molecular informatics

DOI: 10.1002/minf.200((full DOI will be filled in by the editorial staff))

Learning the structure of random forest models in QSAR modelling: Predicting molecular Solubility

Søren H. Welling(1,2), Line KH Clemmensen(1), Per B. Brockhoff(1,2) and Hanne HF Refsgaard(1,2).

Dedication((optional))

Abstract: Random forest (RF) models are used in QSPR models to predict solubility by the molecular structure. Non-linear models, such as RF, have been difficult to interpret as the model of many trees each of many nodes is far too complex to comprehend. Instead a model can be understood as a high-dimensional mapping structure which can be decomposed into a series of main effects and interactions. With feature contributions, and a newly developed tool it is possible to produce 2D and 3D visualizations to browse the model structure. We have built a model of 12 standard molecular descriptors on a very cited data set of 1200 molecules and illustrated how a RF model fit weigh the information to produce predictions of solubility. It appears that interactions between used descriptors have a minor contribution on solubility prediction accuracy. The exemplified particular RF model fit can be boiled down to a series non-linear transfer functions, one for each descriptor, and some minor interactions. Moreover, the error of making such a specific generalization can be quantified. The proposed tool will likely be useful to interpret many other RF based QSAR models.

Keywords: keyword 1, keyword 2, keyword 3, keyword 4, keyword 5

1 Introduction

Quantitative structural activity/property relationship (QSAR/QSPR) models have been used to perform solubility predictions, and have e.g. been used in the pharmaceutical industry to select drug candidates for oral delivery. Insufficient solubility is likely lead to lower bioavailability [10]. Related, QSAR models have been used to estimate impact of pesticides on aquatic environment as a function estimated molecular octane/water partition coefficients (logP) [11].

QSAR models represent an empirical approach to establish a relationship between measured properties such as molecular solubility and a numerical description of molecules. Molecular formulas, SMILES or connection tables are graph representations of connected atoms by different types of bonds[cite]. These representations can be encoded to produce numerical descriptions, such as *molecular weight* or *ratio of rotatable bonds.* Molecular descriptors can make use of physicochemical theoretical calculations to estimate internal partial charges between atoms to predict e.g. polarity of the molecule [8,PEOE]. Prediction by other empirical derived models of logP(*SlogP*) or molar refractivity(*SMR*), can be reused to predict solubility [4(SlogP/SMR)]. Molecular descriptors, based on atomic contributions or functional group contributions, will naively view the molecule as a simple sum of its atoms or functional groups. Scores for each type of atom or functional group are fitted to explain a data set of measured logP or molar refractivity. Other descriptors such as Kierhall can quantify how branched the molecular graph is[cite]. Finally encodings can perform a 2D or 3D force field simulations predicting an energy favourable conformation of the molecule (MFA,dipole[cite]).

[section on the previous models and descriptors from ESOL, huuskonnen, Delaney,]

Multiple linear regression (MLR) has been used to find a linear relationship between molecular descriptors and the predicted property. Often within a narrow selection of related molecules [cite sulphonate prediction] or when the molecular descriptors are well designed, linear models will perform well[zheng]. The last couple of decades, non-linear models such as support vector machines, neural nets and random forest have improved the prediction performance[cite some review]. These algorithmic models do not rule out unspecified non-linear relationships and neither interactions.

[1] *Department of Applied Mathmatics and Computer Science, Technical University of Denmark, Matematiktorvet, Building 324, 2800 Kgs. Lyngby, Denmark*

[2] *Novo Nordisk Global Research, Novo Nordisk Park 1, 2760 Maaloev, Denmark*
**e-mail:HARE@NOVONORDISK , +45 3075 0367*

Supporting Information for this article is available on the WWW under www.molinf.com

[Continue with new models, new palmer, laura, bergstrom]

Ideally by improving molecular descriptors a complex non-linear regression model would not be needed. In practice it is difficult to adapt to an unknown non-linearity, especially when high. A successful RF model structure should not only be considered as a black-box structure, but it should inspire to new and better feature engineering. A deeper insight of the trained model structure of RF could improve our understanding of predicted molecular solubility and point to further improvements of molecular descriptors.

1.2 Article, aim, goal, approach

We will demonstrate how a RF model structure can be systematically deconstructed and visualized to describe the learned QSAR between molecular descriptors and solubility. In a previous article[cite], we used the concept of feature contributions[kuzmin, anna] to illustrate how a random forest model was able to predict a specific biological activity of molecules in a cell-culture model. Hereafter we investigated the topology of feature contributions and made a new tool, forest floor, to visualize and understand the structure of random forest models[cite]. In this article we build a conventional QSPR solubility model based on a highly cited and reused data set[((huskonnen)), Delaney, palmer, bergs-trøm, laura] to discuss how RF utilize this information of molecular descriptors to produce predictions of solubility. [specify athours contribution, hertil gør palmer normalt så langt går vi videere]

[Beskriv composition af artikel]

2 Method

2.1 Introduction to random forest regression

A random forest model[leo] is a bootstrap aggregate ensemble model (bagging). It consists of hundreds or thousands of individual decision trees, whom are aggregated to form a joint robust, yet adaptive, ensemble model. Growing each decision tree starts with drawing *N* samples from the training set with replacement. Hereby, in average a .631 fraction out of N training set molecules are sampled to the root node of a tree at least once. These samples for a given tree are the inbag samples. The root node has a node prediction defined as the average measured solubility of the molecules in the node. The mean square error (mse) of the root node prediction can be reduced by splitting into two daughter nodes. Molecular descriptor are used to search for a splitting rule. A splitting rule (larger or equal to a value by one molecular descriptor) will split the node into two daughter nodes. One daughter node will have a higher molecular solubility and one with a lower than the parent node. The best split will lower mse of predicted solubility in the daughter nodes the most. By default, only a random third of descriptors are evaluated in any node to ensure not only one dominant molecular descriptor greedily is used first. Every node is recursively split until node size reaches 5 or less. Then the node is designated as terminal node. To perform predictions, new samples are passed down the tree according to the split rules. The terminal nodes will make up the possible solubility predictions of the tree. These almost fully grown trees are likely low biased, as the potential model structure is very flexible. Though individually, a tree has a high variance as each prediction is based on only 5 or less samples. To counter the instability of each tree, the bootstrapping and only evaluating a random third of descriptors in every node ensure low correlation between trees. When trees are less correlated and the variance is random and symmetrical, the learned structure will be amplified and the variance will be averaged out [breiman].

