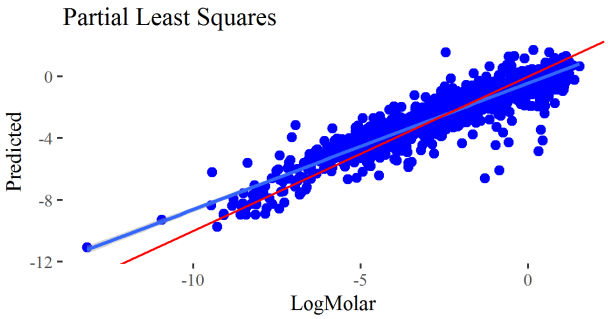
property	mlr	pls	svm	kNN	gbm
Log P	0.8125	0.8098	0.9052	0.8546	0.9234

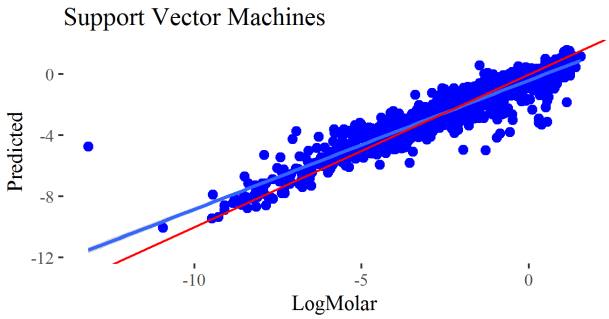
5.2.b Log S

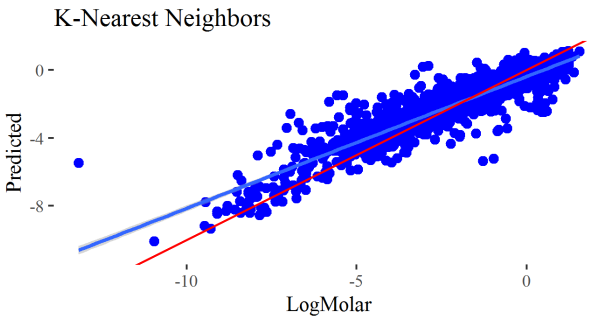
property	set	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Log S	train	3149	-12.060	-3.825	-2.284	-2.580	-1.007	1.581
	test	1063	-13.172	-3.929	-2.284	-2.553	-0.945	1.541

