Classification of biodegradable materials using QSAR modeling with uncertainty estimation

Werickson Fortunato de Carvalho Rocha[a] and David Allan Sheen[b,*]

[a]*Division of Chemical Metrology, National Institute of Metrology, Quality and Technology -INMETRO, 25250-020 Duque de Caxias, RJ, Brazil*

[b]*Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA*

[*]Corresponding Author, Email: david.sheen@nist.gov
David A. Sheen
National Institute of Standards and Technology
Mail Stop 8320
Gaithersburg, MD 20899
Tel: 301 975 2603

Classification of biodegradable materials using QSAR modeling with uncertainty estimation

The ability to determine the biodegradability of chemicals without resorting to expensive tests is ecologically and economically desirable. Models based on quantitative structure-activity relations (QSAR) provide some promise in this direction. However, QSAR models in the literature rarely provide uncertainty estimates in more detail than aggregated statistics such as the sensitivity and specificity of the model's predictions. Almost never is there a means of assessing the uncertainty in an individual prediction. Without an uncertainty estimate, it is impossible to assess the trustworthiness of any particular prediction, which leaves the model with a low utility for regulatory purposes. In the present work, a QSAR model with uncertainty estimates is used to predict biodegradability for a set of substances from a publicly available data set. Separation was performed using a partial least squares discriminant analysis model, and the uncertainty was estimated using bootstrapping. The uncertainty prediction allows for confidence intervals to be assigned to any of the model's predictions, allowing for a more complete assessment of the model that would be possible through a traditional statistical analysis. The results presented here are broadly applicable to other areas of modeling as well, because the calculation of the uncertainty will clearly demonstrate where additional tests are needed.

Keywords: partial least squares discriminant analysis; uncertainty estimation; bootstrap; machine learning; biodegradable materials; QSAR

1 Introduction

In recent years, several countries around the world have recognized the need to reduce the amount of non-biodegradable materials used to intensify measures for the environment and encourage the recycling of materials. An example was the signing of the Treaty of Paris by 175 countries in April 2016 for the reduction of carbon dioxide emissions and other greenhouse gases. By this initiative, countries compromise to establish their own targets for the reduction of greenhouse gases which implies indirectly the reduction of the consumption of non-biodegradable materials. This is because the decrease provides a smaller amount of material sent to landfills, which

reduces the production of greenhouse gases [1-5]. Another factor that contributes to it is the increasing use of biodegradable materials due to the results of research related to discovery and production of new materials [1, 6-11], as well as the use of alternative non-toxic and biodegradable source of energy as biodiesel [12-15].

Several countries in the world have agencies and regulations responsible for the use of chemical substances and evaluation of their potential impacts on both human health and the environment, including the Environmental Protection Agency (EPA), National Health Surveillance Agency (ANVISA), and the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH). REACH, a regulation of the European Chemicals Agency of the European Union, is particularly notable because it promotes alternative methods for the hazard assessment of substances in order to reduce the number of tests on animals. Such alternative methods include the biodegradability predictions of chemicals from quantitative structure-activity relationship (QSAR) models.

Biodegradability is fundamental to the assessment of environmental exposure and risk from chemical products. QSAR models can be used to pursue both regulatory and chemical design goals. In the literature, various QSAR models have been investigated that are intended to predict the ready biodegradability of different substances [16-22]. Other authors have examined different methods of selecting molecular descriptors [23, 24] and the use of different machine learning algorithms [22, 25]. In all cases, the model's performance was quantified using aggregate statistics such as sensitivity, specificity, and correlation coefficients [26]. However, it is rarely reported by what method, if indeed at all, uncertainty in an individual model output is quantified. Uncertainty in this context means the range of values that can be reasonably attributed to an analytical result, considering the level of confidence [27-30]. Without an

estimate of the individual prediction uncertainty, the results of these models are not complete.

The objective of this work is to calculate the uncertainty of the predictions of the classification of a QSAR model using the residual bootstrap method to predict the ready biodegradability of chemicals using literature data [31]. The uncertainty then provides an estimate of the reliability of the PLS model's predictions.

2 Theoretical Background

2.1 Partial least squares discriminant analysis estimation of a QSAR model

Partial least squares regression discriminant (PLS-DA) is a classification method in multivariate analyses that combines the properties of partial least squares regression with the discrimination power of classification techniques [32]. This method searches for latent variables that are a linear combination of the independent variables **X** which have the maximum covariance with the dependent variables **Y** [33-37].

The general underlying PLS-DA model is given by

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}, \tag{1}$$

and

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F}, \tag{2}$$

where **X** is the matrix of independent variables, in this case the molecular descriptors; **Y** is the matrix of dependent variables, which has values of either 0 or 1 to indicate to which class the corresponding sample belongs. **T** and **U** are orthogonal score matrices of **X** and **Y** respectively. **P** and Q are the corresponding loadings matrices that describe the latent variables, and **E** and **F** are the residual terms.

