University of Utrecht

Thesis Msc — Artificial Intelligence

Chemical Similarity Screening With Machine Learning and Active Learning Using Physical Chemical Properties

Author:
Bendik Flåt Aas
(SolisID:6602630)

Supervisors:
Ass.-Prof. Dr. Habil.
Georg Krempl
dr. Albert Wong
dr. Emiel Rorije
Co-reader:
Ass.-Prof. Dr. M. Thijs
Van Ommen

June 17, 2020

Universiteit Utrecht

Contents

1 Abstract

Toxicology is a field plagued by lack of experimental data and labels, as assessment of chemical toxicity is a time-consuming and costly process, all the while release of new substances grow in number. This further strengthens the need for robust screening tools and models for classification. Chemical similarity screening traditionally include a two-dimensional fingerprint representation of a chemical sub-structure, in which a distance measure between fingerprints determines similarity. This approach neglects potential importance ordering for sub-structures. The novelty of the approach presented in this paper aims to model so-called persistent, bioaccumulative and toxic(PBT) substances based on their physical chemical properties, and whether such an approach is an improvement over related fingerprint based approaches. Aims further include to inspect whether feature importance match *a priori* expert expectation, and whether the results could be improved by application of active machine learning. Two baseline machine learning models were fit to naive and filtered physical chemical data in the form of Random Forests and Support Vector Machine. The best performing model achieved a 94.28%classification accuracy, and was also able to pick up on existing legal guideline thresholds for substance evaluation. Further hypothesis of expert feature importance was showed to be true, with added importance for features previously not considered. Further utilizing a curious machine learning algorithm named Active Learning, it was shown that a similar accuracy could be achieved with 40-50% less data used, with a demo for interactive annotation with a chemical expert that could serve as a cross-referencing check on expert chemical evaluation. Albeit in need of further confirming data, the main contribution of this paper is the novel approach of using physio-chemical data, showing the value of utilizing machine learning algorithms as a tool for the classification of harmful chemicals.

Key Words: *Toxicology, Machine Learning, Chemical Similarity, Random Forest, Support Vector Machine, Active Learning, PBT substances*

2 Introduction

Chemical usage permeates the products we both consume and develop in modern society, such as agricultural pesticides, plastics, food additives and cosmetics. However, these same chemicals may carry adverse effects for health and the environment, and thus require evaluation prior to their usage. Screening potential toxic chemicals is a time consuming and rigorous process, and at the same time, more are released on the markets every year. How can artificial intelligence help speed up the process of classifying new chemicals? Technological advances have created new opportunities for research in general. First, new measurement methods and storage technologies have stimulated the collection and availability of data in general. Second, improved computing power has made it feasible for machine learning methods to be implemented in practice. These advances are also important for the field of 'predictive toxicology'. This predictive work is an important task within the RIVM, the Dutch national institute which concerns itself among other things with the evaluation of chemical substances in terms of their (public) health and environmental effects. The process of evaluation has always been complicated by a lack of experimental and observational data from which the effect can be directly inferred(That is, many substances have not been directly assessed in terms of their health effects.) However, when we assume that the toxicity of a substance can be predicted well by considering other substances with similar characteristics, we might be able to support this evaluation with evidence from data. This is also where the aforementioned advances can play an important role. More specifically, in recent years increasingly large databases containing toxicological parameters and endpoints such as a chemical's environmental fate has become available. It is in the interest of the RIVM to investigate the value of using new databases and machine learning techniques in the field of predictive toxicology, both as novel avenues of studying the chemical space and as added tools for substance evaluation. A further matter of salience is the legislative nature of the task, where any model that may be used for further legal justification in a chemical toxicity ruling. Thus, the methods would require a certain level of transparancy.

The goal in this paper is to investigate the capabilities of using machine learning for the classification of so-called persistant, bioaccumulative and toxic (PBT) substances. PBT substances are chemicals that do not easily degrade in their immediate released environment, or further bioaccumulate in biological systems such as humans or fish. More information about this can be found in the related works for the uninitiated reader. The project aims to explore the following research questions: Is there an added benefit of modelling based on physio-chemical properties of substances over the more traditionally used structure-activity analysis and two-dimensional binary fingerprint comparison. Further, as expert evaluation of chemical substances involve an *a priori* expectation of variable importance, does a naive modelling approach match these expectations, and if not, to what extent does ranked feature importance for a machine learning model deviate from expert ranking? What further modelling options are well-suited for such a task? Are they explainable? Further, a point of interest is to inspect the relative impact of persistence and bioaccumulation features, and to what extent the different PBT criteria play in determining the outcome of a label. This includes modelling approaches that includes for the nuance of different labels beyond a binary approach. The notion of similarity then lies in which way a substance's relative impact on its environment shares characteristics with those that are known to be harmful, and those that are known to be safe. Finally, in a field where experimental data can be lacking or insufficiently labeled, how can active machine learning be used to increase the quality of prediction and data sets with selective sampling?

To begin with, a background on both the field of toxicology and concepts of machine learning will be introduced, as well as the related work done in the field, as the target audience for this thesis is mixed. The methods to be used for the paper will then be presented, along with criteria of model evaluation. Further, the data to be used in the paper will be described, along with the physiochemical properties and their explanations for the sake of clarity. Finally, results will be presented along with a discussion and making remarks for improvement and limitations before reaching a conclusion on stated research questions.

3 Related work

The following sections include background information on work done within predictive toxicology, methods used, and a review of literature on the active learning approach and its intuition.

3.1 Toxicology

3.1.1 Background

The field of toxicology is a scientific field dedicated to the study and evaluation of substances that have adverse effects on biological systems and the environment. It carries a staunch overlap with the fields of chemistry, pharmacology and medicine, with a focus on rigorous study and judgement of said substances and their toxic effects. This judgement is of a legislative sort, thus within toxicology, there is a need for robust and quality models for classification of toxic chemicals that have various affects in biological systems. A matter of salience here is that chemicals are widely used in the production of products that we frequently use and permeates our daily lives, which in turn carry effects on the environment, our health and well being. Within food production alone there is food additives, pesticides, fertilizers, and further in commercial products like toys, furniture, plastics and prescriptive drugs. The degree to which how much a chemical substance can be present in a given product is mediated by law. These laws are determined by carefully studied evaluations and testing of substances. To get a sense of scope to the problem, The American Environmental Protection Agency(EPA) estimate that there is about 2000 new chemicals introduced to the market each year[40][1], however only a few of these are processed at a time under the new Frank R. Lautenberg Chemical Safety for the 21st Century Act signed in 2016[5], thus rendering the endeavour of evaluating and classifying chemicals a time-consuming and costly process, all the while being a pressing matter. The EPA has made this a point of prioritization, and have developed a research program dedicated to utilize an already existing list of structural classes and chemical subgroups to be used in future classification and evaluation. The program is entitled "ToxCast"[15]. This problem, however, is not unique to the United States.

The EU has taken its own regulatory steps in controlling and measuring chemical usage. Headed by initiatives such as REACH[17] and CLP[18] directives under the EU and European Chemicals Agency(ECHA)[20], European manufacturers and exporters of chemical substances need to classify and label their chemicals through strict guidelines. This labeling includes a thorough account of its environmental effects and its toxic features. Controlling these procedures is no small task, as the European Chemical Industry Council measure the chemical industry to be the fourth largest in Europe according to their 2018 industrial report[10]. A salient matter is the set of categories that a given substance might

[1] last accessed December 17th, 2019.

fall into. Chemicals denoted as Substances of Very High Concern(SVHC) are banned under the directives of aforementioned initiatives and are not available for production or distribution. There are further sub-categories to the substances that are categorized as SVHC that carries different properties that are of concern. Some substances have so called CMT properties for short, meaning that they are carcinogenic, mutagenic, and reprotoxic. They are known to have chronic effects on health and are often grouped together as they carry similar classifications and legal action. Further, there are the mentioned PBT substances. An important note here is that the label of toxic is not necessarily directly derived, rather, it is something that is concluded based on two other priors, namely how persistent it is, how bio-accumulative it is, or a combination of the two. Some substances are thus not consider toxic as they are neither bio-accumulative or persistent, however a toxic chemical can be non-persistent yet bio-accumulative and vice versa. If above a certain threshold of persistence or bioaccumulation, substances can be labeled as PBT/vPvB, or very persistent and very bio-accumulative. Finally, there are substances that are hormonal disruptors(ED), in that they may change the hormonal balance in systems that are driven by hormones. Outcome of significant hormonal disruption may lead to birth defects and adverse development problems and disorders. This paper is more concerned with PBT substances and the development of a screening tool for such substances.

3.1.2 Machine Learning in Toxicology

The use of machine learning models in chemistry and toxicology is a growing toolbox of models and approaches that span a wide variety of techniques, ranging from molecular drug-target interaction to toxicity classification. These so-called *in silico* studies attempt at giving detailed accounts of chemical properties and interactions on a computational level, whereas historically chemicals have been studied *in vitro*, or in careful and rigorous lab environments. As highlighted in the book "Computational Toxicology—A State of the Science Mini Review"[23], as the field of toxicology gets increasingly acquainted with state-of-the-art modelling techniques, one important aspect that computational modelling within toxicology provides is scale. Scale in directions and breadth of approaches across biological organization levels, drug-complexities and dosage discrimination to name a few. These models often end up being informative supplements to a field that is drive by legislative decision making, thus making effective screening tools worthwhile in segmenting elements of chemical components for further scrutiny.

There are numerous in silico models that have been developed, such as the prediction of complex chemical reactions at the mechanistic level using machine learning[24], where the authors model chemical interaction on a molecular level using a two-step machine learning algorithm. There are drug-drug interaction models that study adverse affects in drug administration, or multi-drug administration in clinical trials in a world with increasing healthcare costs[11]. An overarching goal for these models is to study not only the effects of chemical compounds in the body, but increase the quality and effectiveness of healthcare or the safety of chemical use, exemplified further by the prediction of chemical acute oral toxicity using a variety of classification methods[29]. More recent applications include state-of-the-art neural nets to model Quantitative Structur-Activity Relationships(QSAR) - in which the goal is to capture molecular activity and reactivity in a predictive manner[54] for thousands of data entries.

Not only does the machine learning approach increase the rate and efficiency of testing, but may also allow for a reduction on the number of tests that need to be made, and consequently reduce harm on test subjects. A motivating example here is from the world of cosmetics. People's eyes or skin are subject to exposure of chemicals in various ways, and the safety of a chemical is often studied through testing of animals. One such test is The Draize Eye Test[52]. The Draize test was developed to study effects found in chemicals used in products like cosmetics, such as eye-irritation or corrosive traces. Testing is done through the use of rabbits, in which the eyes of the rabbits are subjected to the chemical being tested to check for adverse outcomes. Not only are the rabbits subjected to numerous tests, but also repeated tests for chemicals deemed to be similar to previously studied ones, which in certain instances can amount to 90 repeated rabbit eye tests[51] to determine proper labels for a given cosmetic. Further, the cost of in vivo tests of chemicals is estimated to exceed 68,000 entries under the REACH legislation, rendering the usefulness of predictive models stronger. Studies on these substances have been made, in which it was found a hypothetical estimate of 54 million animals was in demand for extensive testing of effects over all possible categories ranging from respiratory irritation to developmental neurotoxicity[37]. The overall cost amounted to a hypothetical estimate of 9.5 billion euro. Notably, findings like these contributed to the EU 7$^{\text{th}}$ Amendment to the Cosmetic Directive[19], effectively banning animal testing for new cosmetic ingredients, further adding weight to the need for more efficient modeling tools.

With this in mind, an initial analysis of publicly available REACH Draize Eye Irritation test data was made, in which the data dossiers dated from 2008-2014 were data mined and inspected in 2016[32]. The dossiers contained 9,782 Draize Eye Tests that had been performed on 3,420 unique substances, pointing to the high amount of repeated tests mentioned earlier. The authors then assessed the reproducibility of these tests, in which there was 10% estimate chance of classifying a previously labeled irritant as a non-irritant according to the relevant classification criteria under the UN GHS system[2]. The most reproducible tests where outcomes that were labeled as negative at 94% and severe eye irritants at 73%. Reproducibility was determined by a probability estimate of one outcome of a Draize test would give the same outcome in a subsequent Draize test. Having established a notion of reproducibility, the authors explored whether other classification criteria under the UN GHS could be used to expand upon the quality of prediction towards eye irritation with a differently constructed data set. Overall their classification methods had valuable results that called for further exploration, leading the same authors to data mine REACH data on skin sensitization from the same time period[31]. The work done in both of these studies accumulated to the development of a machine learning algorithm for toxicological big data using Read-Across Structure Activity Relationships(RASAR)[33]. Since the range of possible molecular activity is so vast, it is hard to establish or derive complex rules for chemical structures by both human and computational means. Read-across approaches like RASAR utilizes a pragmatic approach, in which a case-by-case comparison to similar chemicals is done to determine structural relationships. To construct the RASAR, they utilized both supervised and unsupervised learning steps. For the unsupervised step, an exhaustive comparison of the distance between one chemical to another is done for the entire set. After the similarities have been constructed, local graphs were made for each component that describes the distance to other chemicals surrounding the one of particular interest using Jaccard similarity.

The final unsupervised step was to apply an aggregate function to the local graph to generate feature vectors, and based on these vectors, KNNs would create n-dimensional vectors based on the number of times n number of labels would occur in the k closest neighbours. The authors further used these vectors in a supervised learning step using simple RASAR and random forests with data fusion, the data fusion being an extension of the simple RASAR by incorporating more similarity feature vectors to each catalogued compound. Not only did they include positive similarities, but also analog negative similarity feature vectors, and general known hazard feature vectors. In the end, both simpler and more complex models were able to show predictive capability rivaling standardized animal testing in the REACH set they explored in 2016 of roughly 70-82%+ balanced accuracies across the different categories in the UN GHS initiative. The results lend credence towards computational models as a means of both reducing harm and efficiently determining risk. There were however some caveats, among them that the specificities, or true false predictions, were lower in the simple RASAR than they were for repeated animal tests. This was further remedied in the data fusion model when further feature vectors were included. In addition, chemicals are not always on equal footing in chemical complexity and variance, rendering the reproducibility of a given test not independent of the chemical that was being tested. Chemicals that are soluble may be easier to reproduce than those that are insoluble in eye irritation tests, which is a further point of improvement for the future in establishing a framework or rubric of appropriate testing, however a noteworthy improvement was made.

3.1.3 RIVM screening tool

The notion of chemical similarity may also be used to separately classify subcategories of harmful chemicals[50]. Previously mentioned CMT, PBT/vPvB and ED substances are here the target of consideration, in which a range of researched similarity measure sets[53] are paired with a 2D chemical structure representation. This representation can be described as a binary bit-string, also called a fingerprint, in which chemical substructures are either given a 1 if it is present in a given substance, or 0 otherwise. Similarity is again given as an estimate of structural overlap between substances that carry adverse properties with known SVHCs. A total of 957 substances were selected from a Dutch list that is in legislative accordance with SVHC criteria under REACH, the data partition being 546 SVHCs and 411 non-SVHCs. The criteria for a best model selection is based on its balanced accuracy for all possible fingerprint-coefficient pair in the set, which totalled at 112 different measures given the existence of 16 various fingerprint representations and 7 similarity coefficients. The selection of these coefficients were based on work previously done[46].

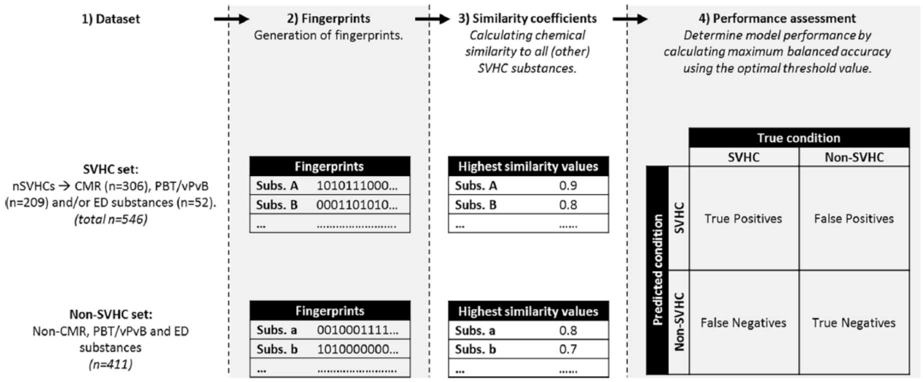

1) Dataset
2) Fingerprints
Generation of fingerprints.
3) Similarity coefficients
Calculating chemical similarity to all (other) SVHC substances.
4) Performance assessment
Determine model performance by calculating maximum balanced accuracy using the optimal threshold value.
SVHC set:
nSVHCs → CMR (n=306), PBT/vPvB (n=209) and/or ED substances (n=52).
(total n=546)
Non-SVHC set:
Non-CMR, PBT/vPvB and ED substances
(n=411)
Fingerprints
Subs. A 1010111000...
Subs. B 0001101010...
...
Highest similarity values
Subs. A 0.9
Subs. B 0.8
...
Fingerprints
Subs. a 0010001111...
Subs. b 1010000000...
...
Highest similarity values
Subs. a 0.8
Subs. b 0.7
...
True condition
SVHC
Non-SVHC
Predicted condition
SVHC
Non-SVHC
True Positives
False Positives
False Negatives
True Negatives

Figure 1: *Experimental setup of screening tool developed at the RIVM From [50]*

A worry, as expressed by the authors in discussing model performance, was whether bias would be introduced for smaller substances in their similarity coefficient with known small SVHC substances. Smaller substances would have a lot of chemical substructures not present, marked as 0, and thus would carry a significant amount of overlap with other small substances for the same reason, rendering substances erroneously as SVHCs. Thus for these substances, when such bias was detected in a model, an asymmetric similarity coefficient served as correction for the problem. Further, a couple of stability tests were done on well-performing models to verify results. The initial check was done by adding any non-relevant substance pertaining to a particular category as a non-SVHC substance. For example the ED model, all substances of CMT and PBT/vPvB that did not show ED characteristics were considered non-ED, and thus added to the non-SVHC data set, a check that only works for subcategory models. As a second robustness check, group representation structures were reduced, overall leading to a reduction in the number of substances for each category. Although no cross-validation was used in the training process, similarity was determined by a leave-one-out methodology in comparing one chemical to all other chemicals with optimized threshold values that exceeded 0.8 on all best-performing models. Noteworthy is it that the threshold is determined through an optimization processes of selecting the best pairing of representation and similarity measure. As such, if a similarity coefficient outputs a 1, the substances, or fingerprints, are identical, whereas in a scenario where the ouput is 0, a total dissimilarity is the case. Results show that for an overrall model including all subcategories, a variety of fingerprints showed up in the top-performing models albeit with a recurring similarity coefficient called simple matching(SM):

$$s = \frac{c+d}{c+a+b+d} \tag{1}$$

In this measurement, a denotes that a structure is present in fingerprint X while not present in fingerprint Y, b the inverse(present in Y, not in X), c present in both fingerprints and finally d in which it is present in neither of the two. This is different from a commonly used similarity coefficient called Jaccard-Tanimoto(JT), which is similar to that of formula (1), only with the removed term d.

The best performing model was a PubChem fingerprint of bit size 881 with a SM coefficient combination. The model was able to identify SVHCs with a balanced accuracy of 84.6% with a 0.985 similarity threshold, however as the authors note, not the best combination for identification of sub-groups. Different couplings provided varying results for the different subgroups, yet all of them showed a higher balanced accuracy than the general model. For example, balanced accuracy for the ED set of substances was at 0.99 with FCFP4-SM coupling. A note to the reader is that different fingerprints are shown to on average have similar retrieval measures[16], so the majority of tweaks done to this model was determining a similarity threshold value. A caveat to the model performance on the ED category is that 52 ED substances were present in the set, and are known to be particularly similar in structure, hence the very strong correlation and predictive performance for these models(all models were above 0.9). Some diversity between groups showed up as well, notably that a MACCS style fingerprint were better for PBT/vPvB substances. Nevertheless, it was proven to be an efficient screening tool for the purpose of detecting structural similarity to known SVHCs.

Work done here naturally is a benchmark for comparison for the findings in this paper, as it carries a staunch overlap in the data used and is a model that is utilized by the RIVM.

The research explored above spans across different levels of chemical organization and abstraction to predict molecular interaction and potential toxicity classification. However, nothing out of the literature points towards the use of physical properties alone to classify potentially harmful substances. The novelty of this approach aims to be a first step in a direction of scrutinizing chemical substances from another angle. Labeling chemicals further comes with some modelling decisions to determine how a chemical by legal standards can be determined. A comparison of such machine learning algorithms show several methodologies as applicable for bioassay data commonly used *in vivo* chemical classification, even when accounting for noise, feature-selection, feature irrelevance and imbalances in labeled data[23]. Models frequently used for classification include classical machine learning models such as a support vector machine(SVM), classification trees, random forests(RF), k-nearest neighbor(KNN) and naïve Bayes(NB). This not only informs modelling decisions for the topic of this paper, but further sheds light on the capabilities of machine learning techniques for the development of potential screening tools within the field of toxicology.

3.2 Active Learning

The following is a brief overlook of the intuitions that go in to active machine learning and its use-cases. This is an part so the reader may catch the intuitions for the related work done. A more in depth description of active learning specifics can be find in 4.3.

3.2.1 Background

In certain machine learning paradigms, data may lack proper labels or complete descriptions. This is a problem within data mining, say, where a large business may have missing or incomplete customer data. The company would nevertheless like to classify or make predictions on customer behavior. A further example is within medical data, where we want to suggest a treatment for a patient based on a narrow subset of features, and we may lack information of other crucial features that are complex and expensive to compute. Data sets might further have an imbalance in the amount of labeled experimental data available, which is a common theme within the field of toxicology described in this paper. As explored earlier, thousands of chemicals remain unclassified whereas only a tiny speck is labeled or may have a legislative instruction. This could pose a challenge when wanting to developing robust models. Active learning is a sub-field of machine learning partly researched to able to deal with such challenges. Active machine learning algorithms aim to reduce otherwise tedious labeling costs and uncertainties by introducing the element of *choice* from the standpoint of the algorithm. In other words, the model chooses what data it trains itself on, and thus may drastically reduce the amount of labeled data needed to acquire performance that rival other passive algorithms[27][45][42][58]. This curious approach is a form of semi-supervised learning that consults an *oracle*, where the oracle is some domain expert, or human annotator. Consulting an oracle works as follows; during training for a classification problem, there might be a specific set of unlabeled instances that are hard to determine a particular label for. They might be on the border of what the machine perceives to be the decision boundary, or in a probabilistic sense, a 50/50 case. The idea is to select for, and label, data points that better relieves 'confusion' from the viewpoint of the model or provide the most information.

Figure 2 shows the intuition behind the AL framework. Each data point has a feature vector denoted $\overrightarrow{x}$ and an appropriate class label $y \in \{1, ..., C\}$, where C denotes the number of possible classes. The set $\mathcal{L}$ of labeled instances are subsequently filled from the set $\mathcal{U}$ of unlabeled instances through optimal selection of instances, denoted $\overrightarrow{x}_{opt}$. This loop continues under some budget or stopping criteria t, where t be after an n amount of queries or a specific incremental improvement threshold for the model.

```
function ACTIVE LEARNING
    L = {}
    model = init_classifier()
    while ($b = 1; b < t; b++$) do
        x* = active_learning(U, model)
        y = ask_oracle(x*)
        U = remove(U, {x*})
        L = append(L, {x*, y})
        model = train_classifier(L)
    End
```

Figure 2: *Active Learning pseudo code[25]*

The active learning algorithm further has a strategy of sampling x* from the instance space and further a query strategy to obtain informative information about them. In light of this, an overview of commonly used sampling strategies and query strategies will can be seen in the methods section. First, we examine some domains of use for active learning.

3.2.2 Related Work

Active learning has a wealth of applicable domain areas, such as a wealth of natural language processing problems[34], text classification[48] and image recognition[47]. Intuitively is the idea of nuances that exist to data within all three domains. Natural language processing and sentiment classification not only concerns the model from a syntactical standpoint, but further in semantical interpretation. These aspects can in turn be affected by culture or available training data, thus active learning can increase the robustness of the models we create and deploy. A plethora of cases utilize SVMs for classification, in which the active learning component involves data points being queried on instances located around the decision boundary. The work done highlights the capability for modelling that can drastically reduce the amount of labeled data needed for matching, or better yet improve the performance of so-called passive learning models for linear separation[7].

Active learning is further used in more recent deep learning projects, such as sentiment classification[57], or investigating named entity recognition within NLP[45], capable of at the very least matching their passive counterparts in performance with just using 25% of original training data. It has been further used in applications of sentiment classification, in which subtle language queues of comments or website reviews are categorized in different temperamental categories or moods. This is useful for businesses that want to somehow data mine areas of improvement through the eyes of the customer. Active learning also incorporates missing values, or active feature acquisition, for problems where data might be sparse or incomplete as mentioned earlier[56]. Here the authors suggests two single-pass operations over the data. First pass tries to impute missing values, second acquire about ones it is least confident about classifying. Incremental feature acquisition does this through a batch of misclassified examples, a few important features at a time.

The use of active learning within the field of toxicology is sparse compared to other domains mentioned, yet present in related fields of pharmacology and biochemistry. The classification of compound sets for structural activity relationships can reduce the cost of much needed experimental data and human expertise by using 10-80% of data that normally would be needed for a passive learning algorithm[27]. The motivation behind the work was the tedious process entailed by classifying thousands of compounds used in drug discovery, where in addition unsupervised clustering methods do not improve the problem. This is because the clustering methods do not take in to account the user-preferences of the experimenter. Thus, by using active learning the size of labeled samples are reduced. One way this paper separates itself from previously mentioned work is that active learning algorithms have in large parts been applied to binary classification problems, while here it was used to determine several sub-clusters as viable options for drug synthesis. To do this, a medicinal chemist compiled a training set for the model. The authors used uncertainty sampling mentioned earlier as a sampling strategy for their active learning algorithm on ten unlabeled substances. This was combined with several SVMs trained on forced binary classifications across all potential class assignments in the problem space. Thus, having paired possible forced binary classifications in the larger multi-class set, they calculate the difference between model m for every k class, where m is the SVM for class k in the larger set of K possible class assignments.

Further application of active learning in pharmacology found that the active learning approach lead to a discrovery of novel protein chemotypes that improved upon structure-activity models[35]. Active learning was here used to obtain optimal bioactive compounds of protein-protein interaction. This lead the authors to focus more on the bioassay data that the structure-activity SAR model compiled for the actively learned data points, effectively increasing drug-target research efficiency.

As with the area of toxicology in general, application of the the active machine learning approach on physical chemical properties has not receieved as much attention, and will further be a novel application in this paper. The overall pattern however is the findings that qualitative selection of training data can either improve or equal otherwise data-hungry passive models.

4 Methods

Modelling methods to be explored in this paper include two baseline models in the form of Random Forests and SVMs. Depending on the better performing model, a further active learning approach to that same model will be used to gauge the added value of an active learning approach, and whether the passive performance can be equalled or surpassed.
The choice of models further reflect the need for some notion of explainability and the overall trend seen with related work done within toxicology. As stated earlier, results could be used in a legislative manner and thus require some level of explanation, which is why briefly the working intuitions of the models will be given below. Should the reader want to seek out more information on machine learning topics, a great start would be ”Artificial Intelligence: A Modern Approach” by Stuart J. Russell and Peter Norvig[39].

4.1 Random Forest

A common task in machine learning or data mining is to build models for prediction of a class of an object based on some of its attributes. The use of the term *object* can be interpreted loosely here: It could be a customer, a transaction, an e-mail message, patient or in this case a chemical. Further, the class of such an object can be a plethora of things, such as:

- spam or not spam in emails
- good or bad credit score of a bank customer
- harmful or not harmful substances

For the sake of clarity, we can follow the example of a credit-scoring model used by the Classification and Regression Trees(CART) authors[9]. The idea of the random forest in this case is to fit a model based on customer data, such as their age, income and marital status. Figure 3 shows a tree fit to credit data on bank customers from table **??**. In the top root node, we see that there's a 50/50 split between good and bad credit scores. Intuitively, if we were to simply predict majority class - or any class in this case - we would get 50% of the cases wrong. This, from the viewpoint of the bank would be a case of malpractice, but would also be a weak classification model. The goal is to reduce the amount of errors one makes by finding the best *variable* that would maximize information about a class label, and further the best *value* split point for that variable. The goal is to reach terminal pure leaf nodes(box shaped figures) where all cases belong to a single class.

In this case, the first check is on whether an individual has an income larger or smaller than 36,000 a year. With the data used for this example, 3 such customers have that income and immediately one could conclude that for any customer, if their income is above 36,000 then they have a good credit score. For the remaining 7 customers, more checks need to be performed to evaluate

Record	age	married?	own house	income	gender	class
1	22	no	no	28,000	male	bad
2	46	no	yes	32,000	female	bad
3	24	yes	yes	24,000	male	bad
4	25	no	no	27,000	male	bad
5	29	yes	yes	32,000	female	bad
6	45	yes	yes	30,000	female	good
7	63	yes	yes	58,000	male	good
8	36	yes	no	52,000	male	good
9	23	no	yes	40,000	female	good
10	50	yes	yes	28,000	female	good

Table 1: *Bank data over 10 customers.*

what the most appropriate label is. The reason for this is that for individuals with an income of lower than 36,000, 5 of these are considered bad whereas two are considered good. Thus, a further check on age is done and finally on marital status to correctly classify all cases in the data set.