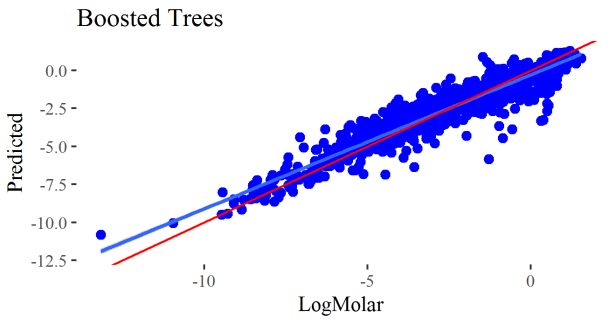








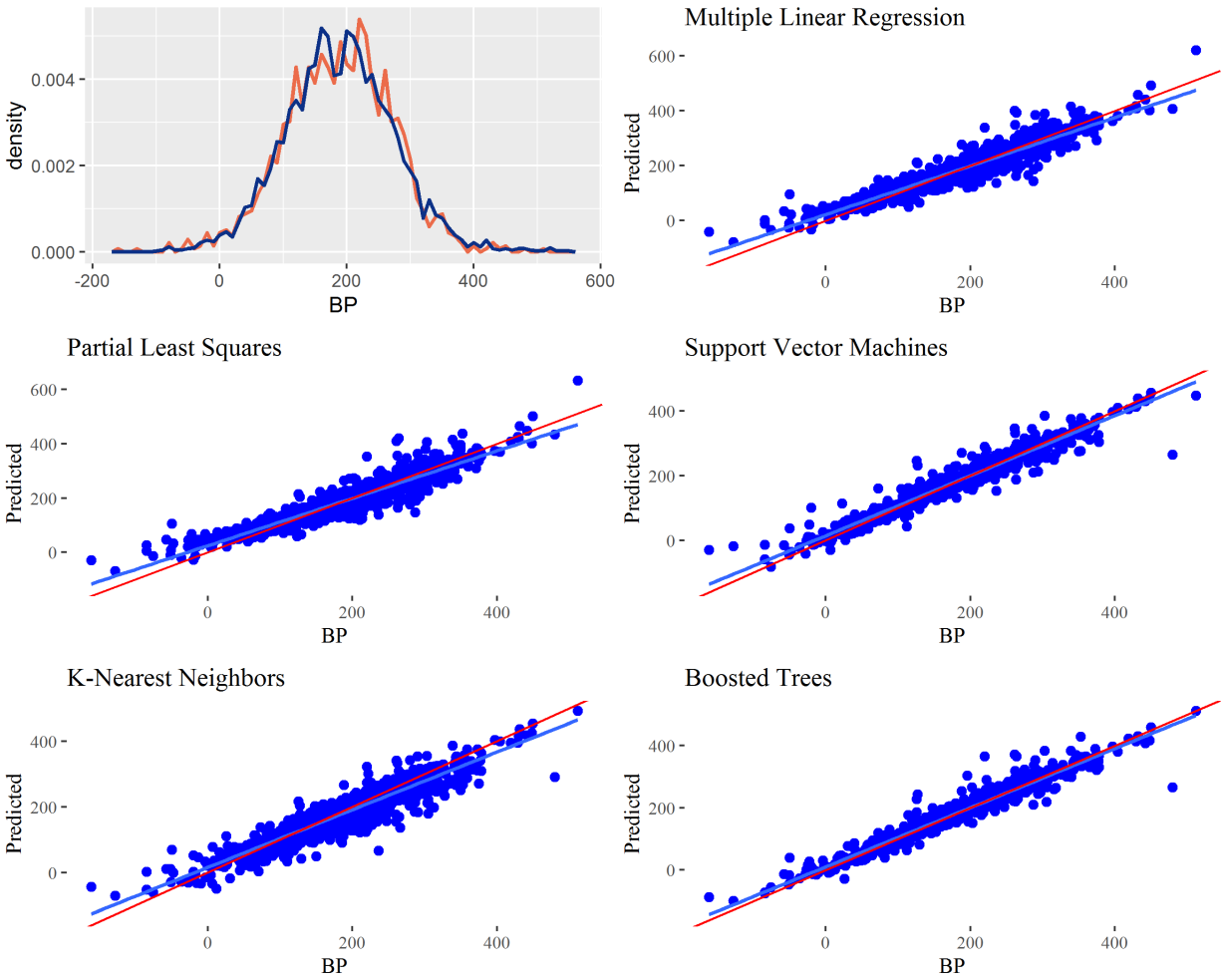




property	mlr	pls	svm	kNN	gbm
Log S	0.8317	0.8315	0.8770	0.8113	0.8924

5.2.c Boiling Point

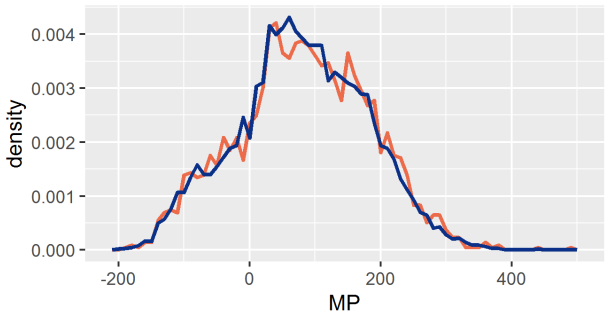
property	set	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
BP	train	4062	-103.7	134.0	189.0	189.3	245.0	548.0
	test	1353	-161.5	131.0	190.5	187.5	244.5	512.0

