Date: December 21, 2018

Technical Report

De : Paul KOWALCZYK

À: Jean-Yves DELANNOY
Alessio TAMBURRO
Pascal METIVIER
Nicolas CUDRE-MAUROUX
Patrick MAESTRO
Jean-christophe GALLAND
Jacques-Aurelien SERGENT
James WILSON
Jean-Pierre MARCHAND
Lydie CAMUS
Antoine EMERY
Jordy BONNET
Laura TURBATU
Mathieu BERTIN

Copie :
Ref :

Pages : 18

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),partial least squares regression (pls), support vector machines (svm), k-nearest neighbors (k-NN), and gradient boosted machines (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.

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 coefficient, 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 QSPR 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)[1] 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 [1], 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.

[1] http://www.oecd.org/

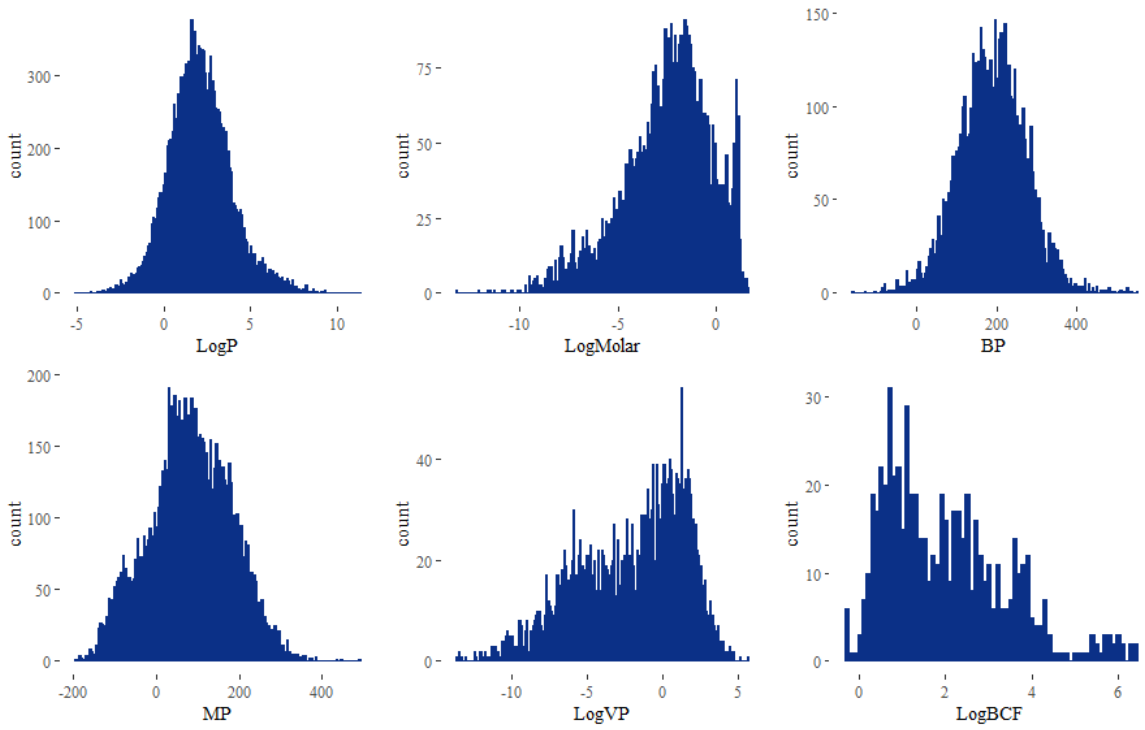
2 Materials & Methods

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 [1] [2]. 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.

[2] (http://www.srcinc.com/what-we-do/environmental/scientific-databases.html#physprop)

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.