The quality of a split s in node t is thus defined as the reduction of impurity a split achieves,

$$\Delta i(s,t) = i(t) - \pi(L)i(L) - \pi(R)i(R) \tag{2}$$

where $\pi(L)$ denotes the proportion of cases sent to the left, and $\pi(R)$ the proportion sent to the right. A standard measure of impurity, and also to be used in this paper, is the Gini-index. For the two-class case, the formula is denoted as

$$i(t) = p(0|t)p(1|t) = p(0|t)(1 - p(0|t)) \tag{3}$$

Where the multi-class generalization is denoted as,

$$i(t) = \sum_j p(j|t)(1 - p(j|t)) \tag{4}$$

Once a tree model has been fit to the data, new unclassified customers are ”dropped” down the tree to get a label. With a random forest, a large collection of such classification trees are gathered, where the accumulated vote across all trees ends up being the final label for an instance.

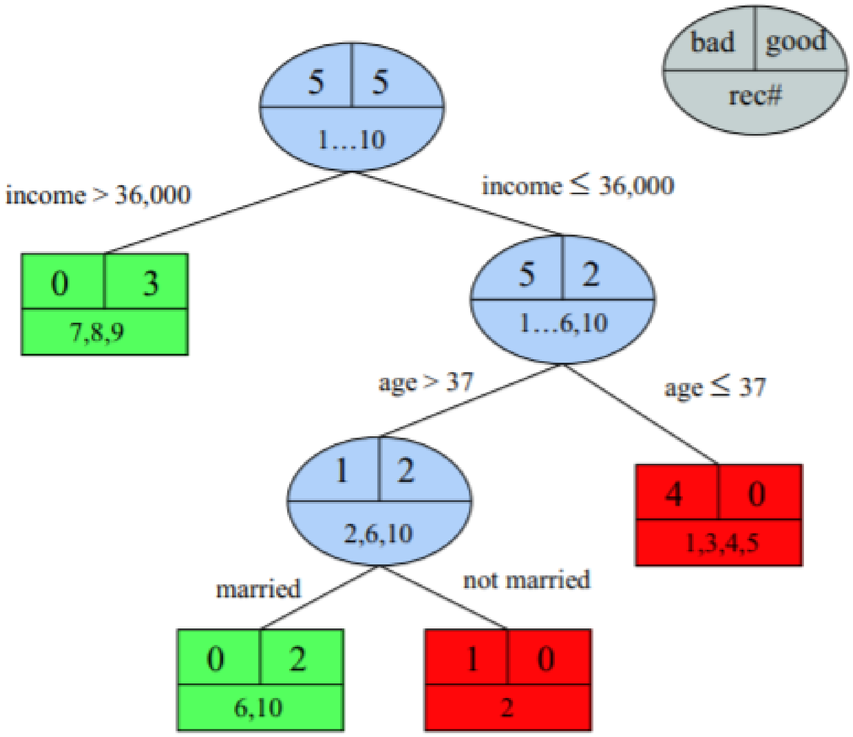


Figure 3: *Tree built for credit scoring data*

Why is this better than a single tree? By constructing several trees and aggregating on their vote, we reduce overall variance, as single trees are prone to model on noise and can overfit on our training data. We also introduce the notion of randomness in two ways. One, the model is trained on bootstrapped samples of the training data(also called bagging). A bootstrap sample from the training data is a sample with as many rows as the training set, where each row in the boostrap sample is selected with replacement from the training rows. A row may appear multiple times here, however this leads to different trees being made for different samples of the data, and as such we reduce the variance. A worry is whether each sample would create identical trees, which would make bagging redundant, however in a random forest algorithm, the randomness is taken one step further. The trees are "decorrelated" in a sense by introducing the condition that in *each split* in the tree construction, only a smaller random subset of columns are considered.

The number of trees we create in a random forest model or the number of columns considered at a split is something called hyperparameters, or conditions for training the model. Subset of columns is often denoted as *mtry*, where the number of mtry dictates the cardinality of variable consideration.

4.2 Support Vector Machines

The task of a SVM is to identify an optimal two dimensional line of separation between classes[8]. In a 2 dimensional problem space, the data can be separated by a single line, however with higher dimensionality we need a higher separating plane. We want to create an optimal decision boundary in a N-dimensional space, where N denotes the number of features that differentiate between distinct classes of objects, much like the example in figure 3. If we look at the example depicted in figure 4, the leftmost image depicts the number of possible decision boundaries that can be placed. The downside however is that some of these boundaries are weaker than others, i.e leave more room for new instances to be misclassified due to how close certain datapoints are to the decision boundary. In an SVM, the goal is to *maximize* the margin between the closest instances of each class, depicted in the figure on the right. Here, the instances for either classes that are the closest to a decision boundary are called the support vectors. The decision boundary that maximizes the distance between the closes samples of both classes is the optimal decision boundary. By maximizing the distance, future classifications can be done with more confidence.

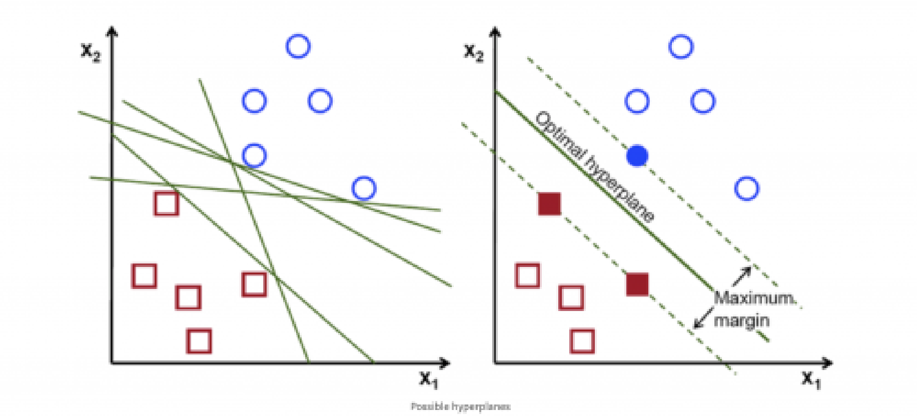


Figure 4: *Figure of linearly separable data. Figure on the left shows hypothetical decision boundary placements in an SVM. Figure on the right shows the application of using support vectors to create a maximum decision boundary*

This decision boundary can be formalized as,

$$D(x) = \sum_{i=1}^{N} w_i p_i(x) + b \tag{5}$$

Where decision boundary $D(x)$ is found by the sum multiplication over feature support vector w_i for each input data point x plus a bias term. More generally for the two dimensional case,

$$w * x + b = 0 \tag{6}$$

where if a data point (x,y) is on the hyperplane, then $w * x + b = 0$. If the data point is not on the hyperplane, then $w * x + b$ could be either positive or negative, or in other words, if it is negative it could be assigned to class 0 and inversely 1 for positive cases.

Not all problems are separable in two-dimensional space, and further does not allow for linear separability, like the problem depicted on the left in figure 5.

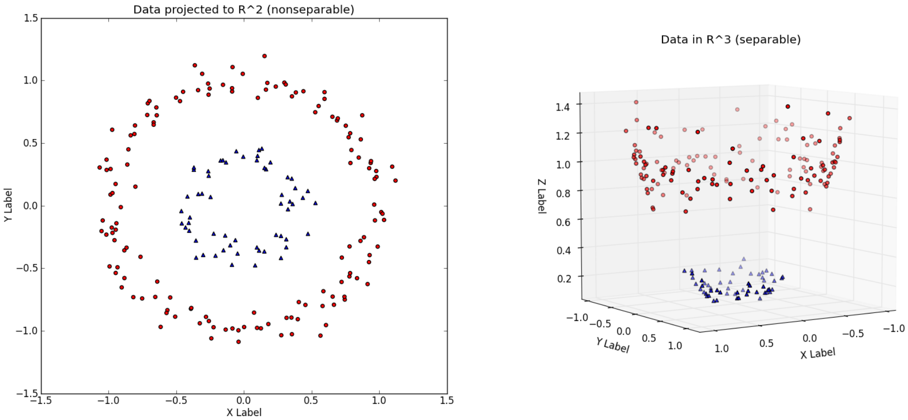


Figure 5: *Image showing how non-linear data can be separated by utilizing a kernel trick to suspend data to a higher dimension for linear separability. Image from [55].*

The solution for this is to suspend the data to a higher dimension to identify an optimal separating hyperplane, where in a N-dimensional feature space, the decision boundaries are computed using a kernel trick. Kernel methods allow for operations done in high-dimensional feature spaces like the project presented in this paper, without having to explicitly calculating expensive data coordinates. Instead of computing the coordinates across all feature dimensions, the kernel method computes the dot product between all pairs of data in the feature space for every class. The kernel is a similarity function

$$k(x_i, x^{'}) \tag{7}$$

where $x^{'}$ denotes an unlabeled instance and x_i all training instances. Thus, for a binary classifier using a kernel, the label is denoted as a weighted sum of similarities;

$$\hat{y} = lbl \sum_{i=1}^{n} w_i y_i k(x_i, x^{'}) \tag{8}$$

Where *lbl* denotes the negative or positive label to be given, and w_i the weights of the trained examples. These weights in a binary setting are the coordinates of input vectors orthogonal to the hyperplane seen in figure 4. The sign or direction of the vector indicates its class assignment. So for any new input point $p(x)$, the sign of the dot product between this point and the calculated support vectors becomes its new class assignment.

Given the nonlinear nature of our problem, a radial basis function will be used to compute 7, denoted as

$$k(x_i, x^{'}) = exp(\frac{||x - x^{'}||^2}{2\sigma^2}) \tag{9}$$

Where $||x - x^{'}||^2$ denotes a squared Euclidean distance between two feature vectors and σ a cut-off parameter for the Gaussian sphere. Increasing gamma means increasing the influence of individual training examples, which in turns affects the contortion and tightness of the decision boundary. This is a subject for tunable hyperparameters for the model in addition to the cost introduced to having a soft margin. The maximized decision boundary accounts for 0 samples of misclassifications inside the margin, i.e no instances are allowed within it. This is also called "hard margin classification". This however can lead to overfitting, as the decision boundary is based on weighted transformations of the support vectors closest to the line, or in other words a subset of datapoints. Thus, the goal is to monitor the amount of samples allowed inside the margin while simultaneously optimizing the fit of our decision boundary by introducing a reasonable level of slack. The cost parameter C arbitrates what the optimal value is for variance reduction. With a low value of C, samples within the margin is not penalized as hard as with higher values of C.

4.3 Active Learning

The following sections is an explanation over frameworks for how an active learning model picks a substance, and subsequently what strategy is in place for picking a particular data point for querying.

4.3.1 Sampling Strategy

A wider survey of active learning identifies three main query sampling scenarios[41]. Scenario (i) is called membership *query synthesis*[6], in which an algorithm may randomly select from all unlabeled instances in the input space, including the ability to query information that the machine finds as a point of interest. (ii) Stream-based selective sampling[13] amends the previous one by being more selective. Stream-based selective sampling regards the acquisition of an unlabeled instance as cost free, or inexpensive, since it first selectively samples a set of unlabeled instances, and then individually for each one decides whether to request a label or discard it. Finally, (iii) involves a pool-based sampling[28], in which a larger collection is gathered in a large set that is thought to be static, though not strictly necessary. From here, there is a measure of quality of information gained by selecting an instance from this set. The instance that would garner the most amount of unvertainty reduction is selected from the pool. The model then retrains and repeats the process until some performance metric is reached. The main difference between the pool-based and the stream-based approach is that in the former case, entire sets are under consideration, where as in the latter case instances are sequentially dealt with.

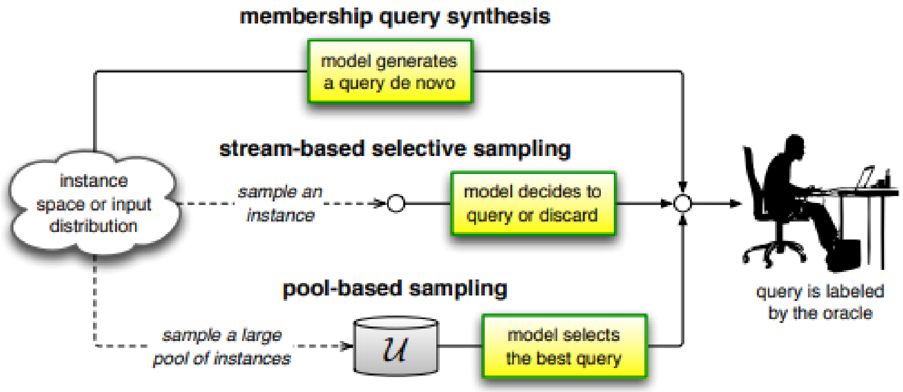

membership query synthesis
model generates a query de novo
stream-based selective sampling
instance space or input distribution
sample an instance
model decides to query or discard
query is labeled by the oracle
pool-based sampling
sample a large pool of instances
$\mathcal{U}$
model selects the best query

Figure 6: *Overview of the sampling process in an AL framework [41]*

4.3.2 Query strategies

Having determined a method of sampling, what follows is a query strategy with a specific goal in mind. By evaluating data points, the goal is to maximize the yield of information in selecting a specific instance. Most common here is *uncertainty sampling*[28]. As the name implies, the intuition here is that the active learner pays particular interest to instances is the most uncertain about, and requests further information about them. In a probabilistic sense, if there is an unlabeled instance that has a posterior probability for a given label of 0.5, ideally the learner will select this for a query. In other words, the machine will query instances that it is *least* confident about labeling, formulated as follows:

$$x^*_{LC} = \underset{x}{argmax}\ 1 - P_\theta(\hat{y}|x), \tag{10}$$

where X_A^* denotes the best query under a query algorithm A. $\hat{y} = argmax_y \, P_\theta(y|x)$ denotes the class with the highest posterior probability in the set. A downside to this however is that only the label with the highest likelihood is here considered, leading to a loss of information about other classes. A more general uncertainty strategy then uses marginal sampling, which takes in to account the posterior difference between the first- and secondnmost label:

$$x_M^* = \underset{x}{argmin} \; P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x), \tag{11}$$

The final variation that will be explored in this paper is Entropy sampling[44]:

$$x_H^* = \underset{x}{argmax} \; -\sum_i P_\theta(y_i|x) log P_\theta(y_i|x), \tag{12}$$

where y_i ranges over all possible labels that can be assigned to an instance. Entropy is often used in machine learning as a measure of impurity or uncertainty especially in trees[21], and known for its ease of implementation for multi-class problems. This is not to say that the use is limited to classification problems, however for the purposes of this paper, the application to regression problems will not be further explored. The use of entropy here is particularly mentioned as it commonly associated with the models chosen for the problem in this paper. However, there are other strategies that is worth mentioning for the sake of completeness.

A further strategy is called 'query-by-committee'[43]. The idea behind the strategy is that a 'committee' of models is stored, and instances that garner the highest amount of disagreement is the ones to be queried. The prerequisite here is that all models have been trained on the same labeled set, while carrying competing hypotheses of appropriate labeling. How model disagreement is quantified has been proposed in two ways, first up is the *Vote Entropy*[14]

$$x_{VE}^* = \underset{x}{argmax} \; -\sum_i \frac{V(y_i)}{C} log \frac{V(y_i)}{C}, \tag{13}$$

where as in entropy sampling, $V(y_i)$ ranges over all possible labelings and the cardinality of votes from a model a given labeling carries. This in a way can be seen as a ensemble approach to that of 12. and then through some mathematical function the notion of disagreement is quantified. Another strategy is to consider how big of a change is made to the model if we were to know its label, also called *Expected Model Change*[42](Settles et al., 2008), where the intuition is to query instances that have the most influence on the model. Although this sounds familiar to other strategies, it is frequently applied to gradient-based learning algorithms. Other strategies worth mentioning is that of *Expected Error Reduction* proposed in 2001[38], that introduces the concept of reduction in generalization error for querying a given instance in the input space, albeit at a notable computational cost. Finally there is *Variance Reduction* that aims to reduce the space of error indirectly by narrowing down the variance of the input space, and *density-weighted methods* where much of the yield of information for a given query is heavily influenced by the degree it is representative of the rest of the input space as analysed in[42].

4.4 Experimental approach

The experimental approach in this paper can largely be divided in two. The first approach is a naive one, where all predictor variables are included in the set and deemed to be independent variables. The second approach the data set is filtered, and the dependencies are removed. As noted earlier, some of these dependencies are linear in their transfor-mation, for example half-life in soil in equation 15. Dependant variables can be deemed to not give any further information about the label in such a case, but as seen in equation 14, some of these relationships are quite complex and as such have been included for testing. Data was compiled with no missing values, and further pre-processing included numerical transformation of predictor variables for conversion from the software they were derived from.

The methods used in this paper will further be evaluated on separate metrics than accuracy as it can be deceiving in this case. This is due to the skewed strong class imbalance in the labeled data set, as one would simply be correct 79.8% time by blindly predicting majority class for each substance. Further, a repeated k-fold cross-validation approach will be used, in which the model is trained and subsequently tested on a hold-out set to simulate its capability for generalization. This is done by separating the training data in to k number of folds. When the model trains, it trains on the all folds except the one remaining as a hold-out internal test set. This process is repeated n number of times. Cross-validation is also used due to the size of the data set. A model is only as good as the data it is trained on, and when working with smaller data sets one might not afford creating a separate test set, as for the purpose of this paper is to analyze the fit of a model and not only its classification capabilities. Thus, model evaluation is done by inspecting the balanced accuracy of a model, its receiver operating characteristics and area under the curve(ROC/AUC) metrics, and finally in the case of the random forests - its out-of-bag score. We further want to scrutinize samples of decision trees to determine what variable splits the model performs as described earlier, and inspect its misclassifications.

With this in mind, a 10-fold cross-validation with 3 repeats is used for models tuned on optimally grid searched hyperparameters - such as the number of trees or the cost metric C of the SVM. Hyperparameters are simply the conditions that are in place for model training. Given the skewed distribution of the data set, the values were centered and scaled for the SVM models so as to not let larger values ”dominate” the lower ones. This generally increase the capability of the model to establish a decision boundary between PBT and non-PBT substances while losing some level of interpretability. Further, the RF model was trained on its default settings. This is partly due to the size of the data set but also to qualitatively explore the model fit beyond classification performance as explained earlier. Another metric to evaluate model results is to inspect the value of Cohen's Kappa, where we inspect the agreeableness of the model with the *ground truth* while taking into account correctness by chance[12]. In other words, what sort of improvement does the trained model offer over a model that predicts purely on *expected accuracy*, which in our case is 79.8% due to the imbalance of class frequency. Cohen's kappa is always less than or equal to 1. Thus, if the kappa statistic is 0, the classifier can safely be discarded. As a final note, while other results are interesting, the larger focus will be on scrutinizing the best-performing model while other data can be found in the Appendix below.

5 Data

The data to be used is a list of ECHA and REACH evaluated substances of both concern and non-concern. Such substances include pervasive substances such as glucose, known used or banned pesticides, insecticides and industrial filtering chemicals. A total of 1115 such substances with 27 physical properties have been compiled, where 236 of these are known hazardous SVHCs and the remainder 881 non-SVHCs. These 27 features are physical properties that measure each chemicals' physical manifestation in its released environment and its lifespan. As mentioned earlier, the general toxicity of a substance is an induced judgement based on its relative bioaccumulative and persistence levels. As such, the physical properties calculated is a set of properties measuring both bioaccumulation and persistence respectively.

The selection criteria for the negative substances(non-PBTs) in the dataset is data from readily biodegradable tests of substances, meaning they pass the mark of not being considered persistant. In other words, these chemicals would be dealt with at a waste water treatment plant and subsequently disappear within 5 days of release in to an aquous environment.

5.1 Chemical property representation

As explored previously in the QSAR approaches and machine learning, a large body of toxicological classification has been done by analyzing the structural activity relationships of a given chemical, where among others, two dimensional transformation of a chemical structure coupled with a semi one-hot encoding is deployed as features. A distant measure is further used to capture similarity and arrive at a conclusion. In this project however, this two-dimensional representation is circumvented by measuring the physical impact of different chemical structural components directly. Some of these features are further outputs of experimental degredation models, or combined models of both experimental and expert-solicited estimates, such as the Biowin models. Finally, table 2 shows a subset of the physical properties and their description. The full table can be found in the appendix in table 10. Inspection of the descriptive statistics of the data is given due to the nature of the task which includes exploratory analysis of the data and its distribution.

The physical chemical properties are based on work done to develop a new persistence and bioaccumulation score for a substance[36]. For example, a total of 6 different Biowin metrics are included in the set that measure persistence at different levels from 3-4 training sets. As denoted in table 2, this includes inspection of linear(Biowin1) and non-linear(Biowin2) degradation transformations at different speeds - i.e slow vs not slow degradation. Biowin3 and Biowin4 are both regression models, however the aim of Biowin4 is to reproduce expert estimates of environmental half-life of what is called *primary degradation*, meaning the time it will take to reduce the concentration of the original chemical substance

Property Name	**Description**
kOH (AOPv1.92) atmospheric*	Denotes the rate of atmospheric degradation for a chemical
t1/2 atmosphere in hours	Denotes the half-life time of a substance found in the air measured in hours.
VP (mm Hg)*	Denotes vapor pressure of a substance, measured in millimeters of mercury
VP (Pa)	Denotes vapor pressure of a substance, measured in Pascal
Biowin1	Denotes the linear model output that predicts slow vs not slow degradation.
Biowin2	Denotes the non-linear version that predicts slow vs not slow degradation.
Biowin3	Denotes the estimates of environmental half-life necessary to mineralize a chemical ˜(i.e to turn 50% of the substance in to the ultimate degradation products - namely water and carbon dioxide
t1/2(water)hrs*	Denotes the half-life time of a substance found in water measured in hours.
**Dependant variable*	*dependant variables means that is a compound calculation of other physical properties.* *An example: 1/2 life in soil = 2* 1/2 in water*

Table 2: *Sample of predictor variables used for modelling. Includes the name of the pyshical property and the intuition behind the metric.*

in an environmental area by 50%. This can be due to a simple first transformation step, where the chemical you begin with is transformed into something else, i.e. the parent chemical has disappeared. Biowin3 on the other hand tries to reproduce expert estimate of half-life based on necessary *mineralization*, i.e the time it will take to turn 50% of the substance in to the ultimate degradation products, namely water and carbon dioxide. This measure calculates for more time however, as it takes in to account long transformation times, but also differing stages of transformation depending on chemical complexity.

Biowin5 and Biowin6 are linear and non-linear versions of a model trying to predict the outcome of a test called "ready biodegradability test". If a substance is readily biodegradable, it will quick dissolve and disappear in aquous environment, and therefore not be deemed as a PBT substance. These predictions are probability based, which according to software manual used to derive these properties from The Organisation for Economic Co-operation and Development (OECD), there the differencial is set at 0.5[4]. If a substance is above 0.5, then it is deemed readily biodegradable, and not the case otherwise.

An additional set of features is added in the form of a substance's Long Range Transport Potential(LRTP). As the name suggests, it describes the potential of a substance to spread and transfer itself across distance, additionally from one media to another - such as from air to soil. This is however more of a measure for so called Persistent Organic Pollutants(POPs). Commonly known POPs include DDT(dichloro-diphenyl-trichloroethane), a substance that was the first synthetically developed insecticide from the 1940s [3]. It was quite potent in combating insect-born human diseases like malaria, but because of its very persistent and ability to travel long distances, it's current status remains quite restricted under the treaty commonly referred to as the Stockholm Convention on POPs[30]. LRTP measures were originally left out of the PB-score calculation as overall persistence measures already included persistence in air, as this is also an element to LRTP and did not change results[36]. However, inclusion of POP criteria can further be used to be a POP specific scoring metric.
LRTP(CTD) and LRTP(Pov) can be described as a substance's characteristic travel distance measured in kilometres, while finally LRTP(TE) measures the transfer efficiency of a substance across different emission scenarios.

For the sake of clarity, some of the features in the set are noted as dependent, meaning that they are some alternative representation of a metric - such as vapor pressure measured in pascal or millimetres of mercury - or that they are some transformation that involves another property. An example is the half-life in water metric,

$$\textit{half-life}_{water} = 7300 * e^{-2*Biowin3} \tag{14}$$

which measures the half-life in days for a substance. The base value start at 7300 days(roughly equaling 20 years) multiplied by e to the power of the constant -2 times its Biowin3 value. However some dependant relationships are

criteria	UNEP/UNECE POP	EU PBT	EU vPvB
Persistence	- Water: t1/2 >2 months - Soil: t1/2 6 months sediment: 1/2>6 months - other evidence	- fresh or estuarine surface water: t1/2 >40 days - Marine surface water: t1/2 >60 days, -Soil, or fresh, or estuarine water sediment: t1/2 >120 days - marine sediment: t1/2 >180 days	- Marine, fresh, or estuarine surface water: t1/2 >60 days - Soil, or marine, fresh or estuarine water sediment: t1/2 >180 days
Bioaccumulation	- BCF >5000 or log Kow >5 -monitoring data - Other, e.g. very toxic	- aquatic organisms BCF >2000 or log Kow >4.5	-aquatic organisms BCF >5000 or log Kow >4.5

Table 3: *EU classification and labeling criteria under Directive 67/548/EEC. All points do not need to be fulfilled for a ruling, one condition will suffice.*

stronger than others and much more linear, such as the formula for half-life in soil:

$$\textit{half-life}_{soil} = 2 * \textit{half-life}_{water} \tag{15}$$

Here, as the authors note, a factor of two is a conservative estimate, albeit generally acknowledged that the two measures are connected[36]. The takeaway here is that dependant variables are some transformation of - or includes - other feature metrics but that the dependancy is not completely linear in all cases.

Finally, an interesting point of enquiry will be to inspect whether the ensemble of trees in a random forest model is able to capture legal evaluation guidelines. These legal thresholds are measured by EU directive 67/548/EEC. Table 3 shows the different criteria that can go in to determining POP or PBT status under EU legislation.

5.2 Descriptive statistics

As an initial check on the potential for separability, an examination of the descriptive statistics for the two classes could be useful. We describe the stats of the full dataset of 1115 substances. Tables 4 and 5 show the same subset of features described earlier and the descriptive statistics for their class(complete feature statistics can be found in the appendix). A noteable aspect to the data is the level of dispersion of values within both classes. In general, the mean values of degradation *rates* are higher for non-PBTs compared to that of PBTs, meaning they disappear at a faster rate than harmful substances in a released environment. Further, expected half-lives for harmful substances are higher as they are deemed to be more persistent and bioaccumulative, however if we inspect the atmospheric half-life, the mean values for non-PBTs are substantially higher and denotes a mean half-life of roughly 4.38 billion hours, or roughly 460,000 years. This generally means that a chemical released in the atmosphere will not go away, however due to their other physical characteristics, they do not pose a threat to other mediums or life-forms such as aquatic species, and are thus rendered as subjects of not concern. This data is however quite skewed, and as such the median values are a more robust representation of the central tendency. If we inspect the median of this same half-life, the median value is 14.9 hours, which is quite a staunch difference.

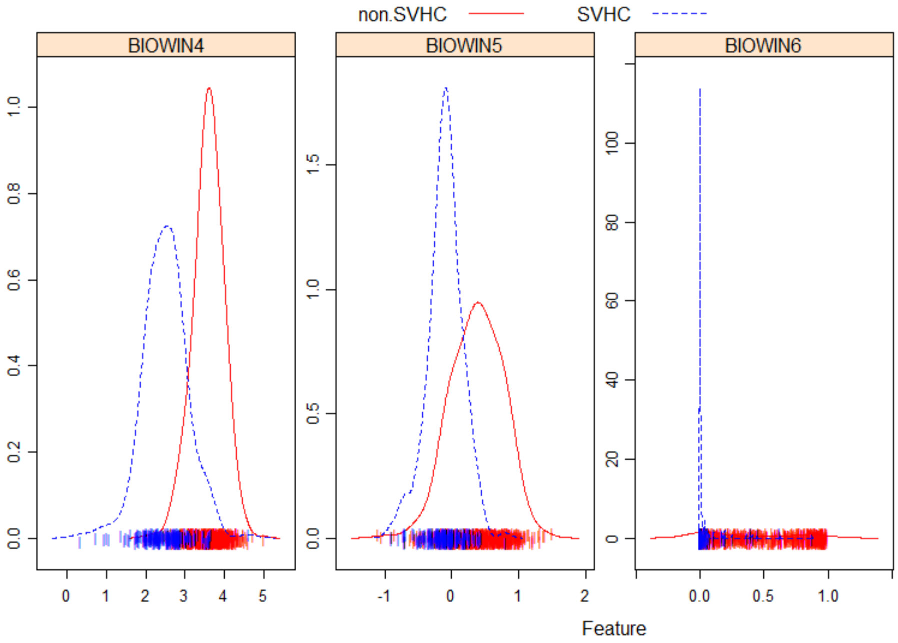


Figure 7: *Feature plot showing a comparison of feature distributions for non-SVHCs and SVHCs. Features ploted here are Biowin4, Biowin5 and Biowin6.*

A visual representation of these stats might shed some light on how strong these differences are. Figure 7 show the distribution for both classes for three different Biowin models. As mentioned earlier, the Biowin metrics are model outputs based on several chemical features, thus carrying similarly scaled distrib-utions. Due to the nature of its derivations, no log transformations were done to the predictor variables. Figure 8 shows the distribution for two LRTP estimates and the molecular weight, where the LRTP potential between harmful and non-harmful substances are quite different. The full distribution for all other features can be found in the appendix. Although there are stark differences between non-harmful and harmful substances for some of these features, the nonlinear nature of label assignment makes it a worthwhile problem. This is due to the nature of assigned labels, where a non-PBT substance might have highly persistent characteristics yet still be deemed safe, or vice-versa that a known PBT substance might not be as bioaccumulative as other substances.