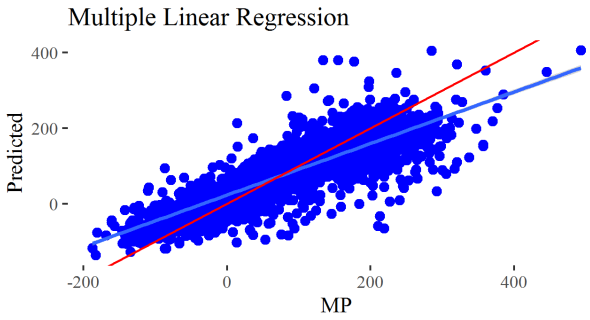


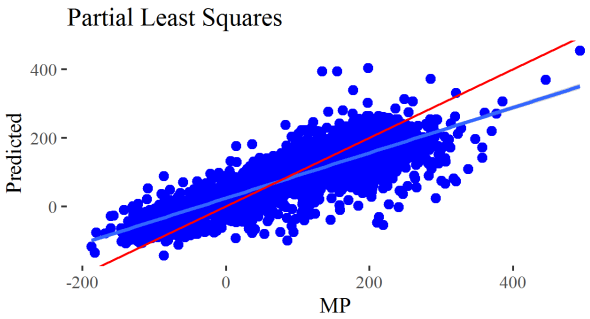
property	mlr	pls	svm	kNN	gbm
BP	0.8978	0.8875	0.9487	0.8829	0.9535

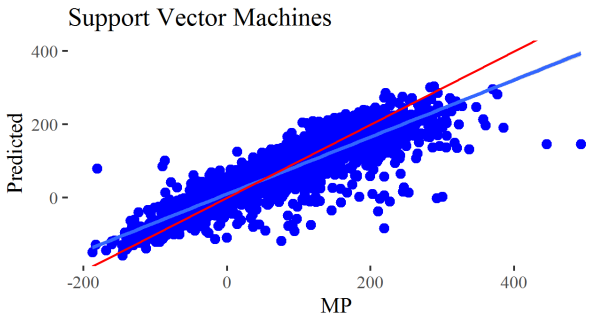
5.2.d Melting Point

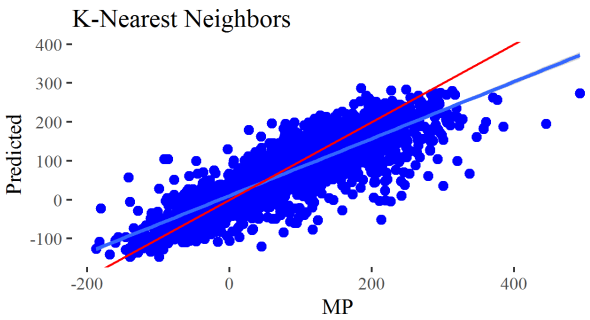
property	set	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
MP	train	6463	-196.00	15.95	79.00	79.61	150.00	437.30
	test	2162	-187.60	16.62	83.00	82.97	156.88	492.50