The **T** scores are orthogonal and estimated as a linear combination of the **X** variables [38, 39] with weighting coefficients **W*** which are obtained by successive optimizations. Then, the **T** matrix can be determined using

$$\mathbf{T} = \mathbf{X}\mathbf{W}^{*} . \tag{3}$$

The **T** scores are then a set of latent variables within **X** that are good predictors of **Y**, assuming that **Y** and **X** are well-described by the same latent variables. Using Equation (3), Equation (2) can be rewritten as

$$\mathbf{Y} = \mathbf{X}\mathbf{W}^{*}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} = \mathbf{X}\boldsymbol{\beta} + \mathbf{F} . \tag{4}$$

A full description of the PLS-DA regression is given by Wold *et al.* [39].

The classification values obtained by the PLS-DA model are real numbers given by Eq. 4, not reading exactly 0 or 1. The results are scattered in a range of values for each class. Thus, it is necessary to establish a threshold value, y_{bound}, to define the class limits. There are several ways to set the threshold, for example, such as Bayes' theorem [40], receiver-operating characteristic (ROC) curves [41], threshold-based classification rule [42, 43] or by establishing confidence limits for each sample classified. These confidence intervals can be calculated by re-sampling techniques, such as bootstrap.

2.2 Bootstrap-based uncertainty estimation

Bootstrap is a test based random sampling with replacement [44, 45] which allows confidence intervals to be placed on a model's predictions based on uncertainties in the input data. In this case, it provides confidence interval of the classification results of substances in a given class.

In this paper, residual bootstrap was used to calculate the uncertainties in the biodegradability prediction of the PLS-DA model. The procedure was originally presented by Almeida *et al.* [33] and will be briefly described.

According to Almeida *et al.*, [33] it is necessary to calculate the residuals of the PLS-DA model using

$$\mathbf{F}^* = \frac{\mathbf{F}}{\sqrt{1 - D_f / N}}, \tag{5}$$

where **F*** is the weighted residual of the model, **F** is the residual term from Equation (2) given by $\mathbf{F} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$, D_f is the number of pseudo degrees of freedom (see [46]), and N is the number of calibration samples (substances, in this case).

Once the residual calculations are complete, the bootstrapping procedure is as follows. First, the substance *a* whose uncertainty is being calculated is removed from the model. A new dependent variable matrix **Y*** is then generated by replacing the remaining **Y** values with the model predicted $\mathbf{Y}_{\text{PLSDA}} = \mathbf{X}\boldsymbol{\beta}$ values. The residuals are assumed to be representative of the uncertainty in the model, and so a new random residual vector $\mathbf{F}^*_{\text{boot}}$ is generated by bootstrapping. The $\mathbf{Y}_{\text{PLSDA}}$ values are perturbed by adding the bootstrapped residual **F***,

$$\mathbf{Y}^* = \mathbf{Y}_{\text{PLSDA}} + \mathbf{F}^*_{\text{boot}}. \tag{6}$$

Then, a new PLS-DA model can be calculated from **Y***, with a new corresponding regression coefficient (β*) and new predictions $\hat{\mathbf{Y}}^* = \mathbf{X}\boldsymbol{\beta}^*$. The confidence interval for substance *a* is estimated based on the difference between the bootstrap predicted values for substance *a*, $\hat{Y}_a = \mathbf{X}_a\boldsymbol{\beta}^*$, and the PLS predicted value $Y_{a,\text{PLSDA}}$, according to

$$\hat{F}_a^* = Y_{a,\text{PLSDA}} - \hat{Y}_a \ . \tag{7}$$

In the case of a 95% confidence interval, the lower bound, denoted c_{low}, is the 2.5 percentile of $\hat{F}_a^*$ and the upper bound, c_{up}, is the 97.5 percentile. More details about bootstrap can be found in the literature [47].

2.3 Uncertainty application and misclassification probability

Calculating the misclassification probability proceeds as follows. First, the classifications $Y_{a,\text{pred}}$ are treated as being normally distributed random variables with mean equal to $Y_{a,\text{PLSDA}}$ and standard deviation $\sigma_a = \tfrac{1}{4}\left(c_{\text{low}} - c_{\text{up}}\right)$. The confidence intervals here are not symmetric but they are close enough for this to be a reasonable approximation. As stated earlier, a given sample a is identified as class 0 if its $Y_{a,\text{PLSDA}}$ value is less than the threshold value y_{bound}. The probability that sample a is class 0, denoted P_0, is equivalent to the probability that $Y_{a,\text{pred}}$ is less than y_{bound}. That probability is then given by the cumulative distribution function for the normal distribution, that is,

$$P_0 = P\left(Y_{a,\text{pred}} \le y_{\text{bound}}\right) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{y_{\text{bound}} - Y_{a,\text{PLSDA}}}{\sqrt{2}\sigma_a}\right)\right]. \tag{8}$$

Likewise, the probability that sample a is class 1, denoted P_1, is equal to $1 - P_0$. The probability of a misclassification, P_{misclass}, can then be determined based on the actual classification of the sample, $Y_{a,}$, using

$$P_{\text{misclass}} = P_{1-Y_a} \ . \tag{9}$$

The misclassification probabilities can then be used to assess the trustworthiness of the model. If a model has large P_{misclass} for the misidentified samples, then using the model

would mean that we would likely make incorrect claims with a high degree of false assurance. Such a model would not be very useful in a regulatory context. Likewise, if the model has large $P_{misclass}$ for the correctly-identified samples, our correct claims would be assigned a low degree of assurance, which is also undesirable for regulation.