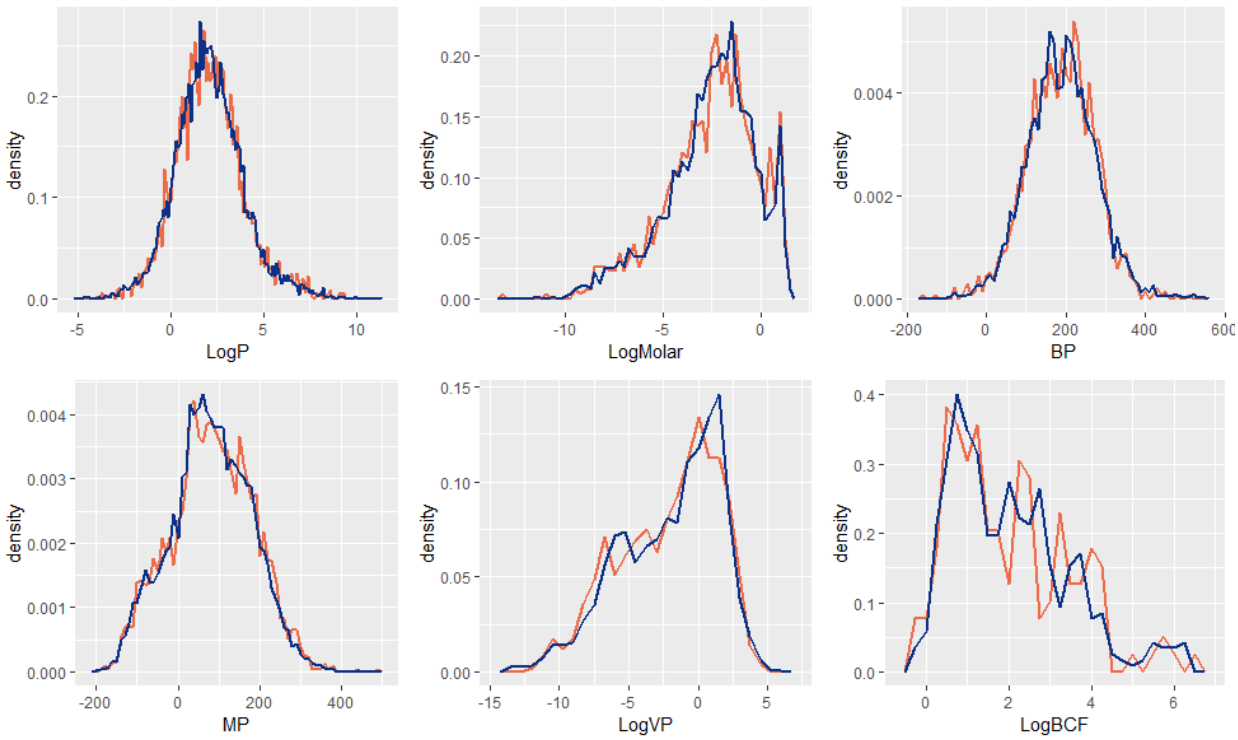
2.2 Selection of Training Sets & Test Sets

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: Property value distributions 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 Chemistry Development Kit (CDK)[2] was used to calculate molecular descriptors. The CDK is a collection of modular Java libraries for processing chemical information. It implements a variety of molecular descriptors, categorized into topological, constitutional, geometric, electronic and hybrid[3]. The CDK was accessed using the R[3] package rcdk[4]. This workflow, so constructed, guarantees reproducibility and scalability. 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.

[3] `https://cdk.github.io/cdk/1.5/docs/api/org/openscience/cdk/qsar/descriptors/molecular/package-summary.html`

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. Each model was subjected to 5-fold cross-validation.

2.5 Model Validation.

The performance of each QSPR model was evaluated by calculating the adjusted R^2 from a linear model of predicted value versus experimental value.

2.6 Applicability Domain

Applicability Domain. The applicability domain (AD) of a QSPR model is the molecule space a model can reliably predict. QSPR models are built using molecules in a training set. The applicability of these models towards reliable predictions is confined to those molecules that are similar to the training molecules. The similarity is measured using the calculated molecular descriptors. A qualitative assessment of the AD was used throughout this study. This assessment is described below.

For each dataset, the reduced set of training set descriptors (non-zero variance, not highly correlated; the # **Descriptors Remaining** column in Table 3) was subjected to a principal component transformation. This same transformation was applied to the reduced set of test set descriptors. The first two principal components of the training set and the test set are plotted. The region(s) occupied by the training set in this 2-dimensional plot represent the applicability domain. To the degree that the test set points fall within this region, the test set falls within the applicability domain. Test set points falling outside this region are outside the applicability domain, and may not be well modeled.

The applicability domain plots for the six data sets are shown in Figure 3. There is considerable complementarity (overlap) in five of the six datasets. It is only for the bioconcentration factor dataset that one observes test dataset observations well outside the applicability domain. One possible consequence will be commented on later in this report.

Figure 3: Applicability domain plots. The applicability domain is represented by the training dataset (cyan points). The test dataset (light red points) is projected *onto* the applicability domain.

3 Results & Discussion

The results are summarized in Table 4. Gradient boosted machines (gbm) outperformed the other four approaches in predicting log P, log S, BP, MP, and BCF, as measured using the adjusted R^2. In the case of VP, the adjusted R^2 for SVM (0.9385) is only slightly better than the adjusted R^2 for gbm (0.9327).