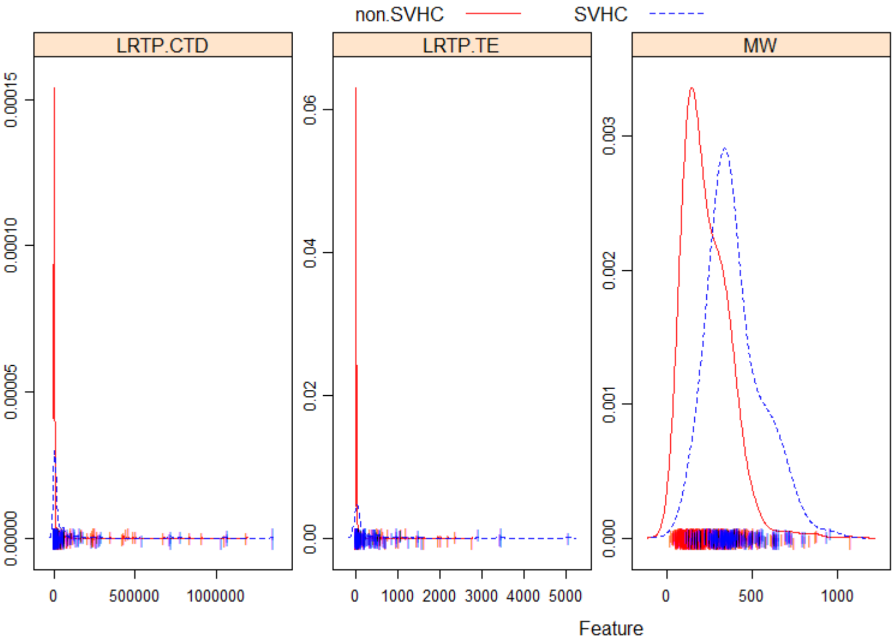


Figure 8: *Feature plot showing the distribution of measures for each class in the binary approach. Plotted here are LRTP CTD, LRTP TE and MW*

Another extension of the level of dispersion to the data is seen in the standard deviations for the features. As examined with the half-lives, standard deviations here can extend into the millions atmospheric degradation in known SVHCs.

Non-PBT	**mean**	**sd**	**median**	**skew**
k OH (AOPv1.92) atmospheric	5.4322e-11	7.6109e-11	2.582e-11	3.392
t1/2 atmosphere in hours	4.3809e+9	1.2988e+11	1.4914e+1	29.546
VP (mm hg, v1.43)	8.2681e+1	1.2400e+3	1.0225e-4	21.788
Biowin1	5.7096e-1	5.7706e-1	0.7005	-2.875
Biowin2	5.9003e-1	4.1113e-1	7.925e-1	-04.235
Biowin3	2.5854	6.2096e-1	2.7091	-1.0847
t1/2(water) in hours	2.7179e+3	1.0897e+4	6.5359e+2	10.680
Biowin4	3.5887	3.9292e-1	3.6061	-29.387

Table 4: *Descriptive statistics over a sample of variables for non-PBT substances.*

PBT	**mean**	**sd**	**median**	**skew**
k OH (AOPv1.92) atmospheric	1.3433e-11	2.9681e-11	7.6e-13	3.974
t1/2 atmosphere in hours	6.1515e+5	4.6662e+6	5.0687e+2	7.420
VP (mm hg, v1.43)	7.8888	1.145436e+2	1.8744e-06	15.078
Biowin1	-2.3600e-1	5.4654e-1	-2.133e-1	-0.590
Biowin2	6.7579e-2	2.2946e-1	0	3.306
Biowin3	1.2906	0.7429	1.3066	-0.463
t1/2(water) in hours	7.6248e+4	6.4696e+5	1.0084e+4	13.613
Biowin4	2.4951	0.5504	2.5299	-0.131

Table 5: *Descriptive statistics over a sample of variables for known PBT substances.*

6 Experimental Results Evaluation

6.1 Binary classification

Table 6 shows the comparative performance between naive and filtered results in both modelling approaches. All models have a balanced accuracy larger than 90%, with the better-performing model being the naive random forest approach. Further, sensitivity metrics are high for all models as well, while lower in specificity. One way to interpret this is that the models overall are good at identifying the non-PBT class of substances but makes more mistakes for the PBT class, especially for the SVM models. This can be due to the nature of SVMs generally being more prone to overfitting when fed an imbalanced dataset and a skewed data distribution. One can further draw attention to the fact that the misclassifications are higher in direction of the majority class, where known pbt substances are classified as not harmful at higher rate than non-pbt substances being considered as pbt.

	Naive Random Forest	**Filtered Random Forest**	**Naive SVM**	**Filtered SVM**
Balanced Accuracy	94.28%	94.18%	90%	91%
Sensitivity	0.9943	0.9920	0.99	0.9857
Specificity	0.8894	0.8936	0.81	0.8371
Kappa	0.9136	0.9112	0.8544	0.8092

Table 6: *Overview of baseline model performances. The naive random forest that includes dependent variables is marginally better than the filtered approach. In totality, the naive approach yields the best results. True positive predictive statistics is high for all models, while specificity metrics are especially lower for the SVM models.*

In addition, evaluation of the kappa metric does not have a closed form rubric of evaluation, however relevant literature denotes that between values 0-0.20, there is slight agreement, values 0.20-0.40 denotes moderate, 0.40-0.60 significant, 0.60-80 good and 0.80-1 near perfect agreement[26]. When taking expected accuracy in to account the interpretation of improvement could change. For example, one might consider a kappa value of 85% a lot better if the expected accuracy is at 50% than at 70%.

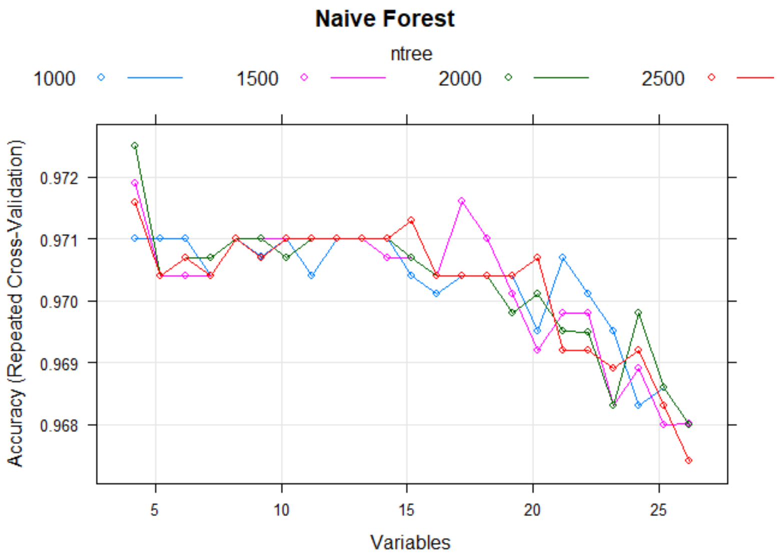


Figure 9: *Naive model performance over different hyperparameters of number of trees and available features for random sampling. Best performanced was achieved with 2000 trees and 4 mtry albeit with marginal differences.*

In light of this, both random forest models prove to be significant results with values marginally exceeding 0.91, where the naive model is indiscriminately better. As mentioned earlier, the RF models are grid searched across different number of trees and *mtry*, denoting the number of random available sampled variables a given tree can access. Standard values for mtry start at the square root of total number of predictor variables(in this case 5), and search up until the max. The grid searched naive model can be seen in figure 9, whereas the filtered model can be seen in figure 10. Optimal parameters for the naive model is at 2000 trees and 4 mtry. Albeit marginal differences, increasing the number of candidate variable samples for the trees leads to a decrease in performance, which could point to an issue when random sampling allows selection of less informative variables for tree construction, however this is inconclusive and could be due to noise. The same optimal number of variables is found for the filtered forest with optimal number of trees at 1500. This decrease of performance is however marginal.

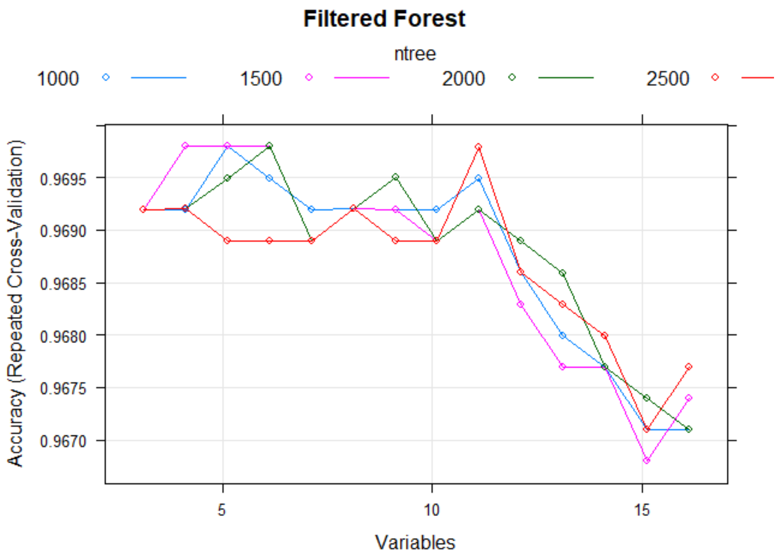


Figure 10: *Filtered model performance over different hyperparameters of number of trees and available features for random sampling.*

Specifics related to the SVM models can be found in the Appendix, as further exploration of the better-performing model will be in focus, however for the sake of insight figure 11 shows the grid-searched model for the filtered SVM approach.

Optimal cost parameter for the model is at 5. As explained earlier, the SVM tries to find the optimally separating hyperplane between classes by maximizing the margin to the support vectors. A low cost parameter would designate that the model looks for a larger margin that allows for more misclassifications, whereas a larger value would seek to narrow the margin to avoid these misclassifications. For this particular model, some 235 support vectors were used to select for the optimal decision boundary and a softer margin. A further gamma value Σ was most optimal at 0.25, where the bias is lowered and the variance is raised. It further means that the model tries to accomodate its training data which lowers its ability for generalization Figure 12 shows this relationship, where a decision boundary for the SVM is fitted to the octanol water coefficent and the Biowin4 metric.

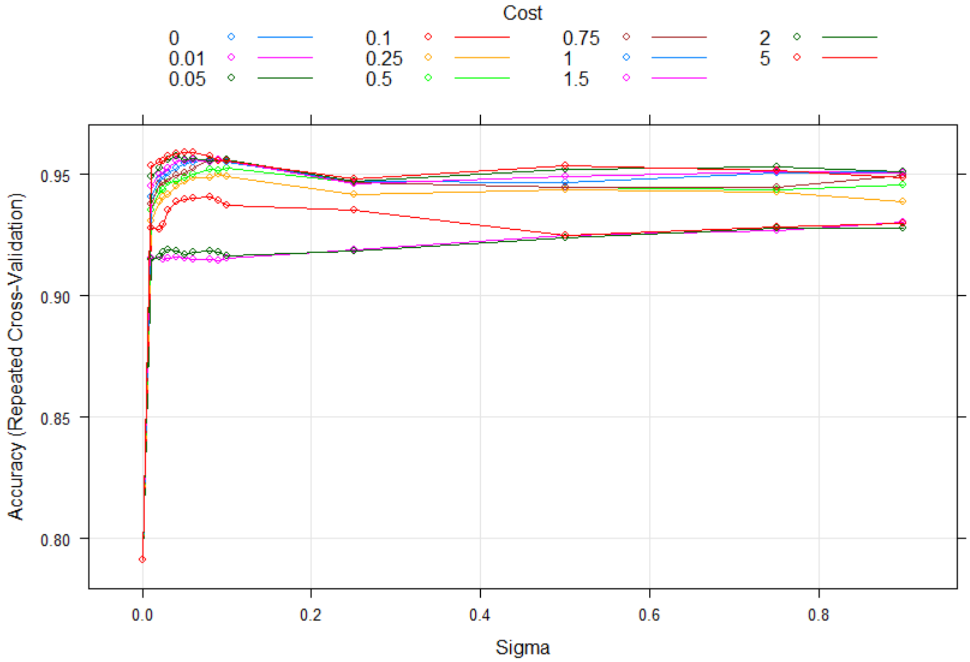


Figure 11: *Grid-searched radial SVM for optimal parameters. Best performance was achieved with cost parameter of 5 and* Σ *at 0.25*

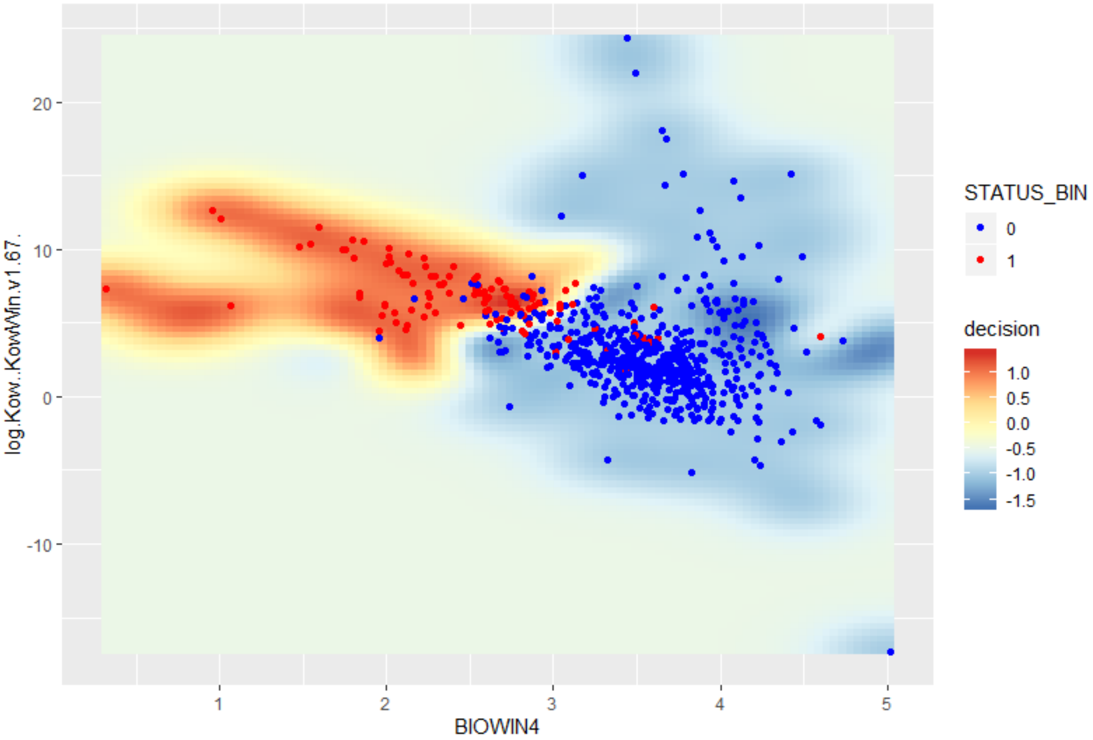


Figure 12: *SVM decision boundary on the octanol water coefficient and the Biowin4 metric. Decision boundary is here designated as a heat map of classification zones, where 0 denotes a non-PBT assignment and 1 a PBT assignment.*

Here one can visually see the relation between the non-PBT label and higher values of Biowin4, with few exceptions for PBT cases. Substance clusters are quite distinct, with a substantial portion of non-PBTs hovering in Biowin4 values 3-4 with near 0 value for log Kow.

We can further inspect the confusion matrix based on a 50% probability threshold of both the the naive and filtered random forest models to qualitatively explore their table numbering,

		Actual	
		Non-PBT	PBT
Pred	Non-PBT	874	26
	PBT	5	209

		Actual	
		Non-PBT	PBT
Pred	Non-PBT	872	25
	PBT	7	210

where the naive random forest model incorrectly classifies 26 PBT substances as non-PBTs, and inversely classifies 5 non-PBTs as PBT substances. For the filtered approach, the false positive rate is lowered and the false negative rate decreased. The shape of the results carries marginal difference, potentially suggesting that the models achieve similar mode fits with few tweaks in favor of the naive approach.

We can further see this in the ROC curve plotted in figure 13. The ROC graph shows performance of a classification model at all thresholds, where the plotted metrics are the sensitivity and specificity thresholds. If our curve was closer to the diagonal, the more useless the test is.

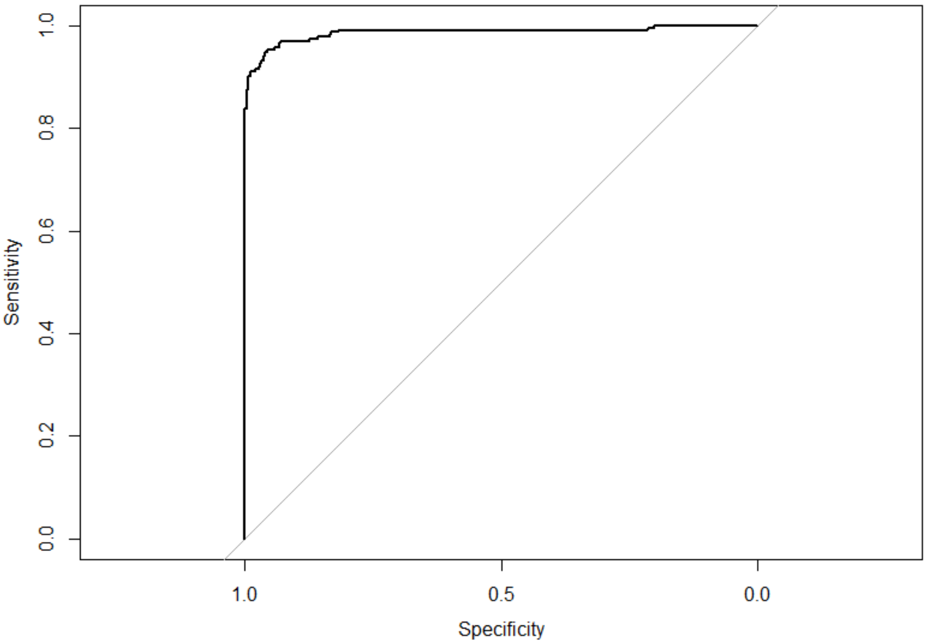


Figure 13: *ROC curve plotting classification model at different thresholds*

Instead of computing all possible classification thresholds found, we inspect the area under the curve(AUC) which computes the aggregate performance across all thresholds in the two dimensional space below the curve. The AUC for the naive model is at 0.987. To illustrate the impact, the AUC metric ranges in values from 0 to 1. A model that gets every prediction correct has an AUC value of 1.

Given the nature of the problem, an argument could be made for prioritizing lower false negative rate, as the potential cost of classifying known toxic chemicals to be safe is higher than that of classifying substances known to be non-harmful as harmful. Thus, one could make an argument that the filtered model is the better model. Partly due to a reduced dimensionality for feature inclusivity, but also for having marginally less false negatives. However as table 7 shows, the nature of these misclassifications not only carry repeated mistakes, but also to what extent how ”wrong” the models are.

Substance	**Actual**	**Predicted**	**Prob non-PBT**	**Prob PBT**
Vinyl Neodecanoate	PBT	non-PBT	0.92	0.072
Flucythrinate	PBT	non-PBT	0.78	0.21
Indene	PBT	non-PBT	0.94	0.05
Methoxychlor	PBT	non-PBT	0.57	0.42

Table 7: *Sample of false negative misclassifications done by the models. Colored cells that are orange denote misclassifications across two or more models, and yellow across all models. Probabilities are from the best performing naive random forest model.*

Vinyl Neodecanoate is a substance used as decorative emulsions in substances like paint. Flucythrinate and Methoxychlor are insecticides, and Indene is a principally used industrially to create thermoplastic resins. Methoxychlor is indeed a banned substance under ECHA legislation from 2002[1]. The table includes probabilities from the naive random forest model assigned to the substances. At a glance, one can see that the model is quite confident in its assessment except for the substance Methoxhychlor which is a bit more uncertain. What seemingly separates these particular substances from the rest? For one, the average Biowin4 and Biowin3 metrics for the four substances are higher than both the PBT and non-PBT. Their average Biowin4 value equals 3.8, while it is at 2.4 for PBT substances, and 3.59 for non-PBTs. This suggests that the rate of natural degradation rate according to the Biowin model is higher than usual. For Biowin3 their average value is at 3.1, while for known PBTs it is 1.29 and 2.59 for non-PBTs. The average molecular weight for these substances is much lower than both classes, with the heavier chemicals belonging to the PBT class. This difference is on an even stronger level for BCF BAF values, where mean PBT values sits at roughly 9100, non-PBTS at 262 and finally this group at 2.881. As discussed earlier, the relative skewness of the data can affect performance in this case.

Nevertheless, a further point of interest is the rather high confidence for classifying Indene as a non-PBT substance and further Naphtalene in the larger table. Indene and Naphtalene are both on a list of substances of very high concern in the Netherlands. The criteria for their classification has been due to them being on a list for poly-aromatic hydrocarbons(i.e more than one aromatic ring). Both Indene and Naphtalene have 2 such rings, however expert opinion are in agreement that both of these are indeed not PBT substances. A report on Naphthalene(whose evaluation can be extended to Indene) from 2018 arrives at this conclusion [49]. In this particular case, the model may be right in its conviction, however there are examples where the confidence of the model is to a certain degree misaligned, as seen with Methoxychlor.

A further argument for the relatively high confidence of the model classification can be seen in figure 14, which is further a reflection of the relative skewness and high separability of the data. Methoxhychlor here is indeed one of few substances that the model seem insecure about.

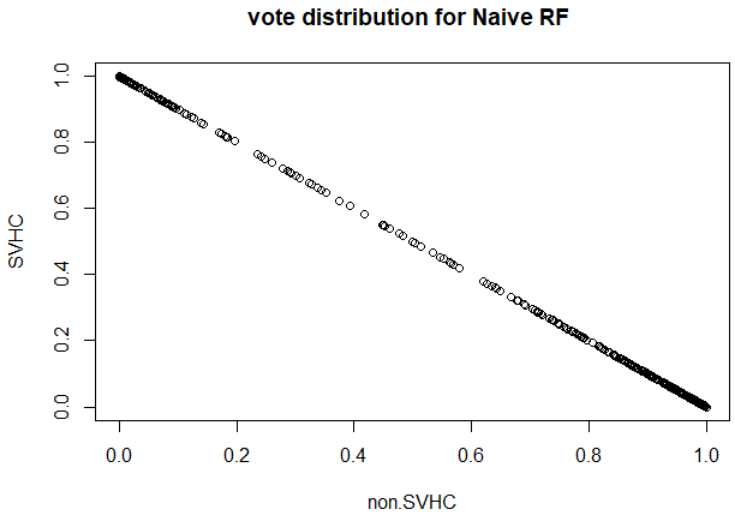


Figure 14: *Vote distribution for the naive random forest model for both substances deemed to be of concern and non-concern.*

It is unclear how large of an impact the inclusion of dependent variables offer over the model where they are filtered out, nevertheless the shape of these outputs are marginally different, suggesting an overlap in important variable ordering.

Figure 15 further shows the variable importance for the naive forest model(filtered variable importance can be found in the appendix). Both a mean decrease in accuracy and mean decrease in gini impurity is reported to cross-check the variable importance. This is done to not only have a more robust understanding of what features are most informative for the model classification, but to manage limitations related to both metrics, as absolute variable importance can be hard to determine. For example, molecular weight has a smaller range of values than the variable for atmospheric half life, thus having fewer candidate splits and becomes a notable problem for only examining gini impurity. However, approximations are useful which is why we examine both. For one, mean decrease in accuracy is an averaged metric over all out-of-bag cross-validated prediction on permuted variables. An intuitive interpretation is that for a given variable it is shown the mean number of misclassifications that would increase if that variable were to be excluded or permutated. We include this further due to the level of disparity of candidate splits between variables that have high variance in range. As such, dropping Biowin4 and BCF BAF would lead to on average 32 more substances to be misclassified. An even higher importance is given to the POP-specific long range transport travel efficiency and characteristic travel distance. This finding is further interesting relating to the work done on PBT and Pop screening by Rorije, Verbruggen[36], where Biowin3 was deemed to be of higher importance for persistance evaluation than Biowin4. Further, LRTP estimates was left out of consideration in this paper, where here their importance rank high with otherwise hypothesis-confirming variables.

Mean decrease in gini impurity reflects the much more local function described earlier, where we determine optimal split variable and subsequently best value for said variable in a given node that increases purity. High values here indicate that the variable is useful and tend to split mixed and unclear nodes towards more pure single class leaf nodes. Again as with the overall accuracy metric Biowin4, the BCF BAF metric and half time of metabolised substances score particularly high, with Biowin4 having almost 5 times the amount of impurity decrease than for example the octanol water coefficient(log koa).

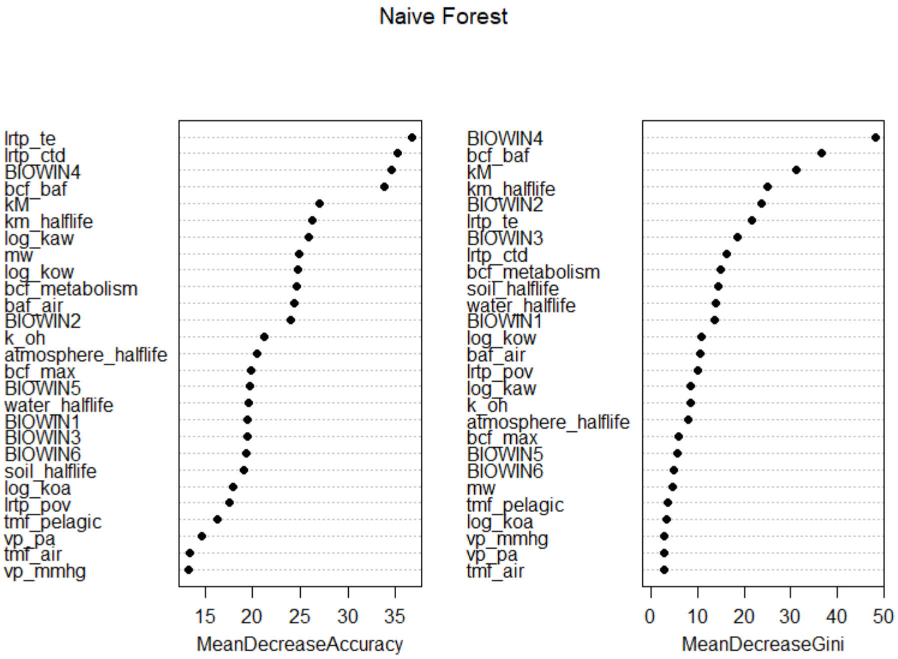


Figure 15: *Variable importance for the naive forest model. Notable importance lies in the Biowin4 measure,the bioconcentration factor for bioaccumulation and depuration rate for aquous species. The ordering is related to the relative decrease of accuracy on cross-validated samples on the left, and Gini impurity a split on this variable offers over the data on the right.*

While important in overall mean accuracy decrease, POP specific criteria of transport efficiency and characteristic travel distance does not score as high for impurity reduction. In fact, it carries similar importance as half lives for soil and water, which generally is considered not to be a clear indicators during expert evaluation of PBT status. A caveat to this however is that the importance for permuted variables of global out-of-bag samples is generally more reliable. Further variables that does not sufficiently gives us more label information can be seen in vapor pressure, trophic magnification rate for air-breathing mammals and octanol air coefficient to name a few.

This ranking further grants insights into what sort of partitions is performed for decision tree construction that maximize the data partitioning. We can exemplify this by examining a sample tree construction for the dataset. Figure 16 shows a sample decision for the naive model.

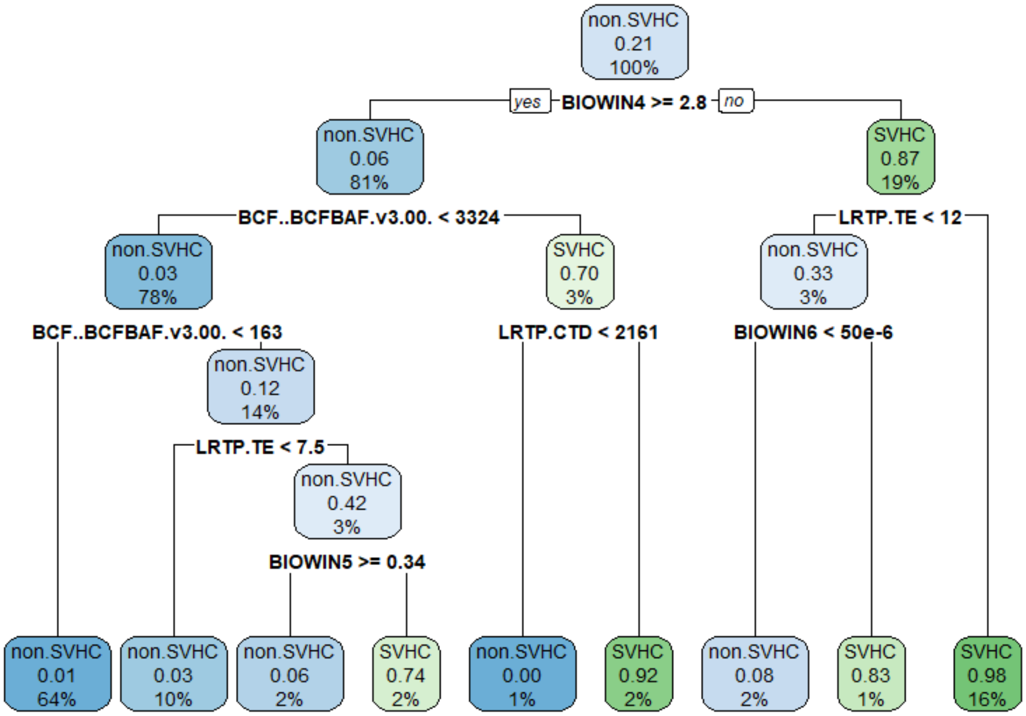


Figure 16: *Sample tree construction and optimal splits for the naive approach. In accordance with variable importance, a staunch part of the data is separated on Biowin4 and BCF BAF alone for non-PBT classification. Conversely, values of Biowin4 <= 2.8 and LRTP TE >12 designates the larger portion of PBT substances.*

The nodes in the tree shows the following ordered information:

- Majority class in the node.
- Probability of classifying positive class(PBT).
- Proportion of data present in the node.