3 Implementation

3.1 Data Sets

In the present work, QSAR models with estimation of uncertainty were explored to discriminate chemicals into two classes: RB (readily biodegradable) and NRB (not readily biodegradable). The data used in this study can be obtained from the publicly-available QSAR biodegradation data set described by Mansouri, et al. [31]. A version of the data set is included in the supplementary information.

The data are of three sets of substances: 837 substances used for the calibration stage (284 RB and 553 NRB), 218 substances for the validation stage (72 RB and 146 NRB) and 670 substances for the external validation stage (479 RB and 191 NRB). All the data have 41 molecular descriptors. According to Mansouri, et al. [31], using only 23 descriptors among the 41 descriptors it is possible to improve the performance of the model. Thus, only these 23 recommended molecular descriptors were used to generate the model (Table 1).

3.2 Procedure and Software

The PLS-DA models from PLS Toolbox 8.0 and from scikit-learn 0.17 were used to analyze the data. Uncertainty in the PLS model predictions was estimated by the residual bootstrap technique. 10^4 bootstrap evaluations were used and the analysis was repeated 15 times to ensure the reliability of the bootstrap results.

A dummy matrix **Y** was created with 0 for readily biodegradable and 1 for not readily biodegradable substances. The optimal number of latent variables for the PLS-DA model was determined by cross-validation using the leave-one-out criterion [40] in order to avoid overfitting or lack of fit. The threshold for the class was calculated using the *plsthres* function from PLS Toolbox [40] and a similar function in Python. The confidence interval estimations for each sample were obtained with residual bootstrap, according Almeida *et al.* [33] and as described in Section 2.

The structural data were preprocessed through autoscaling [40], because the units on each descriptor are different and have different ranges of variation. Calculations were performed in Anaconda Python 4.0.0 and in Matlab R2015b. The results presented here are from Anaconda, but the Matlab results were largely similar. The Python code to conduct the analysis and generate the figures has been included in the supplementary information.

4 Results and Discussion

The performance of the PLS-DA model was rated using the following standard statistical parameters (Table 2): root mean squared error (RMSE), Pearson's correlation coefficient, sensitivity (percentage of true positives, i.e., samples were correctly assigned to the RB class), specificity (percentage of true negatives, i.e., samples that were correctly assigned to the NRB class) and misclassification error, ME, defined as

$$ME = \frac{Y_a - Y_{a,\mathrm{PLSDA}}}{Y_a} \quad . \tag{10}$$

where $Y_{a,\mathrm{PLSDA}}$ represents the PLS-DA predicted class observed, and Y_a denotes the reference class.

The model developed can be said that to be accurate, based on the aggregate statistics (Table 2). In particular, it has Pearson's correlation coefficient values considered high for this dataset (near 0.65) and low error values (represented by RMSE and ME (%) values). The model presented specificity and sensitivity close to 0.8, meaning that the majority of substances were classified according to their correct class.

In addition to the global statistics of the model, we examine the scores and loadings of the PLS-DA model developed (Figure 1). The scores for the first two latent variables of the PLS-DA model show how the calibration and validation sets are separated by the model (Figure 1a) and the loadings shows the influence of each descriptor in the separation of substances (Figure 1b). Most RB substances are in the region of the scores plot where the first and second latent variable are both negative (Figure 1a), while the majority of NRB substances are in other regions. Through the analysis of the graph of loadings plot in Figure 1b, it is possible to explain this separation.

The molecular descriptors related to the presence of oxygen (nO, F03 [C-O], and SDO), LOC, and TI2_L have negative loadings with respect to the second latent variable (Figure 1b), which corresponds to the scores of the RB substances (Figure 1a). These descriptors are therefore likely responsible for the separation of RB substances. Descriptors involving cycles, halogens, and nitrogen have positive loadings with respect to the second latent variable, and the molecular matrix-based descriptors have loadings above a value of about 0.2 with respect to the first latent variable. These descriptors are therefore likely responsible for the separation of the NRB substances, as the NRB substances have scores similar to these descriptors' loadings. The results shown here are consistent with the literature [48, 49], where it has been shown that materials which

have functional groups with oxygen atoms increase biodegradation. On the other hand, the presence of atoms such as nitrogen and halogens decrease biodegradation.

The scores of the external validation set follow the same trend as the separation found for the set of calibration and validation substances (Figure 2), i.e. the RB substances are mainly located on the negative region of the first and second latent variable.

Through analysis of the detailed results of the PLS-DA model (Figure 1 and Figure 2), it is possible to have a general vision of the performance of the model developed according to Mansouri [31]; however it is not possible to estimate the uncertainty of the classification of each sample individually. That is, most substances were classified according to their respective class and some substances were misclassified, but the reliability of that classification is not known. This is the motivation behind the use of residual bootstrapping to calculate the individual classification uncertainties.

