UPPSALA
UNIVERSITET

UPTEC K 19021
Examensarbete 30 hp
Juni 2019

Reaction Conditions Data Mining

The application of Machine Learning towards predicting the future of process development

Samuel Hallinder

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Reaction Conditions Data Mining

Samuel Hallinder

In organic chemistry and especially process chemistry, there is a constant need to develop cost-effective ways to optimize different reaction conditions. With the increased development of Machine Learning (ML) combined with Data Mining (DM) new possibilities arise to reduce time and costs in the field of chemical science. In order to address the need for reduced time-/cost savings in process chemistry, the often-employed Suzuki-Miyaura reaction was studied by such ML and DM methods. A representative dataset containing molecular and structural properties of substrates and product were calculated with open-source toolkits Indigo, Chemistry Development Kit and RDKit available in KNIME. To predict any form of reaction outcomes, catalysts and reaction conditions were ranked based on several binary classification Machine Learning models designed with a Random Forest algorithm. On model lead to a binary classification model performing at a low computational cost. It showed an AUC of 98.5% predicting a reaction to a certain threshold of yield (>=60% and <=40%). A second model encompassed six unique binary classification models and presented an average accuracy of 91.6% to predict a correct catalyst. These six different models were combined to later rank catalysts that are best suited for a new reaction and gave a probability result between 23.6% to 77.3%. The experimental validation was proven to highlight the uncertainty of the performance, were the least suitable (23.6%) catalyst demonstrated best performance. Overall, the models showed a promising correlation to support the synthesis optimization problem and with further adjustment there are great opportunities to obtain a model that can assist chemists in the future.

Handledare: Fernando Huerta
Ämnesgranskare: Peter Brandt
Examinator: Christian Sköld
ISSN: 1650-8297, UPTEC K 19021

TABLE OF CONTENT

1 Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Networks
AUC	Area Under the Curve
CART	Classification And Regression Tree
CDK	Chemistry Development Kit
DM	Data Mining
DR	Dimensional Reduction
ER	Error Rate
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
HPLC	High-Pressure Liquide Chromatography
IPC	In Process Control
HTS	High Throughput Screening
KNIME	Konstanz Information Miner
KNN	K-Nearest Neighbours
LR	Logistic Regression
ML	Machine Learning
MQN	Molecular Quantum Numbers
Mtry	Number of features for deciding best splits
NB	Naïve Bayes
OOB	Out Of Bag
PMML	Predictive Modell Mark-up Language
RF	Random Forest
ROC	Receiver Operating Characteristics
RP	Reversed Phase
SL	Statistical Learning
SVM	Support Vector Machines
PCA	Principal Component Analysis
TN	True Negatives
TP	True Positives
TPR	True Positive Rate
TPSA	Topological Polar Surface Area
VSA	Van der Waals Surface Area
$Pd(dppf)Cl_2$	[1,1´-bis(diphenylphosphino)ferrocene]dichloropalladium(II)
$Pd(dtbpf)Cl_2$	[1,1´-bis(di-tert-butylphosphino)ferrocene]dichloropalladium(II)
$Pd(PPh_3)_2Cl_2$	Bis(triphenylphosphine)palladium(II)dichloride
$Pd(PPh_3)_4$	Tetrakis(triphenylphosphine)palladium(0)
$Pd(t\text{-}Bu3P)_2$	Bis(tri-tert-butylphosphine)palladium(0)
XPhosPdG2	Chloro(2-dicyclohexylphosphino-2′,4′,6′-triisopropyl-1,1′-biphenyl)[2-(2′-amino-1,1′-biphenyl)]palladium(II)

2 Populärvetenskaplig sammanfattning

Inom organisk kemi och processkemi ser man ett ständigt behov av att utveckla, tidseffektivisera och på kostnadseffektiva sätt optimera olika processer. Försökplaneringsmetoder, så som faktoriell design används i en bred utsträckning för att optimera system där matematiska funktioner utnyttjas för att finna ett linjärt samband i utfallet av en reaktion. Detta har demonstrerats vara en tidskrävande process även för de mest skickliga inom området, där flertalet testförsök erfordras för att finna de mest gynnsamma reaktionsparametrarna som krävs för att uppnå ett så högt utbyte av önskad produkt som möjligt.

Suzuki-Miyaura-korskopplingsreaktionen är en av de mest använda reaktionerna i att bilda nya kol-kolbindningar inom medicinsk kemi, men har visats vara tidskrävande att optimera. Dessa reaktioner sker inte spontant i rumstemperatur utan måste utföras med tillförsel av värme tillsammans med närvaro av en metallkatalysator och bas i ett lösningsmedelssystem. Behovet här är att finna en kombination av faktorer så som katalysator, bas samt lösningsmedel som ska vara mest fördelaktig för att erhålla så högt utbyte som möjligt för en specifik reaktion. För att undvika stora kostnader på flertalet optimeringsprocesser har man nu sett en stor utveckling och förhoppning till att kombinera kunskapen inom maskininlärning och datautvinning med den kemivetenskapliga sfären. Med olika algoritmer har man sett en möjlighet att identifiera mönster inom en stor representation av data som tidigare inte varit möjligt. Maskininlärning har de senaste årtionden blivit applicerad i en allt större utsträckning inom läkemedelskemi, bioinformatik, etc. Syftet har då varit att på ett effektivare sätt kunna identifiera möjliga substanser, föreningar eller större komponenter som är av relevans till utvecklingen av nya läkemedelskandidater. Intresset har med detta även väckts inom kemisk syntes och processkemi.

Målet för denna studie har varit med hjälp av maskininlärning skapa ett verktyg för att på ett så kostnadseffektivt och tidsparande sätt optimera förhållandena för en reaktion. Den tidigare nämnda Suzuki-Miyaura-reaktionen valdes i denna studie för att studera det definierade problemet. Molekylära parametrar och strukturella molekylelement beräknades för att modellera utfallen från tidigare dokumenterade reaktioner. En maskininlärningsalgoritm implementerades, som läste in sig på befintliga data för att hitta intressanta mönster som skulle vara av relevans för att modellera det experimentella utfallet. I detta projekt valdes en Random Forest-algoritm, som i tidigare studier visats ha en hög säkerhet i olika binära klassificeringsfrågor. Första klassificeringsfrågan som konstruerades i detta projekt var om man kan uppfatta någon korrelation med de beräknade molekylära och strukturella egenskaperna för de olika komponenterna och utbytet för reaktionen. Om så var fallet, skulle det finnas en möjlighet att då förutsäga vilken katalysator som skulle kunna vara mest pålitlig att presentera ett högt utbyte för den specifika reaktionen.

Flertalet binära klassificeringsmodeller konstruerades och den mest framgångsrika modellen i att identifiera om en reaktion kommer ge ett högt eller lågt produktutbyte resulterades i att ge en säkerhet på 93.5%. Vidare undersöktes om det var möjligt att man på något likande sätt skulle kunna förutsäga vilken katalysator som skulle vara mest gynnsam för en reaktion i att få ett högt produktutbyte. Här valdes enbart sex stycken katalysatorer som alla använts i en mängd olika reaktioner; från 1 000 till 25 000. Unika binära klassificeringsmodeller konstruerades likt den framgångsrika modellen, och presenterade med ca 92% säkerhet rätt katalysator för de dokumenterade reaktionerna i test datasetet.

De två modellerna kombinerades för att senare rangordna de katalysatorer som är mest till minst lämpad för en ny reaktion, resultaten spred sig mellan 23.6% upp till 77.3% för de sex katalysatorerna. Dessa resultat validerades senare genom en experimentell analys där olika katalysatorer testades i laboratoriet och jämfördes med resultaten från modellen. De minst lämpade katalysatorerna visade sig prestera bättre medan den högst trovärdig utifrån modellens förutsägelse presenterade ett mindre önskvärt resultat. Detta motbevisade modellens säkerhet att ge ett representativt förslag av katalysator till en ny reaktion. Osäkerheten som uppstår beror generellt på de komplikationer som tillkommer i att förutse utfallet av en reaktion. Den beskrivning som erhålls från tidigare dokumenterade resultat tillsammans med de deskriptorer som modellen använder sig av har visats vara en osäkerhetsfaktor. Eftersom ingen direkt information ges kring de centrerade bindningar som bildas eller bryts i reaktionen så medför det en problematik i modellens förmåga att förutsäga reaktiviteten. Denna studie har dock visat att en korrelation existerar, vilket ger en god grund för framtida forskning. Metoden i sig är inte fel, med optimering och justering finns det stora möjligheter att erhålla en procedur som kan assistera kemister i att optimera reaktioner på ett tidseffektivt och ekonomiskt sett i framtiden.

3 Introduction

3.1 Background

A common challenge organic chemist encounters daily is the optimisation process of reaction conditions, which is one of the most time-consuming processes even for a person skilled in the art of synthetical organic chemistry and can take up to several weeks or even months to execute in the most efficient way. Different techniques have been used to confront these types of problems and factorial designs, D-optimal designs as well as High Throughput Screening (HTS) [1] [2] [3], are a few methods that have been successfully used to find the most preferable conditions at the minimal cost for a specific reaction. The techniques of HTS and combinatorial synthesis have contributed to gather and store large information of molecules into databases, and have made up for the evolution of the big data era in organic chemistry [4]. The field of drug discovery and process design have gained a lot from this to detect new target molecules as well as optimize several chemical processes.[5] [6].