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 for each model built in this study. **mlr**: multiple linear regression; **pls**: partial least squares regression; **svm**: support vector machine; **kNN**: k-nearest neighbors; **gbm**: gradient boosted machines. These machine learning algorithms are briefly described in the Supplement.

The bioconcentration factor dataset is the *least* accurately predicted across *all* machine learning models. This is particularly true for multiple linear regression ($R^2 = 0.0668$) and partial least squares ($R^2 = 0.0398$). These two machine learning algorithms are sensitive to outliers. The applicability domain (AD) plot for this dataset (shown in Figure 3) shows observations in the test dataset falling outside the applicability domain, *i.e.*, molecules that are *dis*similar to the training set of molecules. If a test compound is considered outside a model's applicability domain it cannot be associated with a reliable prediction, as is shown with the bioconcentration factor dataset.

4 Conclusions

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). The predictive accuracy of the models, as measured by the adjusted R^2 between observed and predicted values for a hold-out test set, ranges from an adjusted $R^2 = 0.8085$ (bioconcentration factor) to an adjusted $R^2 = 0.9535$ (boiling point).

The data required to initiate a machine learning campaign is minimal. That data is

1. a representation of the molecule, or, information (*e.g.*, compound name, CAS number, ...) that could be used to generate the representation, and

2. the endpoint (outcome) associated with the molecule.

An example is presented in Table 5. Only one of **SMILES** or **NAME** is required. In this example, **Log BCF** is the endpoint.

SMILES	NAME	Log BCF
NC(N)=S	thiourea	0.3
Oc1c(Cl)cccc1Cl	2,6-dichlorophenol	1.08
CN(C)c1ccccc1	N,N-dimethylaniline	0.9
...	...	...

Table 5: Example input for ecotoxicological property prediction.

There are opportunities available to Solvay colleagues:

- Colleagues interested in using one, a few, or all of the models presented in this report to predict outcomes for their molecules should contact the author. Today, these models are provided as a service. Work is underway to build an interactive application that would afford colleagues the opportunity to run these models from their desktop/laptop.
- Colleagues interested in building a predictive model for data they have in-hand should also contact the author. As mentioned, requirements are minimal: (1) a representation of the molecule, and (2) the outcome associated with that molecule.

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:

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

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.