The bootstrapping process allows us to attach classification uncertainties and misclassification probabilities to the PLS-DA model results. The PLS-DA-predicted classifications, $\mathbf{Y}_{\mathrm{PLSDA}}$, can be plotted along with the corresponding confidence intervals and compared the threshold value, y_{bound}, (Figure 3). In particular, the samples are ordered by the probability of classification as NRB, P_0, (Equation 8). This information is used to calculate the misclassification probabilities P_{misclass} (Equation 9), plotted with respect to the $\mathbf{Y}_{\mathrm{PLSDA}}$ values and also including y_{bound} (Figure 4). Substances with confidence intervals above y_{bound} were classified as RB, which corresponds to $P_0 << 1$. Those substances with confidence intervals below y_{bound} were classified as NRB, corresponding to $1 - P_0 << 1$. Likewise, these are the substances that have P_{misclass} close to either 0 or 1, because the model allows us to make confident assertions about how

these substances should be classified. Those substances with confidence intervals that include y_{bound} could not be confidently classified in either group. Such substances have a P_0 and therefore a $P_{misclass}$ approaching 0.5. In these cases, the model does not permit a confident assertion about the classification of the substance. Here we have a new type of results caused by the uncertainty of calculation, that is, substances that are not possible to classify into any of the two classes.

Through the uncertainty calculation, it is possible to question the classification of a particular substance. An example of this is the classification of the substance 200 belonging to the validation set in the NRB class (Figure 3b). If the prediction the uncertainty had not been calculated, this substance would be considered to belong to the NRB class, however due to the greater rigor imposed by the uncertainty calculation, this substance cannot be classified in any of the two classes, because of its uncertainty intersects with limit between the classes

Similar results can also be seen, for example, with the validation substance 190 (Figure 3b) and the external validation substances 287, 288, 444, and 451 (Figure 3c), which likewise cannot be classified in either of the two classes. This type of result provides a more rigorous classification of substances. Indeed, when the uncertainty of the result of classification is not calculated, there are only two types of classifications possible: correctly classified substances and incorrectly classified substances.

5 Conclusions

A partial least squares discriminant analysis (PLS-DA) model was used to determine the biodegradability of substances based on quantitative structure-activity relations (QSAR). In addition, the classification uncertainty for this model was estimated using bootstrapping. Traditional modeling allows the substances to be distinguished into two

classes (readily and not readily biodegradable). Considering the uncertainty in classification allows for a third classification, those substances about which no confident statement can be made.

The uncertainty analysis methodology used here permits a more in-depth evaluation of the QSAR model that would be possible using the standard statistical parameters. A standard analysis would allow some conclusion about the accuracy of the model as a whole, but it would not allow any statement about the reliability of any particular prediction. The uncertainty analysis, by contrast, allows for an evaluation of the precision of the model's predictions, thereby allowing us to say that the model cannot confidently classify certain substances. Estimating the uncertainty makes it possible to obtain a conclusion that is more reliable and complete. These results highlight the challenges associated with developing reliable and easily applied acceptability criteria for the regulatory use of QSAR models, and it is hoped that a more widespread adoption of uncertainty analysis in these models will help to address some of these challenges.

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

References:

[1] A.V. Colling, L.B. Oliveira, M.M. Reis, N.T. da Cruz, and J.D. Hunt, *Brazilian recycling potential: Energy consumption and green house hases reduction*, Renew. Sust. Energ. Rev. 59 (2016), pp. 544-549.

[2] D. Allinson, K.N. Irvine, J.L. Edmondson, A. Tiwary, G. Hill, J. Morris, M. Bell, Z.G. Davies, S.K. Firth, J. Fisher, K.J. Gaston, J.R. Leake, N. McHugh, A. Namdeo, M. Rylatt, and K. Lomas, *Measurement and analysis of household carbon: The case of a UK city*, Appl. Energy 164 (2016), pp. 871-881.

[3] G. Lazzerini, S. Lucchetti, and F.P. Nicese, *Green House Gases(GHG) emissions from the ornamental plant nursery industry: a Life Cycle Assessment(LCA) approach in a nursery district in central Italy*, J. Clean Prod. 112 (2016), pp. 4022-4030.

[4] A.D. Adam, and G. Apaydin, *Grid connected solar photovoltaic system as a tool for green house gas emission reduction in Turkey*, Renew. Sust. Energ. Rev. 53 (2016), pp. 1086-1091.

[5] S.W. Goh, J.S. Zhang, Y. Liu, and A.G. Fane, *Membrane Distillation Bioreactor (MDBR) - A lower Green-House-Gas (GHG) option for industrial wastewater reclamation*, Chemosphere 140 (2015), pp. 129-142.

[6] L.H. Meng, C.C. Gao, L. Yu, G.P. Simon, H.S. Liu, and L. Chen, *Biodegradable composites of poly(butylene succinate-co-butylene adipate) reinforced by poly(lactic acid) fibers*, J. Appl. Polym. Sci. 133 (2016), p. 6.