The knowledge of Statistical Learning (SL), Machine Learning (ML) and Artificial Intelligence (AI) have influenced chemists to gather information and discover patterns in various databases. A broad range of different machine learning algorithms have later been used to either obtain vital information from databases or to forecast future estimations derived from available data. Logistic Regression (LR) [4] [7] [8], k-Nearest Neighbours (kNN) [6] [8] [9], Support Vector Machines (SVM) [4] [6] [9] [10] [11], Artificial Neural Networks (ANN) [4] [6] [12] [13] [14], Naïve Bayes (NB) [11] [12] and Random Forest (RF) [4] [6] [8] [11] [15] are a few widely used algorithms in this field of research. The ability to influence and help to develop future new drug candidates, or to assist organic chemists to make reactions more effective, shorten time to optimized conditions, or even predict the outcome for specific reactions, is why these algorithms are one of the most acknowledged techniques today.

Some studies dedicated to combine ML techniques with synthetic chemistry predictions have been disclosed in the past decade. For example, Coley CW et al. [13] made an attempt to design a software tool to assist chemists in predicting reaction outcomes, by using a neural network model to score and rank candidate products with high prediction accuracy. The final estimate of selecting the major product with only one rank was obtained at a 71.8% accuracy. With three number of candidate ranks, the accuracy of desired product was within the top three gave 86.7% and with five, 90.8%. The incorrect documented predictions were notably related to the lack of information from descriptors on chemical reactivity.

In another study performed by Skoraczyñski G et al.[4] different machine learning algorithms, such as LR, SVM, ANN and different types of RF were constructed. The models were trained with a large set of different reactions and descriptors, predicting a binary classification of yields respectively classification of reaction times. The main objective was to analyze if currently available descriptors are enough for forecasting reaction outcomes. Usage of molecular descriptors from RDKit (open source toolkit for calulating molecular descriptors) together with parameters from experiments, solvents and temperature, made it possible to obtain a prediction accuracy of 65% for yield respectively 75% for reaction times.

Aires de Sousa J, et al.[11] published a study with a sligtly different approach, were the objective was to design a system to predict reaction conditions for Michael reactions. 198 different Michael reactions with a number of different set of descriptors were used to train a couple of ML algorithms (SVM, NB and RF) to predict best combinations of conditions. The model that predicted the best for each solvent and catalyst tested was the RF model, after a 3-fold cross validation. The accuracy for different solvents and catalysts varied between 70 and 100% [11]. A similar approach was defined for a toxicity prediction study [16].

Using large sets of reactions and/or a large number of descriptors has been proven to be an immense burden on computational resources. Robin Gebuer et al. [15] published a study where the investigation was performed analyzing the major challenges, using big data in machine learning. The RF algorithm was introduced as an example, were different parameters were presented to be tuned for optimizing the model design to minimize the expense of calculations.

In summary, the Random Forest algorithm has been a good approach in constructing different model designs for receiving fast and solid predictive performances in machine learning modelling. Therefore, it was chosen as primary algorithm for this study.

3.2 Aim of thesis

Due to a high interest and desire to identify synthesis procedures that optimize time-consuming steps, the aim of this project was to *"obtain a machine learning tool to rank the best reaction conditions for a specific reaction using only molecular properties"*. The reaction of choice was the Suzuki coupling since this is one of the most used transformations in medicinal chemistry. All codes & algorithms are publicly available in the open-source software KNIME [17] which was used to apply the widely used Random Forest algorithm for model creation.

4 Theory

In this section the underlying theory of Machine learning and its application will be described.

4.1 Machine Learning

Machine learning is a field in computer science that focuses on the development of different methods to automatically detect patterns in various datasets and use the uncovered patterns to create predictions for future data. To obtain this information, numerous distinctive types of algorithms and methods are readily available to apply for different tasks. ML is based on two strategies of learning, supervised and unsupervised. Different algorithms are constructed under each learning approach. [18] [19]

4.2 Unsupervised Learning

Unsupervised Learning is defined as the descriptive approach. It has been proved to find patterns within an input X_i of the p-dimensionally vector space and present useful information of that set [20]. A dataset is defined as, $p = \{(X_i)\}_{i=1,\ldots,n}^{T}$ where it is possible to find implicit correlations and variance within vectors in the p-dimensional space. Clustering and Principal Component Analysis (PCA)[6] are a few known unsupervised methods that are widely used to acquire information of the features of a data set. The principles with both methods are to find similarities within parameters and place them into separate groups, or, place different thresholds to separate them from each other. This is very useful in large data sets problems[18], where the interests is to find information on poorly presented descriptors.

4.3 Supervised Learning

Supervised learning is described as a predictive approach where a computer can learn what has been presented by different algorithms, thereafter to predict future events based on that information [20]. It is defined as $D = \{(X_i, Y_i)\}_{i=1,\ldots,n}^{B}$, were D is the set for training and B is the number of training examples. X_i Is defined as a vector of distinctive features or numbers in the p-dimensional space, $X_i \in \mathbb{R}^p$. Features/properties/descriptors are a representation of a value e.g. the topological surface area (TPSA) which is a feature/descriptor that presents information of the molecular shape of a molecule. In general, can it be anything that will influence the categorical or nominal feature which is defined as the output variable [18], Y , $Y_i \in \{1, \ldots, C\}$. For classification tasks Y_i is of a categorical value. For regression tasks Y_i is defined as a real value, $Y_i \in \mathbb{R}^p$.

4.3.1 Classification

In this project, a classification task was of interest where a categorical value was used to describe the problem. Defined as $Y_i \in \{1, \ldots, C\}$, were C presents number of possible outcomes to a specific task of interest. The aim in classification tasks is to define the problem as a function [18], when $Y_i = f(X_i)$. The function (f), describes the problem with a representation of a fixed table of data points (X_i), named training set. This function is thereafter applied to another set of input data to predict classification outcomes. An ML algorithm will learn to find the patterns in the training data set and give an estimation of the performance to predict the desired y_i. The performance is later measured by the generalisation error, which gives an estimation of the performance. A prediction variable is made, $\hat{Y}_i = g(\hat{X}_i)$, that approximates the function of the training set to give a prediction of the output. An independent test set is later defined as another function, to validate the model's predictability. Accuracy measures is calculated by dividing all correct predictions to all predictions from the results of the test set. Which is a decent way to get an understanding of the model performance.

4.3.2 Random Forest, CART and OOB

There can be a few or multiple algorithms that are of best fit for solving different tasks for each problem. In this project, the supervised learning model, Random Forest algorithm, was implemented to first predict two yield classes and thereafter different catalysts classes. The RF is well known for its predictive accuracy, flexibility and can be used for both classification and regression tasks. It was firstly mentioned by Leo Breiman and defined as an ML algorithm that can be found under the branch of ensemble learning methods [21]. The theory of RF is closely correlated with the Classification And Regression Tree (CART) and the theory of bagging.

CART algorithms often refer as if a tree is drawn upside down. A statement is created in form of a question. For example: "*Do the molecule follow the Lipinski rule of five?*"

When using this algorithm to solve this classification task. Different questions are formed, in finding an acceptable representation what corresponds to the Lipinski rule of five. The CART uses a recursive binary splitting method, which means that all data are within the field to predict the outcome. Each feature is divided into distinct branches to acquire an information on the outcome and is referred as a candidate split. All candidate splits are calculated according to accuracy of choosing the right class, and this follows until no further splits is possible within a feature. The final decision from all final splits is combined and an average of all results is obtained on how accurate it can predict the molecules appearances to follow the rule.

The CART algorithm is simple to understand and interpret [18]. It works with both numerical and categorical features and can define multiclass tasks [20]. In addition, it also has the possibility to handle noisy parameters remarkably well, hence no need of a heavily curation performance. However, the main drawback with a CART algorithm in general is that it can easily create trees that overfit the data. The most apparent problem is the algorithm´s sensitivity with unstable data, which is observed when minor changes appear. The creation of that tree will be affected, and the final estimation would provide a negative impact to the final prediction. In most cases a high variance is shown within the final estimation and a low prediction is presented. One solution to this problem is to lower the variance for estimators by implement bagging. The short definition of bagging is to create decorrelated [21], noisy decision trees by random selection of descriptors and then average them to gain higher accuracy.

In binary classifications tasks and ensemble trees algorithms is the Gini Index often selected as the split criterion defined as $\mathrm{Gini}\ (\mathrm{T}) = 1 - \sum_{\mathrm{i}=1}^{\mathrm{k}}(\mathrm{p}(\mathrm{c_i}))^2 - \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}(\mathrm{t_i}) \sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{p}\left(\mathrm{c_j}\middle|\mathrm{t_i}\right)(1 - \left(\mathrm{c_j}\middle|\mathrm{t_i}\right))$. [18] [20] T is the test used in a node. $\mathrm{c_i}$ Is assigned as a categorical class and the estimated probability in which a feature is in that specific class is defined as $\mathrm{p}\left(\mathrm{c_j}\middle|\mathrm{t}\right)$. In a node is all examples assigned to be in that class, therefore the value is set to 1 and all other set to 0. The variance is combined from each class k and calculated together from $\sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{p}\left(\mathrm{c_j}\middle|\mathrm{t_i}\right)(1 - \left(\mathrm{c_j}\middle|\mathrm{t_i}\right))$, n is defined as the number of splits that are needed to get better "purer" splits then the one before. The test or defined question that maximizes the function above will be the best split that is going to be selected for that particular internal node.