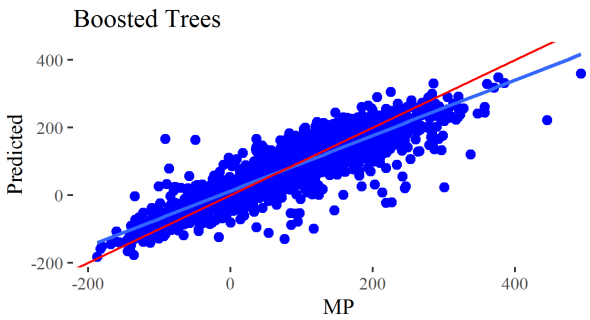








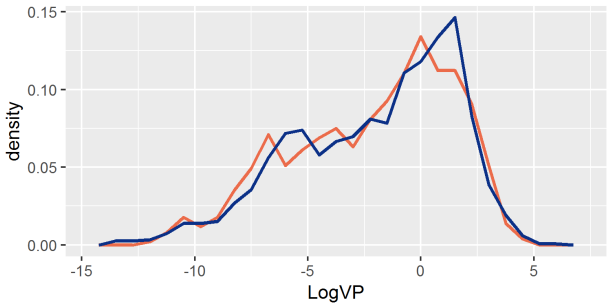


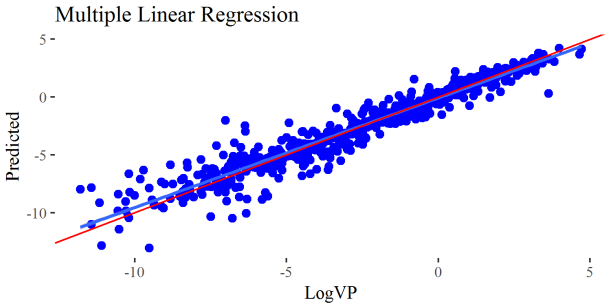


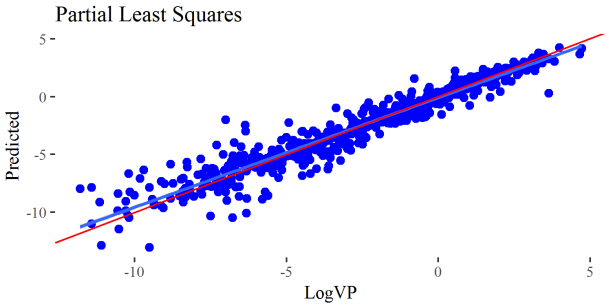
property	mlr	pls	svm	kNN	gbm
MP	0.6849	0.6656	0.7841	0.7253	0.8106

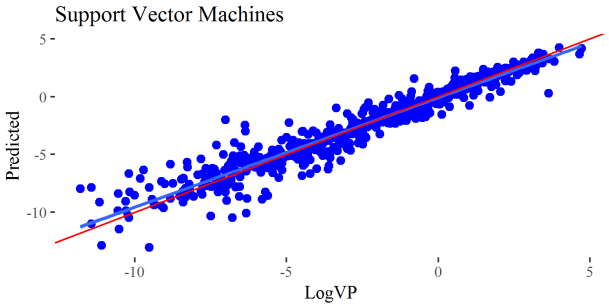
5.2.e Vapor Pressure

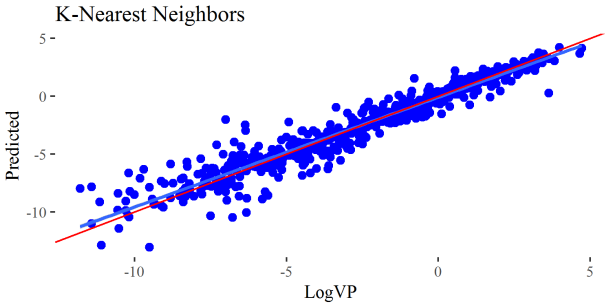
property	set	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
VP	train	2024	-13.678	-4.770	-1.222	-2.005	0.919	5.668
	test	677	-11.796	-4.737	-1.396	-2.142	0.732	4.717