From the root node, a check on Biowin4>= 2.8 sends 81% of the data left and 19% of the data right. As reflective of the model variable importance seen in figure 15, two subsequent splits on BCF BAF of <3324 and <163 makes out the majority evaluation of non-PBTs, where a 10% portion of the observations has a larger value for BCF BAF than 163, however with a transport efficiency lower than 7.5 kilometers. These two leftmost terminal nodes captures the majority of the data with only 1% and 3% predicted accuracy for being PBT. The split points that the model identifies as evaluation thresholds between classes is in close accordance for ordinary classification guidelines seen in figure 3. The threshold found seem to be located somewhere in between the thresholds determined for EU PBT and vPvB substances respectively.

Conversely, the rightmost terminal node contains 16% of observations in the data, however with a probability of PBT membership at 98%, the two splits on Biowin4 and LRTP TE includes a majority of PBT substances present in our data set.

6.1.1 Reductionist binary model

In light of the rather strong separation for the data in Biowin measures and BCF BAF alone, one may wonder how model performance changes if these were excluded. Although counter to the idea of further improving classification performance, this is partly done to check the impact of dimensionality reduction, but also to explore modelling capabilities of non Biowin measures and other physical properties of PBT classification.

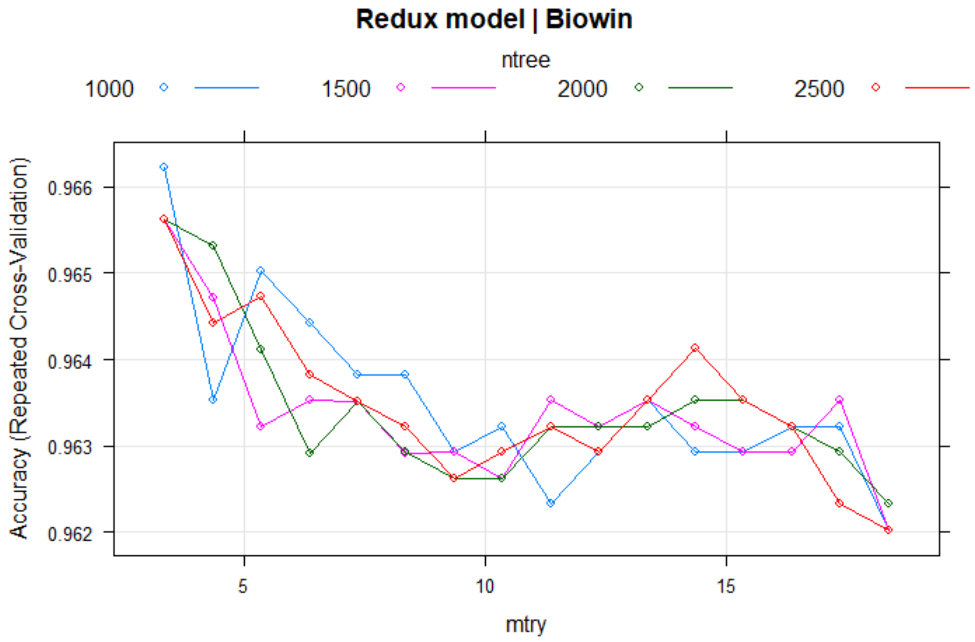


Figure 17: *Cross-validated grid searched model for reductionist approach. Optimal values were 4 for mtry coupled with 1000 trees.*

Grid-searched hyperparameters show optimal performance for 1000 trees and with variable sampling at 4 again, as with previous models. Furthermore, the same trend of increasing variance by allowing for higher cardinality of sampling leads to marginal performance decrease.

If we inspect the confusion matrix for the reduced model, its balanced accuracy is 93%, which is close to the performance of the benchmark models, but does not improve upon initial results. A matter of salience further is the fact that the general shape of the confusion matrix is the same, and that the increase in rate of false negatives is the most notable difference. This could further be seen as an argument of the skewness of our data and subsequent ease of separation for a select few variables, however this is not conclusive.

		Actual		
		non-PBT	PBT	Total
Prediction	non-PBT	873	30	903
	PBT	7	205	212
	Total	880	235	1115

Cohen's Kappa is further estimated at 0.89 that renders it an improvement over models for expected accuracy similar to that of both the naive and filtered SVM.

Further, the reductionist model's ROC curve can be seen in figure 18, where the model is still capable of accurately matching model prediction with the ground truth.

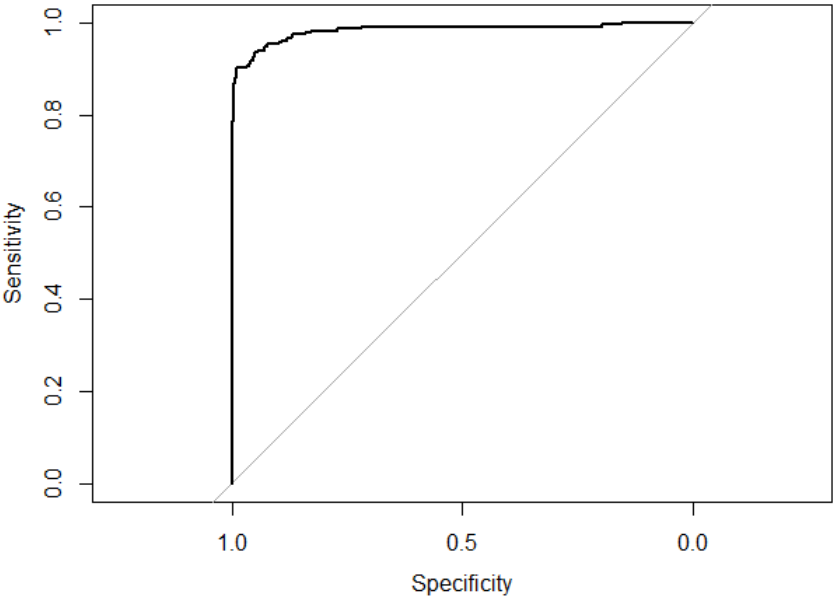


Figure 18: *ROC curve for reductionist model, plotting different thresholds for sensitivity and specificity for the model. Area under the curve is calculated to be at 0.983*

It can further be interesting to inspect variable importance change when the model is stripped of otherwise informative metrics for bioconcentration and persistence. Figure 19 shows this distribution. Here the local impurity decrease is high for variables previously not considered as important, among others half life in water and soil. The more global variable importance for permutated variables for OOB prediction remains high for POP-specific criteria, with soil and water half-lives are in close proximity of features such as molecular weight, bioconcentration factor for air-breathing mammals and flat depuration rate for a substance.

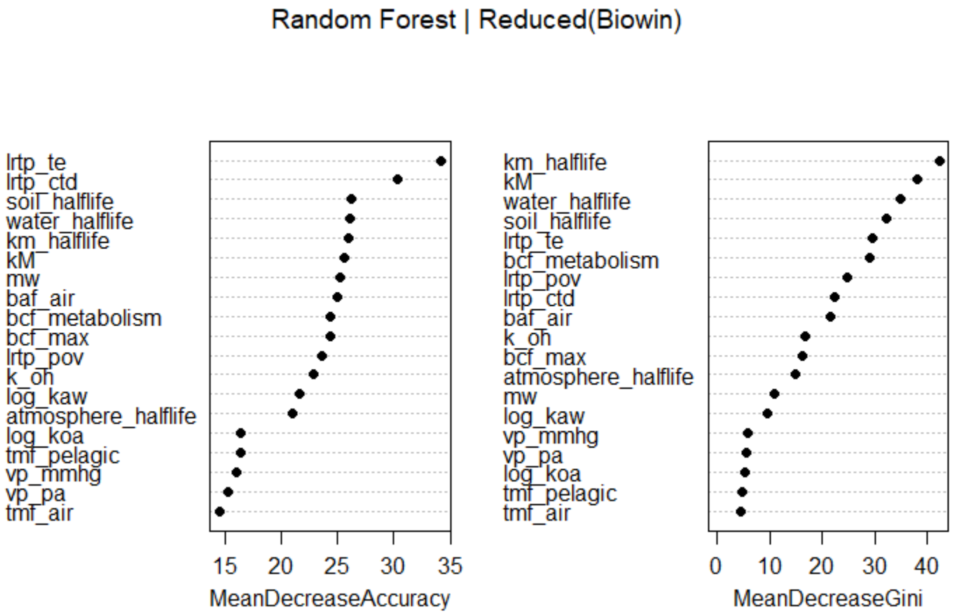


Figure 19: *Variable importance for the reduced model where Biowin measures and BCF BAF is removed.*

Examining sample tree fit to the reduced model further highlights adaptability to find new optimal splits. This can be seen in figure 20.

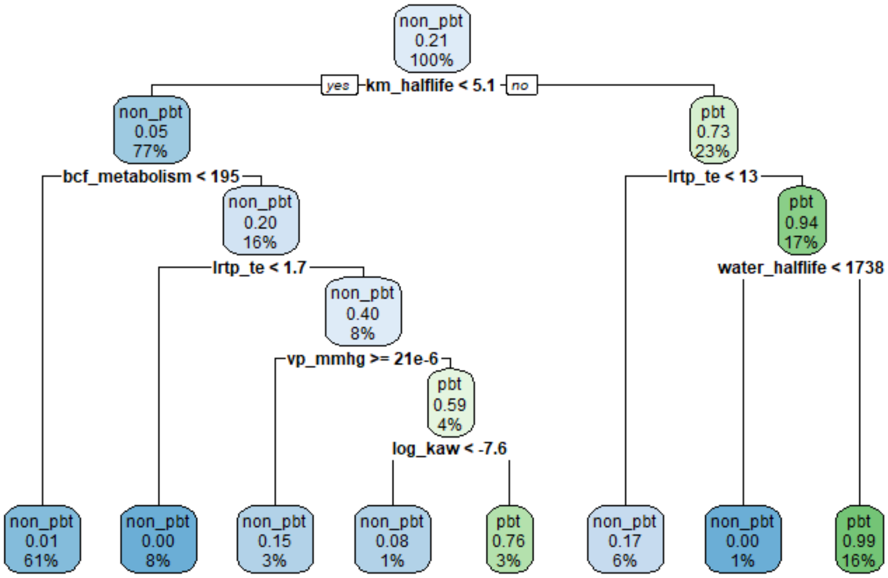


Figure 20: *Sample tree fit for reduced data set. 77% of the data can be predicted to be non-PBT with depuration rate in aquous species of less than 5.1 days.*

Depuration rate half-life in aquous species for less than 5 days provides the largest split on the data, with a predicted accuracy for being non-pbt is achieved with a subsequent split on bioconcentration factor that is adjusted for metabolism at less than 195. Conversely, known PBTs intuitively have a higher depuration rate, and as seen in figure 16, checks on LRTP efficiency separates the data further.

Another point of interest is the observation that a large portion of PBTs seem to have a higher half life than 1738 hours, which roughly translates to around 70 days. To illustrate the high level of persistence of these substances, one can compare this to table 3 thresholds for both EU PBT and vPvB criteria, which in exceeds the threshold for vPvB half-life in marine, fresh or estuarine surface water at 60 days.

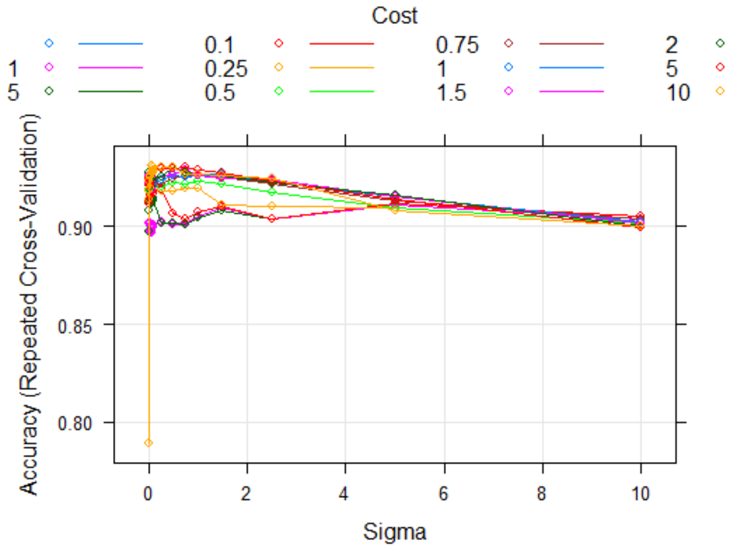


Figure 21: *Grid-searched SVM for the reduced model. Optimal hyperparameters were selected for gamma at 0.1 and cost of 10.*

As with the previous results, a reduced model does not improve classification using SVMs, where in fact performance drops to a balanced accuracy of 86%. Figure 21 shows the grid-searched model and its optimal parameters. Optimal parameters in this case is a gamma value of 0.1, which effectively translates to the model having very low bias towards its training data, but its variance will be higher for generalization. The further cost for a misclassification is at 10, meaning that the optimal hyperplane has a hard margin. Further stats relating to sensitivity is at 0.98, with specificity dropping lower than previous models to 0.74, a further decrease in capability of identifiying PBTs.

6.2 Subclass classification

One hypothesis to test is whether the nuance of labeling PBTs and non-PBTs has an effect for classification. As stated earlier, a major goal of the project is to explore avenues for PBT classification that may give insight into the differential impact of a substance being persistent, bioaccumulative, or a combination of the two that leads to its toxicity. The fact that there are different thresholds for substance evaluation is further an argument to support this hypothesis. Characterization of what a PBT substance is, or what further makes it a vPvB substance can be thought to carry substantial overlap, as the latter category can be concluded to be a more persistent or bioaccumulative extension of the former, thus the binary approach can be said to encapsulate this and not improve upon classification, however as an avenue of research of PBT and non-PBT evaluation boundaries for such categories, it can be a worthwhile endeavour to examine.

The modelling approach is performed in a similar way as the ordinary binary setting, where a grid-search of optimal hyperparameters is used to obtain optimal model fit. This further includes conditions from section 6, like the use of cross-validation. A fit of both naive and filtered models for subclass classification was performed. There was 6 possible labels in the subclass approach:

- Not P (could be B)
- Not B (could be P))
- not PBT
- PBT
- PBT/vPvB
- vPvB

Included is the notion that although a substance might not be persistent, it can be shown to be bioaccumulative and as such be a PBT substance. This same reasoning applies to non-bioaccumulativity and persistence. Table 8 shows the full performance across all models.

	Naive Random Forest	**Filtered Random Forest**	**Naive SVM**	**Filtered SVM**
Balanced Accuracy	83.1%	83.4%	66%	67%
Cohen's Kappa	0.73	0.73	0.35	0.39

Table 8: *Overview of results across all models for subclass classification.*

In terms of classification performance, the best performing model is the filtered random forest with a balanced accuracy of %83.4, which is marginally better than its naive counterpart, however both have an evaluated kappa value at 0.73. For both SVM models, balanced accuracy for the naive and filtered approach is at 66% and 67% respectively. Cohen's kappa further designates the models to have a score of 0.35 and 0.39, rendering the models not too reliable improvements above models selecting for expected chance. Specifics for these models can be found in the appendix.

The grid-searched filtered model can be seen in figure 50. Optimal parameters were selected to be at 1000 trees with 16 available variables for random sampling.

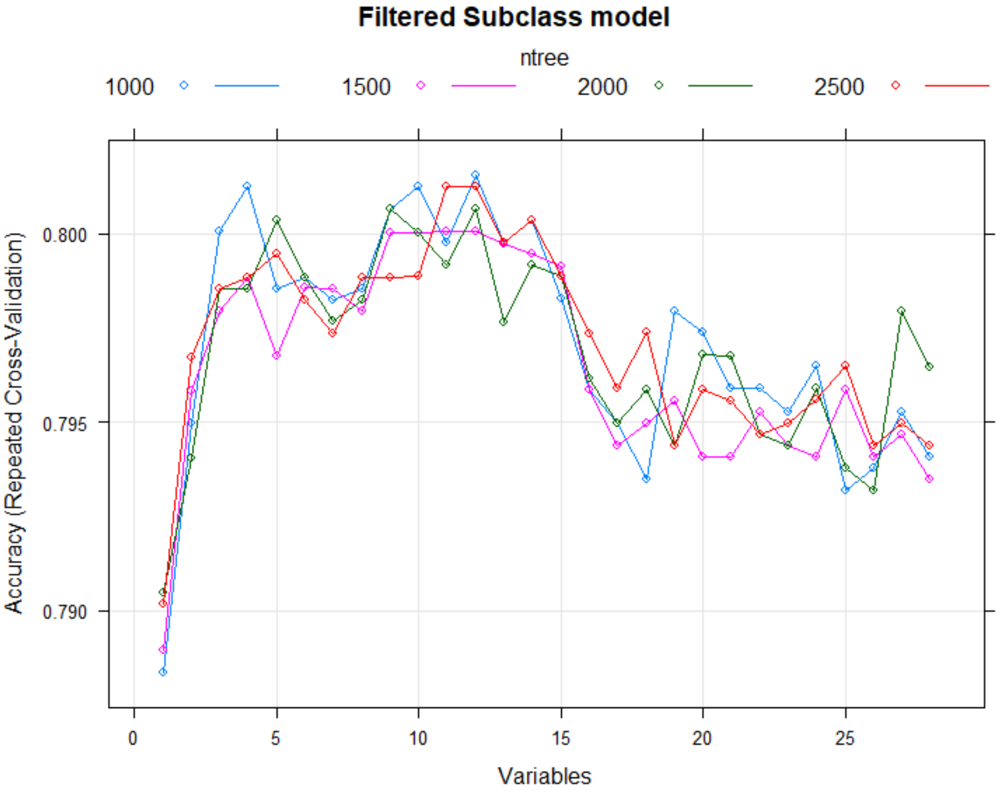


Figure 22: *Grid-searched filtered model for subclass classification. Optimal parameters where chosen to be at 1000 trees and 12 mtry.*

We can further inspect the confusion matrix of the filtered model below, where most of the errors are on either end of the label spectrum. Some 76 substance designated as either not B or not P substances are predicted to be the counterpart. Further, the misclassifications are very local to relating label assignments, where few cases are predicted to be on the other end of the spectrum, such as non B or non P substances rarily was predicted to be vPvB or egregious PBT substances, however with some caveats.

Actual

Predicted	Not-B	Not-PP	Not-PBT	PBT	PBT/vPvB	vPvB
Not-B	156	41	25	5	1	0
Not-P	34	203	10	6	0	0
Not-PBT	11	12	374	2	0	0
PBT	10	3	2	63	7	12
PBT/vPvB	1	0	2	7	16	15
vPvB	0	0	2	5	15	78

	Not B	**Not P**	**Not PBT**	**PBT**	**PBT/vPvB**	**vPvB**
Sensitivity	0.68	0.80	0.94	0.64	0.42	0.78
Specificity	0.93	0.93	0.94	0.97	0.97	0.97

Table 9: *Sensitivity and Specificity measures for the filtered subclass model.*

The model further has high specificity metrics, where it correctly identifies the true negatives in the set. Sensitivity on the other hand is quite a bit lower, as reflected by the confusion matrix. We can inspect the ROC curve for the model that demonstrates the relationship of different thresholds for sensitivity and specificity in figure 23. Note that this is an averaged curve over all folds for each class, where the curve is calculated on a one vs all basis, such as classification for not B versus Not P, Not PBT, PBT, PBT/vPvB, vPvB. This is performed on every class and averaged.

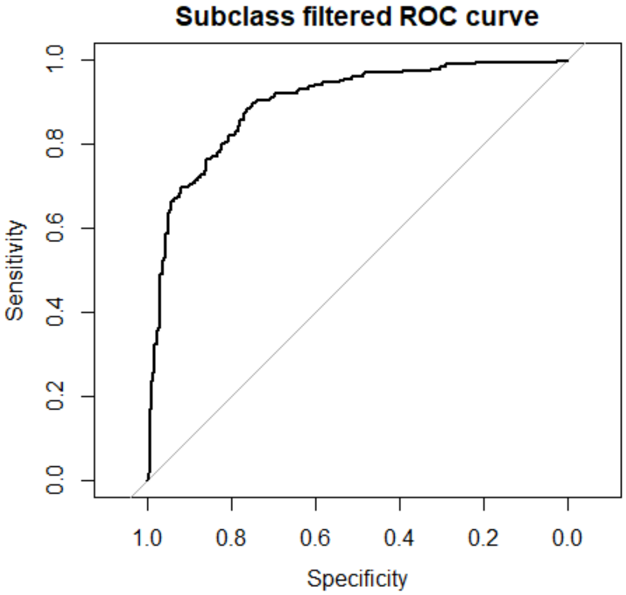


Figure 23: *ROC curve for the filtered subclass model plotting different thresholds for sensitivity and specificity trade-offs.*

Lowering the threshold for sensitivity increases our true negative detection rate, whereas if we increase the threshold, a mark of performance decrease starts at a threshold of around 0.7 sensitivity. Area under the curve is calculated at 0.81 , denoting that the model is capable of identifying the correct class assignments, however to a lesser extent than the strict binary approach, leading to a preliminary conclusion that dividing the problem into its subcategorical representation does not improve upon the screening model. However, it can from a research standpoint be interesting to look at the variable importance of the approach.

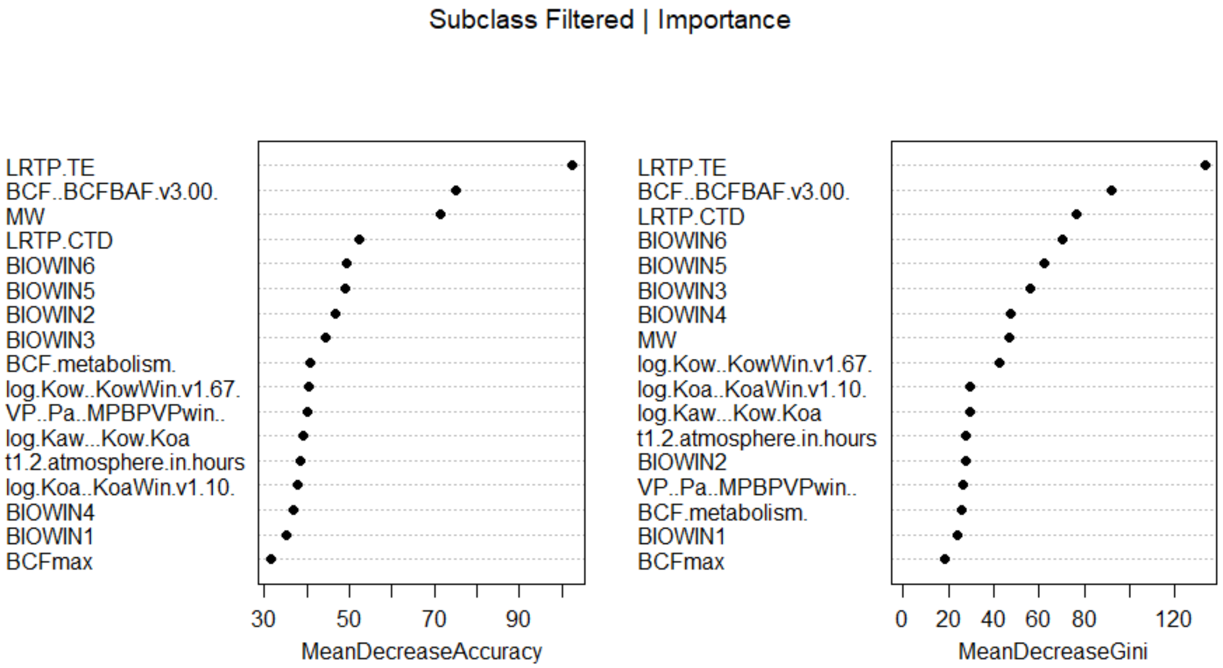


Figure 24: *Variable importance for the filtered subclass model.*

The global variable importance for the permutated variables for out-of-bag predictions follow similar importance structure we saw in figure 15, where metrics for LRTP overall transport efficiency scores high for the local impurity reduction as well. Further reccurency is found in importance of the Bioconcentration factor and Biowin estimates. Biowin4's overall importance drops lower in the subclass approach, where exclusion leads to on average some 38 more misclassifi-cations if excluded. This can be contrasted to LRTP TE exlusion leading to just over 100 substances. This again could be an indication of the strong separability of certain features in the set. This can further be examplified by the heightened impact of molecular weight, which ordinarily might not be too informative in expert evaluation, albeit an indicator for a potential PBT label. A sample decision tree of fit for the data can be seen in figure 25.

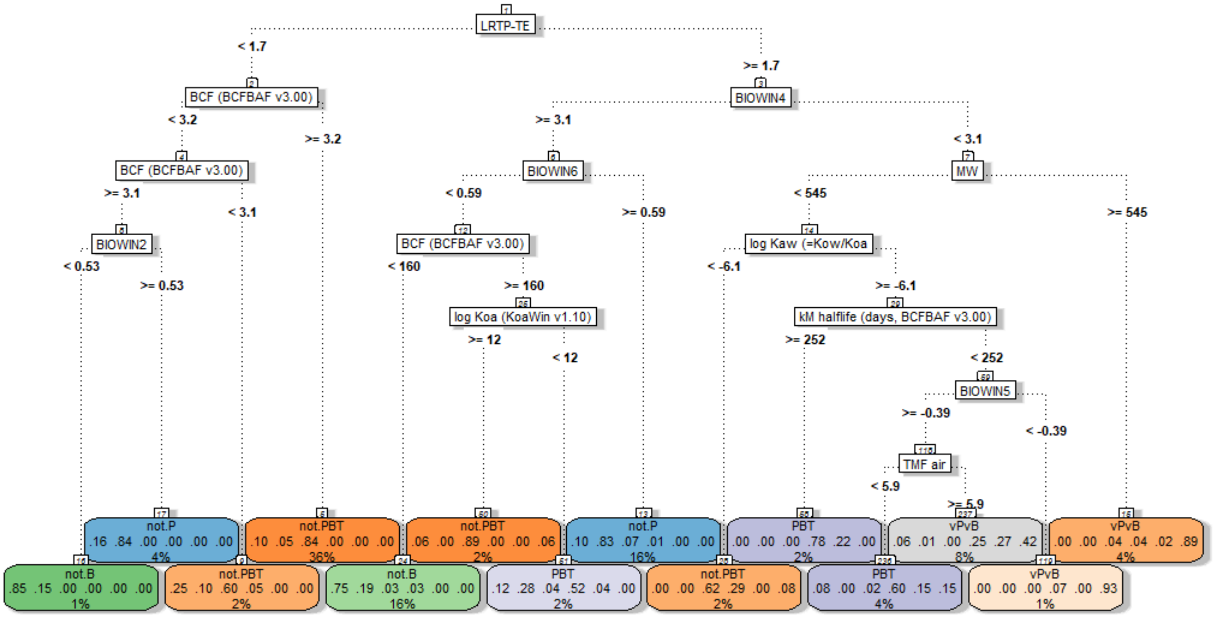


Figure 25: *sample tree fit for the filtered subclass approach*

Denoted are the optimal splits point for the different variables, and final cells show majority class, probability distribution across classes(in order) and again proportion of cases that would fall into the chain of variable splits. Major proportion of non-PBT subcategory labels can be observed to be classified through splits at LRTP TE, BCF BAF, Biowin6 and Biowin2 metrics. Conversely, for PBT subclasses the splits that are most informative are biowin4 measures, MW, Log Kaw and km half-life.

6.3 Added Value of Active Learning

In light of the best performing naive binary model, the question remains whether one can achieve similar results while allowing for explicit data selection from the viewpoint of the model. To reiterate, this can be a valuable approach for a field like toxicology were experimental data and labels are expensive to obtain, or if they exist, is sparse. This further informs the selection of our query strategy framework for the model, where a pool-based approach is selected, where we mask a subset of data for training and put the rest in the unlabeled pool U. The model trains and selects for new data points Until the performance metric or budget is met, all the while reporting error progression.