[7] R.G. Wang, T.G. Ren, Y.X. Bai, Y.Z. Wang, J.F. Chen, L.Q. Zhang, and X.Y. Zhao, *One-pot synthesis of biodegradable and linear poly(ester amide)s based on renewable resources*, J. Appl. Polym. Sci. 133 (2016), p. 6.

[8] X.Y. Peng, Y.X. Zhang, Y. Chen, S. Li, and B. He, *Synthesis and crystallization of well-defined biodegradable miktoarm star PEG-PCL-PLLA copolymer*, Mater. Lett. 171 (2016), pp. 83-86.

[9] F.C. Ma, S. Chen, P. Liu, F. Geng, W. Li, X.K. Liu, D.H. He, and D. Pan, *Improvement of beta-TCP/PLLA biodegradable material by surface modification with stearic acid*, Mater. Sci. Eng. C-Mater. Biol. Appl. 62 (2016), pp. 407-413.

[10] Y.H. Wu, X.G. Luo, W. Li, R. Song, J. Li, Y. Li, B. Li, and S.L. Liu, *Green and biodegradable composite films with novel antimicrobial performance based on cellulose*, Food Chem. 197 (2016), pp. 250-256.

[11] P.K. Qi, Y. Yang, S. Zhao, J. Wang, X.Y. Li, Q.F. Tu, Z.L. Yang, and N. Huang, *Improvement of corrosion resistance and biocompatibility of biodegradable metallic vascular stent via plasma allylamine polymerized coating*, Mater. Des. 96 (2016), pp. 341-349.

[12] S. Soltani, U. Rashid, R. Yunus, and Y.H. Taufiq-Yap, *Biodiesel production in the presence of sulfonated mesoporous ZnAl2O4 catalyst via esterification of palm fatty acid distillate (PFAD)*, Fuel 178 (2016), pp. 253-262.

[13] J. Ahmad, S. Yusup, A. Bokhari, and R.N.M. Kamil, *Biodiesel Production from the High Free Fatty Acid "Hevea brasiliensis" and Fuel Properties Characterization*, in *Process and Advanced Materials Engineering*, I. Ahmed ed., Trans Tech Publications Ltd, Stafa-Zurich, 2014, pp. 897-900.

[14] S. Chattopadhyay, and R. Sen, *Fuel properties, engine performance and environmental benefits of biodiesel produced by a green process*, Appl. Energy 105 (2013), pp. 319-326.

[15] L.V. Rasmussen, K. Rasmussen, and T.B. Bruun, *Impacts of Jatropha-based biodiesel production on above and below-ground carbon stocks: A case study from Mozambique*, Energ. Policy 51 (2012), pp. 728-736.
[16] A. Fernandez, R. Rallo, and F. Giralt, *Prioritization of in silico models and molecular descriptors for the assessment of ready biodegradability*, Environ. Res. 142 (2015), pp. 161-168.
[17] L. Ceriani, E. Papa, S. Kovarich, R. Boethling, and P. Gramatica, *Modeling ready biodegradability of fragrance materials*, Environ. Toxicol. Chem. 34 (2015), pp. 1224-1231.
[18] P. Xu, W.C. Ma, H.J. Han, S.Y. Jia, and B.L. Hou, *Quantitative structure-biodegradability relationships for biokinetic parameter of polycyclic aromatic hydrocarbons*, J. Environ. Sci. 30 (2015), pp. 180-185.
[19] R. Boethling, *Comparison of ready biodegradation estimation methods for fragrance materials*, Sci. Total Environ. 497 (2014), pp. 60-67.
[20] A. Lombardo, F. Pizzo, E. Benfenati, A. Manganaro, T. Ferrari, and G. Gini, *A new in silico classification model for ready biodegradability, based on molecular fragments*, Chemosphere 108 (2014), pp. 10-16.
[21] A. Sabljic, and Y. Nakagawa, *Biodegradation and Quantitative Structure-Activity Relationship (QSAR)*, in *Non-First Order Degradation and Time-Dependent Sorption of Organic Chemicals in Soil*, W. L. Chen, A. Sabljic, S. A. Cryer and R. S. Kookana eds., American Chemical Society, Washington, 2014, pp. 57-84.
[22] S. Vorberg, and I.V. Tetko, *Modeling the biodegradability of chemical compounds using the Online CHEmical Modeling Environment (OCHEM)*, Mol. Inf. 33 (2014), pp. 73-85.
[23] M.P. Gonzalez, C. Teran, L. Saiz-Urra, and M. Teijeira, *Variable selection methods in QSAR: an overview*, Current topics in medicinal chemistry 8 (2008), pp. 1606-27.
[24] A. Yasri, and D. Hartsough, *Toward an optimal procedure for variable selection and QSAR model building*, J. Chem. Inf. Model. 41 (2001), pp. 1218-27.
[25] S. Gupta, and J. Aires-De-Sousa, *Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness*, Mol. Divers. 11 (2007), pp. 23-36.
[26] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second ed., Springer-Verlag, 2009.
[27] W. Bich, *Error, uncertainty and probability*, in *Metrology and Physical Constants*, E. Bava, M. Kuhne and A. M. Rossi eds., 2013, pp. 47-73.
[28] W. Bich, *Revision of the 'Guide to the Expression of Uncertainty in Measurement'. Why and how*, Metrologia 51 (2014), pp. S155-S158.
[29] W. Bich, M.G. Cox, R. Dybkaer, C. Elster, W.T. Estler, B. Hibbert, H. Imai, W. Kool, C. Michotte, L. Nielsen, L. Pendrill, S. Sidney, A.M.H. van der Veen, and W. Woger, *Revision of the 'Guide to the Expression of Uncertainty in Measurement'*, Metrologia 49 (2012), pp. 702-705.
[30] W. Bich, M.G. Cox, and P.M. Harris, *Evolution of the 'Guide to the Expression of Uncertainty in Measurement'*, Metrologia 43 (2006), pp. S161-S166.
[31] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, and V. Consonni, *Quantitative Structure–Activity Relationship Models for Ready Biodegradability of Chemicals*, J. Chem. Inf. Model. 53 (2013), pp. 867-878.
[32] D. Ballabio, and V. Consonni, *Classification tools in chemistry. Part 1: linear models. PLS-DA*, Anal. Method. 5 (2013), pp. 3790-3798.