For every tree, a set of molecules will be out-of-bag (OOB) in contrary to inbag, when used to grow the tree. To estimate the accuracy of the model fit, it must be cross validated. Any sample will be OOB in roundly one third of the trees in the model and thee tree can independently predict this sample. The cross validated prediction error is the expected performance of the model if new molecules were predicted, assuming these molecules were drawn independently from the same population as the training set. OOB cross validation is faster and yields comparable estimates as 5-fold cross validation [wessivik]. Variable importance (VI) can be used to order molecular descriptors by usefulness to the model. VI is the decrease of OOB cross validation performance (mse) if a given molecular descriptor, after growing trees, but before predicting OOB samples, was permuted (random shuffled) [caroline strobl]. VI can be used for variable selection or as in the visualizations of this paper, to bring the attention to the most useful variables first. Random forest and feature contributions can also be used for probabilistic classification[mig, anna]. In a QSAR context mainly regression is used.

2.2 Decomposing a RF model with feature contributions

The applied methodology, forest floor, does not visualize directly the decision trees of the random forest. With hundreds or thousands of trees, it is intractable for a user to comprehend the overall structure of a trained RF model by inspecting the trees. Instead the model can be understood as the learned mapping function (*f*), that maps from a feature space of molecular descriptors (*X*) to a physicochemical target ($\hat{\boldsymbol{y}}$). *X* has as many dimensions as features in the model. The geometrical shape of the model mapping can neither be visualized nor comprehended directly, as the mapping is likely non-linear and high dimensional. Instead, projections or decompositions are needed to visualize the structure with only 2-3 dimensions. Feature contributions [kuzmin, anna] serve as a particular useful decomposition of the prediction for each descriptor, which assist to choose the optimal visualization of the model structure.

A random forest algorithm ($\boldsymbol{g}$) when trained on a data set of *N* solubility measurements $\boldsymbol{y_i}, \boldsymbol{i} \in \{\boldsymbol{1}, \ldots, \boldsymbol{N}\}$ and encoded molecular features ($\boldsymbol{X_i}$) adjusted with a set of parameters ($\boldsymbol{\omega}$) will yield a model fit ($\boldsymbol{f}$). This model fit maps from any point in feature space ($\boldsymbol{X}$) of molecular descriptors to a predicted solubility scale ($\hat{\boldsymbol{y}}$). This mapping can be understood as a high dimensional geometrical structure. A decomposition is used to visualize and navigate what model structure connects $\boldsymbol{X}$ and $\hat{\boldsymbol{y}}$ in 2D or 3D visualizations. The simplest and perhaps adequately correct decomposition splits the solubility prediction into separate effects with one unique function to explain each molecular descriptor.

$$\boxed{\begin{matrix}\textit{Training}\\ \boldsymbol{g(X_i, y_i, \omega)}\end{matrix}} \rightarrow \boxed{\begin{matrix}\textit{Model fit}\\ \boldsymbol{\hat{y}_i = f(X_i)}\end{matrix}} \approx \boxed{\begin{matrix}\textit{Main effects only}\\ \boldsymbol{f_1(x_{1i}) + f_2(x_{i2}) + \cdots + f_j(x_{ij}) + \bar{y}}\end{matrix}} \approx \boxed{\begin{matrix}2^{nd}\textit{ order interactions}\\ \boldsymbol{f_{1,2}(x_{1i}, x_{2i}) + \cdots + \bar{y}}\end{matrix}}$$

2D structures *3D structures*

Figure x: The Random forest algorithm is a function that when given a data set and training parameters will output a model fit. This model structure can as a start be interpreted as concisting of main effects only and visualized in 2D. Any deviation from a main effect only can be visualized as a 2nd order interaction.

[up to 2nd order interactions, include. Make dash line boxes to indicateconcept formula]

Hereby the model fit $\boldsymbol{f}$ can be simplified to series of additive functions $\boldsymbol{f_1}$+$\boldsymbol{f_2}$+…, which separately can be plotted in 2D. Feature contributions are used to estimate such additive functions and allows an isolated interpretation of each molecular descriptor.

2.3 Computation of feature contributions?

Every root-, intermediary- or terminal node of a decision tree is an individual prediction. When a parent node is split by a given variable, the daughter nodes will each receive some of the inbag samples and hereby construct two new predictions. A local increment is the change of node predictions from a parent node to a daughter node. For any sample, the RF prediction is simply the sum of all its encountered local increments divided by number of trees plus the grand mean of the training set. Feature contributions are constructed by the same local increments divided by number of trees, but feature contributions are summed separately for each sample by each variable. Thus a feature contribution can be understood as the average change of prediction for one sample molecule due to the information of one specific molecular descriptor - given all other molecular descriptors. *Given all*, means in practice that any interaction structure is preserved in the feature contributions.

When computing feature contributions for a training set, the yielded feature contributions can be arranged as a matrix with same dimensions as the molecular descriptor training matrix X_{ij}. Feature contributions can be denoted F_{ij}. Any prediction $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ can be split into separate contributions attributed each of the molecular descriptors plus the grand mean of all solubility measurements ($\bar{\boldsymbol{y}}$).

$$\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{F_{ij}} + \bar{\boldsymbol{y}}$$

To estimate the most accurate RF model structure it is most efficient to use any available training sample. To visualize the model structure it is also preferable to use all training predictions to compute feature contributions. Just as training predictions of a RF model can be out-of-bag cross-validated, so can feature contributions. Cross validated feature contributions yields fewer random ripples in the visualized model structure. These random ripples arise from the inherent overfitting of individual decision trees. [forestFloor]

2.4 Plotting, quantifying goodness-of-visualization and identifying latent interactions.

The first way to plot feature contributions for a given molecular descriptor is as a function of the corresponding descriptor values, and this function can be fitted with an estimator. For this purpose, we suggest an estimator based on leave-one-out k-nearest neighbour Gaussian distance weighting, as it can fit most RF model structures and produces a fast cross-validation.

$$\boldsymbol{E(F_{.j}, X_{.j}) \rightarrow f_j(X_{ij}) = \hat{F}_{ij}}$$

Hereby is obtained, a 2-axes plot of feature contributions (y-axis) versus the corresponding molecular descriptor values (x-axis) plus a fitted line describing the trending main effect not considering any interactions. See Figure 1 of the result section as an example. In Figure 1 the y-axis is feature contributions for any molecule in training set by a specific molecular descriptor (x-axis).