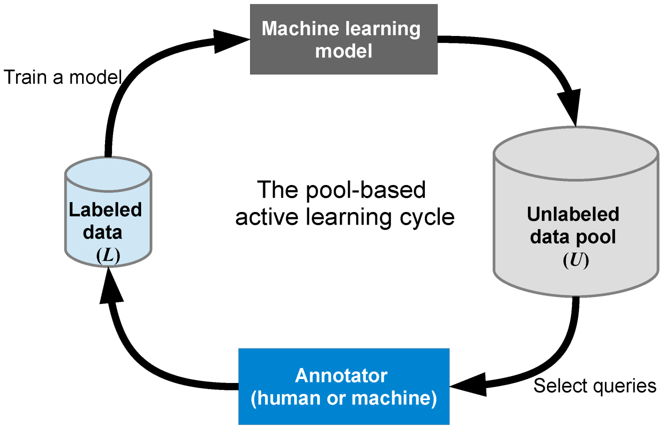

Machine learning model
Unlabeled data pool (*U*)
Select queries
Annotator (human or machine)
Labeled data (*L*)
Train a model
The pool-based active learning cycle

Figure 26: *Overview of the active learning loop with pool-based sampling[22].*

The active learning was explored in two ways:

- Automated labeling
- Interactive labeling

In the automated approach, we have the model continue the active learning loop until it has equaled the error of the binary naive random forest model, whereas in the interactive approach, the model will select for queries and ask a hand-picked label based on the properties of the data it is insecure about labeling. This is done to demo the capabilities of expert analysis and solicitation for a model-quantified uncertainty for label assignment. The active learner inherited the hyperparameters of the binary naive model of 1500 trees and 4 available variables for sampling. An initial training set of 100 substances was selected for the initial active learning loop. Figure 27 shows the performance for the initial single learning loop with the different uncertainty sampling strategies, where the model obtains similar performance to the passive model after some 486 queries, or in other words 47.2% of the total data.

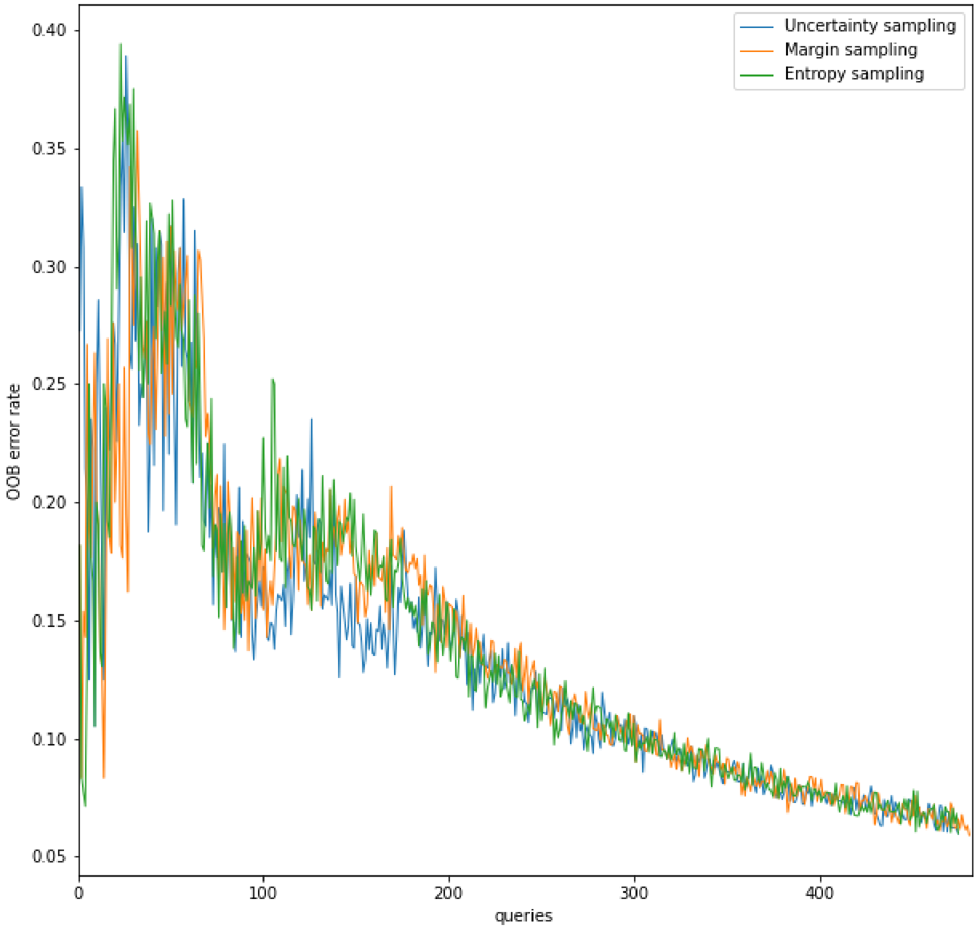


Figure 27: *Performance of the active learning algorithm across uncertainty, margin and entropy sampling strategies.*

Inspecting this graph, one can further see that initially when the training set is low, subsequent queries and re-training comes with a certain level of noise and outliers before the error converges. This can be seen in the way the curve with various frequency oscillates in error early on in the query process. As more data is acquired for learning, the error stabilizes and converges roughly around 200 obtained labels in this particular case. This relationship can further be highlighted in the progression curves for the model during the active learning loop for a single strategy. Figure 32 plots the progression using margin sampling, where model selects for absolute minimal difference between label candidates.

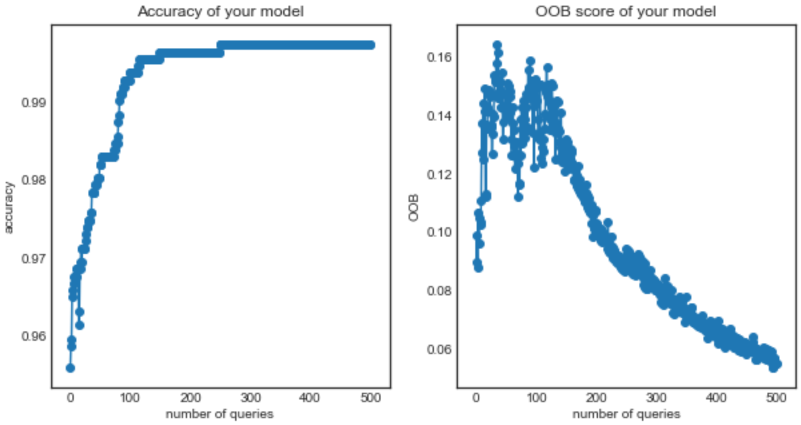


Figure 28: *Training progression for the active learning algorithm using margin sampling.*

As a test and a proof of concept of the information gain from the active learning loop, one could hypothesize that by the time the model is finished training, it has selected for the most difficult instances in the data set, like Methoxychlor in table 7. One can compile the remaining non-queried substances in U as a sample test set to verify this being the case. The confusion matrix can be found below,

		Actual	
		Non-PBT	PBT
Pred	Non-PBT	482	0
	PBT	3	103

where the balanced accuracy is at 99%, indicating that the most informative samples have been included for in the training data. This further includes 103 out of total 236 PBTs, and 485 out of 880 non-PBTs in the overall dataset. The further misclassifications are all false positives;

- Benzenemethanol, 2,4-Dichloro-
- Benzenemethanesulfonyl chloride
- Dioctyl Sebacate

A salient remark here is that this is by no means representative of a external validation set, and is a mere highlight of the fact that the approach of selecting for optimal training information works as intended.

One can further inspect the feature importance of the active learning model in figure 29,

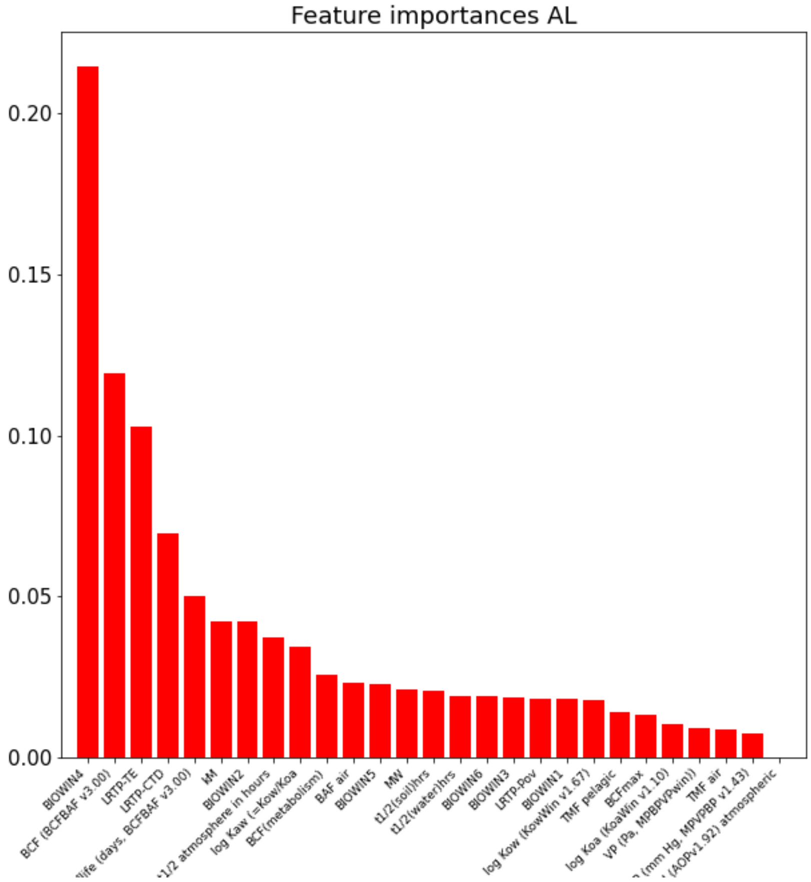
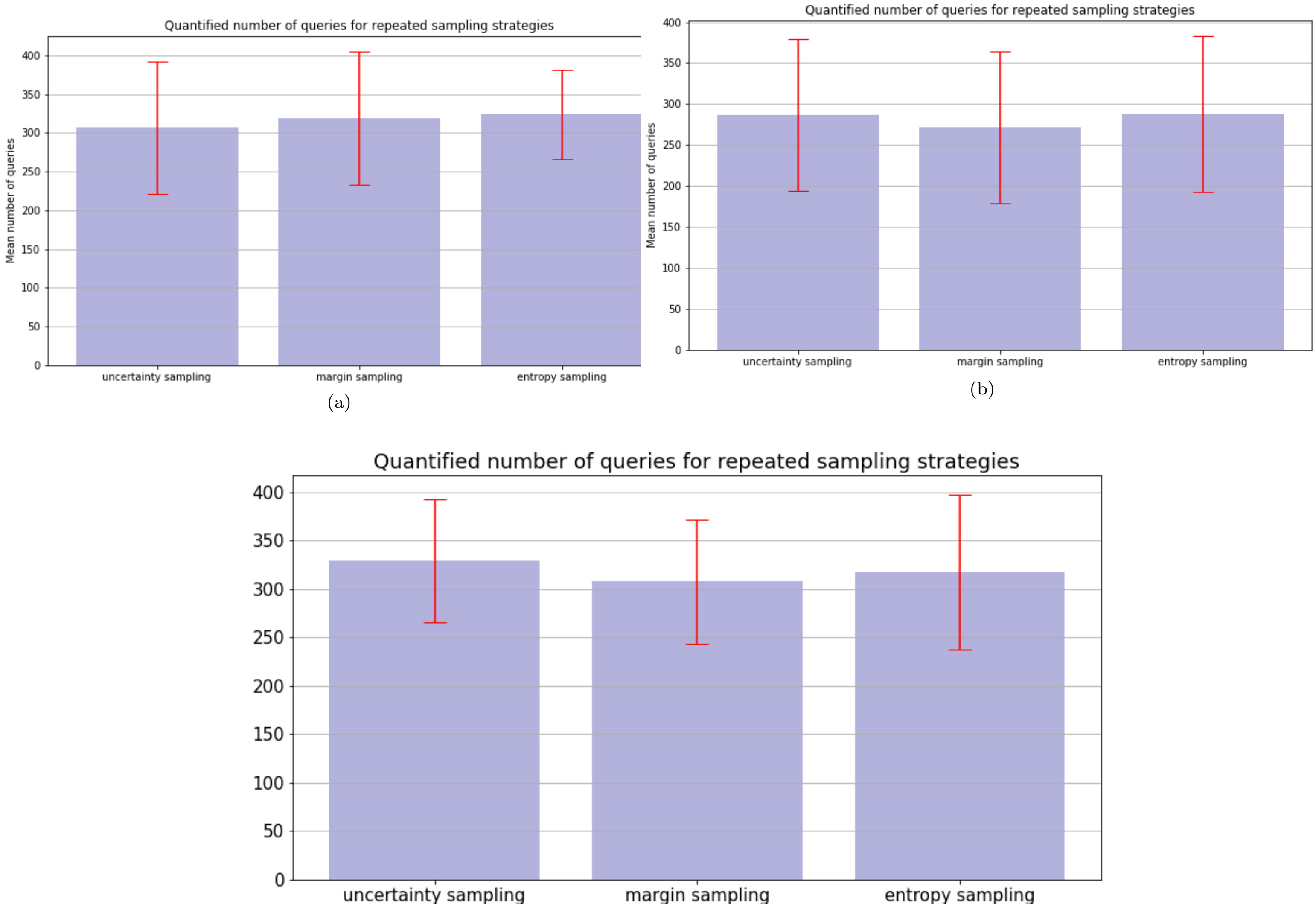


Figure 29: *Variable importance for the active learning model with margin sampling*

where as seen with the naive passive model, Biowin4, LRTP TE, LRTP CTD and BCF BAF measures again make out the four most important features.

The result of the initial test loop however is not too reliable. The reason for this is the notion of randomness in initial training set allocation. For one, one can hypothesize that the larger the initial training set is, the less queries is ultimately needed to achieve the performance of the benchmark. This can simply be due to the fact that the model starts with more information available to fit the model, however the queries performed by the model is at any point in time a reflection of its current training set distribution, thus if by chance the initial training set automatically includes for data points that describes the larger data distribution, one can expect even fewer queries needed. A related point is that these sets are in turn randomized from the larger data space, and as such one can not be sure whether the sample drawn is a representative training set or has an appropriate label distribution.

This was explored by creating an experimental loop of 100 iterations. Within each iteration, three different forest models are fitted each with their own uncertainty sampling strategy and performance criteria(the binary benchmark). Further, the initial training sets for the models in each iteration is randomized for an appropriate size. The active learning process then starts and the models train and query until the criteria is met before the next iteration starts. Models are subsequently reset and their performance data is stored. Initial train sets are further reinitialized before the next iteration begins. This loop was repeated for sizes 50, 100, and 150 initial training indices.



(a)

(b)

(c)

Figure 30: *Experimental loop performance for average queries needed to equal the benchmark model across three training set sizes.*

Figure 30 shows the overall performance across the different training set sizes across query strategies, where the mean number of queries and standard deviation is plotted for a full experimental loop. Figure (a) denotes results for initial train size of 50, (b) denotes an initial size of 150 and (c) denotes a size of 100. Albeit with marginal difference, one can conclude that the decrease in initial training set size leads to an increase in total number of queries. This is highlighted by the fact that on average difference between (b) and (c), where average number of queries is in the range of 300-350 for an initial set of 100, however this drops to just below 300 when we increase the size of the set to 150.

A further highlight to the impact of the initial set can be seen in the red deviation bars plotted for each strategy, which denotes the standard deviation for number of queries that were used to achieve target results. As an example, for some entropy sampling training sets, the model spent closer to 400 queries in a loop or down to 250 to achieve target performance for some loop iterations. As seen in fig 27, progressively adding data leads to noisy oscillation before error convergence happens. We can plot the experimental loop results as in figure 27. The idea is to plot the mean error and standard deviation for error progression throughout the experimental querying process.

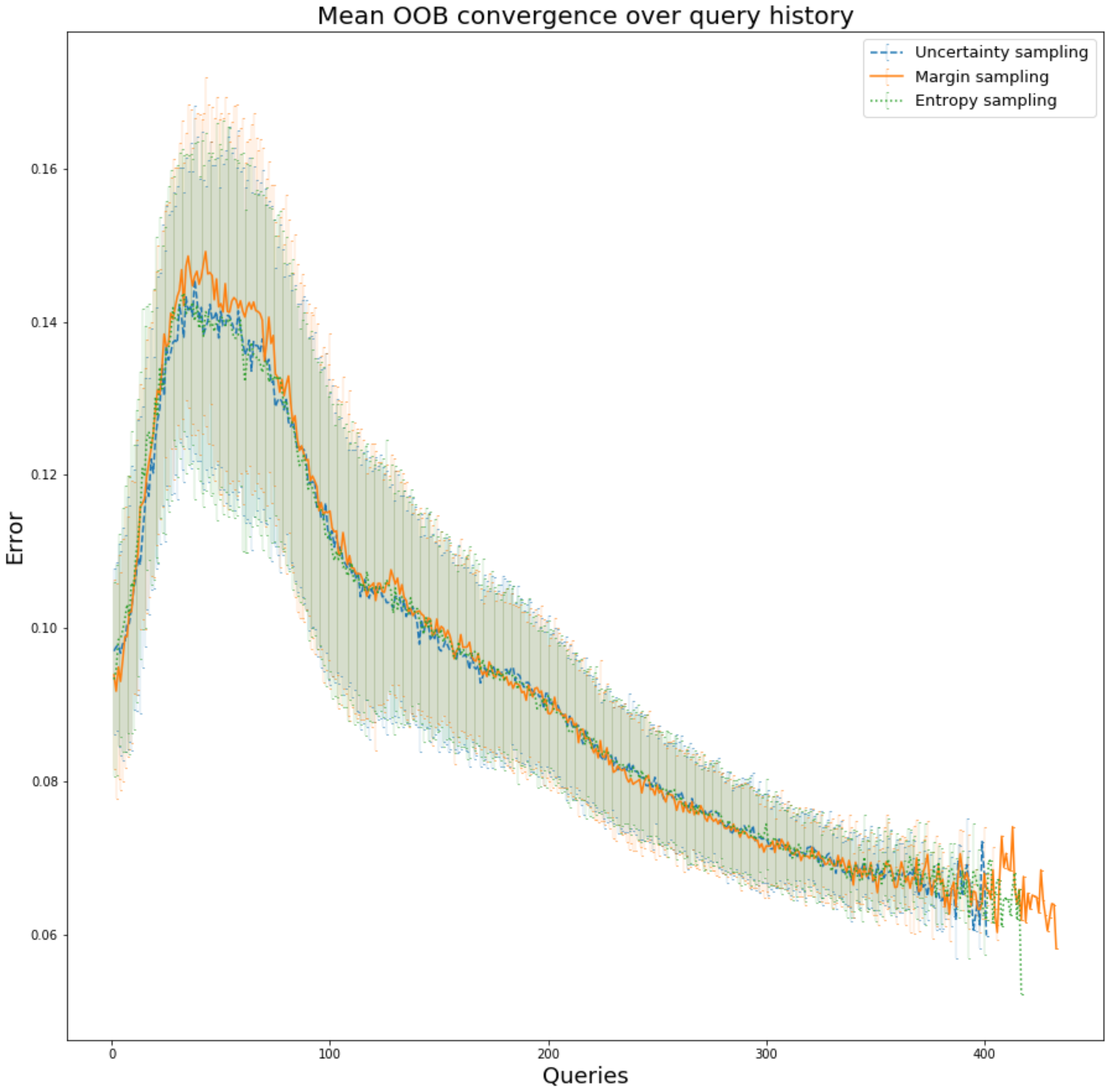


Figure 31: *Mean error convergence for querying strategies throughout the looped experiment.*

Figure 31 shows the expanded results of figure 30, where the mean error progression for the different query strategies is shown on the y axis, and where x indicate the number of queries done in the given iteration. Standard deviation at every query index in the experimental loop is further plotted for each strategy. As seen in figure 27, standard deviations here too show a noisy initial stat for the querying process for all strategies, which could further point toward the model acquires data points that are anomalies or outliers based on its current training set distribution. This error further converges once more data is acquired. This particular plot further shows the progression for each strategy starting with 50 initial substances, where margin sampling on average needed more queries than uncertainty and entropy sampling.

The active learning improvement can further be quantified by plotting the learning curve of a cross-validated random forest model over different training sets to examine the bias and variance trade-off. In other words, what is the learning progression for a passive model that slowly increases its training data against a model that learns on handpicked datapoints from the active learning procedure. Throughout the process, the trained model is tested on a validation set for that respective training set size using 10-fold cross-validation.

The active learning training indices were selected based on an initial set combined with the subsequent queried instances when the model was given a budget of 500 queries. This can be seen in figure 32.

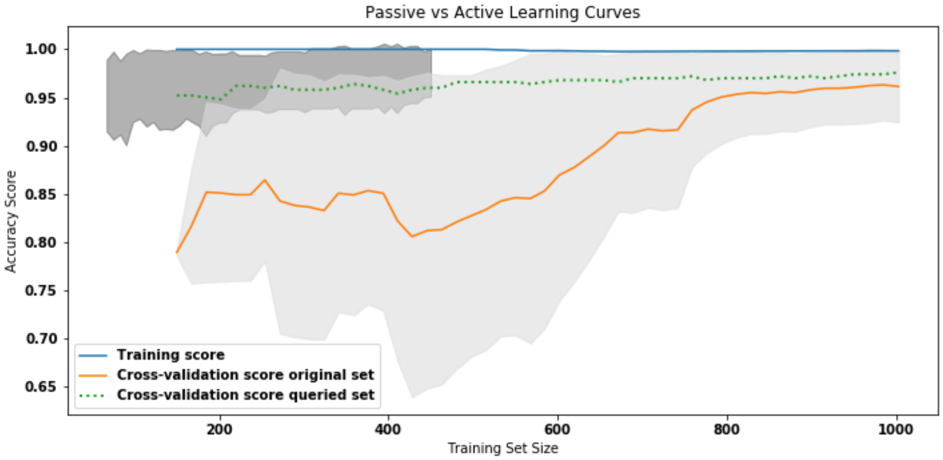


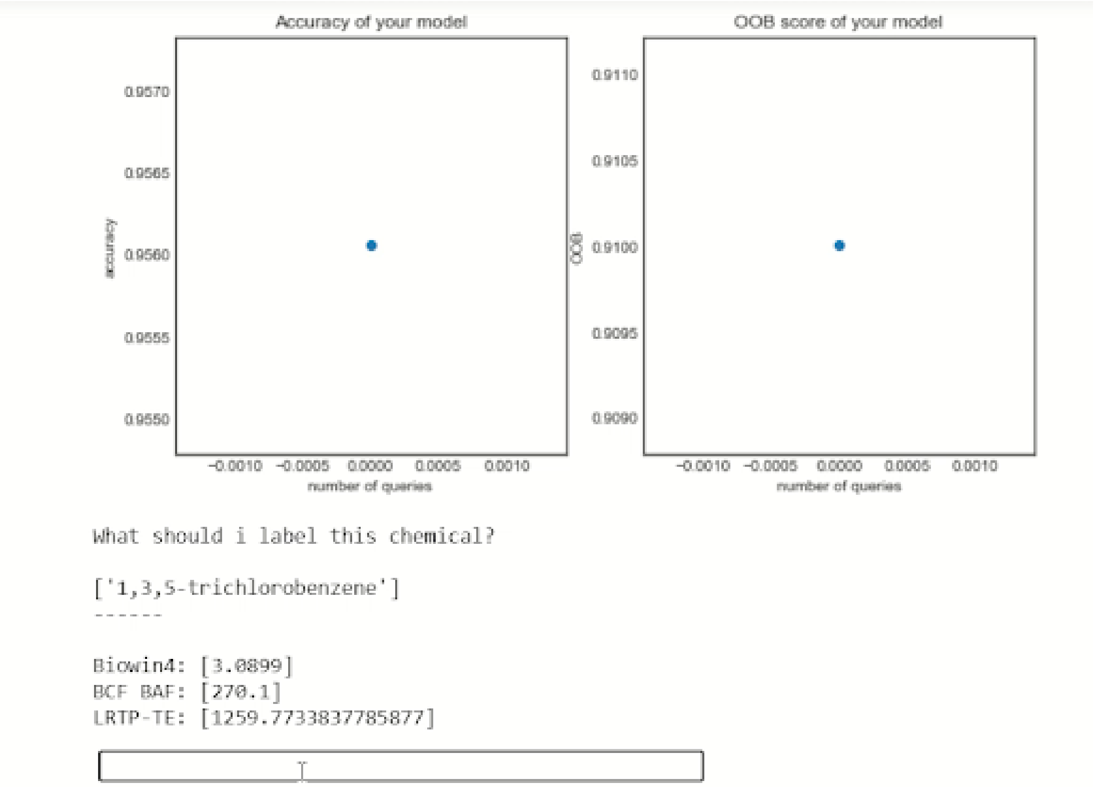
Figure 32: *Model learning curve for for original training set versus a queried training set. Shaded polygons indicate the variance range for the models.*

A first observation is that for a passive model approach, the increase in training data does not only *not* improve accuracy scores for the model, but leads to an increase in variance for a training set size of 150. One could expect that the variance was the same at 150 as it would be at a size of 200, however this could be due to the additional data added to the set is noisy and does not match ultimately what is tried to be predicted. The size of the variance remains rather large for sizes 250-650 before converging as expected. Conversely, the queried data set both increase accuracy scores and have lower variance denoted by the dark-shaded polygon. For the sake of clarity, the training curve that is plotted is singular here as the training score for both models were the same and thus it was rendered redundant for the inclusion of both. This further lends credence to the hypothesis that the valuable data in the overall data set has been selected for in the queried set.

6.3.1 Interactive Learning demo

As a demo for what is possible, some conceptualization of interactive labeling was explored. As noted earlier, active learning allows for label acquisition using a domain expert. This usually would come in a form where the model queries an instance from the larger pool U in this case and subsequently has the expert annotate the data point before adding it to its larger training set. This requires some elaboration on what information an expert would need to make a decision. The model could for example report the SMILES code for a substance, which is a structural description in the form of a string. However, the activity of decoding the descriptors could be industrious and time consuming. Another way would be to report the CAS registration number which is unique to any chemical, which would facilitate an easy search in a chemical database, say. This would further require that experimental data and label exists for the chemical, which is not a guarantee within the area of toxicology.

The approach attempted for this paper includes the model reporting the name of the chemical, and subsequent what has earlier been established to be important features for the model. This is due to the fact that the feature importance described in the passive approach was in accordance with expert intuition, and it allows for the required data for appropriate labeling to be explicitly reported in the graphical user interface. As a demo, figure 33 shows what this could look like. The initial training set size for this particular model was at 100 substances.



```
What should i label this chemical?

['1,3,5-trichlorobenzene']
------

Biowin4: [3.0899]
BCF BAF: [270.1]
LRTP-TE: [1259.7733837785877]
```

Figure 33: *Screen grab from Interactive initial learning annotation demo with an expert, where the model requests labels for data it is unsure about and the data points' feature characteristics.*

Here the active learning sequence has just started, as the performance graphs indicate. The first query comes in the form of '1,3,5-Trichlorobenzene', an industrial chemical or termite preparation and insecticide. The model asks what to label the chemical and supplies its Biowin4, BCF BAF and LRTP-TE metrics. Based on the metrics, the expert makes a decision and types either the number 0, 1 or 2. Typing 0 denotes the chemical as a non-PBT, 1 as a PBT and 2 discards the instance from consideration. This leads it to delete the query from the unlabeled pool, and the sequence continues. This is done to allow for expert uncertainty, as giving either wrong or right chemical assignment will influence the classification capabilities of the model, but further the subsequent queries the framework makes. This is continued until the assigned budget of queries is fulfilled, however as seen in the more quantified automated approach, having an expert go through over 400 queried substance is unrealistic. The value lies in the possibilities of - while being wary of noisy data points - being able to get a sense of the impact on the larger predictive space by adding a given substance, and further enables cross-examination of expert opinion.

7 Discussion

Results presented in this paper lend credence to some of the initially stated hypotheses. For one, using machine learning with physical chemical properties to classify harmful substances is shown to be a novel and useful approach in PBT screening. Further, a cross-validated balanced accuracy of 94.2% for the naive random forest model is further a testament to the predictive capabilities of the approach. Indeed, all models presented exceed performance of existing screening tools for PBT substances by the RIVM[50], with a balanced accuracy >90%. However, as seen with the reductionist model in which otherwise important predictor variables are excluded, performance metrics are quite high at 93% balanced accuracy. This further leads to the question whether the model fit is good due to the separability of classes or whether the data used for modelling purposes is representative of the wider chemical space. For example, known PBTs in the data set are substances that are known to be particularly egregious in their toxic affects, while confirmed non-PBTs includes substance on the other end of the spectrum, like Glucose. This in turn can be reflexive of the rather stark difference in feature distribution for either class, as the data might be constructed from the outer edges of either category.

An addendum to this point is if one takes in to account the fact that as stated earlier, the field is plagued by lack of experimental data and labels, thus the existing labels for known PBTs are substance that have been selected for screening based on the *assumption* that they could be troublesome and subsequently tested, whereas the known non-PBTs are data from readily biodegradable tests, which includes a number of commercially used pesticides. Thus an argument could be made that the data distribution is indeed representative, and that the labels themselves need verification. The reason for this is the nature of the misclassifications performed by the model, in which it "correctly" gets the labels wrong for substances like Naphtalene, Indene[49] and according to expert evaluation Vinyl Neodecanoate. There are however evidence for the counterclaim, namely that the models gets substances like Methoxychlor incorrect.

One can further make an argument that the approach in this paper has a much more narrow area of focus on PBTs than the more generalizable RIVM tool which utilized the binary fingerprint approach. This per say is not a limitation, as models like these allow for qualitative exploration of the PBT space. In other words, two-dimensional structural fingerprints generalizes better to other substance groups like CMTs and EDs more than modelling on physical properties does, but nevertheless do not offer a ranked ordering of substructures for a given outcome, as a substance either has a substructure or it does not. The approach in this paper not only is a benefit of being an added tool for substance evaluation, but can further capture additional information that related work on fingerprint approaches do not offer. This can be further highlighted by the fact that model outputs confirmed *a priori* expectations of RIVM experts but with added important feature discovery such as POP-specific criteria, or the relative heightened importance of Biowin4 over Biowin3 metrics, which in earlier reports were left out or not deemed as important[36]. As seen further with sample tree fits, variable splits seemed to pick up on European PBT evaluation guidelines, while saliently noting that the data was not collected based on these same guidelines.