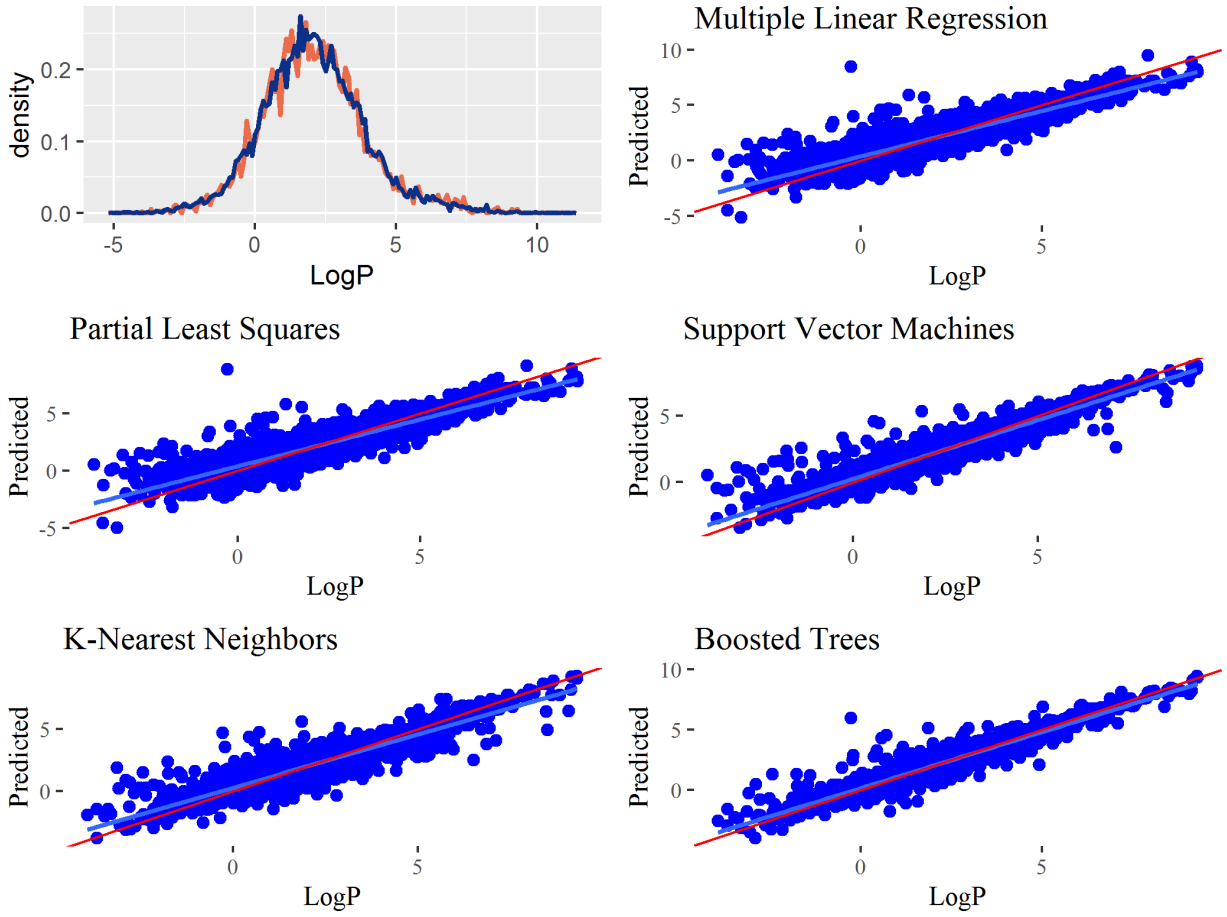
5.1.e Gradient Boosted Machines

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

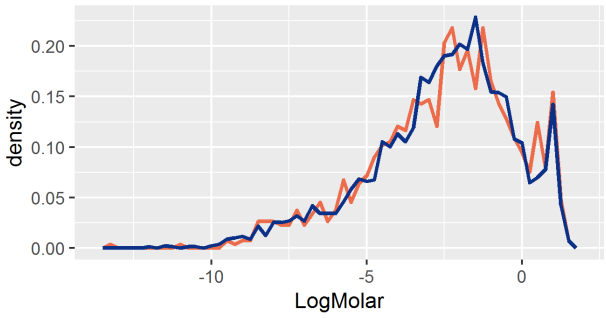
property	set	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Log P	train	10520	-5.080	0.890	2.000	2.075	3.160	11.290
	test	3506	-3.950	0.910	2.000	2.096	3.170	9.300

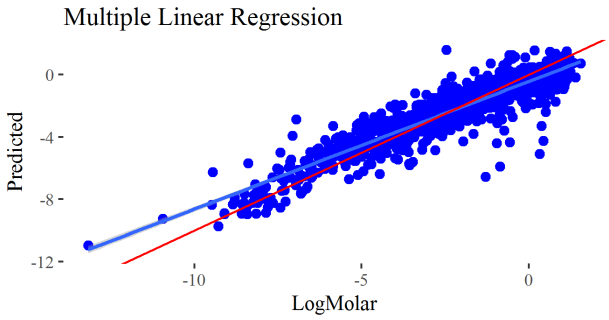


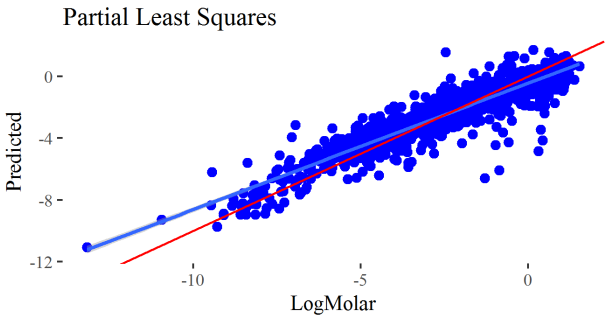
property	mlr	pls	svm	kNN	gbm
Log P	0.8125	0.8098	0.9052	0.8546	0.9234