The fitted line may be an inadequate description, as a random forest model possibly may also have captured one or more interaction effects related to this molecular descriptor. The cross-validated explained variance of the feature contributions (R^2) by the fitted estimator quantified how well the 2D visualization describes the descriptor effect as a main effect only.

$$\boldsymbol{R^2_{f_j} = 1 - \frac{\sum_{i=1}^{N}(F_{ij} - \hat{F}_{ij})^2}{\sum_{i=1}^{N}(F_{ij})^2}}$$

If the explained variance is e.g. only 50%, one may choose to find a better context to understand the feature contribution. A broader context can be plotted as a 3D plot where the feature contribution e.g. can be plotted as a function of the two descriptors, e.g. by the first and second descriptor

j={1,2}. Again the feature contributions can be fitted with an estimator and the goodness-of-fit can be quantified.

$$\boldsymbol{E}\left(\boldsymbol{F}_{\boldsymbol{i},(\mathbf{1},\mathbf{2})}, \boldsymbol{X}_{\boldsymbol{i},(\mathbf{1},\mathbf{2})}\right) \rightarrow \boldsymbol{f}_{\boldsymbol{j}}\left(\boldsymbol{X}_{\boldsymbol{i},(\mathbf{1},\mathbf{2})}\right) = \widehat{\boldsymbol{F}}_{\boldsymbol{ij}}$$

In the 3D plot the estimated fit will no longer be a line but a surface, see the fitted surfaces in Figure 2. Unexplained variance of the estimated surface may remain; perhaps a 4D visualization is needed to explain an interaction between 3 molecular descriptors. Fortunately, we observe for random forest models in several data sets, that main effects tend to dominate over second order effects, which tend to dominate over higher order effects[cite me]. Thus, visualizing a model structure in 2D and 3D is likely adequate for most practical purposes.

Colour gradients can be used to provide one extra dimension. The molecule samples in a visualisation can be assigned to a colour gradient reflecting a latent variable to visually identify possible local or global interactions. A local interaction is understood as an interaction effect only learned in a smaller confined part of the model structure. A local interaction for a group of molecules can be highlighted with a colour pattern. In Figure 4 in result section such a highlighting is used to visualize the local model structure for 57 polychlorinated bi-phenyl molecules .

2.5 Software implementations

All visualizations in this article were produced with the R package forestFloor (1.8.9) [cite forestFloor cran]. The supplementary file of this article contains scripts to reproduce the model and visualizations of this paper. The forestFloor package depends on the rgl[Duncan, version] package to produce 3D visualizations, the kknn[cite] package for function estimators and the Rcpp [eddelbuettel, versoion] package to integrate functions implemented in C++ with the R environment. The RF models were trained with randomForest packae [liaw, version]. All packages are available from the CRAN repository [cite cran].

2.6 Data set and molecular descriptors

A public data set by Huuskonnen *et al*[3] was chosen because it is well cited and as it has been reused in many other datasets [palmer,Delaney,bergstrøm,wiisinger,Laura]. Training set and test set were merged in to on single data set of 1256 molecules. SMILES were imported to the software with the application MOE [cite] and sequentially pre-processed with the following functions: 'wash' (simulating an ideal solubilised molecular form), 'partial charges MMFFA96x' calculating the electron densities necessary for a number of descriptor algorithms, and finally 'energy minimize' relaxing the molecule in the minimum 3D state as suggested by [palmer]. To limit the scope of this article, only a small selection of 12 common and useful descriptors identified by Palmer *et al*[palmer] were used. The full data set with descriptors is provided in supplementary materials.

3 Results

3.1 Visualising main effects

A default random forest model of 2000 trees and mtry=4 was trained on the data set. Mean test error of 20 repeated 10-fold was 0.636(+/-0.004) sd? and r^2_s = .903(+/-0.001). Which was a similar performance as [palmer, huskonnen, laura?, hou?]. With the default RF model, out-of-bag feature contributions for every molecule were plotted as a function of the respective features/descriptors (main effect plots). *SlogP* was the most important descriptor by variable permutation importance and plotted first in upper left corner, followed by other descriptors in a decreasing order. A negative linear relation with solubility contribution was observed. High *SlogP* yielded negative contribution to solubility. A flattening of the main effect curve was observed in both ends. Fitted lines and calculated explained variance hereof described how well each molecular descriptor could be regarded as a main effect. The explained feature contribution variance by fitted main effect lines ranged from 90%-87% for the first 7 most important descriptors. Hereafter declined the explained main effect to range .71 to .48 explained variance. And the least important descriptor was only explained 15% as a main effect. Hence, the latter 5 variables were poorly described as main effects, where at the same time less influential for the model prediction deemed on the variable importance and as seen in Figure 1 the absence of feature contribution variance. Thus, overall to visualize the entire random forest model fit as strictly additive explained by the sum 12 main effect estimators explained 89% of the cross validated predictions. Thus to view these descriptors as contributing individually additively to the prediction of solubility would be a fair generalization of this particular instance of a random forest model fit.