Results surrounding the active learning improvement remains inconclusive for classification improvement, but has the added benefit of reducing the data requirement by active selection of informative training samples, as highlighted by the difference in learning curves for passive models utilizing sub sampled training sets of the larger data sphere, and the actively queried set. One can interpret this as saying that selective acquisition of a data set reduce variance, but potentially increase the bias. Further, that the selected data can be deemed to be the most informative data points in the overall data set. The question however is if the data points are the most informative in terms of covering the wider chemical space, or whether they are the most informative based purely on label probability, given the strategies used. An argument for the former is the fact that the final distribution for the queried set incorporated roughly 45-50% of either class.

The error progression as seen in figure 31 further does not favor active learning for model training, as the model selects for noisy data points due to the notion of quantifying uncertainty. If the characteristics of substance is far removed from the distribution of the training set that is being quried, one can potentially expect the model to be sensitive to outliers. As discussed earlier, the overall representativeness of the data is a discussion point here as well.

Finally, albeit shown as a proof of concept in this paper, active learning could be used for expert solicitation of substance labels. This comes with a few assumptions. For one, expert evaluation would need to be sufficiently sturdy, as the solicited label has an effect on the subsequent performance of the model, which is a problem as expert evaluation is not always uniform and faultless. A second assumption is that the information provided for the annotator is sufficient and complete, and finally that the selected substance for the query *is* the optimal selection and derived from an optimal strategy. Assuming these are covered, then the interative labeling concept can be expanded to explore model change for data selection, but also questions for why a certain substance is being queried based on the current membership of the training data used. Albeit inconclusive, this could be an approach in which outliers are identified and scrutinized further, and what characteristics these substances have that separates them or make them ”uncertain” for the model.

7.1 Limitations

In light of the discussion on the data, the size of the training data asserts that the results need further verification, especially in form of a more robust external validation set. Although cross-validation was here used due to the cost of setting aside a large chunk of an already smaller data set, a validation set would include for substances that do not come from the original training distribution, and hence would serve as a better evaluation for generalization. The need for results verification is also due to the lack of comparable benchmarks for comparison in work done in the field, as most of them are oriented in the previously mentioned fingerprint and structural-activity approach, which in turn makes the overall gain of the results over other methods inconclusive.

A further limitation is the width and breadth of applied models for the subject. Albeit that the binary classification included two benchmarks of comparison, the nature of the strong imbalance of labaled data affects classifiers like the SVM which are prone to overfitting and class imbalance, as highlighted by the increased rate of false negatives.

Limitations surrounding the active learning approach can be seen in the behaviour of query strategies, in which uncertainty, margin and entropy sampling generally behaves the same way for binary problems. In other words, all three of them tries to select for samples around a 0.5 probability decision boundary. As the comparison with random sampling shows, the data distribution of the training set makes the active learning model sensitive to outliers, albeit optimal for selecting for informative data points as seen in the plotted learning curves for validation performance.

Albeit being a demo, the interactive labeling framework lacked proper testing as experts were not properly available, subsequently leading to formulation of important information needed discussed in theory only. Further, the ability to discard substances that are of uncertainty from the viewpoint of an expert or human annotator can be thought of as an overall loss of information, given that budgets in interactive settings will be smaller and subsequently making each query more important in terms of information gain. This however is a limitation that can be extended to expert opinion in general.

7.2 Future work

Some of the challenges and questions raised by the findings in the paper can be interesting points of research for the future, both to solve issues left by the constraints of time, and also to serve as a robustness checks for these findings. These are some of the ideas that i encountered that would be of use:

1. Studying the impact of experimental class weights for the benchmark classification models. Adding more data in a field like toxicology undoubtedly will maintain a certain level of labeled imbalance, thus experimenting with different class weights can give further weighting towards reduction of false negatives in a field where classifying otherwise harmful substances as safe would have more dire consequences than false positives. This coupled

with an expansion of other classification models, like the Naïve Bayes classifier could strengthen benchmark results and serve as a cross-reference for variable importance.

2. As structural similarity screening largely involves 2D binary bit strings for representations of chemical features, one could examine what a combination of both physio-chemical properties and binary bit strings would have for classification accuracy. A matter of salience would be keeping in mind the so-called curse of dimensionality, where the number of features we have exceeds the number of training samples, thus introducing a problem of modeling on data of higher dimensions that does not match observations in lower dimensions. Binary bit strings can reach sizes of thousands, thus by introducing binary fingerprints as a feature set one could explore the importance of different molecular sub-structures. This in turn can be optimized by crafting a unary fingerprint that combines different bit string approaches.

3. In light of point 2, one could extend the methods used in this paper to substance groups like EDs or CMTs, albeit with different feature sets(or fingerprints) for a machine learning algorithm, and to investigate what sort of feature engineering could be explored for such a task.

4. Further work on active learning can be done to better quantify improvement. For one, the same model used for training and testing is used for label acquisition, however one could investigate whether there are models that are better suited for label acquisition, apart from the random forest approach used here. In essence, one could separately use a model that obtains a data set, and subsequently have another model train and test on it. This can further be expanded under different query strategies and sampling techniques. For example, quantifiying uncertainty of an substance do not need to be performed by one model or uncetainty metric, but one could introduce a query-by-committee (formula 13) framework, where labels that garner the *highest* amount of disagreement is the one that ultimately gets annotated.

5. subclass classification of substances in active learning was not further pursued to the performance of the passive approach, however one could look into the possibility of utilizing active learning for subclass classification using a multi-class probabilistic active learning framework(McPAL[25]), where querying of a substance takes in to account its posterior probability, the reliability of this posterior and also the expected gain of its selection.

6. Finally, it can be interesting from a research perspective to use interactive substance labeling as a cross-reference test for expert opinion. This can be done to investigate the accuracy of human annotation, but also quantify the level of agreement among experts. Factors here can include developing a proper graphical user interface that allows for a database lookup while offering the expert to select for columns based on chemical name or structure.

8 Conclusion

In this thesis, an attempt at exploring the gain of utilizing machine learning for PBT substance classification was done for the Dutch national institute for public health and environment. One of the main contributions of the work done has been the introduction of ranked importance for physio-chemical properties along with the ability to screen and classify PBT substances on said properties, all the while honoring the need for explanatory models. These models were two benchmarks fit on filtered and naive feature inclusion, with a subsequent active learning improvement for the better-performing model. In the end, the naive RF model obtained the best results, further showing improvement over existing screening tools for PBT substances, with an additional capability of identifying legal guidelines for assessment. Further contribution came in the form of active learning, in which handpicked data collection based on a notion of uncertainty sampling managed to reduce variance over a passive model approach, albeit inconclusive for model training improvement. This approach further introduced a demo for the potential for expert solicitation, paving the way for qualitatively examine impact of expert evaluation.

Though further data is needed to solidify the performance of the best performing model, in conclusion the thesis establishes the gain of adding machine learning as an added tool for efficient risk assessment beyond predictive ability, and to further provide width and breadth to the study of substances of high concern.

9 Acknowledgements

I would first like to thank my thesis advisor Ass.-Prof. Dr. Habil. Georg Krempl for not only introducing me to the interesting concepts of active learning, but for providing continuous guidance and tips during a busy schedule and circumstance. Further heartfelt thanks goes out to the numerous individuals who has supported the project and the writing process. Special thanks to Dr. Emiel Rorije(RIVM), for continuous enthusiasm and always being available in explaining the nuances of risk assessment and PBT substance evaluation throughout the internship. Also someone who continuously has held the door open and provided invaluable guidance is Dr. Albert Wong(RIVM), whos influence on the project approach and thesis input shaped what the final project looks like. I would also like to lend thanks to Rob Beffers(RIVM) and Roel Schreurs(RIVM), for all the followup, tips, support and keeping me on my toes in terms of planning and execution.

And finally, a special thanks to family and friends who have remained patient and supportive throughout the writing process.

References

[1] Methoxychlor - substance information. *Journal ECHA, European Chemical Agency.*

[2] Un ghs - globally harmonized system of classification and labeling of chemicals, Dec 2015.

[3] Ddt - a brief history and status. *Environmental Protection Agency, United States.* , published 2017, Aug 2017.

[4] The oecd qsar toolbox, *updated 2020, April*, Apr 2020.

[5] American congress. H.r.2576 - frank r. lautenberg chemical safety for the 21st century act, 2016. `https://www.congress.gov/bill/114th-congress/house-bill/2576`.

[6] D. Angluin. Queries and concept learning. 1988.

[7] M.-F. Balcan and P. Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pages 288–316, 2013.

[8] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 144–152, New York, NY, USA, 1992. Association for Computing Machinery.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and regression trees. 1984.

[10] Cefic. Facts figures of the european chemical industry, 2018. `https://www.apquimica.pt/uploads/fotos_artigos/files/cefic-facts-and-figures-2018-industrial.pdf`.

[11] F. Cheng and Z. Zhao. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association*, 21(e2):e278–e286, 03 2014.

[12] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[13] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[14] I. Dagan and S. P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Machine Learning Proceedings 1995*, pages 150–157. Elsevier, 1995.

[15] D. J. Dix, K. A. Houck, M. T. Martin, A. M. Richard, R. W. Setzer, and R. J. Kavlock. The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals. *Toxicological Sciences*, 95(1):5–12, 09 2006.

[16] J. Duan, S. L. Dixon, J. F. Lowrie, and W. Sherman. Analysis and comparison of 2d fingerprints: insights into database screening performance using eight fingerprint methods. *Journal of Molecular Graphics and Modelling*, 29(2):157–170, 2010.

[17] European council. European chemical agency, 2006.
`https://ec.europa.eu/environment/chemicals/reach/reach_en.htm`.

[18] European council. The classification, labelling and packaging of chemical substances and mixtures, 2015.
`https://ec.europa.eu/environment/chemicals/labelling/index_en.htm`.

[19] European Parliament. Council regulation (EU) amendment of no 76/768/eec, 2003.
`https://op.europa.eu/en/publication-detail/-/publication/60a70768-2dcf-4771-87b2-194ed4ec0012`.

[20] European union. European chemical agency, 2007.
`https://echa.europa.eu/nl/home`.

[21] R. Hwa. Sample selection for statistical parsing. *Computational linguistics*, 30(3):253–276, 2004.

[22] A. Kale. From medium article - depicts active learning pool loop, Aug 2018.

[23] R. J. Kavlock, G. Ankley, J. Blancato, M. Breen, R. Conolly, D. Dix, K. Houck, E. Hubal, R. Judson, J. Rabinowitz, A. Richard, R. W. Setzer, I. Shah, D. Villeneuve, and E. Weber. Computational Toxicology—A State of the Science Mini Review. *Toxicological Sciences*, 103(1):14–27, 12 2007.

[24] M. A. Kayala and P. Baldi. Reactionpredictor: Prediction of complex chemical reactions at the mechanistic level using machine learning. *Journal of Chemical Information and Modeling*, 52(10):2526–2540, 2012. PMID: 22978639.

[25] D. Kottke, G. Krempl, D. Lang, J. Teschner, and M. Spiliopoulou. Multi-class probabilistic active learning. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pages 586–594. IOS Press, 2016.

[26] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[27] T. Lang, F. Flachsenberg, U. von Luxburg, and M. Rarey. Feasibility of active machine learning for multiclass compound classification. *Journal of Chemical Information and Modeling*, 56(1):12–20, 2016. PMID: 26740007.

[28] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, page 3–12, Berlin, Heidelberg, 1994. Springer-Verlag.

[29] X. Li, L. Chen, F. Cheng, Z. Wu, H. Bian, C. Xu, W. Li, G. Liu, X. Shen, and Y. Tang. In silico prediction of chemical acute oral toxicity using multi-classification methods. *Journal of Chemical Information and Modeling*, 54(4):1061–1069, 2014. PMID: 24735213.

[30] D. Lud. *Stockholm Convention (2001)*, pages 1–8. Springer International Publishing, Cham, 2020.

[31] T. Luechtefeld and A. Maertens. Analysis of draize eye irritation testing and its prediction by mining publicly available 2008–2014 reach data. *ALTEX*, 33, 01 2016.

[32] T. Luechtefeld, A. Maertens, D. Russo, C. Rovida, H. Zhu, and T. Hartung. Analysis of publically available skin sensitization data from reach registrations 2008-2014. *ALTEX : Alternativen zu Tierexperimenten*, 33(2):135–148, 2016.

[33] T. Luechtefeld, D. Marsh, C. Rowlands, and T. Hartung. Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility. *Toxicological Sciences*, 165(1):198–212, 07 2018.

[34] F. Olsson. A literature survey of active machine learning in the context of natural language processing. 05 2009.

[35] D. Reker, P. Schneider, and G. Schneider. Multi-objective active machine learning rapidly improves structure–activity models and reveals new protein–protein interaction inhibitors. *Chemical science*, 7(6):3919–3927, 2016.

[36] E. Rorije, E. Verbruggen, A. Hollander, T. Traas, and M. Janssen. Identifying potential pop and pbt substances: Development of a new persistence/bioaccumulation-score. 2011.

[37] C. Rovida and T. Hartung. Re-evaluation of animal numbers and costs for in vivo tests to accomplish reach legislation requirements for chemicals - a report by the transatlantic think tank for toxicology (t(4)). *ALTEX*, 26 3:187–208, 2009.

[38] N. Roy and A. McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.

[39] S. Russell and P. Norvig. Artificial intelligence: a modern approach. 2002.

[40] M. Scialla. It could take centuries for epa to test all the unregulated chemicals under a new landmark bill. *PBS News Hour*, Jun 2016 - (last accessed December 19, 2019).

[41] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[42] B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *Advances in neural information processing systems*, pages 1289–1296, 2008.

[43] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 287–294, New York, NY, USA, 1992. Association for Computing Machinery.

[44] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[45] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*, 2017.

[46] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, and P. Willett. Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling*, 52(11):2884–2901, 2012. PMID: 23078167.

[47] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118, 2001.

[48] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[49] E. United Kingdom. Evaluation report for napthalene. `https://echa.europa.eu/documents/10162/c5cb00e9-0ff4-ac35-3db1-24566967fea8`.

[50] P. N. Wassenaar, E. Rorije, N. M. Janssen, W. J. Peijnenburg, and M. G. Vijver. Chemical similarity to identify potential substances of very high concern–an effective screening method. *Computational Toxicology*, 12:100110, 2019.

[51] C. S. Weil and R. A. Scala. Study of intra- and interlaboratory variability in the results of rabbit eye and skin irritation tests. *Toxicology and Applied Pharmacology*, 19(2):276 – 360, 1971.

[52] K. R. Wilhelmus. The draize eye test. *Survey of Ophthalmology*, 45(6):493 – 515, 2001.

[53] P. Willett. The calculation of molecular structural similarity: Principles and practice. *Molecular Informatics*, 33(6-7):403–413, 2014.

[54] Y. Xu, J. Ma, A. Liaw, R. P. Sheridan, and V. Svetnik. Demystifying multitask deep neural networks for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling*, 57(10):2490–2504, 2017. PMID: 28872869.

[55] V. Yadav. Radial svm example, Nov 2016.

[56] Z. Zheng and B. Padmanabhan. On active learning for data acquisition. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 562–569. IEEE, 2002.

[57] S. Zhou, Q. Chen, and X. Wang. Active deep learning method for semi-supervised sentiment classification. *Neurocomputing*, 120:536–546, 2013.

[58] S. Zomer, M. Del Nogal Sánchez, R. G. Brereton, and J. L. Pérez Pavón. Active learning support vector machines for optimal sample selection in classification. *Journal of Chemometrics*, 18(6):294–305, 2004.

10 Appendix

Appendix A Data

Property Name	Description
k OH (AOPv1.92) atmospheric*	Denotes the rate of atmospheric degradation for a chemical
t1/2 atmosphere in hours	Denotes the half-life time of a substance found in the air measured in hours.
VP (mm Hg)*	Denotes vapor pressure of a substance, measured in millimetres of mercury
VP (Pa)	Denotes vapor pressure of a substance, measured in Pascal
Biowin1	Denotes the linear model output that predicts slow vs not slow degradation.
Biowin2	Denotes the non-linear version that predicts slow vs not slow degradation.
Biowin3	Denotes the estimates of environmental half-life necessary to mineralize a chemical (i.e to turn 50% of the substance in to the ultimate degradation products - namely water and carbon dioxide
t1/2(water)hrs*	Denotes the half-life time of a substance found in water measured in hours.
t1/2(soil)hrs*	Denotes the half-life time of a substance found in soil measured in hours.
Biowin4	Denotes regression model prediction on expert estimate of environmental half-lives for primary degradation
Biowin5	Denotes linear model prediction output of the ”ready biodegradability test”
Biowin6	Denotes non-linear model prediction output of the ”readily biodegradability test”
log Kow	Denotes octanol water coefficient. Describes the ratio of the concentration of a substance in an octanol phase and its concentration in the aqeuous phase. Higher values of log Kow implies a higher potential to bioaccumulate in living organisms. (for example fish).
log Koa	Demotes the octanol air coefficient.
log Kaw	Denotes the air-water partition coefficient, estimated for compounds at 25 degrees
BCF (BCFBAF)*	Denotes the bioconcentration factor, but adjusted for the impact of the hosts' metabolism(aquatic species)
BCFmax*	Denotes a bioconcentration factor for a substance to bioaccumulate in aquatic species
kM halflife (days)*	Denotes the depuration rate or half life of a substance metabolized in aquatic species, measured in days
kM	Denotes the flat depuration rate constant of a substance(affects the half life rate)
BCF(metabolism)*	Denotes the bioconcentration factor, not adjusted for metabolism rate.
TMF air*	Denotes the trophic magnification factor, and how substances may accumulate up the food-chain for air breathing organism, like mammals or birds.
TMF pelagic*	Denotes the trophic magnification factor for pelagic species (in essence, fish that live neither near the surface or bottom of a body of water)
BAF air*	Denotes the bioaccumulation factor for air breathing organisms. Takes in to account the bioconcentration factor of a substace adjusted for metabolism and the TMF.
LRTP-Pov	Denotes the long range transport potential for a substance measured in days. Includes for soil, water and air
LRTP-CTD	Denotes the long range transport potential for a substance, and its characteristic travel distance. Includes for soil, water and air.
LRTP-TE	Denotes the long range transport potential for a substance, and its travel efficiency. Includes for soil, water and air.
MW	Molecular weight of a chemical. Calculated directly from its chemical structure
**Dependant variable*	*dependant variables means that is a compount calculation of other physical properties or different representation.* *An example: 1/2 life in soil = 2* 1/2 in water*

Table 10: Full table of physical chemical properties used for modelling.

Appendix B Descriptive statistics

Non-PBT	**mean**	**sd**	**median**	**min**	**max**	**skew**
k OH (AOPv1.92) atmospheric	5.4322e-11	7.6109e-11	2.582e-11	0	7.38272e-10	3.392
t1/2 atmosphere in hours	4.3809e+9	1.2988e+11	1.4914e+1	5.2159e-1	3.8508e+12	29.546
VP (mm hg, v1.43)	8.2681e+1	1.2400e+3	1.0225e-4	0	3.14e+4	21.788
VP (Pa)	1.0996e+4	1.6492e+5	1.36e-2	0	4.1762e+6	21.788
Biowin1	5.7096e-1	5.7706e-1	0.7005	-3.2787e+0	1.7401e+0	-2.875
Biowin2	5.9003e-1	4.1113e-1	7.925e-1	0	1	-04.235
Biowin3	2.5854	6.2096e-1	2.7091	-1.869e-1	4.2257e+0	-1.0847
t1/2(water) in hours	2.7179e+3	1.0897e+4	6.5359e+2	3.3906e+1	1.8582e+5	10.680
t1/2(soil) in hours	5.4358e+3	2.1794e+4	1.3071e+3	6.78124758575870e+1	3.7165e+5	10.680
Biowin4	3.5887	3.9292e-1	3.6061	1.9534	5.0192	-29.387
Biowin5	3.7521e-1	3.8297e-1	3.866e-1	-1.0966	1.4913	-22.253
Biowin6	3.6991e-1	3.6840e-1	2.361e-1	0	9.984e-1	0.4278
log Kow v1.67	2.8548	2.9785e+0	2.62	-1.728e+1	2.432e+1	1.212
log Koa v1.10	1.4528e+1	3.0834e+2	8.334	-9.99e+2	8.921e+3	27.254
log Kaw	-1.5502	6.7567e+1	-5.549	-3.5428e+1	1.00214e+3	10.460
BCF (BCFBAF) v3.00	2.6255e+2	1.0592e+3	1.5848e+1	6.197e-1	1.4454e+4	8.0524
BCFmax	4.7914e+3	1.1601e+4	7.0722e+1	1	4.8216e+4	2.659
kM 1/2 in days v3.00	1.0260e+3	2.4732e+4	2.3252e-1	7.59e-12	7.1635e+5	27.809
kM	1.824e+1	3.3758e+1	2.981	9.676e-07	125	2.272
BCF (metabolism)	3.2006e+2	1.1707e+3	2.8345e+1	6.6567e-1	1.9204e+4	10.248
TMF air	-7.1528e+4	2.0179e+6	3.6290	-5.9811e+7	9.9088	-29.518
TMF pelagic	-7.8306e+4	2.2094e+6	-1.0695	-6.5487e+7	6.0493	-29.518
BAF air	1.2650e+3	4.0955e+3	1.0427e+2	6.6567e-1	4.0738e+4	6.096
LRTP Pov	3.8064e+2	1.5581e+3	7.3315e+1	2.0095	2.9645e+4	11.134
LRTP CTD	1.6776e+4	8.6295e+4	7.7247e+2	6.269	1.1913e+6	8.875
LRTP TE	3.8918e+1	2.1108e+2	1.0255	1.446e-12	2.7684e+3	8.641
MW	2.3622e+2	1.3178e+2	210.36	27.03	1080.96	1.367

Table 11: Full descriptive statistics of non-PBTs in the dataset. Values have been reduced to a four decimal representation where applicable for ease of interpretation.

PBT	**mean**	**sd**	**median**	**min**	**max**	**skew**
k OH (AOPv1.92) atmospheric	1.3433e-11	2.9681e-11	7.6e-13	0	2.26529e-10	3.974
t1/2 atmosphere in hours	6.1515e+5	4.6662e+6	5.0687e+2	1.6999	3.5999e+7	7.420
VP (mm hg, v1.43)	7.8888	1.145436e+2	1.8744e-06	1.1e-14	1.7541e+3	15.078
VP (Pa)	1.0492e+3	1.5234e+4	2.493e-4	1.48e-12	2.333e+5	15.078
Biowin1	-2.3600e-1	5.4654e-1	-2.133e-1	-2.6882e0	1.4604	-0.590
Biowin2	6.7579e-2	2.2946e-1	0	0	1	3.306
Biowin3	1.2906	0.7429	1.3066	-2.2058	3.6719	-0.463
t1/2(water) in hours	7.6248e+4	6.4696e+5	1.0084e+4	9.9889e+1	9.5443e+6	13.613
t1/2(soil) in hours	1.5249e+5	1.2939e+6	2.0169e+4	1.9977e+2	1.9088e+7	13.613
Biowin4	2.4951	0.5504	2.5299	0.3189	4.6021	-0.131
Biowin5	-0.1026	0.2569	-0.0825	-0.9037	0.8004	-0.286
Biowin6	2.7396e-2	9.5438e-2	8e-04	0	0.874	5.479
log Kow v1.67	6.8017	1.9024	6.79	1.69	12.66	0.246
log Koa v1.10	10.1386	3.2016	9.84	-1.097	18.428	0.014
log Kaw	-3.3684	2.1789	-3.415	-11.025	6.597	0.150
BCF (BCFBAF) v3.00	9.1082e+3	1.3947e+4	4712	1.562	7.44e+4	2.855
BCFmax	2.1662e+4	1.7719e+4	2.1365e+4	3.3323	4.8216e4	0.172
kM 1/2 in days v3.00	2.7157e+2	1.3706e+3	26.7831	3.9927e-2	1.6885e+4	9.582
kM	0.2662	1.2857	0.02588	4.105e-05	17.36	10.806
BF (metabolism)	7.1837e+3	8.4802e+3	3.9536e+3	3.3323	3.0838e+4	1.344
TMF air	0.5138	14.6229	6.4447	-62.3481	9.9088	-2.475
TMF pelagic	-15.6112	25.3736	-6.4600	-118.6003	6.0457	-1.923
BAF air	5.5193e+4	7.9812e+4	1.7510e+4	3.3323	3.0153e+5	1.601
LRTP Pov	1.8185e+6	1.9579e+7	1.2111e+3	12.0010	2.1371e+8	10.632
LRTP CTD	6.4352e+4	1.9795e+5	5.9135e+3	82.2169	1.3499e+6	4.671
LRTP TE	2.2108e+2	5.8323e+2	22.5424	0.0128	5.0405e+3	5.154
MW	4.0016e+2	1.5707e+2	3.6492e+2	1.1616e+2	9.5917e+2	0.863

Table 12: Full descriptive stats over PBT substances in the data set. Values have been reduced to a three decimal representation where applicable for ease of interpretation

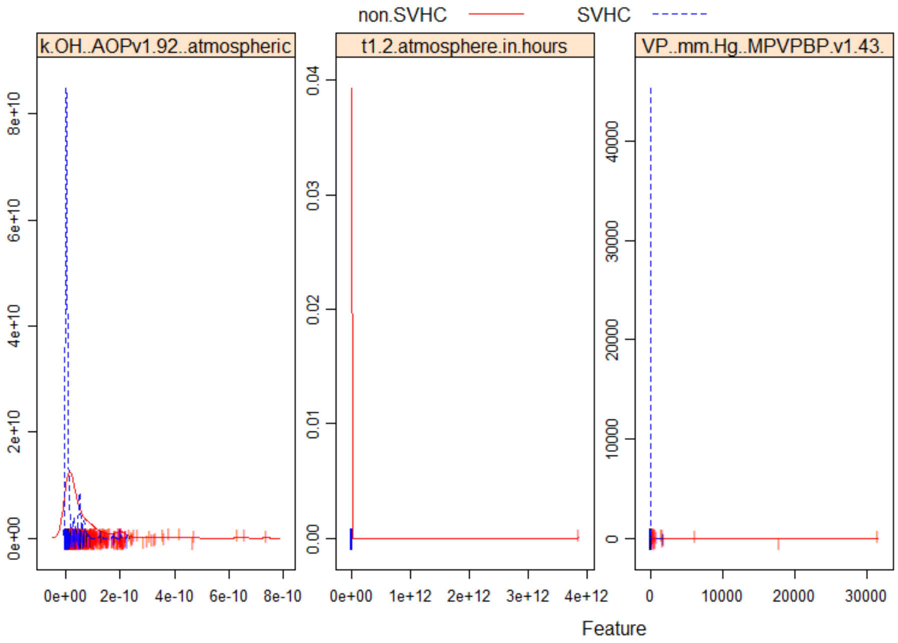


Figure 34: *Feature plot showing a subset distribution of features used for modelling purposes*

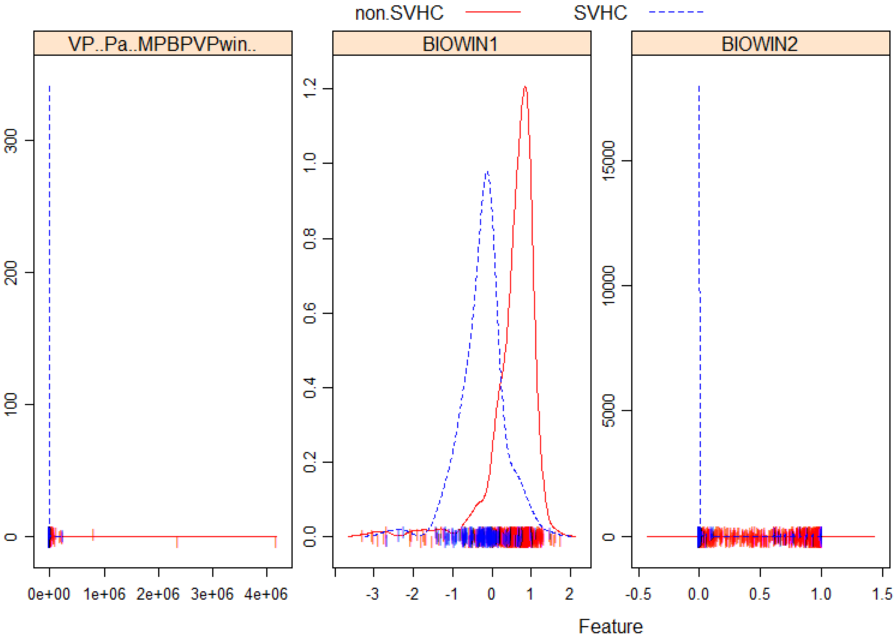


Figure 35: *Feature plot showing a subset distribution of features used for modelling purposes*