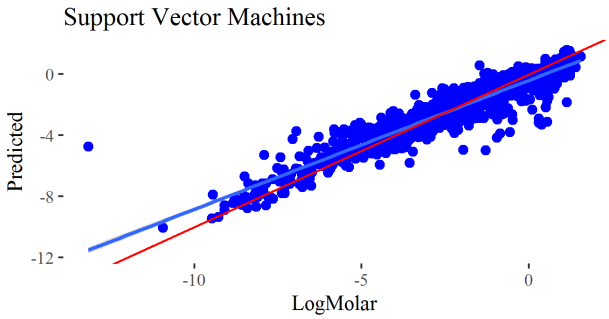
5.2.b Log S

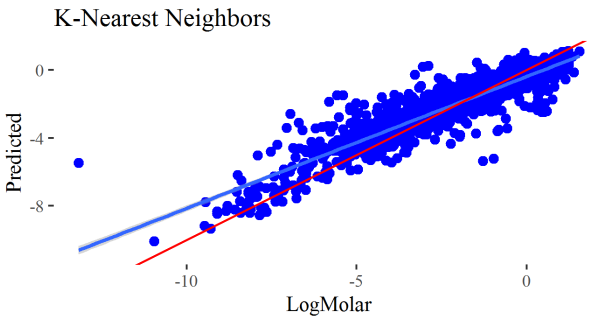
property	set	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Log S	train	3149	-12.060	-3.825	-2.284	-2.580	-1.007	1.581
	test	1063	-13.172	-3.929	-2.284	-2.553	-0.945	1.541