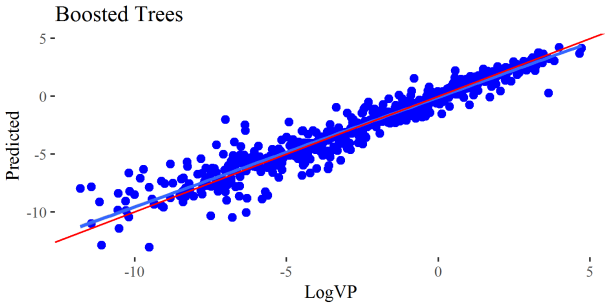








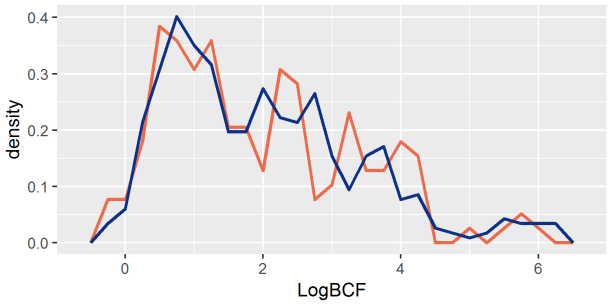


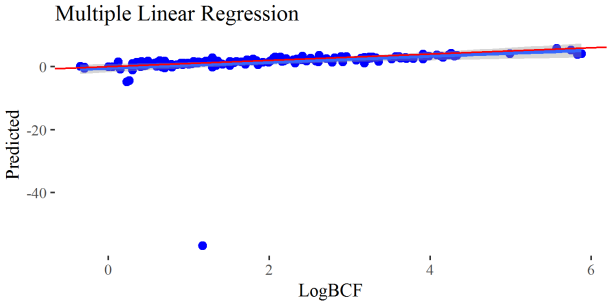


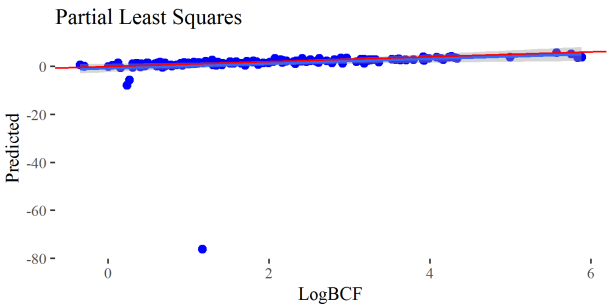
property	mlr	pls	svm	kNN	gbm
VP	0.8834	0.8823	0.9385	0.8662	0.9327

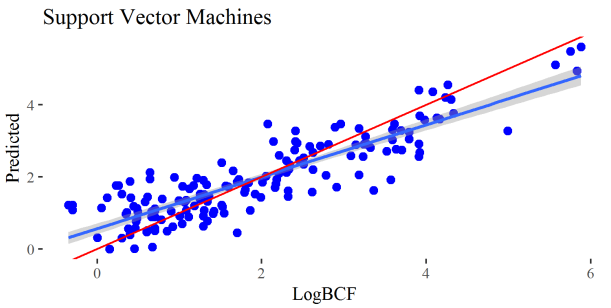
5.2.f Bioconcentration Factor

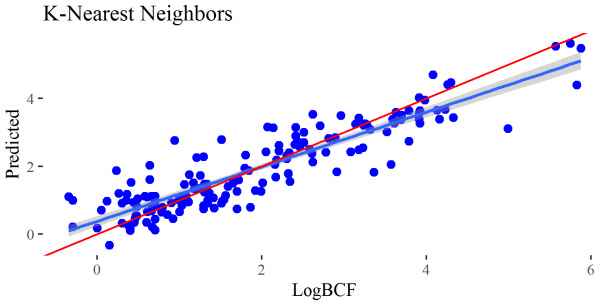
property	set	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
BCF	train	469	-0.300	0.860	1.800	2.006	2.820	6.360
	test	157	-0.350	0.780	1.700	1.990	2.960	6.430

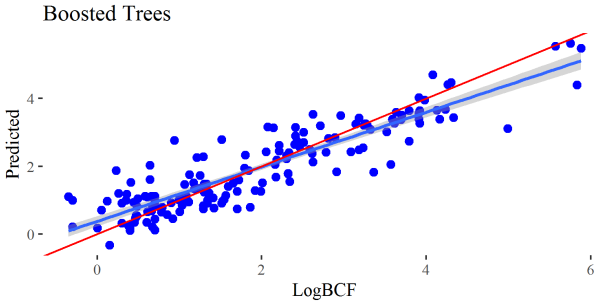












property	mlr	pls	svm	kNN	gbm
BCF	0.0668	0.0398	0.7690	0.6916	0.8085

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)

$$r = \frac{n\sum p_k p_l - \sum p_k \sum p_l}{\sqrt{n\sum p_k^2 - (\sum p_k)^2}\sqrt{n\sum p_l^2 - (\sum p_l)^2}} \tag{8}$$

6 Comparison of the Models

7 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