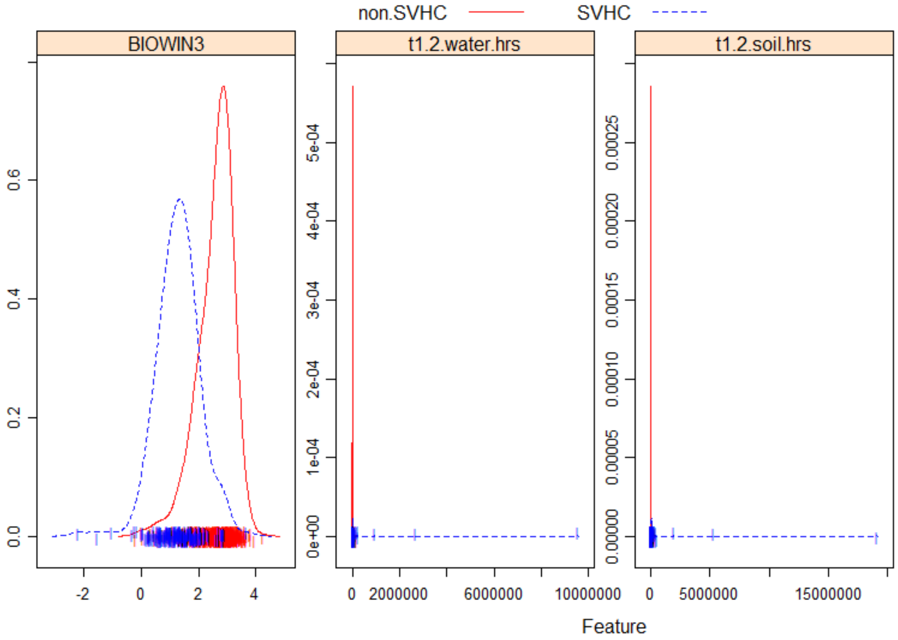


Figure 36: *Feature plot showing a subset distribution of features used for modelling purposes*

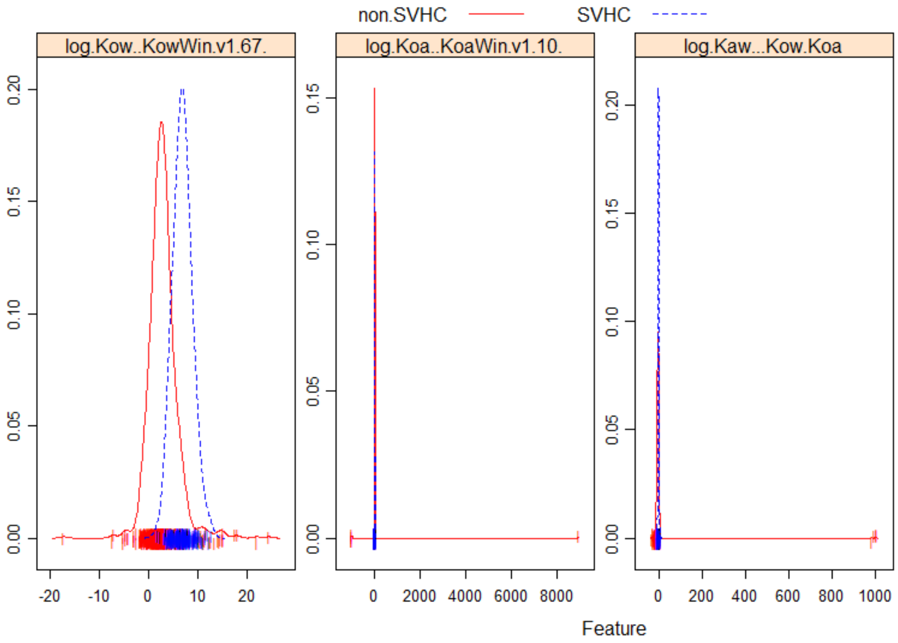


Figure 37: *Feature plot showing a subset distribution of features used for modelling purposes*

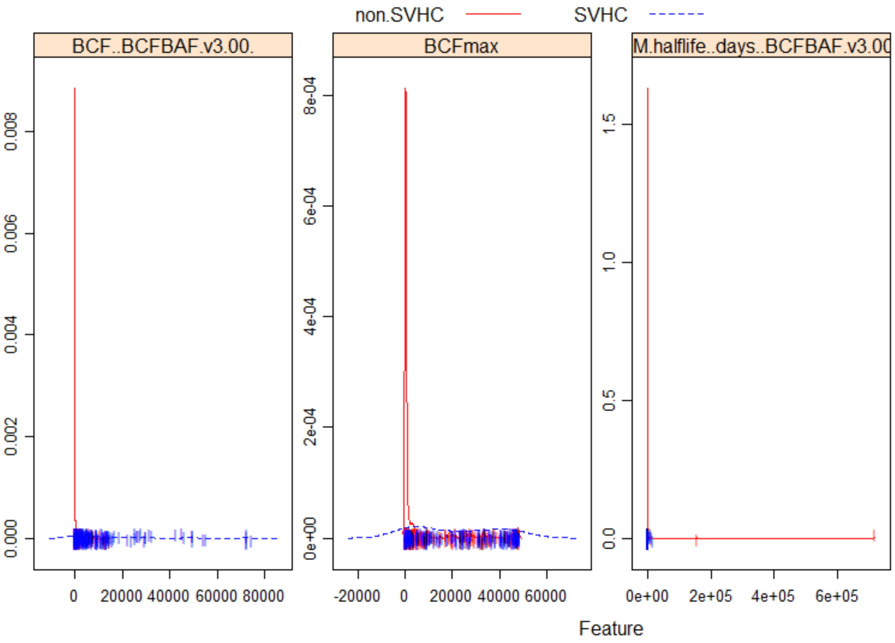


Figure 38: *Feature plot showing a subset distribution of features used for modelling purposes*

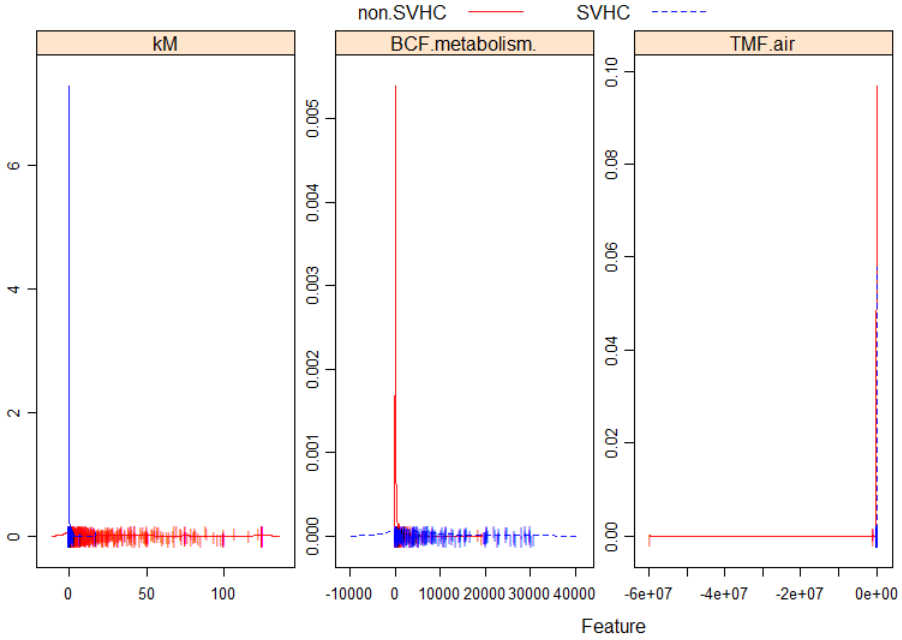


Figure 39: *Feature plot showing a subset distribution of features used for modelling purposes*

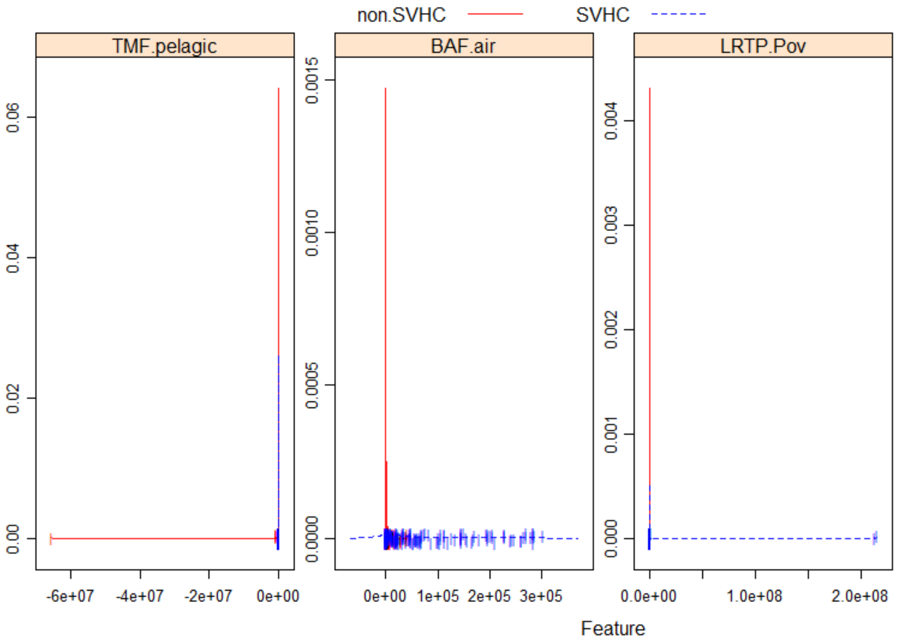


Figure 40: *Feature plot showing a subset distribution of features used for modelling purposes*

Appendix C Results

C.1 Binary Classification

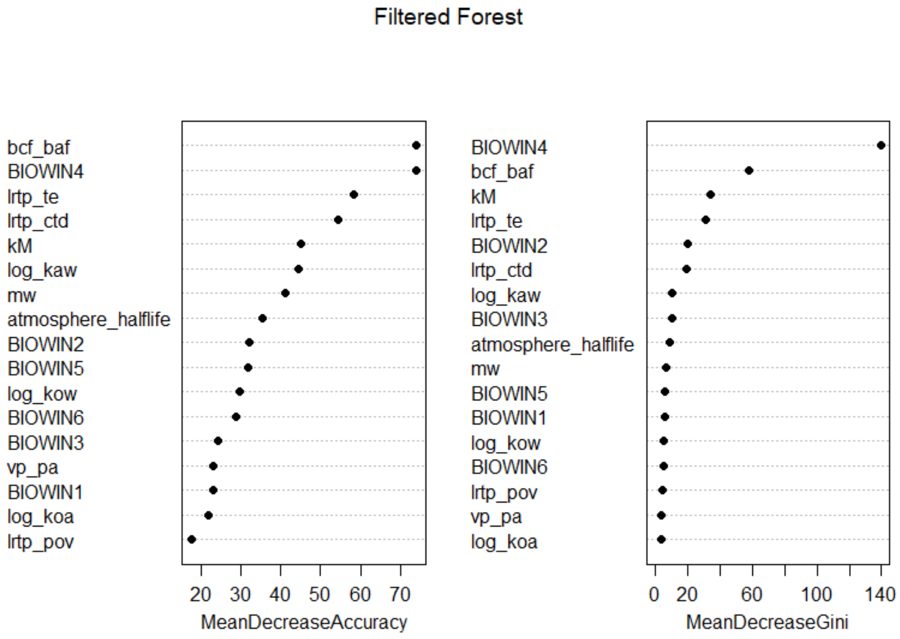


Figure 41: *Variable importance for the filtered forest model. Notable importance lies in the Biowin4 measure,the bioconcentration factor for bioaccumulation and POP specific long range transport potential. The ordering is related to the relative decrease of Gini impurity a split on this variable offers over the data and the overall mean accuracy decrease of variable exclusion.*

Important variables for the filtered random forest are similar to that of the naive one in figure 15. Here too an exclusion of variables like the bioconcentration factor and the Biowin4 model output would lead to an on average over 70 misclassifications. Further, the mean decrease in gini impurity for Biowin4 measures are significantly higher than other variables, which again points to the level of separability of the data based on this one variable alone. A sample decision tree can be seen in figure **??**.

On the other end of the scale, variables such as vapor pressure and LRTP for overall persistency score low in determining the label from the viewpoint of the model. This is further in tune with what experts deem to be important or non-important for chemical evaluation, however interestingly is it that the seeming importance of bioaccumulation metrics being higher than those of persistency.

This however is only a reflection of the data the models are trained on and as such should not be evaluated to be a judgement of overall importance between the two.

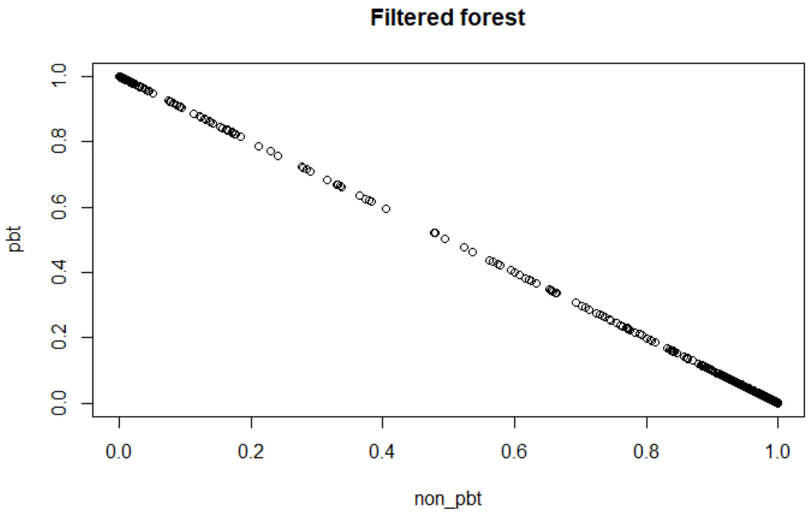


Figure 42: *Vote distribution for the filtered random forest model for both substances deemed to be of concern and non-concern.*

Vote distribution for the cross-validated grid searched filtered model from figure 10. Compared to the vote distribution for the naive approach in figure 14 the votes are much more condensed towards confidence intervals between 0.7-99%.

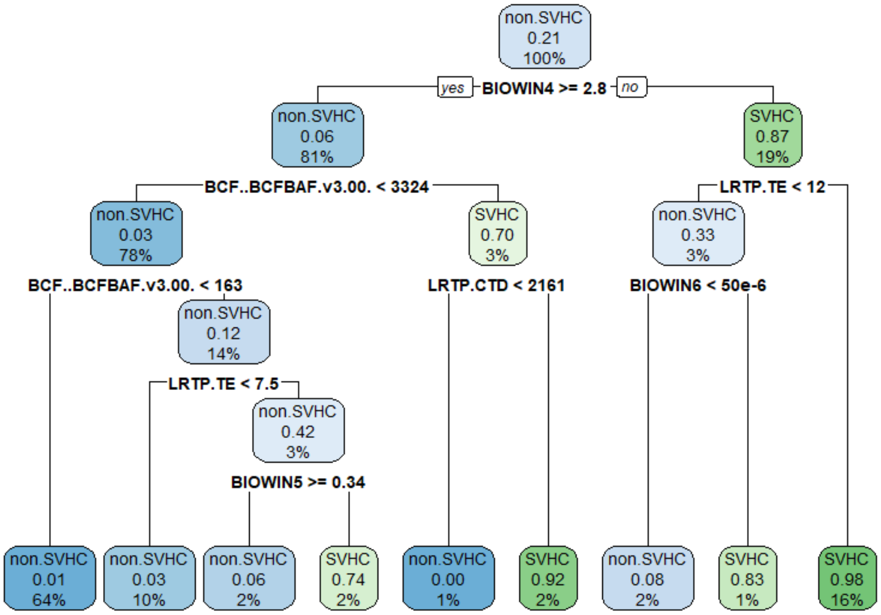


Figure 43: *A sample decision tree based on the filtered data. In accordance with its variable importance, the tree's initial split on Biowin4 at 2.8 separates a large majority of the data.*

Looking at a sample decision tree, one can observe in the root node the majority class, probability of being PBT and proportion of substances that fall in that node. Naturally, no splits have been made and so the root node contains all the data. After the first split, 81% of observations are sent to the left branch with the Biowin4 split at >= 2.8, where out of the 81%, one would make 0.06 error in predicting majority class of non SVHC already. Two further splits on BFC BAF<3324 and subsequently <163 where 64% of the data falls within these conditions with 99% probability of being a non-SVHC.

Conversely, 16% of observations are substances with a Biowin4 value higher than the initial split, and subsequently larger transport efficiency of 12 with probability of being SVHC at 98%.

C.1.1 SVM

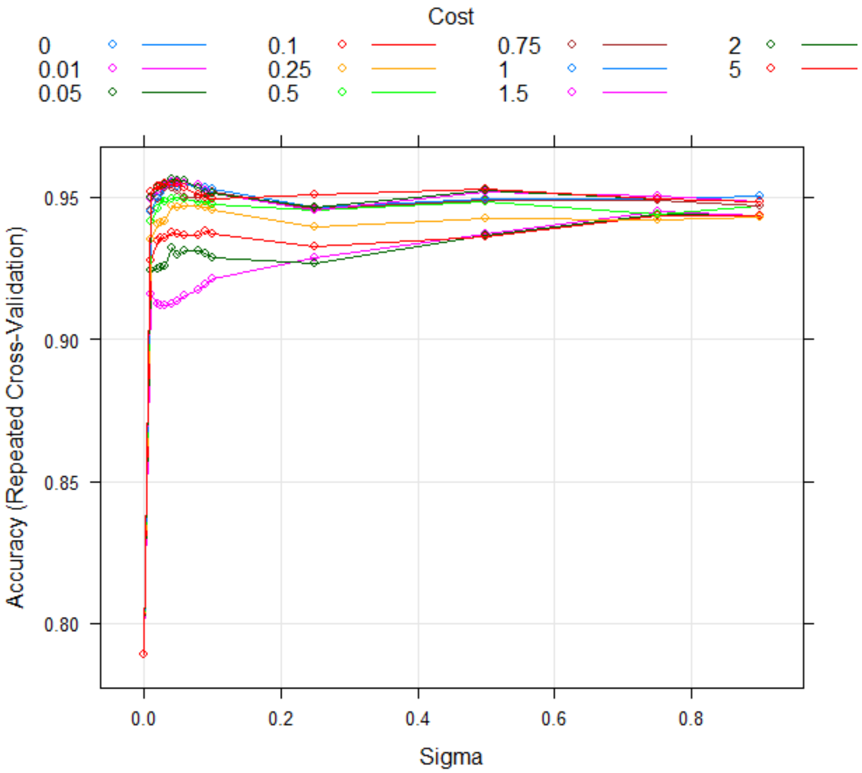


Figure 44: *Cross-validated grid searched naive SVM model, where optimal cost value is 2 and gamma value is 0.04*

The grid searched filtered SVM model has both less cost for misclassifications and lesser degree of gamma than its filtered counterpart. The model penalizes misclassifications to a lesser extent than the filtered model does, implying that the margin is softer for the filtered approach.

C.1.2 Reductionist model

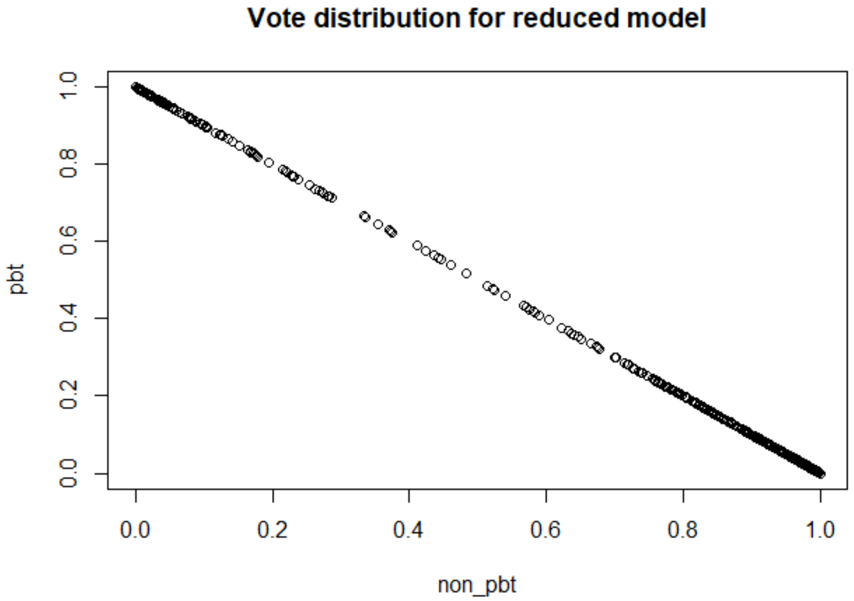


Figure 45: *Vote distribution for the reduced model, where the model's high confidence for classification at either end of the spectrum for non-PBT and PBT.*

Figure **??** shows the vote distribution for both classes, where the confidence of the model remains high albeit having removed variables that highly discriminates between labels.

Figure 46: *Cross-validated grid searched SVM model, where optimal cost is 2 and gamma value of 0.04*

C.2 Subclass classification

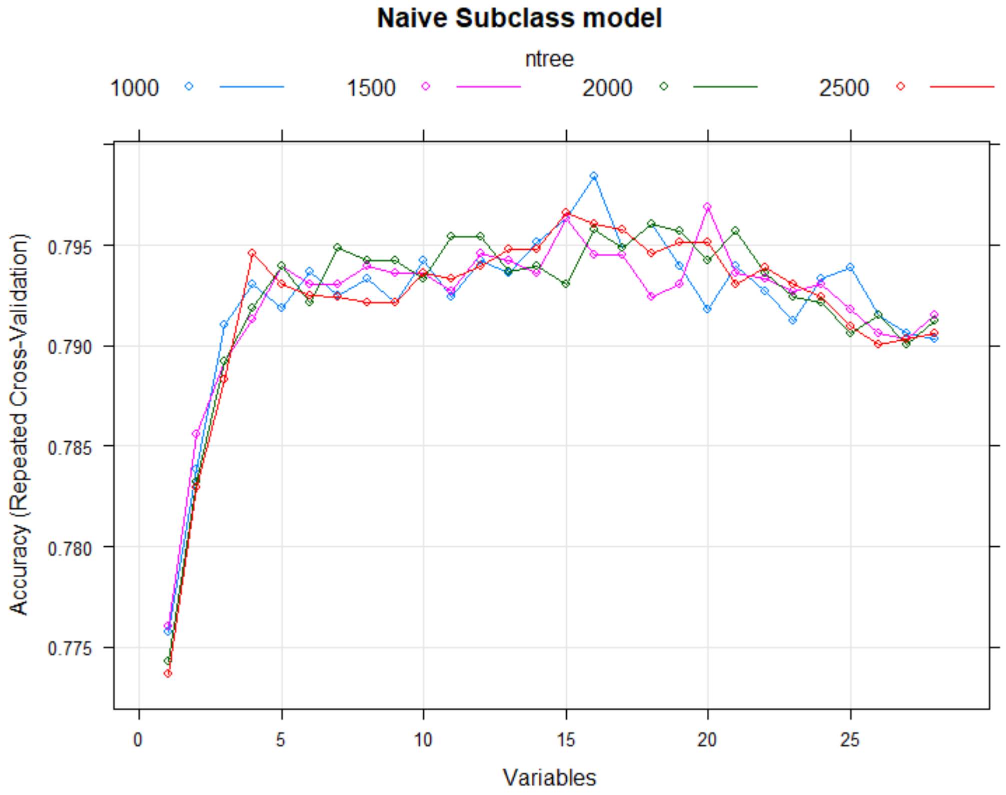


Figure 47: *Grid-searched naive model for subclass classification. Optimal parameters where chosen to be at 1000 trees and 16*

		Actual					
		Not-B	Not-P	Not-PBT	PBT	PBT/vPvB	vPvB
Predicted	Not-B	152	35	8	9	0	0
	Not-P	41	208	12	4	0	0
	Not-PBT	29	8	378	2	1	2
	PBT	5	5	3	62	7	4
	PBT/vPvB	1	0	00	7	15	14
	vPvB	0	0	0	13	14	80

	Not B	**Not P**	**Not PBT**	**PBT**	**PBT/vPvB**	**vPvB**
Sensitivity	0.66	0.81	0.94	0.63	0.39	0.80
Specificity	0.94	0.93	0.94	0.97	0.97	0.97

Table 13: *Sensitivity and Specificity measures for the naive subclass model.*

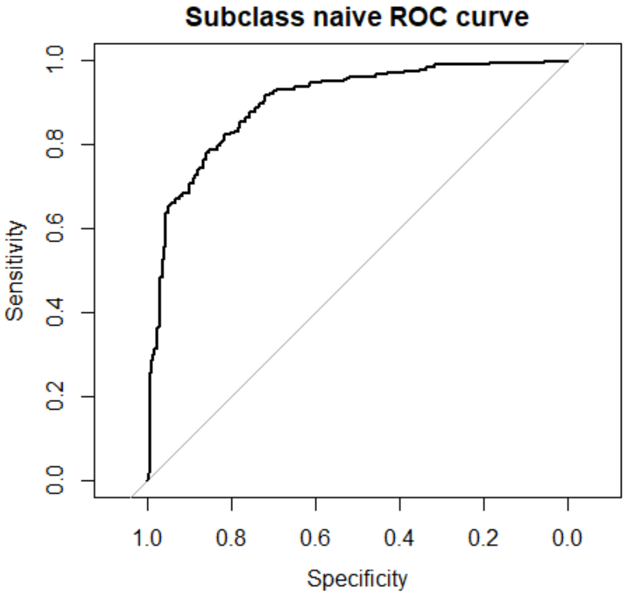


Figure 48: *ROC curve for the naive subclass model, where AUC was calculated to be at 0.80*

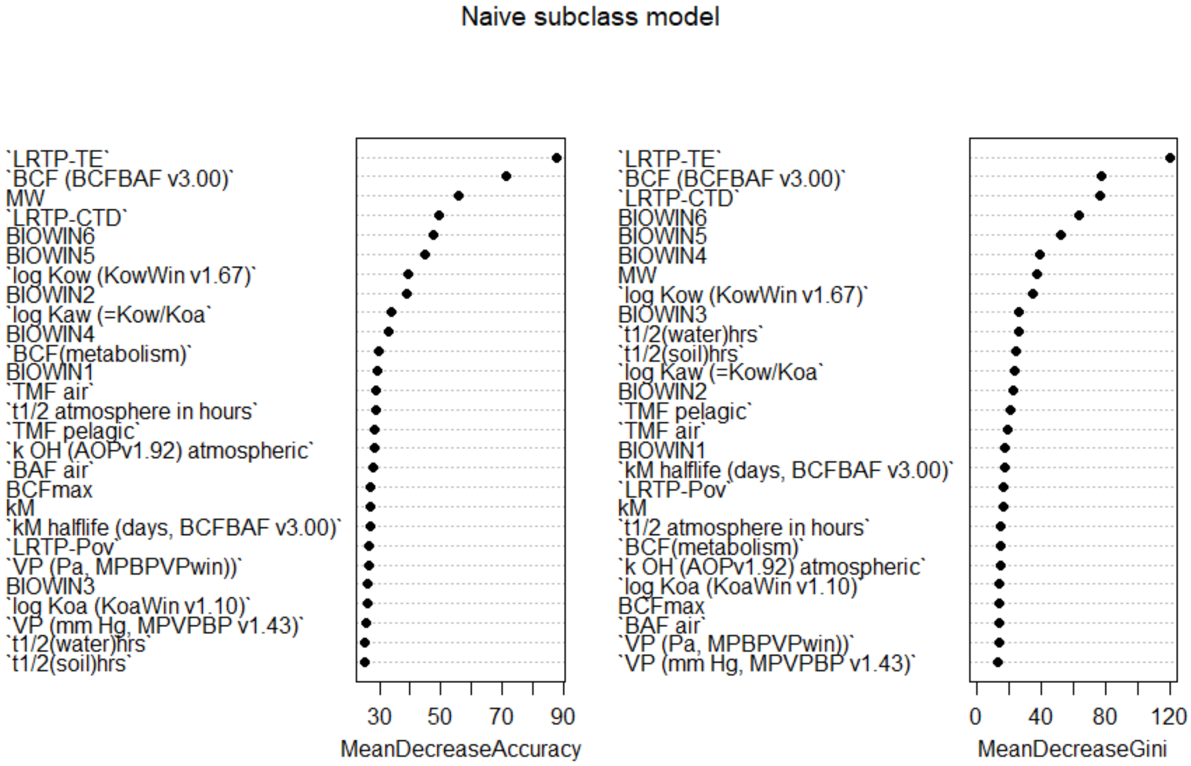


Figure 49: *Variable importance for the naive subclass model*

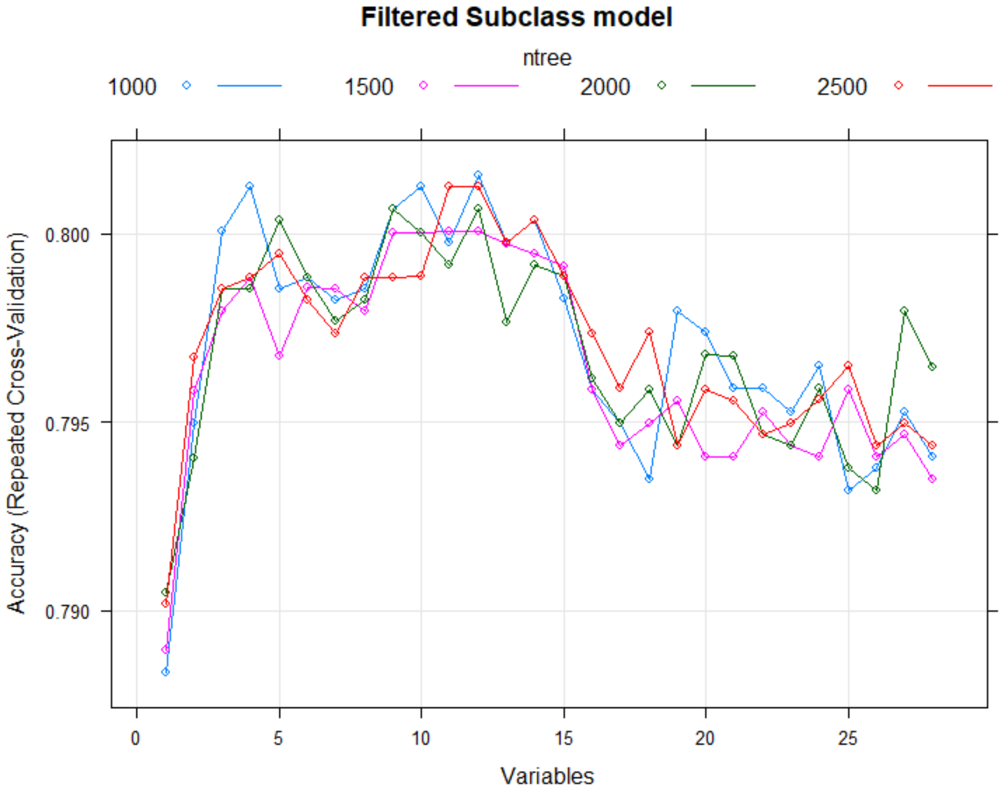


Figure 50: *Grid-searched filtered model for subclass classification. Optimal parameters where chosen to be at 1000 trees and 12*