[33] M.R. de Almeida, D.N. Correa, W.F.C. Rocha, F.J.O. Scafi, and R.J. Poppi, *Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation*, Microchemical Journal 109 (2013), pp. 170-177.
[34] B. Worley, S. Halouska, and R. Powers, *Utilities for quantifying separation in PCA/PLS-DA scores plots*, Anal. Biochem. 433 (2013), pp. 102-104.
[35] J. Xia, and D.S. Wishart, *Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst*, Nat. Protoc. 6 (2011), pp. 743-760.
[36] F.B. Gonzaga, W.F.d.C. Rocha, and D.N. Correa, *Discrimination between authentic and false tax stamps from liquor bottles using laser-induced breakdown spectroscopy and chemometrics*, Spectrochim. Acta B 109 (2015), pp. 24-30.
[37] A.S. Luna, I.C.A. Lima, W.F.C. Rocha, J.R. Araujo, A. Kuznetsov, E.H.M. Ferreira, R. Boque, and J. Ferre, *Classification of soil samples based on Raman spectroscopy and X-ray fluorescence spectrometry combined with chemometric methods and variable selection*, Anal. Method. 6 (2014), pp. 8930-8939.
[38] H. Chun, and S. Keleş, *Sparse partial least squares regression for simultaneous dimension reduction and variable selection*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (2010).
[39] S. Wold, M. Sjöström, and L. Eriksson, *PLS-regression: a basic tool of chemometrics*, Chemometr. Intell. Lab. 58 (2001), pp. 109-130.
[40] Eigenvector, *PLS Toolbox 4.0*, 2006.
[41] M.X. Rodriguez-Alvarez, L. Meira-Machado, E. Abu-Assi, and S. Raposeiras-Roubin, *Nonparametric estimation of time-dependent ROC curves conditional on a continuous covariate*, Stat. Med. 35 (2016), pp. 1090-1102.
[42] W. Aziz, M. Rafique, I. Ahmad, M. Arif, N. Habib, and M.S.A. Nadeem, *Classification of heart rate signals of healthy and pathological subjects using threshold based symbolic entropy*, Acta Biol. Hung. 65 (2014), pp. 252-264.
[43] M. Leila, and S. van de Geer, *On threshold-based classification rules*, Lecture Notes-Monograph Series 42 (2003), pp. 261-280.
[44] R. Beran, *Refining Bootstrap Simultaneous Confidence Sets*, J. Am. Stat. Assoc. 85 (1990), pp. 417-426.
[45] D. Burr, *A comparison of certain bootstrap confidence intervals in the Cox model*, J. Am. Stat. Assoc. 89 (1994), pp. 1290-1302.
[46] H. van der Voet, *Pseudo-degrees of freedom for complex predictive models: the example of partial least squares*, J. Chemometr. 13 (1999), pp. 195-208.
[47] R. Wehrens, H. Putter, and L.M.C. Buydens, *The bootstrap: a tutorial*, Chemometr. Intell. Lab. 54 (2000), pp. 35-52.
[48] R.S. Boethling, *Designing biodegradable chemicals*, in *Designing Safer Chemicals*, American Chemical Society, 1996, pp. 156-171.
[49] S.C. DeVito, and R.L. Garrett, *Designing Safer Chemicals*, Vol. 640, *ACS Symposium Series*, American Chemical Society, 1996.