RF for classification problems is stated as an ensemble of E trees, where $\mathrm{T_E}$ and $\mathrm{X_i}$ is defined as a vector of distinctive features or numbers in the p-dimensional space [18] [20] [21]. The combination of all trees contributes to E number of outputs, $\widehat{\mathrm{Y}}_{\mathrm{E}} = \mathrm{T_E}(\mathrm{X_i})$ where $\widehat{\mathrm{Y}}_{\mathrm{e}},\ \mathrm{e} = 1, \ldots, \mathrm{E}$, is the obtained prediction for eth number of trees. A majority vote is defined from the combined predictions, as $\widehat{\mathrm{Y}}$ (Figure 1) and describe a simplified version of a Random Forest algorithm.

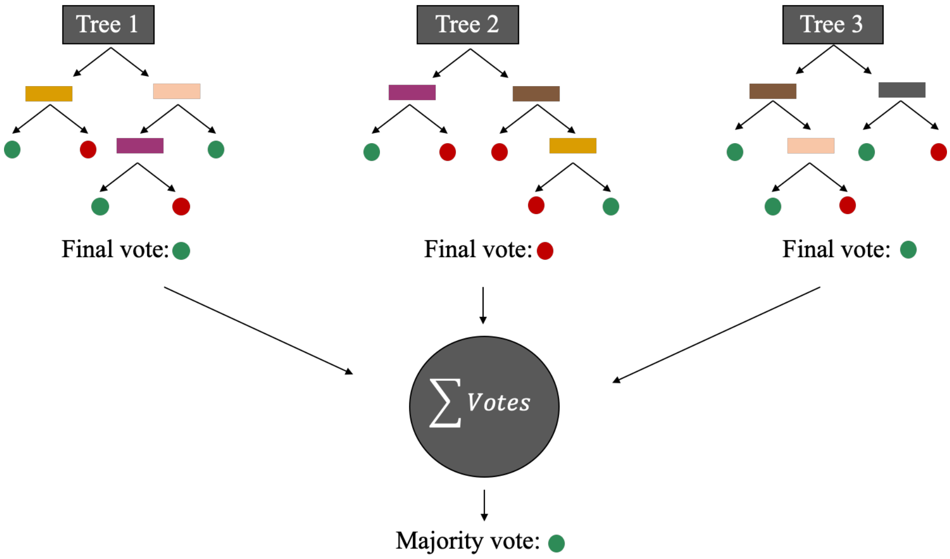


Figure 1: Example of a small Random Forest ensemble of three decision trees. A binary classification problem is exemplified as two colors, red and green. Each decision tree has its own combinations of variables that will be of best choice in finding the most accurate prediction, which is further selected to be best suitable for each defined question. Three number of outputs will present the predictions (final vote) which is obtained from each three. The predictions are combined, and a majority vote is selected as the final prediction for the particular task of interest.

The main differences with a Random Forest algorithm and the earlier bagging approach, that each new training set is drawn with replacements from the original training set. And by random feature selection a new training set is thereafter grown [20]. As described, bagging reduces the variance, but it also gives additional features to the random forest algorithm. It contributes to an ongoing estimate generalization error of the combined decision trees, correlation and strength estimates. All this is returned as the Out-Of-Bag (OOB) estimates, consisting of a test set of OOB samples.

OOB samples are defined as all the samples that were left out during the replacement method while training. Those samples have not been used in training and will be expended as a leave one out type of cross-validation, simultaneously while the model is trained. An error estimate of the model can be obtained by only using the OOB addition. $\text{Error rate (ER)} \approx \text{ER}^{\text{OOB}} = \text{n}^{-1}\sum_{\text{i}=1}^{\text{n}} \text{I}(\hat{\text{Y}}^{\text{OOB}}(\text{X}_\text{i}) \neq \text{Y}_\text{i}$, were $\hat{\text{Y}}^{\text{OOB}}$ is the average prediction from the OOB samples, as a test set and I is an indicator function that defines the wrong prediction of n OOB cases. This type of validation has proved to give similar result as a k-fold cross validation [10].

For measuring model performance, to evaluate how well each class is distinguish by the model is the Receiver Operating Characteristic (ROC) curve of extensively use for classification models [22] [23]. In creation of the ROC-curve, two vital parameters are needed, the first is True Positive Rate (TPR), $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$ .TP is defined as the number of True Positives and FN is the False Negatives. The second parameter is False Positive Rate (FPR), $\text{R} = \frac{\text{FP}}{\text{FP+TN}}$. FP is the number of False Positive outcomes and TN the True Negatives. Together it defines the ROC curve, when TPR are plotted against the FPR.

To compute or evaluate the ROC curve, the Area Under the Curve (AUC) measure is applied [18]. It provides an arrangement of all classification thresholds in a range from 0 to 1. For example, if the designed model predicts zero correct classes an AUC value of 0 will be obtained and if the model predictions are 60% accurate an AUC of 0.6 will be received.

5 Experimental

5.1 Data software

Structural information of reactions was obtained and extracted from the chemical reaction centric database Reaxys (Elsevier) [24]. The open source platform/software used in this project for creation of workflows for curation of data, creation of models and analysis of model predictions performance was The Konstanz Information Miner (KNIME) V3.x.[17]. Molecular properties were calculated in KNIME based on open source toolkits such as Indigo, Chemistry Development Kit (CDK) [25] and RDKit [26]. Default settings of Random Forest algorithm was selected according to Leo Breiman [21]: square root of number of variables for Mtry; sample size, all observations with replacement; node size, selected to one; number of model trees to 100; Gini index as split criterion. Hyperparameters tuned during this project were Mtry and number of trees.

5.2 Laboratory Validation

Two different Suzuki-Miyaura reactions were performed with different sets of catalyst, solvent and base.

5.2.1 Chemicals

Organoborane was provided from RISE Research Institutes of Sweden (Södertälje, Sweden). 1-Bromo-3-nitrobenzene was purchased from Lancaster (Eastgate, White Lund, Morecambe, England), 2-Bromobenzonitrile from Merck (Darmstadt, Germany). Catalysts Pd(dppf)Cl_2, DCM was purchased from CombiPhos, Inc. (Princeton, NJ, U.S.A.), $Pd(PPh_3)_2Cl_2$ from Sigma-Aldrich, Fluka Chemie AG (Buchs, Switzerland) and $Pd(PPh_3)_4$ from Sigma-Aldrich (Darmstadt, Germany). Potassium carbonate was purchased from (Scharlau) Fisher Scientific (Hampton, NH, U.S.A.) and Cesium Fluoride (99%) from Sigma-Aldrich (Darmstadt, Germany). 1-Butanol (99.8%) and 1,4 Dioxane (99.8%) were both purchased from Merck (Darmstadt, Germany).

5.2.2 Instrumentation of High-Performance Liquide Chromatography (HPLC)

The column used for In Process Controls (IPC, i.e. sample taken from reaction) and evaluation of reaction performance, was a Symmetry Shield Reversed-Phase (RP) C8 Column (5 µm, 4.6 x 50 mm) from Waters (Milford, Ma, U.S.A.). Mobil phase composition was with 95% acetonitrile(aq) and 0.1% formic acid. Temperature was set to +40 °C with a pressure at 3 bar. Flow rate was consistent of 2.00 mL/min. with an injection volume of 10 µL. Detection was performed by UV at wavelength of 254 nm.

5.2.3 General experimental procedure

Each Suzuki-Miyaura Cross-Coupling reaction was performed under nitrogen flow at +80 °C. 1 g of organohalide (1-bromo-3-nitrobenzene or 2-bromobenzonitrile), organoborane (provided from RISE) (1.2 equiv.) and base (4.37 equiv.) were mixed and dissolved in a solvent system of 10 mL 1-butanol/H_2O or 1,4-dioxane/H_2O (7/3 ratio). The mixture was stirred at room-temperature with nitrogen flow until full dissolution was observed. Catalyst (0.02 equiv.) was added to start the reaction. IPC's were performed before addition of catalysts and after one hour of addition, with HPLC for monitoring the reaction. After full consumption of starting material, the crude product was extracted and evaporated for crude HPLC analysis. Conversion (%) of product was obtained as final evaluation of the performance of the reaction.

6 Method

This section will present the methodological approach towards the aim, based on the outcomes of three defined milestones:

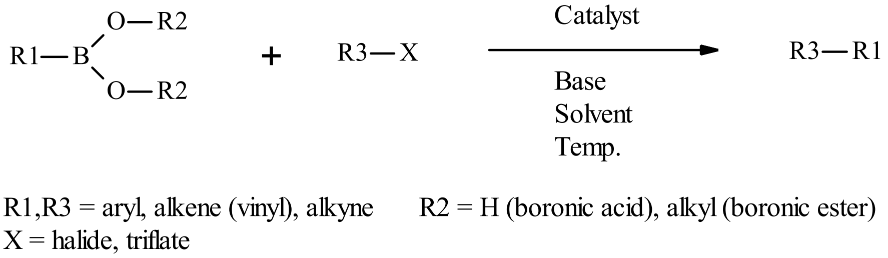
Milestone 1: Design a "local" machine learning model to predict a binary classification task of the Suzuki Miyaura reaction and investigate how to optimize the model in terms of predictability performance.