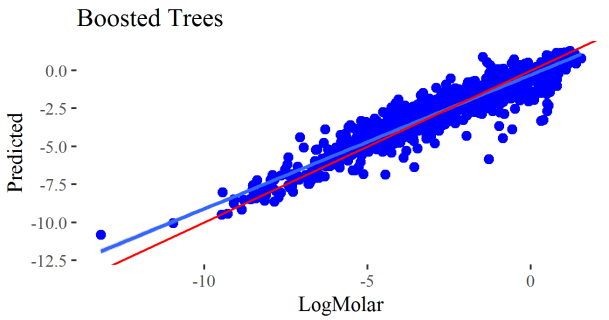








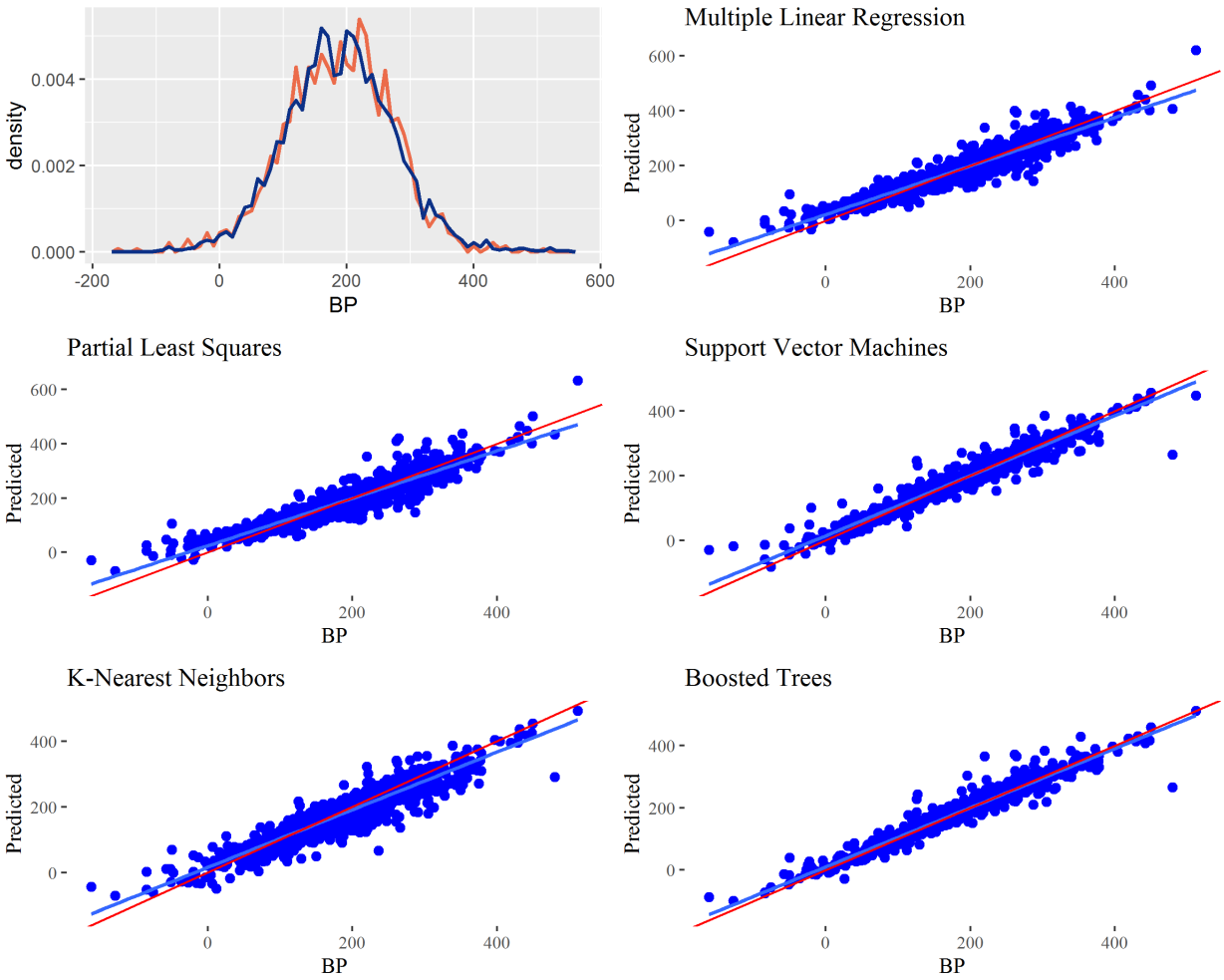




property	mlr	pls	svm	kNN	gbm
Log S	0.8317	0.8315	0.8770	0.8113	0.8924

5.2.c Boiling Point

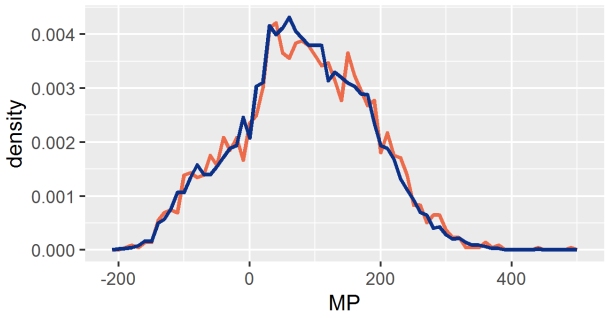
property	set	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
BP	train	4062	-103.7	134.0	189.0	189.3	245.0	548.0
	test	1353	-161.5	131.0	190.5	187.5	244.5	512.0

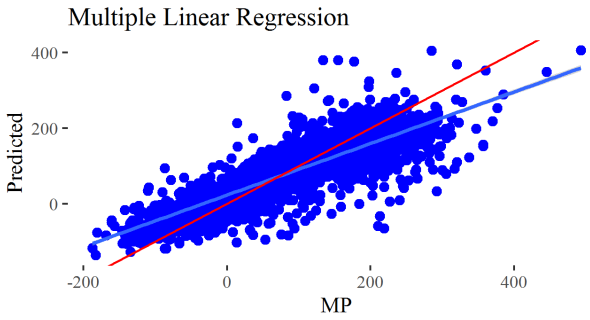


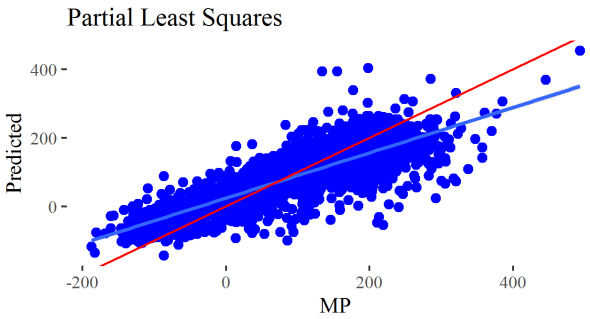
property	mlr	pls	svm	kNN	gbm
BP	0.8978	0.8875	0.9487	0.8829	0.9535

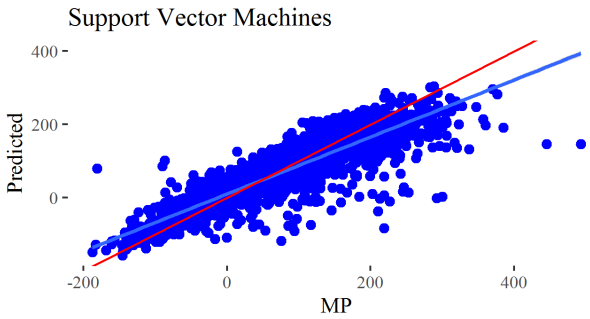
5.2.d Melting Point

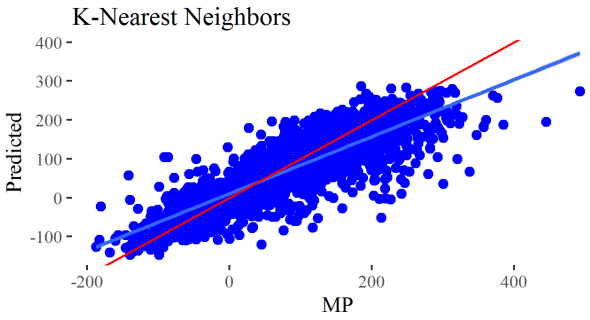
property	set	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
MP	train	6463	-196.00	15.95	79.00	79.61	150.00	437.30
	test	2162	-187.60	16.62	83.00	82.97	156.88	492.50