Table 1. List of molecular descriptors used for the QSAR PLS-DA Model

Symbol	Description	DRAGON block
B01[C-Br]	presence/absence of C−Br at topological distance 1	2D atom pairs
B03[C-Cl]	presence/absence of C−Cl at topological distance 3	2D atom pairs
B04[C-Br]	presence/absence of C−Br at topological distance 4	2D atom pairs
C%	percentage of C atoms	constitutional indices
F03[C-O]	frequency of C−O at topological distance 3	2D atom pairs
F04[C-N]	frequency of C−N at topological distance 4	2D atom pairs
HyWi_B(m)	hyper-Wiener-like index (log function) from Burden matrix weighted by mass	2D matrix-based
LOC	lopping centric index	topological indices
Me	mean atomic Sanderson electronegativity (scaled on Carbon atom)	constitutional indices
Mi	mean first ionization potential (scaled on carbon atom)	constitutional indices
N-073	Ar2NH/Ar3N/Ar2N−Al/R···N···R	atom centered fragments
nArNO2	number of nitro groups (aromatic)	functional group counts
nCIR	number of circuits	ring descriptors
nCRX3	number of CRX3	functional group counts
nN-N	number of N hydrazines	functional group counts
nO	number of oxygen atoms	constitutional indices
Psi_i_1d	intrinsic state pseudoconnectivity index−type 1d	topological indices
SdO	sum of dO E-states	atom-type E-state indices
SM6_L	spectral moment of order 6 from Laplace matrix	2D matrix-based
SpMax_A	leading eigenvalue from adjacency matrix (Lovasz−Pelikan index)	2D matrix-based
SpMax_L	leading eigenvalue from Laplace matrix	2D matrix-based
SpPosA_B(p)	normalized spectral positive sum from Burden matrix weighted by polarizability	2D matrix-based
TI2_L	second Mohar index from Laplace matrix	2D matrix-based

Table 2. Statistical metrics for the QSAR PLS-DA model

	CALIBRATION		VALIDATION		EXTERNAL VALIDATION	
CLASS	RB	NRB	RB	NRB	RB	NRB
N	284	553	72	146	479	191
$N_{MISCLASS}$	34	93	12	19	69	38
ME (%)	11.97	16.81	16.67	13.01	14.405	19.89
TP	0.88028	0.83183	0.8333	0.86986	0.80105	0.85595
FP	0.16817	0.11972	0.13014	0.16667	0.14405	0.19895
TN	0.83183	0.88028	0.86986	0.83333	0.85595	0.80105
FN	0.11972	0.16817	0.16667	0.13014	0.19895	0.14405
SENS	0.880	0.832	0.833	0.870	0.801	0.856
SPEC	0.832	0.880	0.870	0.833	0.856	0.801
R	0.6457		0.6530		0.6063	
RMSE	0.361516		0.356293		0.3676	

N: number of substances in each class.
$N_{misclass}$: number of misclassifed substances.
ME (%): misclassification error;
TP: true positive;
FP: false positive;
TN: true negative;
FN: false negative;
Sens: sensitivity;
Spec: specificity.
R: Pearson's correlation coefficient for calibration, validation, and external validation.
RMSE: root mean square error for calibration, validation, and external validation.

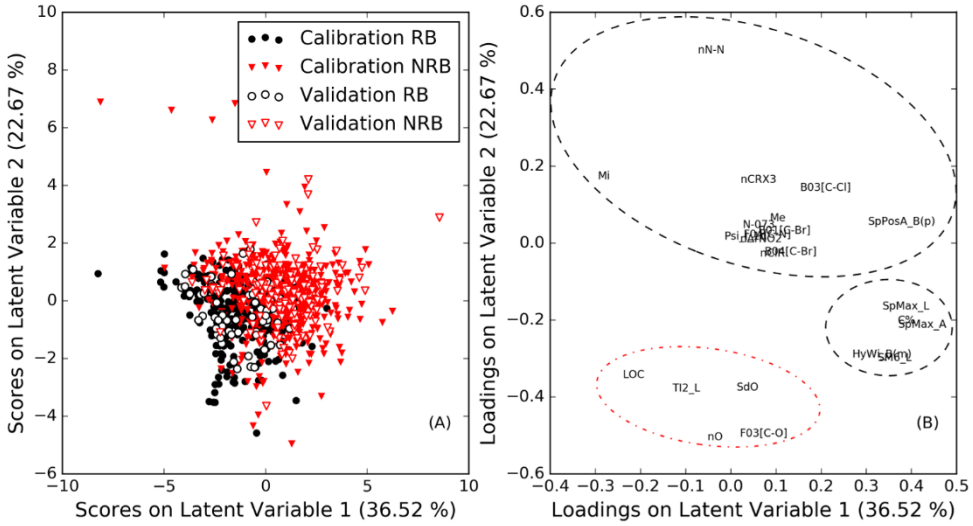

Figure 1. Scores plot (A) and loadings plot (B) with respect to the first and second latent variables of the PLSDA model. Molecular descriptors refer to symbols listed in Table 1. Descriptors responsible for identifying NRB substances are circled with the dashed line and those responsible for identifying RB substances are circles with the dash-dot line.