Milestone 2: Define unique models to predict multiple binary classification tasks of catalysts used in a Suzuki-Miyaura reaction.

Milestone 3: Apply the best model for ranking most preferable catalysts for unknown candidate reactions and test the hypothesis in laboratory to evaluate the model performance.

6.1 Preparation of data

In this project, the widely used cross-coupling reaction Suzuki-Miyaura was studied to create the underlying data table. This reaction is often used for large-scale synthesis of medicinal drugs due to the advantages of mild conditions, using commercially available low-cost reagents [27]. It is a metal catalysed reaction (Scheme 1) most often with palladium. Other metals such as e.g. Ruthenium, Iron and Nickel have also proved to work well. The reaction occurs with the two reactants, organoborane (often boronic acid) and an organohalide or triflate, to form a C-C bond, under basic conditions. [28]


O—R2
R1—B
O—R2
+
R3—X
Catalyst
Base
Solvent
Temp.
R3—R1
R1,R3 = aryl, alkene (vinyl), alkyne
R2 = H (boronic acid), alkyl (boronic ester)
X = halide, triflate

Scheme 1: A generalized reaction scheme of a Suzuki-Miyaura cross coupling. The first component, organoborane, is often shown as a boronic acid but can vary. Reaction takes place with an organohalide, such as iodide, bromide, triflate etc. Reaction conditions, especially catalyst and base, have been shown to be of major impact in the formation of a single C-C bond.

The initial dataset in this project was defined together with structural information, molecular properties and reaction conditions from a total of 100 000 Suzuki-Miyaura cross coupling reactions extracted from Reaxys, (Elsevier) [24] consisting of different borane -, halide/triflate components and products. Each reaction reference obtained information on different yields, solvent systems and catalysts. A KNIME workflow was constructed for data curation, unnecessary information as well as inconsistent data was removed. Duplicates of reactions could have a potentially negatively impact the model performance, by influence the generalization error estimates in predicting a typical classification class. Multi step reactions would give a more complex representation of data to learn, compared to single steps reactions, and would therefore be much harder to describe and predict with only molecular properties and reaction conditions. If two yields were documented for a reaction, the highest yield was decided to remain within the dataset to minimize the uncertainty of the human factor. The 100 most used catalysts were manually highlighted and extracted as a unique value used in the final table. After curation, a total of 94 000 reactions were used for further analysis. For the external validation, 2 000 other Suzuki-Miyaura reactions were extracted from Reaxys and curated as described.

6.2 Descriptor setup

Two distinct descriptor sets were calculated for further implementation and training of the constructed binary classification models. The desired descriptors were primarily based on One-Dimensional (0D/1D) descriptors and Two-Dimensional (2D) descriptors to gather characteristics of individual components.

0D and 1D descriptors are based on molecular formula, and present numerical features e.g. molecular weights, atom counts, etc. These features were extremely fast to calculate and would provide a simplified description of different compounds, compared to 2D descriptors, which are more complex and primarily based on chemical graph theory [6].These latter descriptors provide information on molecules' space (topology), e.g. topological surface area , fragment counts, Zagreb index, etc. These types of descriptors were calculated for each component, containing different open toolkit sources available in KNIME. The first set of descriptors used in the training was a combination of two available toolkits, CDK [25] & Indigo. The second design of descriptors was conducted by RDKit toolkit [26] in KNIME and were calculated for each component of all 94 000 Suzuki Miyaura cross coupling reactions.

6.3 Machine Learning Model design

The desired outcome of this project was to obtain an assisting tool based on the Random Forest algorithm to predict best reaction conditions, e.g. yield and catalyst. To address the main goal of this thesis, a large set of different models were constructed. The strategical approach was to try and obtain the best model, based on a set of several optimization processes. In addition, to meet the defined objectives, different investigations were performed by using "sufficient" numbers of descriptors without losing model accuracy. The main approach was thereafter to optimize the Random Forest model, analysing number of tree models that are enough for obtaining less error misclassification, as well as numbers of descriptors/properties for obtaining the best split at each node while building the forest.

6.3.1 Milestone 1

The approach towards designing a "local" Random Forest model for predicting a binary classification task and investigate and optimize the model to improve predictability measures will be described in this section.

6.3.1.1 First attempt of model design for prediction of yield classes

A first binary classification of yield was performed with the Random Forest algorithm, constructed with the two classes defined as yield ≥60% and ≤40%, to obtain milestone 1. To improve and create a more diverse training set, i.e. increase differences of "good" and "bad" classes, reactions between 60 and 40 percent[1] were filtered out. The random forest model was initially constructed with default values recommended in KNIME, based on the theory of Random Forest mentioned by Leo Breiman [21]. Each RF model was trained with a number of 100 models of decision trees with bagging (Figure 2). Number of descriptors used for each split was randomly obtained by square root of the total numbers of descriptors/properties.

[1] This could also be performed with 65/35 or 70/30 to weed out reactions that might be higher or lower due to experimental errors. Therefore, was it set to 60/40 hence the large interest of very good reactions over bad ones.

Random Forest Model

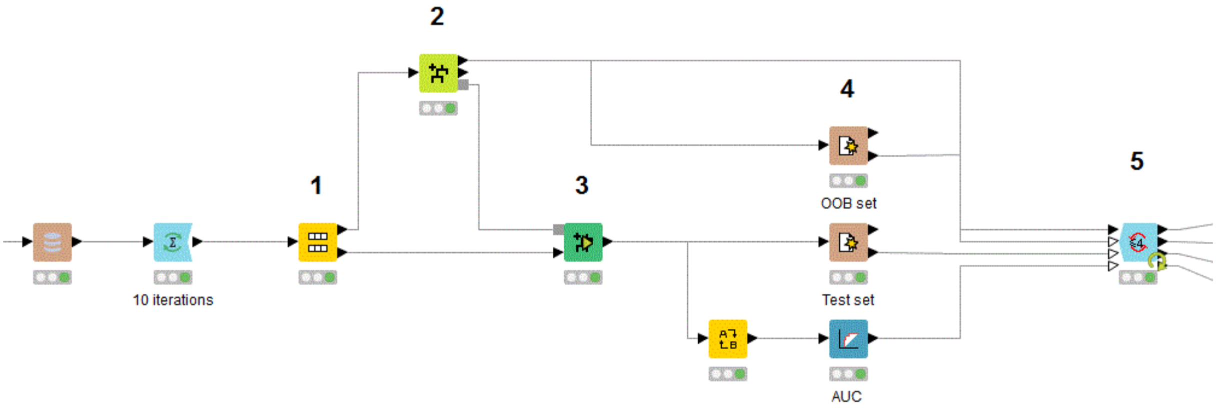


Figure 2: Design of workflow for Random forest model. 1) The curated table was divided into training and test set, 80/20, based on random sampling selection. 2) A tree ensemble learner was implemented, and default settings were selected for the used Random Forest algorithm. 3) Classification predictor was applied to obtain prediction estimates of yield classes on the independent test set. 4) To understand the predictive performances, Cohen´s Kappa, accuracies, specificity and sensitivity measures was obtained respectively AUC measures to get an understanding of the predictive performances of OOB samples and the independent test set. 5) The process was repeated 10 times to obtain an understanding of consistency and model strength.

In addition to the calculated descriptors, different solvent combination and catalyst was defined as a unique number. Each unique number of catalysts and solvent was implemented and combined with the two descriptor sets, to quantifying and comparing experimental performances with each other.

Both data sets were evaluated before learning through a 10-fold cross validation. Each set with an error rate lower that 0, 1% was not accepted, due to the impact of being too biased for a set of reactions. Two different models were created and trained and each respective descriptor sets. From the full dataset, 80% was selected to training and 20% as independent testing the trained model. Evaluation of prediction accuracy was performed from a ROC-curve were AUC measures were obtained.

6.3.1.2 *Optimization to minimize the number of descriptors*

An additional workflow was constructed, to investigate correlation and variance within the two different descriptors sets (Figure 3). The table of descriptors was normalized to a range of 0 – 1, highly correlated descriptors with correlation above 80%, respectively descriptors with a variance below 0.5%, were removed.

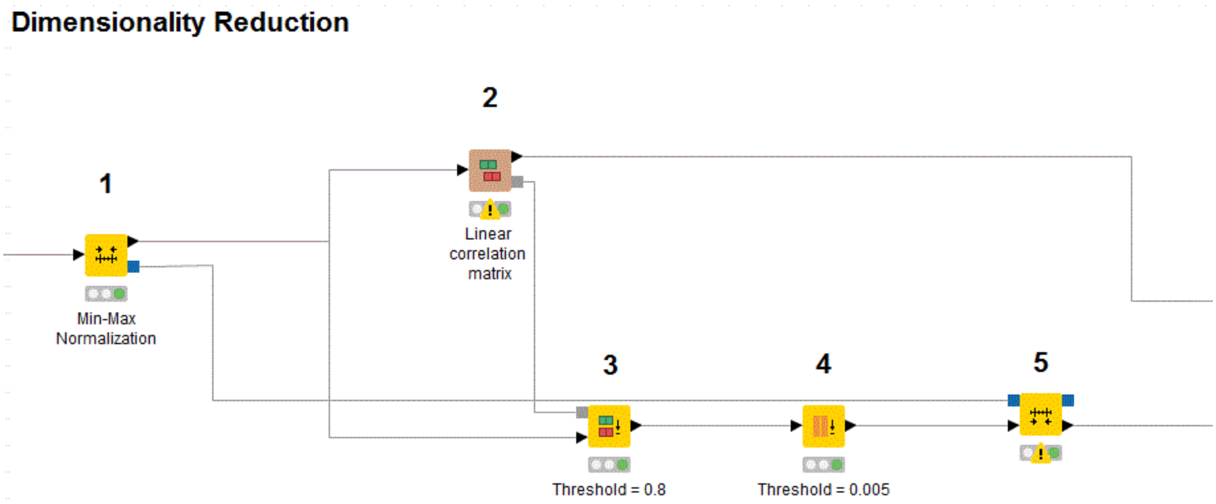


Figure 3: Design of dimensionality reduction (DR) workflow. 1) All input of data, i.e. descriptors, was normalized for further evaluation. 2) A linear correlation matrix was implemented. 3) A correlation threshold was set to 80%, only one of the highly correlated descriptors remained within the table. 4) Descriptors with a significantly low variance (0.5%), will appear to be constant through all reactions and would not have any important impact of building the RF and was later removed. 5) All remaining descriptors were thereafter denormalized to further train the Random Forest model.