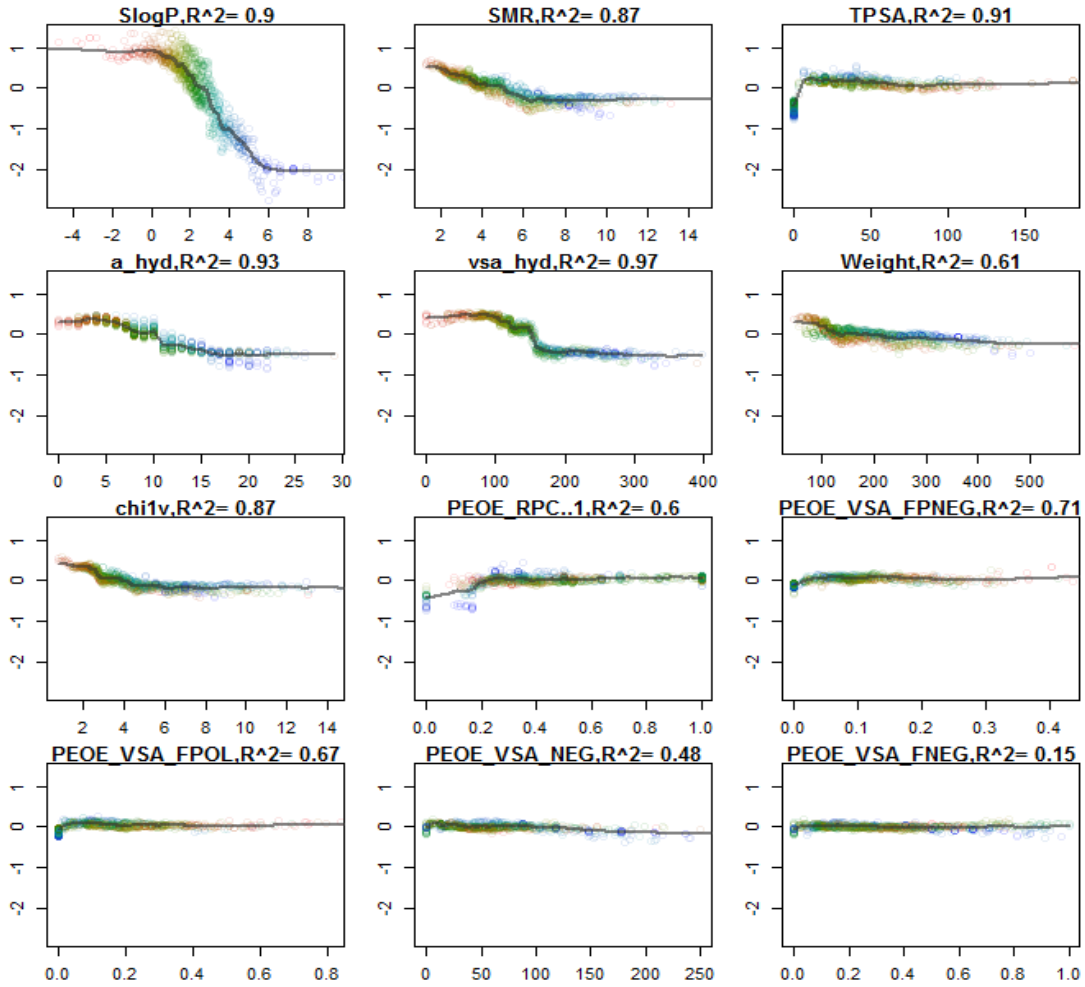


Figure 1. *Main-effect illustration of the 12 descriptors ordered by variable importance. Each molecule is represented once in each plot as a point of a specific colour. Point colour is by defined SlogP descriptor value of each molecule, corresponding to horizontal colour gradient in SlogP plots. A horizontal/diagonal gradient indicates local interactions with* SlogP. *Black lines + R^2 values are estimated fits, a strictly non-interaction interpretation of molecular descriptor effect, as described in equation 1.*

3.2 Identifying and visualizing interactions

To step beyond a strictly main effect interpretation, interaction effects must be identified. A colour gradient (red-yellow-green-teal-blue) horizontally aligned with the *SlogP* axis was used to characterize *SlogP* value of molecules in all other plots. Each molecule will have the exact same colour in all plots. Correlations and interactions with *SlogP* were visually highlighted with this colour gradient. Molecular descriptors correlating with *SlogP*, reproduced the colour gradient horizontally as observed for *SMR*, *PEOE_VSA_NEG*, *vsa_hyd*, *a_hyd*, *Weight*, *chi1v*). Other descriptors *TPSA* and *PEOE_VSA_FPOL* showed a reversed horizontal colour gradient as these descriptors negatively correlated with *SlogP* within the data set R_p~0.5. For all descriptors, deviations from fitted main effect lines were observed. Thus, the variance of each individual feature contribution could not entirely be explained by the descriptor alone. Molecules with specific *SlogP* values indicated by colour gradient were observed to deviate from the fitted lines in specific patterns. Hence, such deviations from a pure main effect could be explained by the many upstream decision splits by the *SlogP* or other correlated descriptors. In Figure 1 a low *Weight*(<120 Dalton) was attributed to a positive contribution to solubility, only when *SlogP*<1.5 (red/yellow). Molecules with high (*SlogP*>4, blue) had a feature contribution near zero for any molecular weight. Only 61% of the feature contribution variance of *Weight* was explained by the fitted main effect line. The remaining variance was thus attributed to interactions, such as the interaction with *SlogP* identified with the colour gradient. *Weight* was a descriptor with medium importance, yet poorly explained as a main effect. Hence, it was found as needed to elucidate the model contribution of *Weight* further. Figure 2 depicts in 3D the feature contributions of *Weight* for every molecule plotted by *Weight* and *SlogP*.

Again the interaction effect between *SlogP* and *Weight* could be observed. The fitted surface, explains the contribution of *Weight* (z-axis) as a main effect by *Weight* (x-axis) itself and as an interaction by *SlogP* (y-axis). This fit increased the explained feature contribution variance to 90%. In figure 2, it

was observed that there were no examples of molecules with low *Weight* and low *SlogP*. Thus this part the RF model structure is extrapolated and the model structure is less likely to be predictive for any such molecule. That the boiling point of small apolar molecules (e.g. propane, halothane etc.) is far below room temperature likely explains no such learning examples exist.