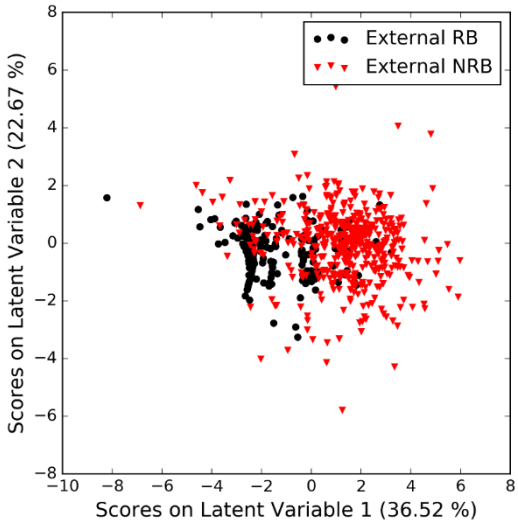

Figure 2. Scores plot with respect to the first and second latent variables of the PLS-DA model for external validation substances.

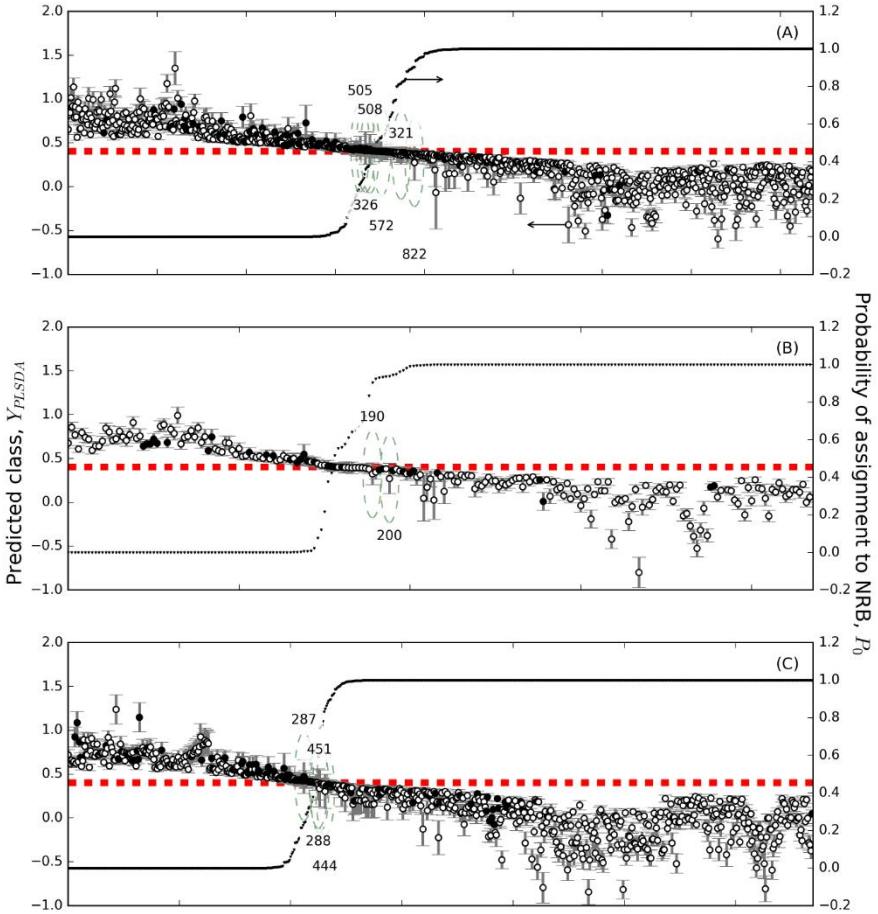

Figure 3. Classes predicted by the PLS-DA model for the calibration, validation, and external validation sets, with confidence intervals for all substances as estimated by the residual bootstrap method. The class boundary y_{bound} is shown with a dashed line and the probability of assignment to the NRB class is shown with dots. Predicted RB substances have confidence intervals above y_{bound} and predicted NRB substances have confidence intervals below y_{bound}. Substances that are correctly classified are shown with open circles and those that are incorrectly classified are shown with filled circles. (A) Calibration substances, (B) Validation substances, and (C) External Validation substances.

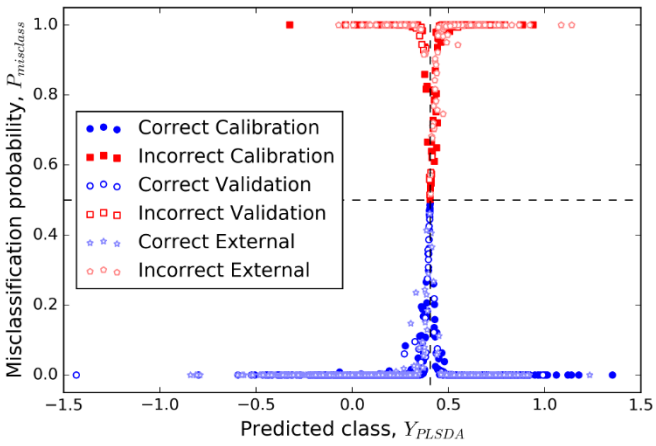

Figure 4. Misclassification probabilities with respect to classifications predicted by the PLS-DA model. The vertical dashed line indicates y_{bound}, with RB substances to the right and NRB substances to the left. The horizontal dashed line indicates $P_{\text{misclass}} = 0.5$, corresponding to the boundary between correctly-classified and misclassified substances.