After the dimensionality reduction was a workflow to obtain variable importance measures designed (Figure 4). A recursively defined column loop was conducted over all descriptors, where error rates were calculated for each model missing a specific one. A comparison of the variable importance for each type of descriptor was performed, comparing 0D/1D against 2D.

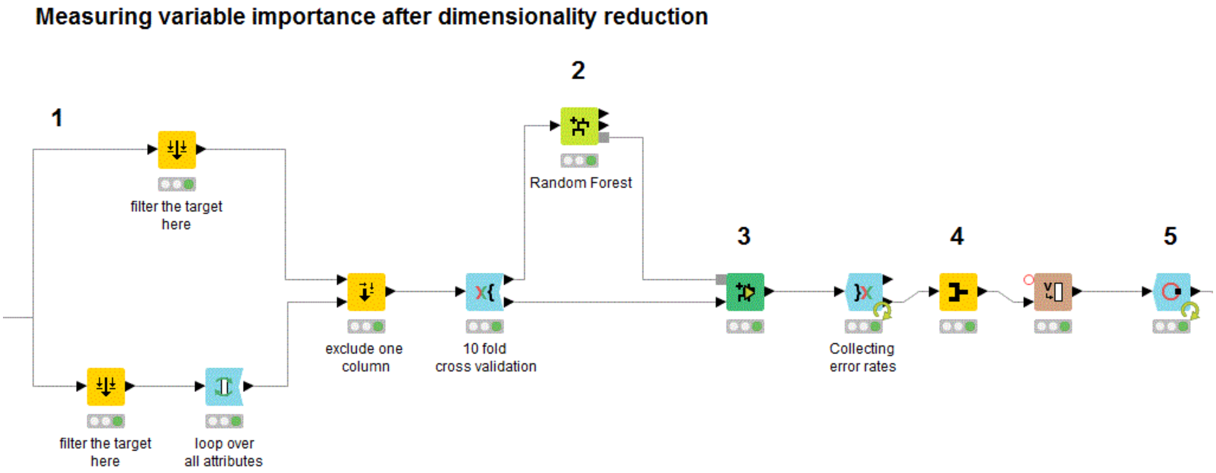


Figure 4: Design of variable importance workflow, 1) all descriptors after DR were selected (numerical values); a column list loop was implemented for selecting one specific descriptor that would be excluding before execution. 2) A 10-fold cross validation was processed, to obtain error rates (percentages of selecting wrong class) of the RF performances of predicting correct class in the independent test sets. 4) A mean error rate was calculated, obtaining information how well the designed algorithm without one specific descriptor would perform. 5) The workflow was recursively performed, until acquiring enough information on the importance of all defined descriptors.

6.3.1.3 *Optimization of model design*

To develop a strong and accurate predicting tool, of e.g. yield classes, the most efficient way was to optimize the training phase of the machine learning algorithm. The Random Forest algorithm contains different parameters that can be tuned to optimize prediction accuracy, defined as hyperparameters [29]. The goal was to investigate in which parameter value would obtain strong and accurate predictions with less computational costs. The first hyperparameter to optimize was Mtry: a new workflow was constructed with eight similar models with different Mtry values. Error rates were obtained from each model and later analysed against the importance of number of Mtry. A similar approach was constructed for investigating the importance of number of trees, consisting of eight Random Forest models with 2 000-, 1 000-, 500-, 100-, 50-, 10-, 5- and 1 tree. Error rate in percent was obtained and plotted against the number of trees, which impact was further analysed to predict its accuracy.

6.3.1.4 *Final Random Forest model of a binary classification of yield*

The final RF model was created based on results from the optimization processes. Initially, each dataset was evaluated through a 10-fold cross validation in an order to avoid a too biased model. Each RF model was corrected and trained using 500 models of decision trees with bagging and random selection of features. The number of Mtry was set to the optimized value for each model and ROC and AUC measures were used for benchmarking the model performance. The final model design was merged with several workflows to later obtain information on AUC for the defined task (Figure 5).

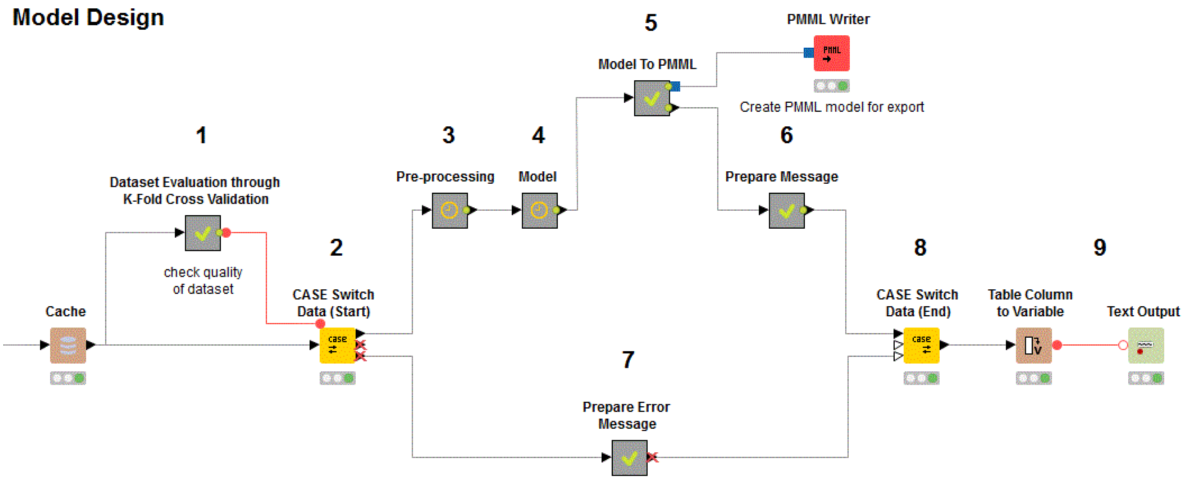


Figure 5: Design of final Random Forest model. 1) The final established dataset was evaluated with a 10-fold cross validation to observe no overfitting was present. 2) If error measures were lower than 0, 1 %, the dataset was rejected. 3) The accepted dataset was later pre-processed, unnecessary features (X_i) were removed for achieving faster calculations in training of RF model. 4) Optimized model conditions for training the data of the RF model were implemented and AUC measure was presented for evaluation. 5) The model was converted to a Predictive Model Mark-up Language (PMML) for export. 6) The obtained results were later performed in a message, with information on obtained AUC for the defined task. 7) If the model was not accepted, an error massages would been performed instead as final text message. 8) Obtained import of data is later concatenated into a final estimation. 9) A final clear-text message was presented for obtaining a final understanding of the model performances.

An external validation set was prepared with approximately 2 000 new Suzuki Miyaura reactions, extracted from Reaxys and curated with the designed workflow. Identical reactions compared to training set were removed and descriptors were calculated. The final RF model was later used in predicting yield classes and estimating model performance. To validate the model, AUC and accuracies were evaluated.

6.3.2 Milestone 2

MS2 encompasses the approach to further define a model for prediction of one or multiple combined binary classifications of the most used catalysts in a Suzuki-Miyaura reaction.