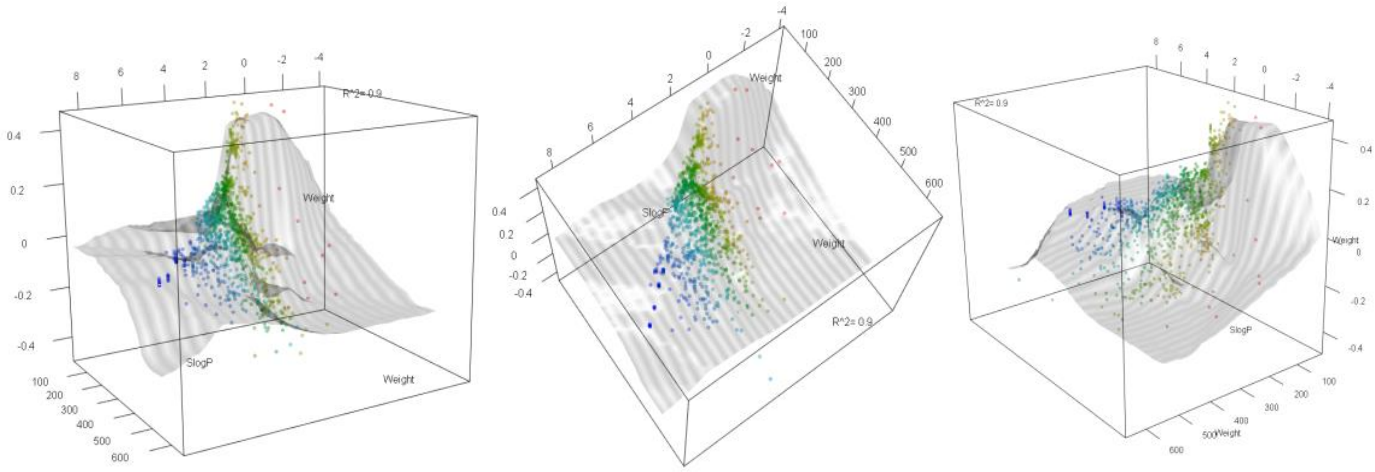


Figure 2. Feature contributions of *Weight* (z-axis) versus feature values *Weight* (X-axis) and feature values *SlogP* (Y-axis). Surface visualizes the fitted estimator, which describes 90% of the variance. Colour gradient parallel to *SlogP* axis as in figure 1. Image visualizes an interaction where *Weight* contributes most to solubility prediction when *SlogP* is negative.

The *SMR* feature was the second most important feature. The main effect of *SMR* feature contribution (molecular refraction by atom contributions) was explained 87%. When viewed as an interaction with *logS*, 95% of the feature contribution variance was explained. *SMR* is intended to approximate the polarizability of molecules, such that these e.g. can form induced dipoles in polar solvents and obtain an energy favourable charged interaction with water[cite]. Such an effect may have been anticipated to contribute in general positively to solubility, but in fact as main effect SMR contribute negatively to solubility. As molar refractivity is the 'polarizability per molecule', this measure was highly correlated with Weight (r_p =.93). If the SMR feature was divided by Weight and the RF model was refitted. The *SMR* feature dropped to the 11th most important feature and the main effect was flat.

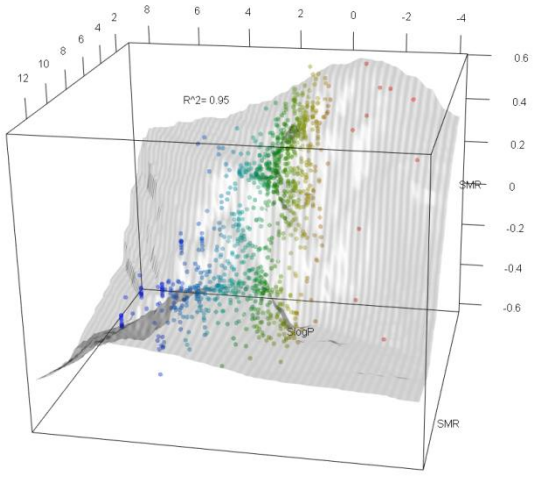


Figure 3. Interaction plot of <u>*SMR*</u> feature contribution as function *SMR* and *SlogP*. This fitted estimator describes 95% of variance of the feature contributions of *SMR*.

3.2 Identifying local effects

In main effect plot figure 1, a distinct group of 57 molecules with low *logS* showed distinct interactions in *SlogP, SMR, TPSA and PEOE_RPC..1*. The group of molecules can be identified in figure 2 middle plot, as having a perfect linear relationship between logS and SMR (r_p=1). In figure 4, the position of these molecules in the model structure was highlighted by colouring any other molecule black. The observed interactions was for *SlogP* a flattening of the negative contribution to solubility of molecules with *SlogP* above 5 whereas ~15 non PCP molecule with SlogP >5 were predicted decreasing soluble as a function of *SlogP*. For SMR a linear reduction in solubility as contrary to the general main effect.

Furthermore these molecules were not only on a line but only placed on 10 different steps with equal distance between them. The molecules were isolated and showed in table 3.

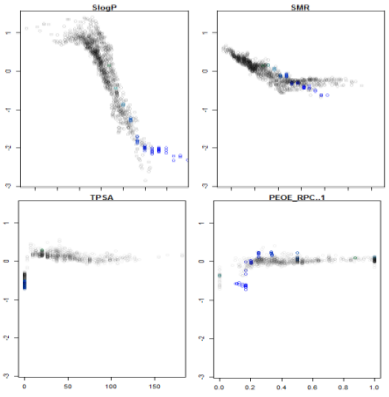


Figure 4. Highlighted feature contributions of PCB molecules.