C.2.1 SVM

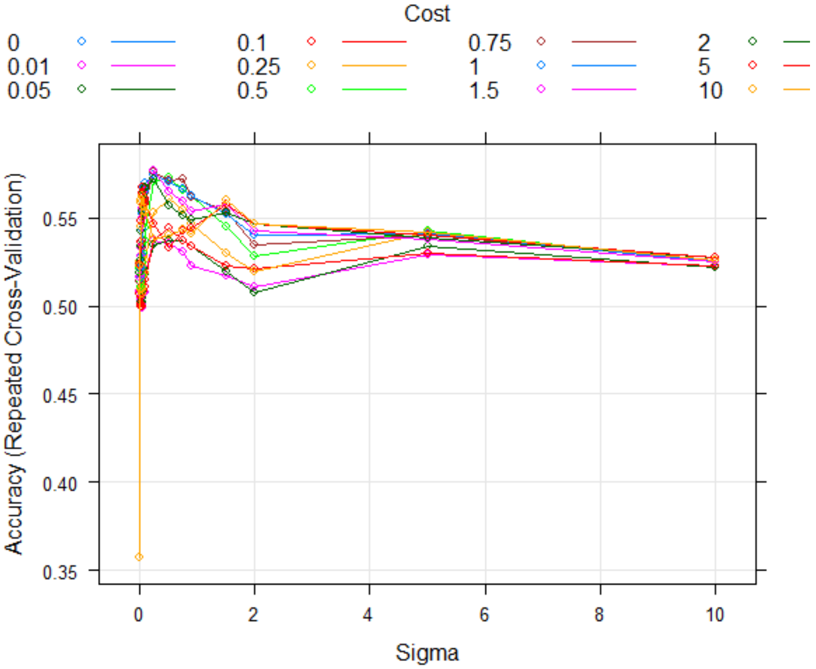


Figure 51: *Grid-searched filtered subclass SVM model. Optimal gamma found to be 0.25 and C value of 1.5*

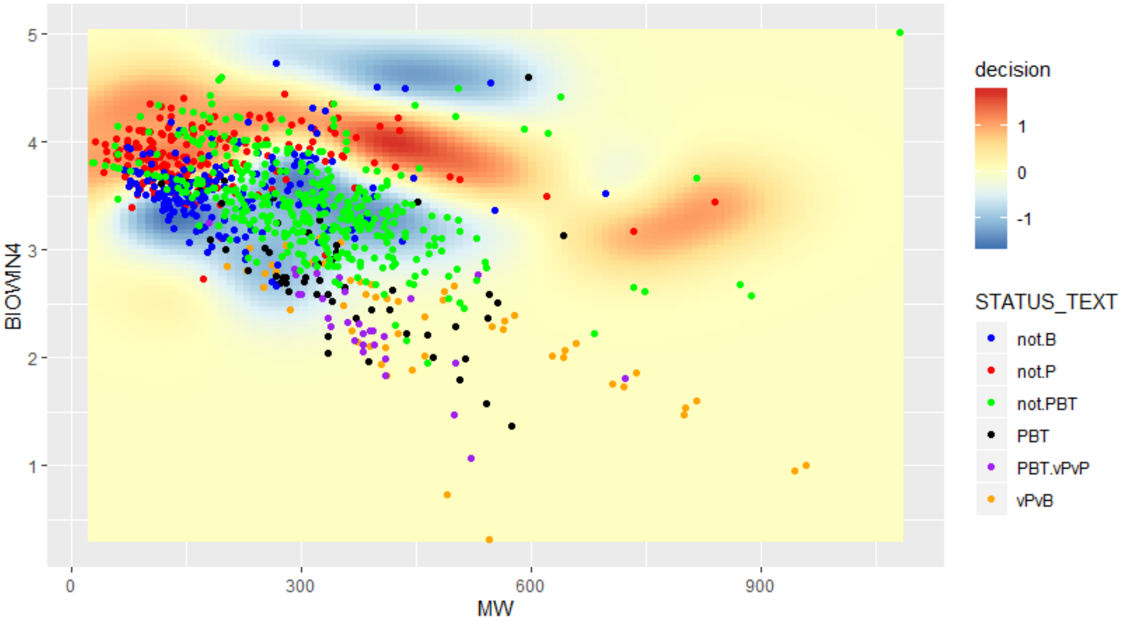


Figure 52: *SVM boundary plot for the subclass model for Biowin4 and molecular weight. Notable is the clustered overlap for subcategories of either binary class*

C.3 Subclasses P & B

Further exploration was done on subcategories of persistence and bioaccumulation separately. This was to explore the potential individual impact of either criteria, as a combination of the two as explained earlier is determining toxicologial categorization. A substance might be P but not B, and still be considered to be PBT. Target variables for persistence includes not P, P and vP, and conversely not B, B and vB for bioaccumulation. Data sets for either class are here smaller, as not every substance in the original set had an individual P or B evaluated label. Thus, some 488 substances were collected for known persistence labels in the set, and 468 bioaccumulation substances.

C.3.1 Subclass P

The following table is the confusion matrix for subclass P classification with a balanced accuracy of 89%.

		Actual		
		Not-P	P	vP
Predicted	Not-B	294	11	0
	P	9	95	23
	vP	0	18	88

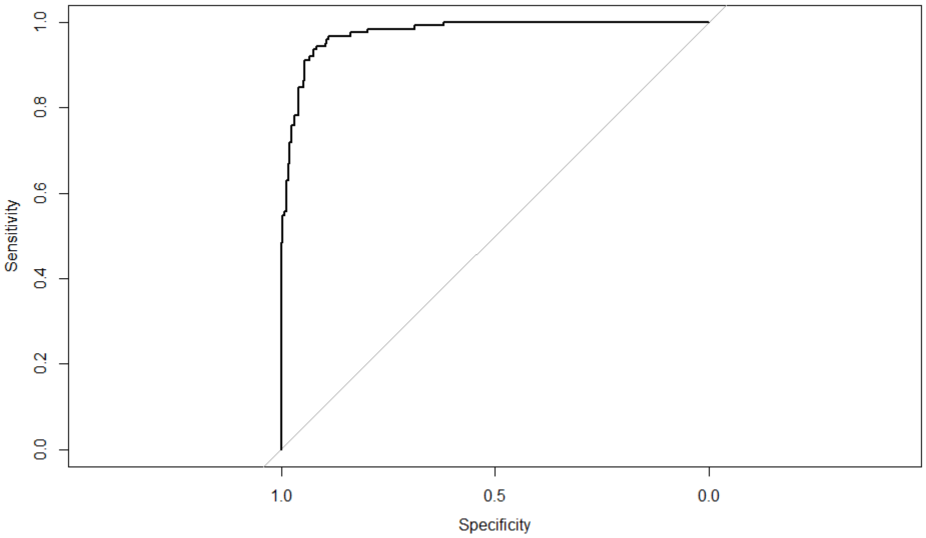


Figure 53: *ROC plot for sensitivity/specificity thresholds for persistence the model, where AUC calculated at 0.97*

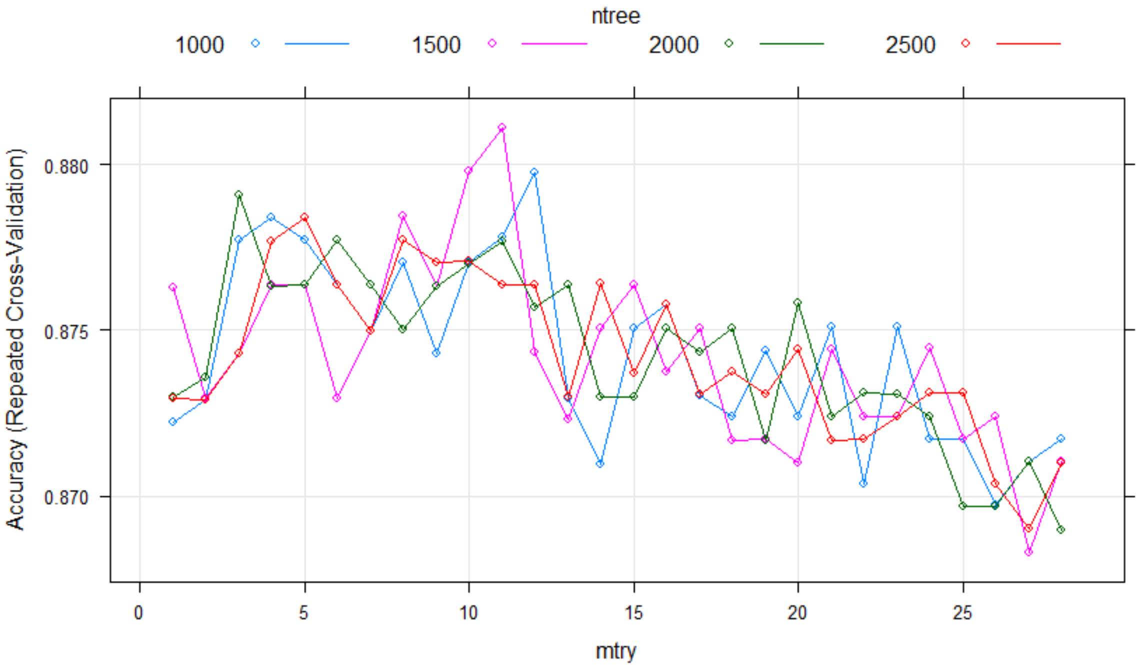


Figure 54: *Grid-searched random forest model for persistence, where optimal trees found to be 1500 and mtry at 11.*

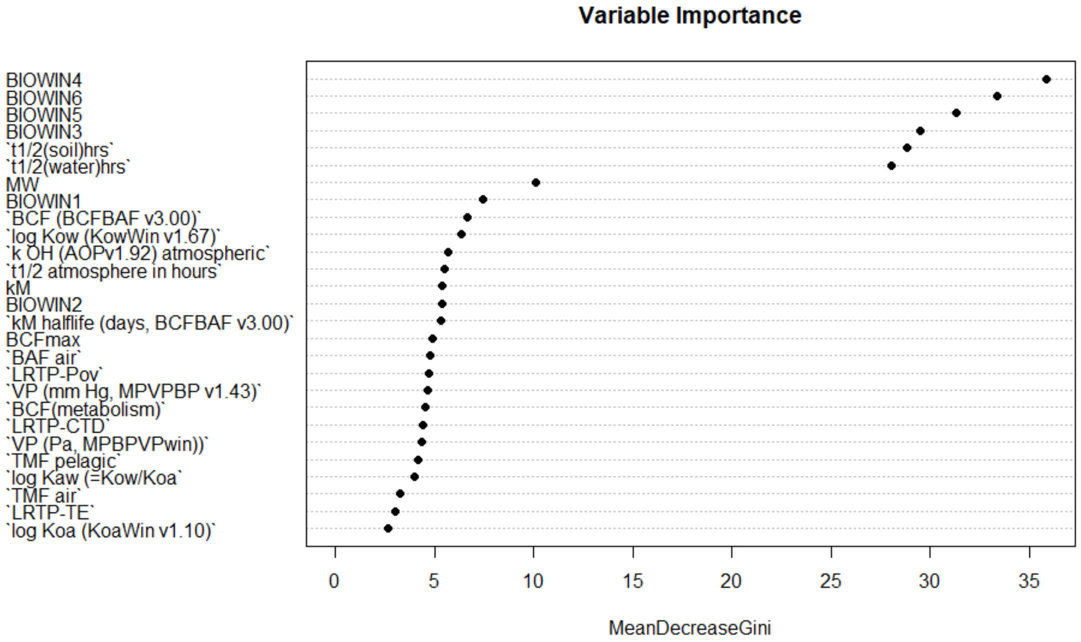


Figure 55: *Variable importance for the persistence model showing gini impurity decrease*

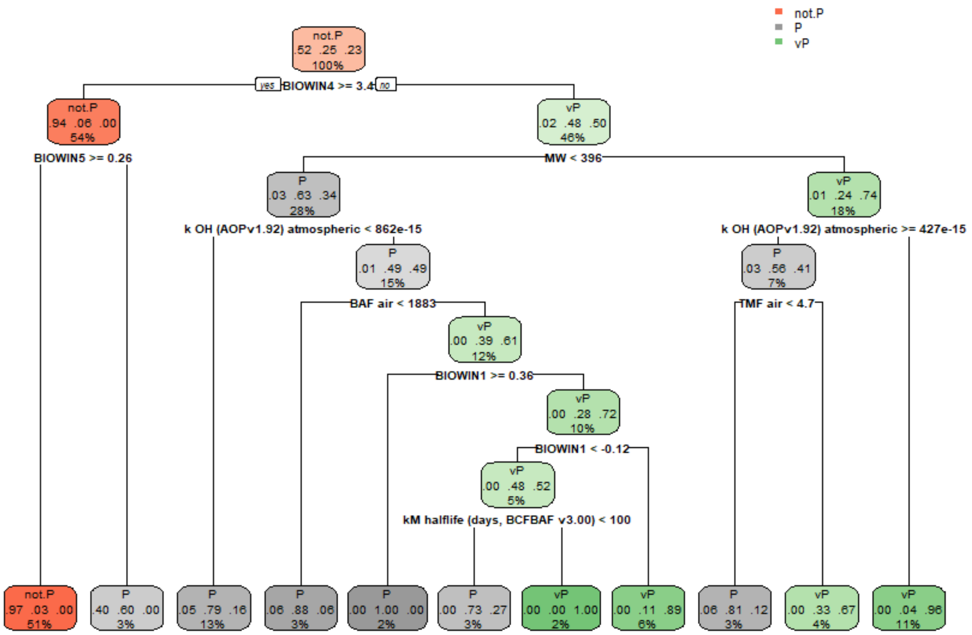


Figure 56: *Decision tree sample for persistence model fit.*

C.3.2 Subclass B

The following table is the confusion matrix for subclass B classification with a balanced accuracy of 89.2%.

		Actual		
		Not-B	B	vB
Predicted	Not-B	221	10	0
	B	17	94	21
	vB	0	21	88

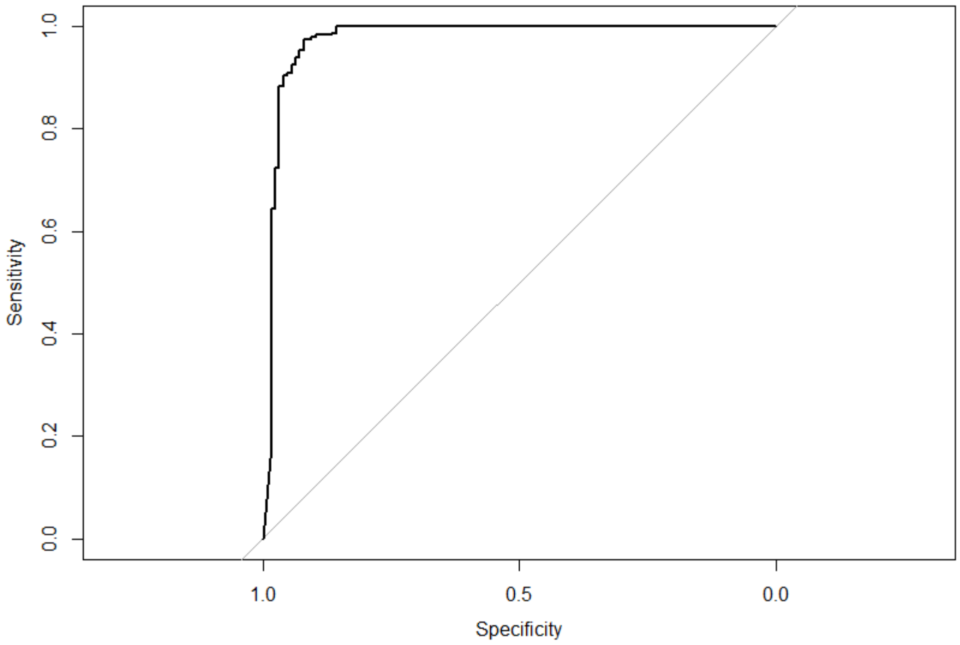


Figure 57: *ROC curve for subclass B, where we plot for different thresholds of sensitivity and specificity tradeoffs. AUC metrics calculated to be 0.97*

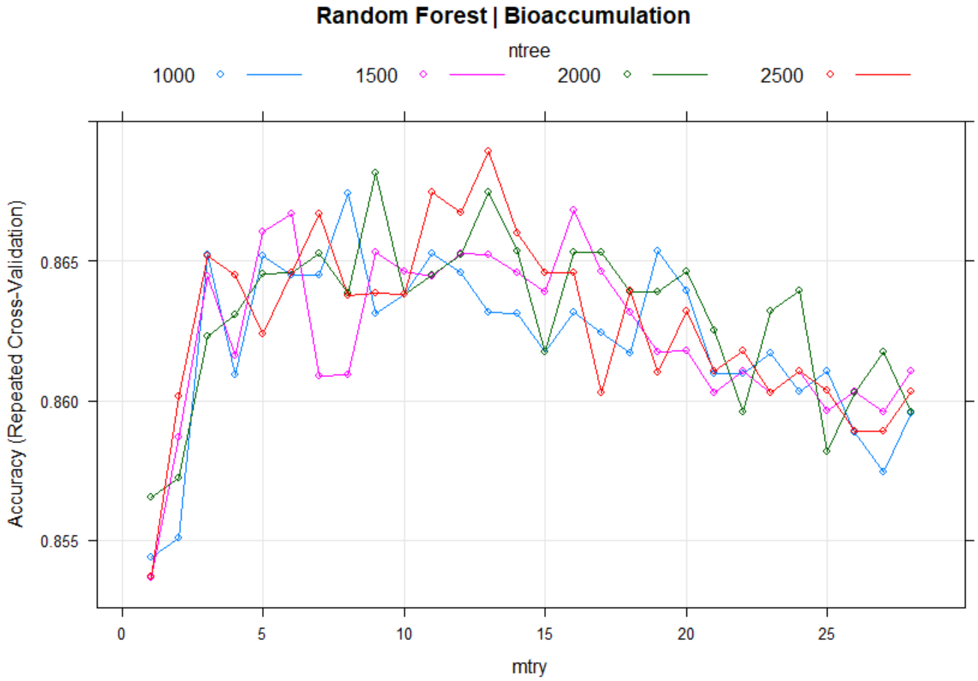


Figure 58: *Grid-searched random forest model for bioaccumulation classification. Optimal hyperparameters included 2500 trees and 13 mtry.*

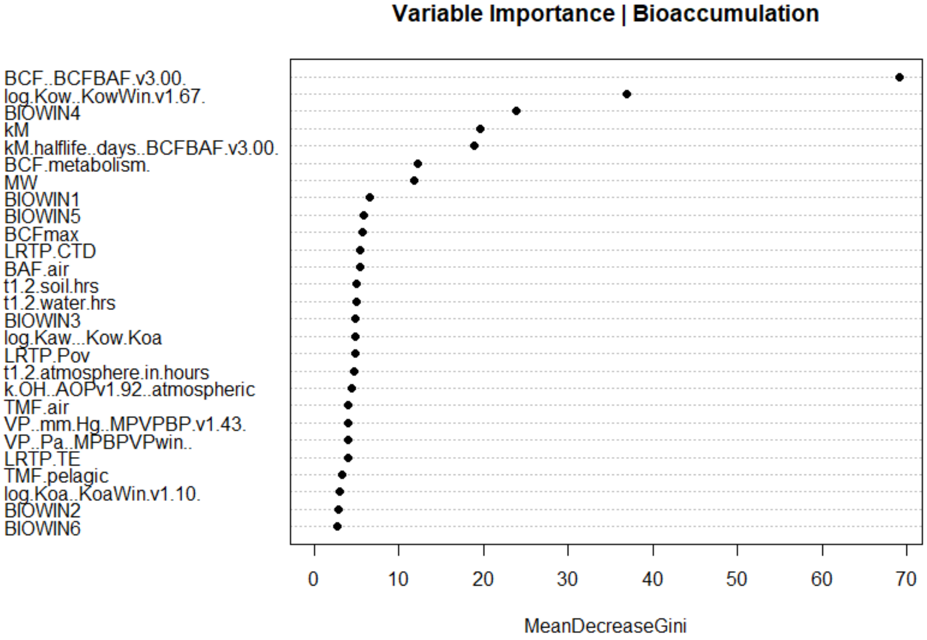


Figure 59: *Variable importance for bioaccumulation model.*

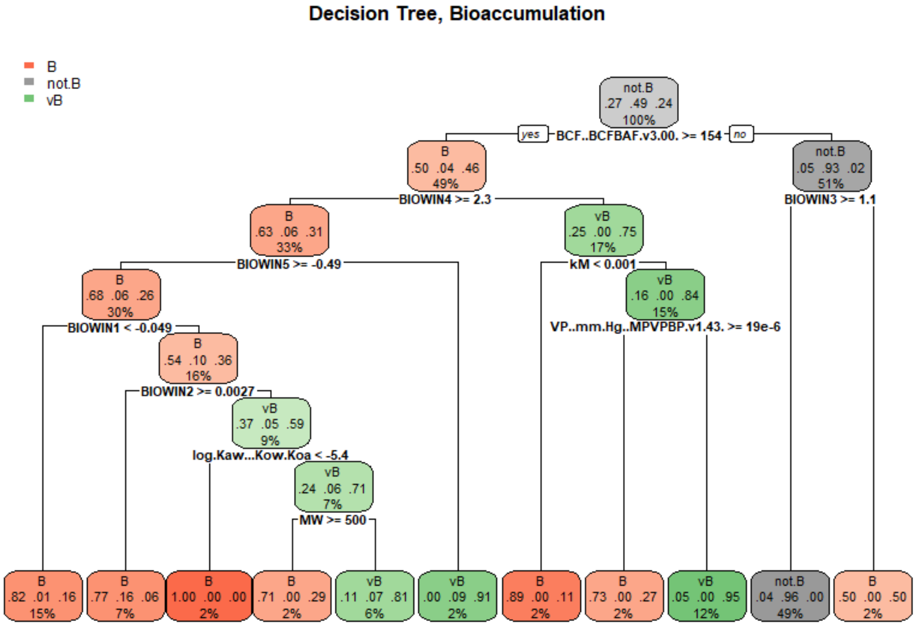


Figure 60: *Decision tree sample for bioaccumulation model fit.*

C.4 Misclassifications

Table 14 shows a full overview over all misclassifications made by the cross-validated binary models. As seen earlier in their individual confusion matrices, model difference can be marginal and as such a lot of misclassifications are duplicated. It is a point of interest nonetheless to inspect what the nature of these misclassifications can be and which ones are deemed to be particularly difficult to label. Colored cells denote;

- Classified wrong by one model
- Classified wrong by two or more models
- Classified wrong by all models

CAS number	Name	Reference	Prediction
000067-48-1	(15) Choline chloride	non-pbt	pbt
003846-71-7	(266) 2-benzotriazol-2-yl-4,6-di-tert-butylphenol	pbt	non-pbt
000143-50-0	(274) chlordecone	pbt	non-pbt
002385-85-5	(288) mirex	pbt	non-pbt
000118-74-1	(285) hexachlorobenzene	pbt	non-pbt
000087-68-3	(286) hexachlorobuta-1,3-diene	pbt	non-pbt
000608-93-5	(291) pentachlorobenzene	pbt	non-pbt
000129-00-0	(292) pyrene	pbt	non-pbt
vPvB-56 (350)		pbt	non-pbt
004904-61-4	1,5,9 cyclododecatriene	pbt	non-pbt
vPvB-59 (353)		pbt	non-pbt
025637-99-4	(368) 1,3,5,7,9,11-hexabromocyclododecane	pbt	non-pbt
000087-61-6	(366) 1,2,3-trichlorobenzene	pbt	non-pbt
000091-57-6	(390) 2-methylnaphthalene	pbt	non-pbt
000578-95-0	(397) 9(10H)acridone	pbt	non-pbt
000083-32-9	(398) acenaphthene	pbt	non-pbt
000208-96-8	(399) acenaphthylene	pbt	non-pbt
000260-94-6	(400) acridine	pbt	non-pbt
000120-12-7	(401) anthracene	pbt	non-pbt
000225-11-6	(402) benz[a]acridine	pbt	non-pbt
000225-51-4	(403)benz[c]acridine	pbt	non-pbt
023593-75-1	(404) clotrimazole	pbt	non-pbt
000294-62-2	(405) cyclododecane	pbt	non-pbt
051000-52-3	(413) vinyl neodecanoate	pbt	non-pbt
002104-64-5	(414) O-ethyl O-4-nitrophenyl phenylphosphonothioate	pbt	non-pbt
000229-87-8	(415) phenanthridine	pbt	non-pbt
070124-77-5	(416) flucythrinate	pbt	non-pbt
000086-73-7	(417) fluorene	pbt	non-pbt
000335-57-9	(420) hexadecafluoroheptane	pbt	non-pbt
000095-13-6	(421) indene	pbt	non-pbt
000119-65-3	(423) isoquinoline	pbt	non-pbt
000072-43-5	(424) methoxychlor	pbt	non-pbt
000793-24-8	(427) N-1,3-dimethylbutyl-N'-phenyl-p-phenylenediamine	pbt	non-pbt
000091-20-3	(428) naphthalene	pbt	non-pbt
000087-86-5	(432) pentachlorophenol	pbt	non-pbt
000382-21-8	(433) perfluoroisobutylene	pbt	non-pbt
124495-18-7	(435) quinoxyfen	pbt	non-pbt
000056-35-9	(438) bis(tributyltin) oxide	pbt	non-pbt
000603-35-0	(439) triphenylphosphine	pbt	non-pbt
001582-09-8	(440) trifluralin	pbt	non-pbt
000126-72-7	(441) tris(2,3-dibromopropyl) phosphate	pbt	non-pbt
PBT-11		pbt	non-pbt
013116-53-5	(595) PROPANE, 1,2,2,3-TETRACHLORO-	non-pbt	pbt
000091-20-3	(631) Naphthalene	non-pbt	pbt
000879-39-0	(662) 1,2,3,4-Tetrachloro-5-nitrobenzene	non-pbt	pbt
000126-72-7	(668) TRIS(2,3-DIBROMOPROPYL) PHOSPHATE	non-pbt	pbt
085535-84-8	(681) SCCPs	pbt	non-pbt
000085-01-8	(682) phenanthrene	pbt	non-pbt
000058-89-9	(683) gamma-HCH, Lindane, Hexachlorocyclohexaan (HCH)	pbt	non-pbt
000319-84-6	(684) alpha-HCH	pbt	non-pbt
000319-85-7	(685) beta-HCH	pbt	non-pbt
000319-86-8	(686) delta-HCH	pbt	non-pbt
000608-73-1	(687) techn. HCH	pbt	non-pbt
034482-99-0	(698)Fletazepam	pbt	non-pbt
001763-23-1	(712) 1-Octanesulfonic acid, 1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,8-heptadecafluoro-	pbt	non-pbt
000074-85-1	(846) Ethylene	non-pbt	pbt
153233-91-1	(847) Etoxazole	non-pbt	pbt
067306-00-7	(857) Fenpropidin	non-pbt	pbt
139968-49-3	(922) Metaflumizone	non-pbt	pbt
019666-30-9	(947) Oxadiazon	non-pbt	pbt
096489-71-3	(982) Pyridaben	non-pbt	pbt
179101-81-6	(983) Pyridalyl	non-pbt	pbt
118134-30-8	(1006) Spiroxamine	non-pbt	pbt
102851-06-9	(1010) tau-Fluvalinate	non-pbt	pbt
120068-37-3	(1048) fipronil	non-pbt	pbt
122453-73-0	(1049) 4-bromo-2-(4-chlorophenyl)-1-ethoxy methyl-5-trifluoromethylpyrrole-3-carbonitrile (Chlorfenapyr)	non-pbt	pbt
122454-29-9	(1050) Tralopyril	non-pbt	pbt
064359-81-5	(1089) 4,5-Dichloro-2-octylisothiazol-3(2H)-one (4,5-Dichloro-2-octyl-2H-isothiazol-3-one (DCOIT))	non-pbt	pbt
080844-07-1	(1109) etofenprox	non-pbt	pbt
082657-04-3	(1110) Bifenthrin	non-pbt	pbt

Table 14: *Overview of all misclassifications over all four models. Non-colored rows are chemicals misclassified a single time, orange rows for 2 or more classifiers and yellow denotes misclassification across all models.*