6.3.2.1 Random Forest model for binary classification of catalysts

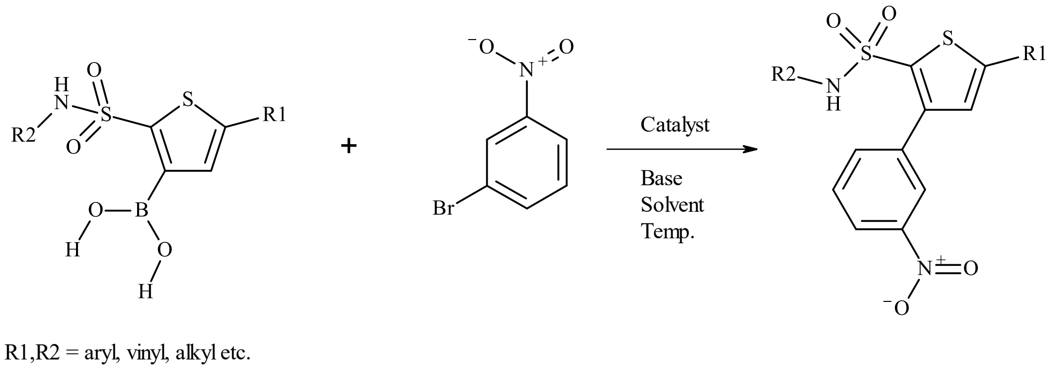
To obtain a tool/procedure in assisting chemists to deciding best reaction conditions, one critical criterion was to find an appropriate catalyst for the desired reaction. The next task was to investigate if the Random Forest algorithm would be of an appropriate choice in predicting multiple binary classifications, represented by a large set of molecular formula and graph-based descriptors. To encounter the defined objectives, 6 different binary classification models were constructed and ranked according to appearances within all 94 000 reactions. The six catalysts selected were confirmed to be commercially available and followed to obtain an active phosphine ligand. In detail, each desired catalyst was given an isolated class versus all others. Descriptors used in each table were selected after DR workflow performance and the creation of 12 models. Yield and solvent information were set to distinctive numbers and later combined with the descriptor tables. For defining a narrower prediction (lowest misclassification error) of the catalysts, each model was optimized after tuning of the two hyperparameters. Further evaluation of the performances was obtained by AUC measures and an external validation set was tested on four randomly chosen models, based on a set consisting of 1000 Suzuki Miyaura reactions extracted from Reaxys.

6.3.3 Milestone 3

This encompassed the description of the method to define a final model for ranking of most preferable catalysts for unknown candidate reactions, as well as the experimental validation of the obtained predictive performance.

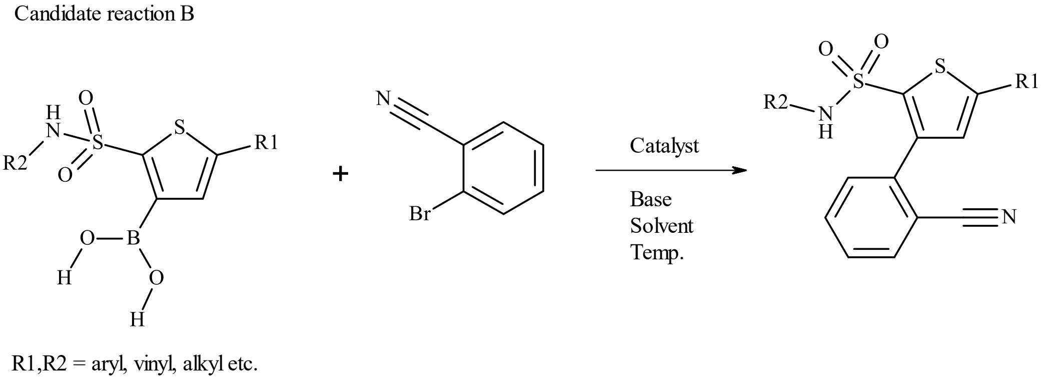
6.3.3.1 Ranking candidate catalyst

To obtain a model that would have the possibility to rank the most preferable catalyst for a specific reaction. The main idea was to combine the methods defined in 6.3.1 and 6.3.2 to create a model that would be able to present catalysts that will have a higher possibility of presenting a high yield. Each catalyst model was later ranked based on probability score. Molecular properties were calculated on each component for each candidate reaction (Scheme 2 & 3) to obtain a representation of data that would be possible to use in prediction of best catalyst for that reaction. Each representation with computed descriptors of both candidate reactions was curated to fit the RDKit based model design and an additional column of desired yields was implemented as a unique column, ranging 80 – 100%, to be a priority for the different catalyst models.



Scheme 2: General reaction scheme for candidate reaction A, the organoborane component selected was a boronic acid with high polarity. The desired C-C coupling product is created using 1-bromo-3-nitrobenzene as halide.

Each candidate reaction was later implemented within each model and ranked in obtaining a yield above a certain threshold. Candidate A (Scheme 2) was first tested and predictability scores was obtained for each catalyst. The approach was later performed on candidate reaction B (Scheme 3). All scores were evaluated, and three predicted catalysts were selected for further validation in laboratory.



Scheme 3: General reaction scheme for candidate reaction B. Identical organoborane component selected as in A, with 2-bromobenzonitrile as halide, forming the desired C-C coupling product.

6.3.3.2 Experimental analysis on candidate reactions

Each reaction was performed under predetermined conditions to validate the ML model performance of ranking best catalyst for a specific reaction. Solvent and base were selected to be optimal for each reaction and was later adjusted based on experimental outcomes. Initially five different conditions for candidate reaction A were tested, with n-butanol as solvent and potassium carbonate as base, using "best" and "worst" catalysts as obtained from the ranking model. Three new reactions were performed using 1,4-dioxane as solvent and potassium carbonate as base, evaluated together with all three catalysts.

The selection of conditions for candidate reaction B was selected according to the structural similarities between the two candidate reactions, therefore was the best conditions for candidate reaction A selected as the first test. To increase the reactivity and to obtain better conversion of desired product, the selection of base was of major impact. In the Suzuki-Miyaura reaction, a base is used to initiate the transfer of the aryl or alkyl group from the organoborane to the catalyst-halide complex. Caesium fluoride was selected and is known to work well when base-sensitive substrates are used, to reduce the formation of beta-hydrogen elimination by-products [27]. Several in process controls were performed with a HPLC analysis for monitoring the reactions. The chromatogram of starting materials was processed for later on comparison with each IPC´s on crude reaction mixtures.

6.3.3.3 Evaluation performance of obtained results from experiments

HPLC was used for analysing obtained product versus by-product. The conversion was calculated from processed chromatograms according to intensity of absorbance. This was further used as support for evaluating the defined model performance of ranking best catalysts for a given reaction. The combined intensity in percentage for desired product was later defined and a preunderstanding of the reaction outcomes, e.g. highest possible obtained yield was obtained.

7 Result and Discussion

In this study, several Random Forest models were constructed for predictions of binary classification tasks. This section will present the achievements toward the milestones and final aim of thesis.

7.1 Milestone 1

The first goal was to design a “local” machine learning model to predict a binary classification task and perform an optimization study to improve the accuracy of the model.

7.1.1 First model for prediction of a binary classification of yield

To predict if a reaction would result in > 60% or < 40% yield, interesting result was found with the Random Forest algorithm. This, with an accuracy of 92.5% of the out of bag samples through the CDK and Indigo descriptor set (Table 1). The result from the independent test set was similar to OOB set, indicating that OOB samples can be a valid performance estimation for further analysis. Together with the true positive rates and false positive rates, a ROC-curve was plotted to acquire more accurate prediction estimation. The AUC measure from the ROC-curve was 98% (Figure 6).

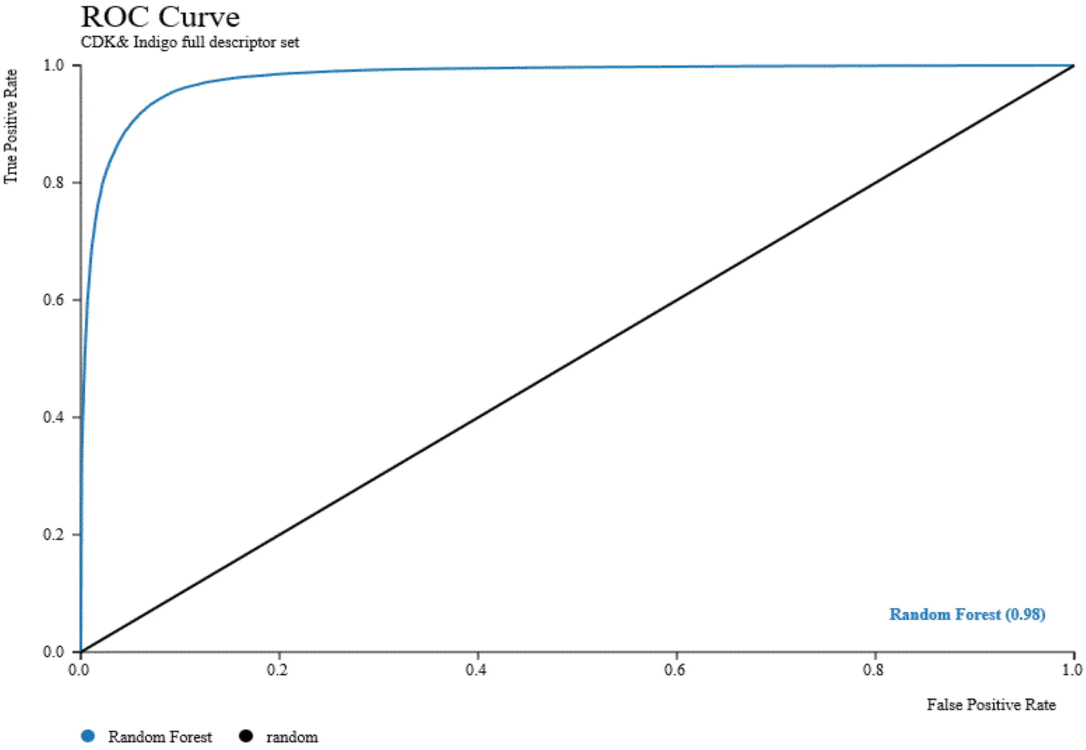


Figure 6: Plotted ROC-curve with AUC measure for the CDK & Indigo method with full descriptor set to evaluate the predictability performance of the model. True positive rates (TPR) were plotted against the false positive rates (FPR) to create the ROC curve. AUC was measured to 0.98 in predicting the correct binary classification task of the two yield classes with Random Forest as ML algorithm.