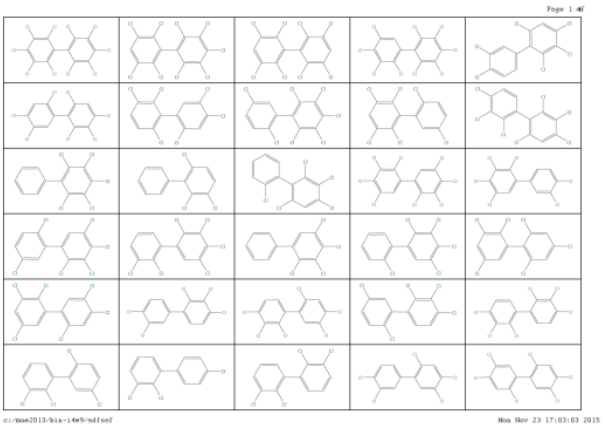


Table 1: Depiction of 36 PCB molecules. What kind of table would be fine here?

It showed that all molecules were polychlorinated biphenyl compounds (PCB). As both *SlogP* and *SMR* are defined by atomic contributions, all PCB with the same amount of substituted chloride atoms will have same *SlogP* and *SMR* values. In fact only two features *chi1v* and *PEOE_RPC..1* produced unique feature values for PCB's with same amount of chloride atoms. But these differences in values were minute, and they were more likely to arise from a non-deterministic convergence algorithm estimating the partial charges[cite method]. Also there appeared to be no obvious relationship between these two features and solubility beyond the number of chloride atoms. Moreover the random forest model fit did not seem to capture any relationship related to substitution pattern, as the OOB cross-validated predictions for these PCB with equal amounts of chloride atoms did not correlate with the actual solubility. Predictions ranged only 0.12 logS units for PCB with same amount of chloride atomes, where the predictions ranged 1.3 logS units. Thus the random forest model was unable with the 12 selected features to predict the relationship between PCB substitution pattern and solubility.

3.2 Model structure is affected by training parameters

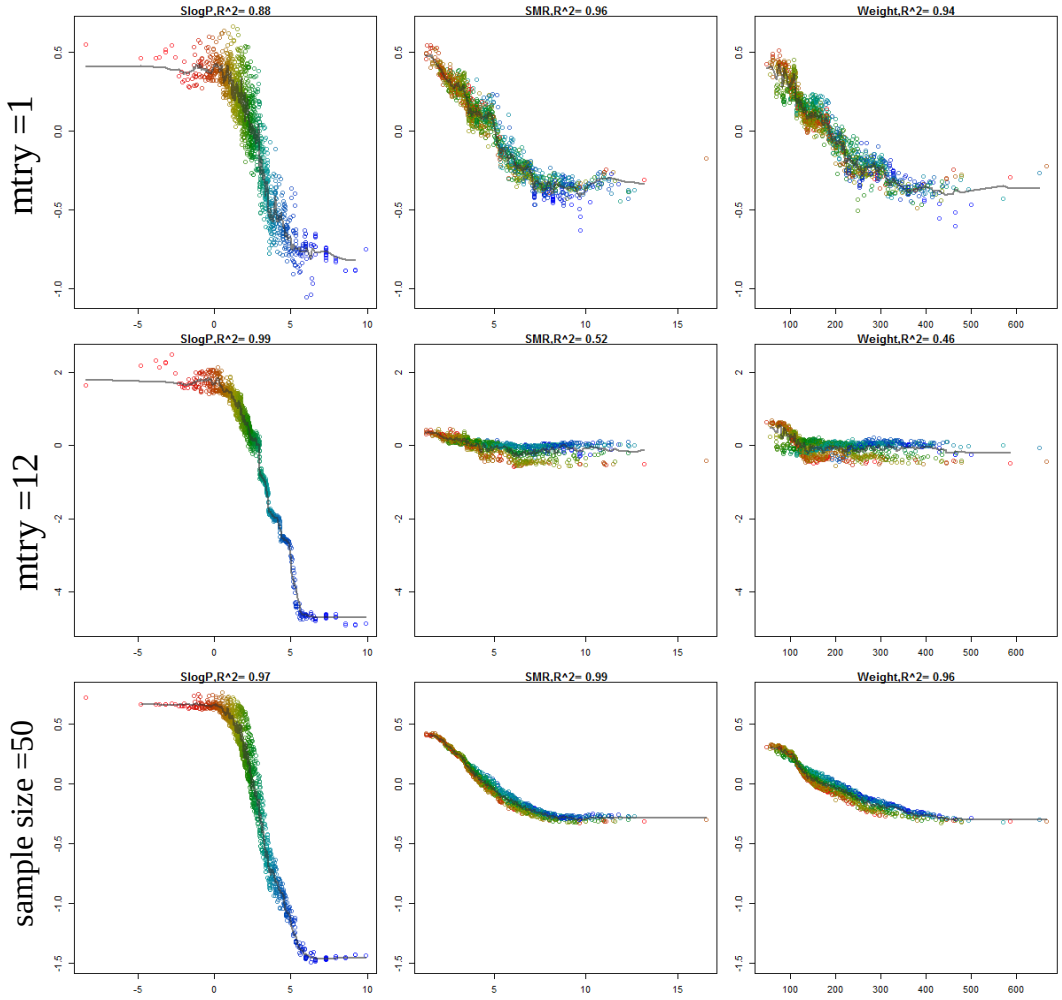


Figure 5: Model structure varies with the parameters. Low mtry(1), uniform use of features. All variables have main effect and interaction effects. High mtry(12,all), algorithm will greedily use best feature first, other features are mainly used for interaction effects. Low sample size(50), smoothens model structure, interactions reduced, model approaches strictly additive model. Sample size is by default 1250 and mtry is by default 4.

Discussion [unfinished]

Choosing a correct set of low dimensional visualizations to account for a complex model structure is not necessarily fully attainable [friedman]. Forest floor can identify and quantify the residuals of any visualization, such that the depiction of the RF model structure can iteratively be elaborated until a sufficiently correct depiction has been attained. Any high dimensional structure cannot be visualized in two or three dimension. In a regression context, a main effect requires 2 dimensions, a 2nd order (interaction) effect requires 3 dimensions and a 3rd order requires 4 dimensions. That said it is possible to understand the 3D structure of a DNA helix from a 2D drawing, and likewise the 4D Kleinn-bottle structure in form a 3D representation. RF is a relatively shallow model and 3rd or higher order interactions or seems almost absent.