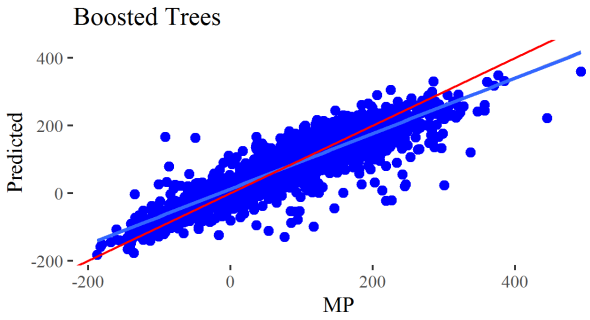








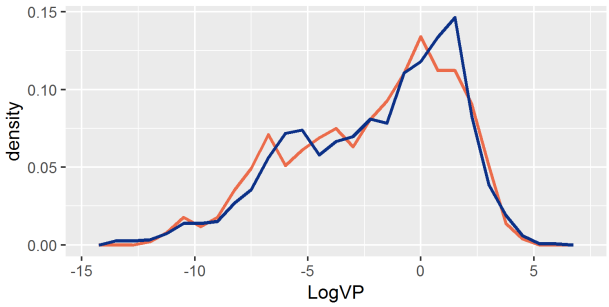


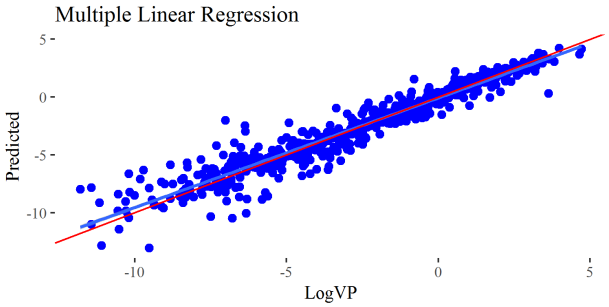


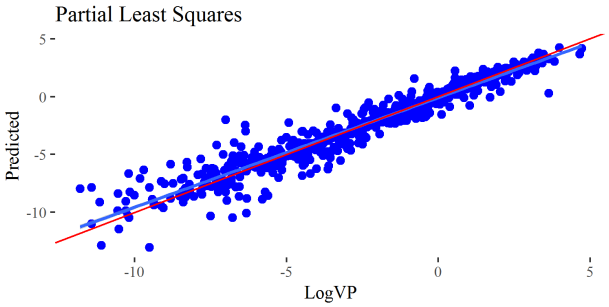
property	mlr	pls	svm	kNN	gbm
MP	0.6849	0.6656	0.7841	0.7253	0.8106

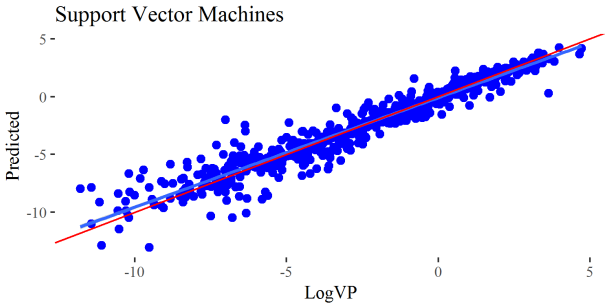
5.2.e Vapor Pressure

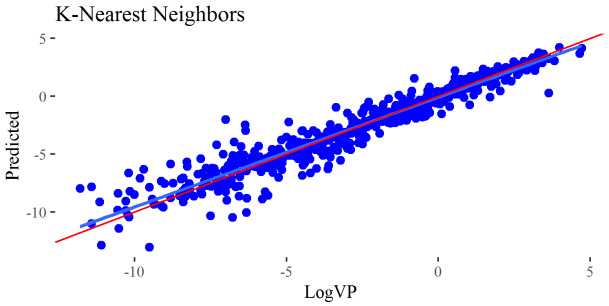
property	set	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
VP	train	2024	-13.678	-4.770	-1.222	-2.005	0.919	5.668
	test	677	-11.796	-4.737	-1.396	-2.142	0.732	4.717