With the RDKit method, comparable results with an increase of 1% in accuracy and 0.2% for AUC was shown (Table 1). The design of the Random Forest algorithm and the Gini index ability for finding best features for each node was a key factor behind the model´s performance. One concern was the known problem with the presence of unwanted parameters. These parameters are known to contribute with a negative effect in the construction of the ML models, e.g. leading to a low predictability performance.

Another concern was if the final model would correspond too precisely to the training set and fail to produce reliable predictions in the future. This phenomenon refers to the theory of overfitting. With a 10-fold cross validation sequence the dataset was thus evaluated, a difference obtained after each "out of sample testing", indicating there is no direct overfitting.

Table 1: Random Forest performance estimates for both models with respective full descriptor sets after a 10-fold cross validation of the binary classification, if yield was >=60% or <=40% with default settings of parameters on the Random Forest algorithm as defined in 5.1. Sensitivity, Specificity, Cohen´s kappa, accuracies and AUC measures was obtained for evaluation of model performance.

Methods	*Descriptor toolkit*	*Set*	*Sensitivity*	*Specificity*	*Cohen's kappa*	*Accuracy (%)*	*AUC*
RF model Default settings (5.1)	CDK & Indigo	Out of bag	0.903	0.946	0.849	92.5	-
		Test	0.904	0.948	0.852	92.6	0.98
	RDKit	Out of bag	0.912	0.95	0.862	93.1	-
		Test	0.914	0.953	0.868	93.4	0.982

The RF algorithm is ordinally defined to avoid the existence of this problem. Under the construction of each decision tree in the formation of the forest, the process of selecting random samples of features would play a sizeable part in avoiding the problem. Each tree was selected and constructed with a certain number of random subsets of features defining one specific tree. The number of trees will minimize the error due to biases and variance and will each time present different models, thereby avoiding overfitting. In the study by Skoraczyñski G et al.[4], best prediction performance was shown by a binary classification of yield to only 65%. The differences of this approach opposing theirs was to first define a "local model" for a specific type of reaction, the Suzuki Miyaura reaction, instead of a "global model" that uses a large set of different types of reactions, as Heck, Negishi etc.

The main idea of this project was to find a non-obvious correlation between molecular properties and reaction conditions as well as reaction outcomes. The problem to describe the reactivity of a reaction has been a large problem, and molecular properties have been proved to not give such information [4]. Hence, they will only present information on specific molecules, not an entire reaction. The use of a large set of different reactions with molecular descriptors/properties for each component of those reactions, has been shown in this project to be enough to find a pattern to mimic good (> 60%) or bad (< 40%) reactions. The obtained results in this study emphazizes that the "local model" approach was more preferable way to encounter these types of tasks and is of interest for further analysis. An interesting approach for further advances was to investigate if there is a possibility to remove unnecessary parameters to improve the execution and reduce number of calculations.

7.1.2 Optimization to minimize the number of descriptors and model conditions

Each defined tree in a Random Forest makes splits of descriptor values into different branches, which continues until a final criterion is reached. For making good splits, an important requirement was to study the variance within the descriptors over the full representation of data. If not, a large number of unnecessary splits would be performed. Highly correlated descriptors will not show any differences under splitting. The appropriate choice was to keep one of them to minimize the number of unnecessary splits.

To optimize and find best number of descriptors, the defined criterion was used in creating the final dimensionality reduction (DR) method. The descriptors/properties used in this project originate from medicinal chemistry to describe certain possible correlations between molecules. To obtain good information of these types of molecules in a Suzuki Miyaura reaction and obtain a model with less computational costs the two datasets were constructed with these types of descriptors.

In comparison of the two different sets of molecular properties, the RDKit toolkit proved undoubtedly to present a better representation for each component in a reaction. The remaining molecular properties with the combinatorial descriptor set of CDK and Indigo, obtained information based on product and organoboranes and few for different halides/triflates. Graph-based descriptors such as TPSA and Petitjean number and count-based properties such as number of rings and atom numbers remained after the DR workflow.

This was not an unexpected observation, descriptors based on topology (2D) presenting in theory a larger variance (more unique) and lower correlation. The representation of 2D properties was based on calculations of connectivity measures, which might play a larger part in the predictivity performance of the model. Molecular formula- and fragment-based descriptors (0D/1D) are more likely to present values with a high correlation and were removed by the defined criterions. The remaining descriptors with RDKit toolkit, was primarily based on Van der Waals surface Area (VSA) and molecular quantum number (MQN) for the different components in the reactions. Greater information on number of rings and atom numbers was received from MQNs. Description of hydrophobicity and hydrophilicity was received from VSA descriptors.

To gain an understanding for each final descriptor selected and the predictability performances of the two sets, a variable importance study was performed. The obtained results for each molecular property were plotted against the misclassification error, i.e. the possibility of not being able to predict correct class (Figure 7). CDK and Indigo descriptors contributed with a distribution in error rate between 8.1% to 9.75%. 2D descriptors, e.g. sp^3 character, Petitjean number and TPSA etc. proved to contribute to stronger predictions and better model performance. Count-based descriptors were presented as envisioned with a higher similarity between components. Lower misclassification error measures were shown and are of less importance in prediction of yield. The 2D molecular descriptors are more likely to present a component in a more descriptive way and are of a substantial importance, which is also observed with the lower range of misclassification error.

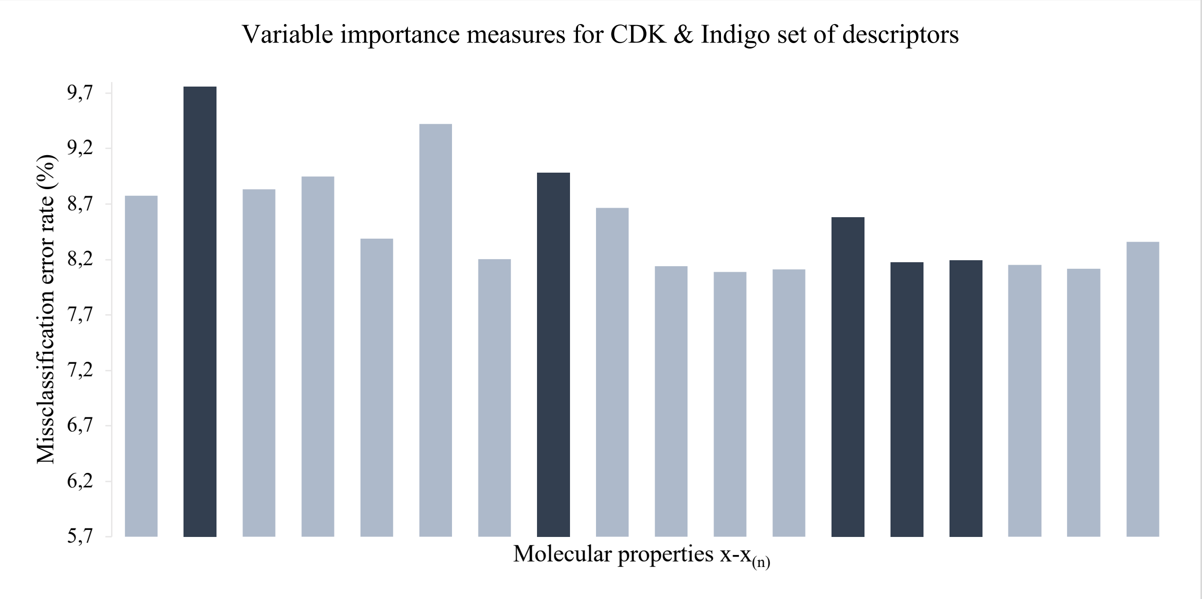
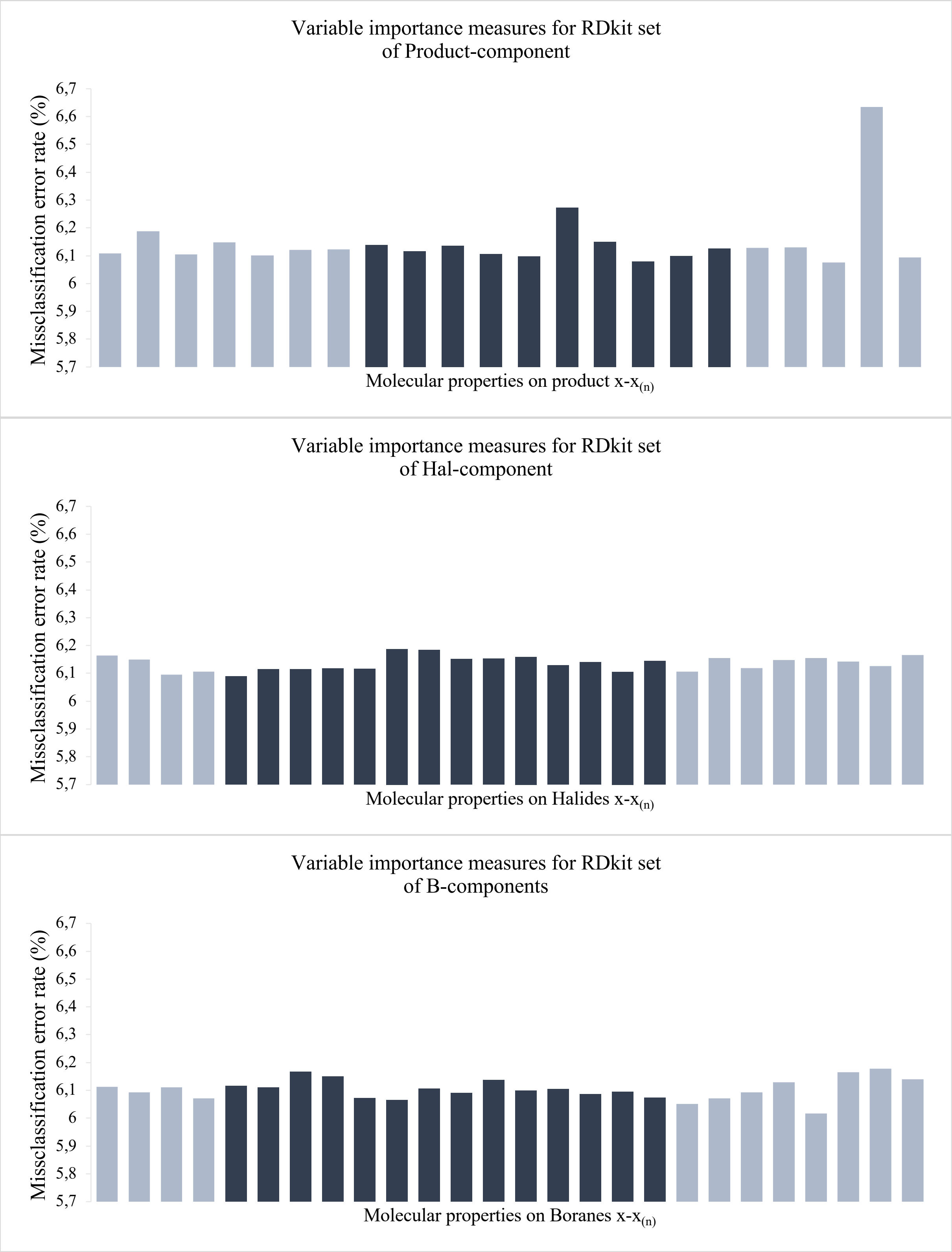