This presented methodology of decomposing effects by descriptors, estimating main effects and interactions effects is one representation of the model structure. Another representation such as the actual trained ensemble of decision trees is concise but is too complex to lend itself to a clear interpretation. Another representation, such partial dependence plots can e.g. in 3D describe an interaction effect between two variables. But classic PD plots are not guaranteed to well generalize the overall high dimensional structure, nor do they point to the location of potential sizeable latent interactions. Thus, the forest floor is a methodology that provides the investigator means to browse the model structure of a random forest model and quantify how well a given low dimensional representation, as a series of visualisations, describe the overall structure.

With

3.1 discussion of other methods

With another method to visualize a mapping such partial dependence plots, to uncover hidden interactions and avoid to extrapolation is more difficult.

Today, mainly variable importance [palmer, laura, others] is used in conjunction with random forest models to interpret the model. Variable importance describes the loss of cross-validated predictive performance when each variable in turns were permuted. VI only approximates the usefulness of each molecular descriptor. VI does not outline how each descriptor is used by the model.

[insert in result section] A group of PCB molecules were identified as to elicit a distinctive interaction pattern. With the 12 selected molecular descriptors, was the chloride substitution pattern of this PCP molecules not learned. *SlogP* and *SMR* the most important descriptors are e.g. themselves based predictions on *logP* and molar refractivity for 10.000 measured molecules. Predictions are based summing empirical derived scores for each atom in the molecule. Atoms are categorized by atom number and type bonding to neighbouring atoms. Thus for PCB molecules having the same number of substituted chloride atoms all scores will be exactly alike. [maybe two extra sentrences of why neither other descriptors has any clue of this effect.] Ghavami et al. [6] produced a regression model only to predict solubility of PCB molecules and found that 90 percent of the variance of PCB log solubility can be attributed the number chloride atoms in a linear regression model. Introducing counts of ortho-, meta- and para configuration contributed to explain up to 97% cross-validated variance of the log solubility of PCB molecules. As the PCB molecules collapse to only extending a string of connecting points in the feature space, where each point concist of PCB molecules with same amount chloride atoms, the sampling density around these PCB molecules is high. Thus, is the random forest model able to fit a very specific structure accounting for the solubility variance related to chloride atoms in PCB molecules. If predicting the solubility of a random molecule, it would be unlikely to fall within the small sub feature space of PCB molecules. If it did fall within this subspace, the learned relationship from PCB's would dominate the prediction of the RF model.

First Main Text Paragraph----without indentation.

Main Text Paragraph----with indentation.

((Insert schemes above the captions. **Note:** Please do **not** combine scheme and caption in a textbox or frame))

Scheme 1. Scheme Caption.

Main Text Paragraph----with indentation.

((Insert figures above the captions. **Note:** Please do **not** combine figure and caption in a textbox or frame))

Figure 1. Figure Caption.

Main Text Paragraph----with indentation.

Table 1. Table Caption. ((**Note:** Please do **not** include the table in a textbox or frame))

Head 1[a]	Head 2	Head 3[b]	Head 4[c]	Head 5
Column 1	Column 2	Column 3	Column 4	Column 5
Column 1	Column 2	Column 3	Column 4	Column 5

[a] Table Footnote.
[b] …

3 Conclusions

First Main Text Paragraph----without indentation.
Main Text Paragraph----with indentation.

Acknowledgements

Acknowledgements Text.

References

[1] ((Reference 1, Example for **Journals**))a) A. Author, B. Coauthor, *Mol. Inf.* **2009**, *1*, 1-10; b) A. Author, B. Coauthor, *Angew. Chem.* **2006**, *118*, 1-5; *Angew. Chem. Int. Ed.* **2006**, *45*, 1-5.

[2] ((Reference 2, Example for **Books**))J. W. Grate, G. C. Frye, in *Sensors Update*, Vol. 2 (Eds: H. Baltes, W. Göpel, J. Hesse), Wiley-VCH, Weinheim **1996**, pp. 10-20.))

[3]

[1]
ESOL: Estimating Aqueous Solubility Directly from Molecular Structure
John S. Delaney*
Syngenta, Jealott's Hill International Research Centre, Bracknell, Berkshire, RG42 6EY, United Kingdom
Received October 29, 2003

[2]
Global and Local Computational Models for Aqueous Solubility Prediction of Drug-Like Molecules
Christel A. S. Bergström ,† Carola M. Wassvik ,† Ulf Norinder ,†‡ Kristina Luthman ,§* and Per Artursson †
Center for Pharmaceutical Informatics, Department of Pharmacy, Uppsala University, Uppsala Biomedical Center, P.O. Box 580, SE-751 23 Uppsala, Sweden, Department of Medicinal Chemistry, AstraZeneca R&D, SE−151 85 Södertälje, Sweden, and Department of Chemistry, Medicinal Chemistry, Göteborg University, SE−412 96 Göteborg, Sweden
J. Chem. Inf. Comput. Sci., 2004, 44 (4), pp 1477–1488
DOI: 10.1021/ci049909h
Publication Date (Web): June 23, 2004
Copyright © 2004 American Chemical Society

[3]
Random Forest Models To Predict Aqueous Solubility
David S. Palmer , Noel M. O'Boyle ,† Robert C. Glen , and John B. O. Mitchell *
Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom
J. Chem. Inf. Model., 2007, 47 (1), pp 150–158
DOI: 10.1021/ci060164k
Publication Date (Web): December 2, 2006
Copyright © 2007 American Chemical Society

[4] Wildman, S.A., Crippen, G.M.; Prediction of Physiochemical Parameters by Atomic Contributions; J. Chem. Inf. Comput. Sci. 39 No. 5 (1999) 868–873.

[6] Ertl, P., Rohde, B., Selzer, P.; Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties; J. Med. Chem. 43 (2000) 3714–3717.

[7] [Cruciani 2000]Cruciani, G., Crivori, P., Carrupt, P.-A., Testa, B.; Molecular Fields in Quantitative Structure-Permeation Relationships: the VolSurf Approach; J. Mol. Struct. (Theochem) 503 (2000) 17–30.