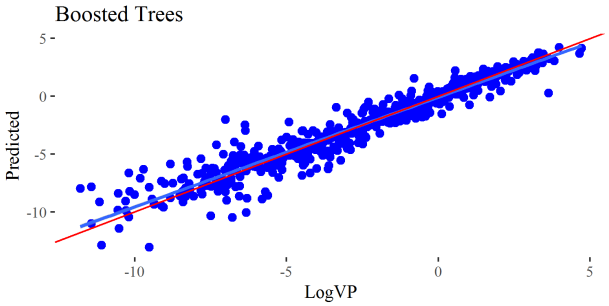








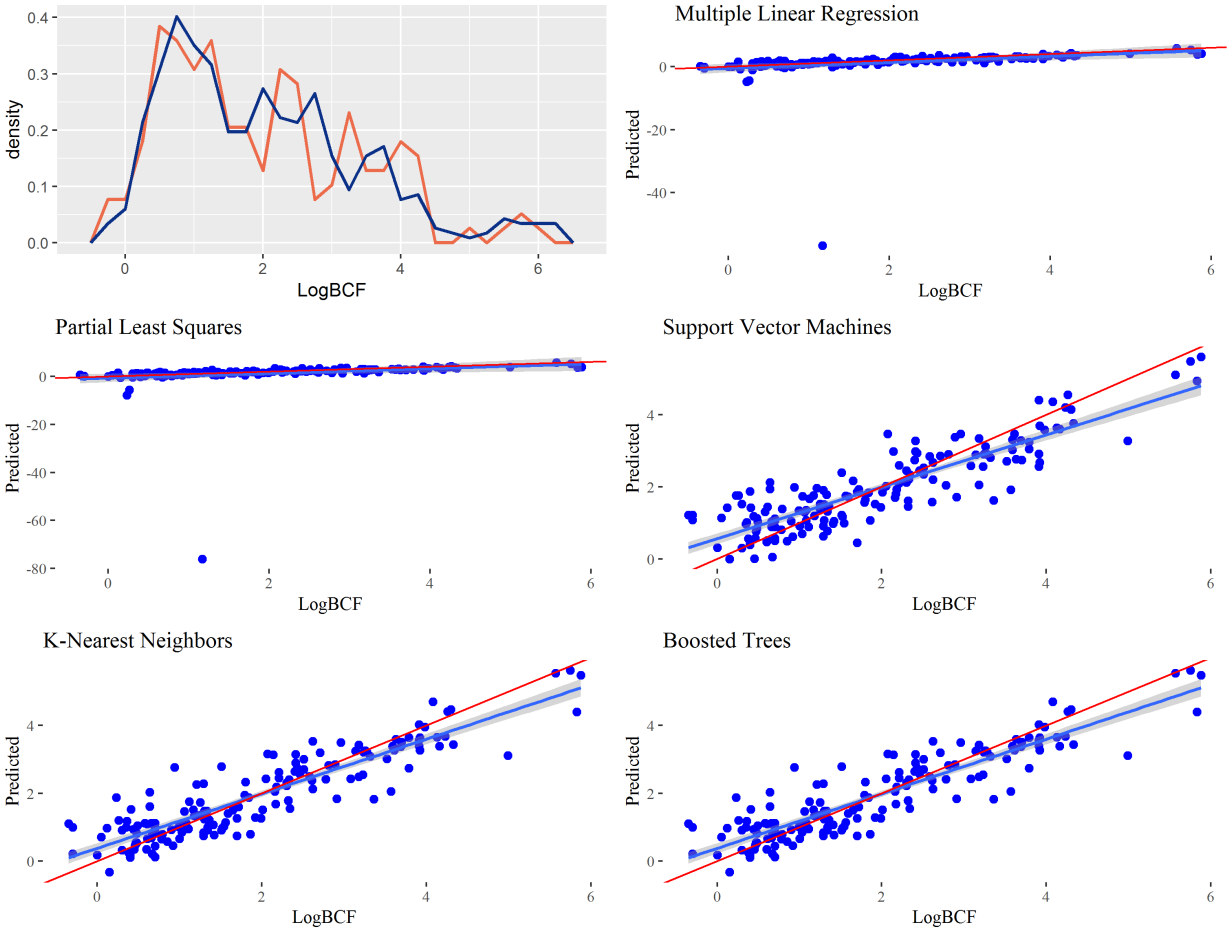




property	mlr	pls	svm	kNN	gbm
VP	0.8834	0.8823	0.9385	0.8662	0.9327

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

References

[1] PH Howard and William Meylan. Physprop database. *Syracuse Research Corp., Syracuse, NY*, 2000.

[2] Christoph Steinbeck, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen. The chemistry development kit (cdk): An open-source java library for chemo-and bioinformatics. *Journal of chemical information and computer sciences*, 43(2):493–500, 2003.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[4] Rajarshi Guha and Miguel Rojas Cherto. rcdk: Integrating the cdk with r. 2017.