[8] Gasteiger, J., Marsili, M.; Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges; Tetrahedron 36 (1980) 3219.,

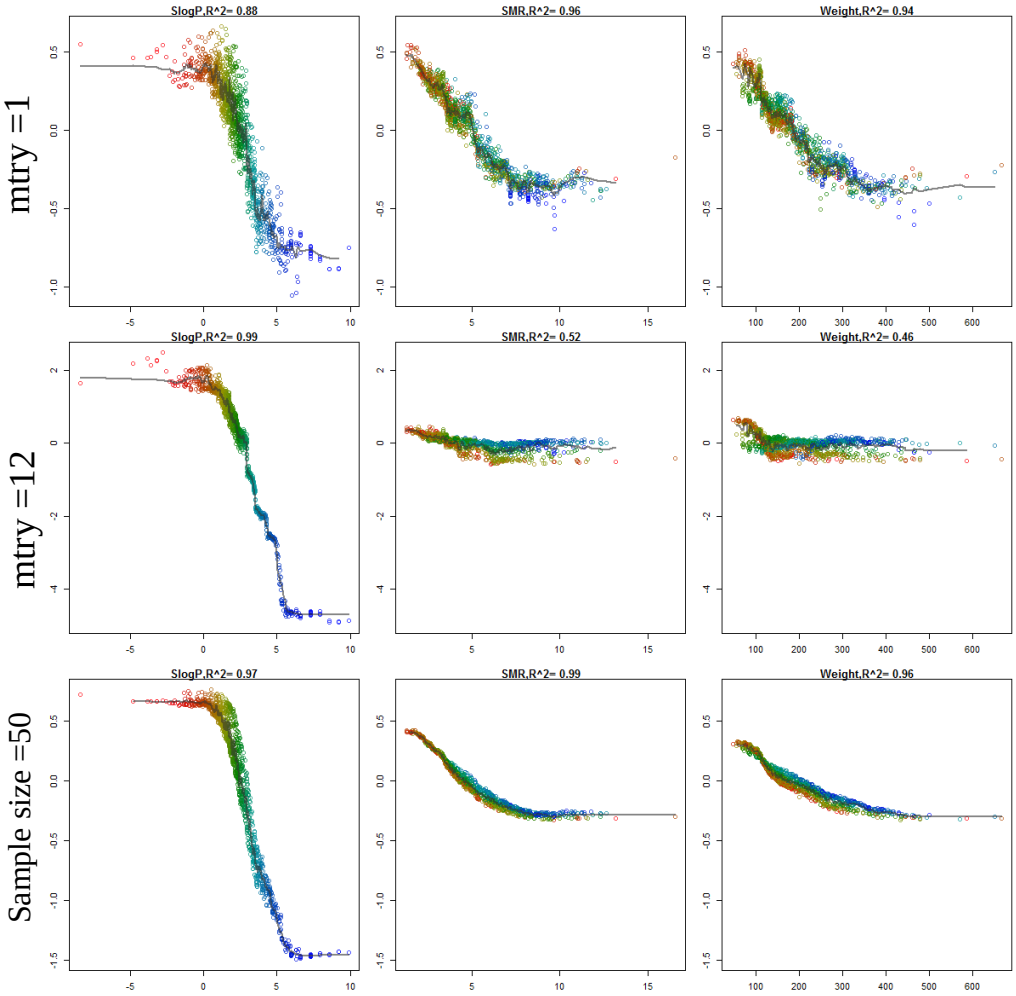
[9] Prediction of drug solubility from structure. William L. Jorgensena, , , Erin M. Duffyb . doi:10.1016/S0169-409X(02)00008-X

[10] Yu, L. X.; Amidon, G. L.; Polli, J. E.; Zhao, H.; Mehta, M. U.; Conner, D. P.; Shah, V. P.; Lesko, L. J.; Chen, M.-L.; Lee, V. H. Biopharmaceutics classification system: the scientific basis for biowaiver extensions. Pharm. Res. 2002, 19 (7), 921−925

[11] OVERVIEW OF DATA AND CONCEPTUAL APPROACHES FOR DERIVATION OF QUANTITATIVE STRUCTURE–ACTIVITY RELATIONSHIPS FOR ECOTOXICOLOGICAL EFFECTS OF ORGANIC CHEMICALS STEVEN P. BRADBURY,† CHRISTINE L. RUSSOM,*† GERALD T. ANKLEY,† T. WAYNE SCHULTZ,‡ and JOHN D. WALKER§ †U.S. Environmental Protection Agency, National Health and Environmental Effect

Received: ((will be filled in by the editorial staff))
Accepted: ((will be filled in by the editorial staff))
Published online: ((will be filled in by the editorial staff))

t showed that all molecules were polychlorinated biphenyl compounds. As both SlogP and SMR are computed by atomic contributions, all PCB with the same amount of substituted chloride atoms will have same SlogP and SMR values. In fact only two features chi1v and PEOE_RPC..1 produced different feature values for PCB's with same amount of chloride atoms. But these differences in values were minute, and they were more likely to arise from a non-deterministic convergence algorithm defining the partial charges. Also there appeared to no relationship between these features, chloirde substitution configuration and actual logS. More over the OOB cross-validated predictions for these PCB varied only 0.12 units where the average variation with PCB with equal many chlorides was 1.3. And this OOB cross validated variation within PCB with equal amounts of chloride did not correlate with actual solubility. Thus the random forest model was unable with the 12 selected features to predict the relationship between PCB substitution configuration and solubility. 90 Percent of the variance of PCB solubility can be attributed the number chloride atoms or any other Ghavami et al. [6] showed how counting ortho meta and para configuration contribute to explain upto 97% cross-validated variance.


mtry =1
mtry =12
Sample size =50
SlogP,R^2= 0.88
SMR,R^2= 0.96
Weight,R^2= 0.94
SlogP,R^2= 0.99
SMR,R^2= 0.52
Weight,R^2= 0.46
SlogP,R^2= 0.97
SMR,R^2= 0.99
Weight,R^2= 0.96