Figure 7: Variable importance measures distribution for CDK & Indigo descriptors after dimensionality reduction. Dark blue bars correspond to 2D descriptors; light blue bars correspond to 0D and 1D descriptors. Showing the highest obtained misclassification error rates of the different parameters, was presented to not exceed 10% and the lowest error rate was presented to 8.1%.

RDKit descriptors with default model conditions, presented in general lower misclassification error rate relative to CDK and Indigo (Figure 8). This supports the hypothesis, that RDKit calculated descriptors presenting a better representation of data. The distribution within all descriptors was obtained between 6% to maximum 6.6% in misclassification error rate. The molecular property that appears of most significant importance was molecular quantum number 35 for products, presented as a light blue bar in figure 6. MQN numbers are categorized into different branches, such as atom count, polarity counts, and topology counts etc. Number 35 represents the number of 5-membered rings present in products [30].



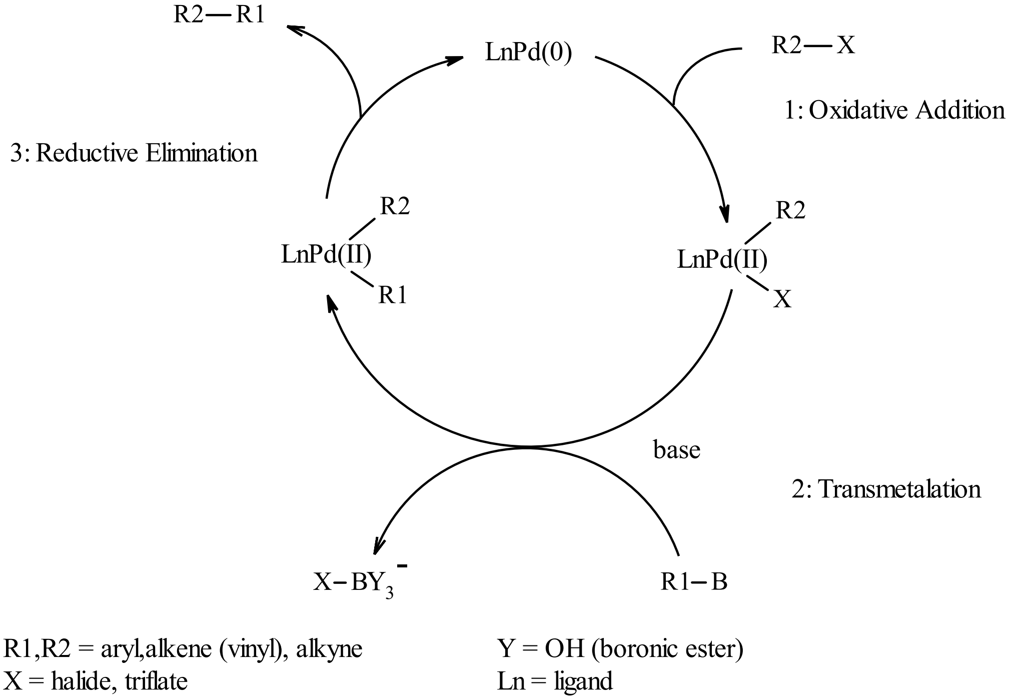
Variable importance measures for RDkit set of Product-component

Variable importance measures for RDkit set of Hal-component

Variable importance measures for RDkit set of B-components

Figure 8: Variable importance measures distribution for all RDKit descriptors after dimensionality reduction. Dark blue bars correspond to 2D descriptors; light blue bars correspond to 0D and 1D descriptors. Showing the highest obtained misclassification error rates of the different parameters, was presented to not exceed 7% and the lowest error rate was presented to 6%.

The catalytic cycle in a Suzuki Miyaura cross coupling reaction follows three steps: 1) oxidative addition (rate determining), 2) transmetalation and 3) reductive elimination (Scheme 4). The position of a five membered ring could have a huge impact in full conversion of starting materials, due to the importance of the rate determine step. If a five membered ring was placed within the boronic component (less electron rich) the possibilities to interfere with the electrophilic halide/triflate in activate the catalysts could be sufficient for having a negative impact on the desired reaction outcome.



Scheme 4: Generalized catalytic cycle of a Suzuki-Miyaura cross coupling reaction.1) Oxidative addition, is the rate determining step were the palladium catalyst together with the halide or triflate forms the first organopalladium intermediate. 2) Transmetalation, is the second step in the cycle were the organoborane reacts with the organopalladium in a participation of a base towards creating the second organopalladium intermediate. 3) Reductive elimination is the last step of the cycle were the desired C-C formation is obtained and the originated palladium catalyst is reestablished to complete the cycle.

Descriptors such as MQN 35 might present too generalized information on how many five-membered rings that are located in the product and will not give the exact information where it is located. It will not give information if it would be presented as a substituent on the boronic/halide component or as the main part of that component, it could be questioned that the model will not capture that information. The MQN 35 has been shown to correlate with reaction outcome for this particular dataset of Suzuki-Miyaura reactions, it might be too bold to suggest that it would be of notable importance in detecting a pattern to predict if a reaction would be of immense success or not. Within this dataset, it seems to be one of the most favourable descriptors to define the best pattern for the yield.

Simplified molecular properties and structural elements that these different datasets are constructed with presents generalized information of the molecules. The model performance in predicting a specific outcome would depend on a number of randomly chosen descriptors best fitted for the Suzuki-Miyaura reactions. The simplified descriptors would tend to coincide with other structural information and conditions to be favourable for the Random Forest model, finding the most preferable pattern to obtain the best predictive results for this particular reaction.

Another concern would be the possibility of beta-hydrogen elimination instead of the desired trans-metalation after the oxidative addition, leading to a higher amount of undesired product. Even though some parameters could have an impact on the predictive performance of a reaction outcome, are some of the descriptors less important to be of an immense impact if a certain reaction will occur or not. From a chemical point of view, would the reactivity of a reaction not be dependent on e.g. molecular weight or element count, to define the possibilities in producing the desired product. There are more parameters that one must take under consideration in producing a certain amount of product with a specific transformation reaction, such as the Suzuki- Miyaura reaction.

However, in these types of models might such of different "unnecessary" parameters be crucial in finding patterns for specific outcomes. The information from the literature will be extremely important to obtain a good representation of data. Molecular properties will not give enough information to predict an outcome of a reaction and the size of extracted reactions will play a key role in the model's ability to find patterns towards the goal. From a statistical point of view were the results more acceptable. A non-obvious correlation was presented in predicting a binary classification of yield with the selected simplified molecular descriptors. For further development was all descriptors selected for each dataset. The importance measures for each molecular property did not deviate significantly more than 0.1% and an improvement would not be reached by randomly remove molecular properties

The optimization of a ML model does not only refer to the representation of data. For obtaining "best" model conditions, different hyperparameters were tuned to receive optimal predictive performance with the model. The approach was to find conditions with a high accurate predictability performance at lowest computational cost, meaning with as few calculations as possible. The best Mtry observed for CDK and Indigo set was with 10 features after a 10-fold cross validation (Figure 9). If only three or less features were selected would the misclassification rate be of a considerable importance in predicting the desired class. The approach was to select number of Mtrys´ in obtaining lowest and most stable misclassification error rates, and with as few numbers of features as possible.

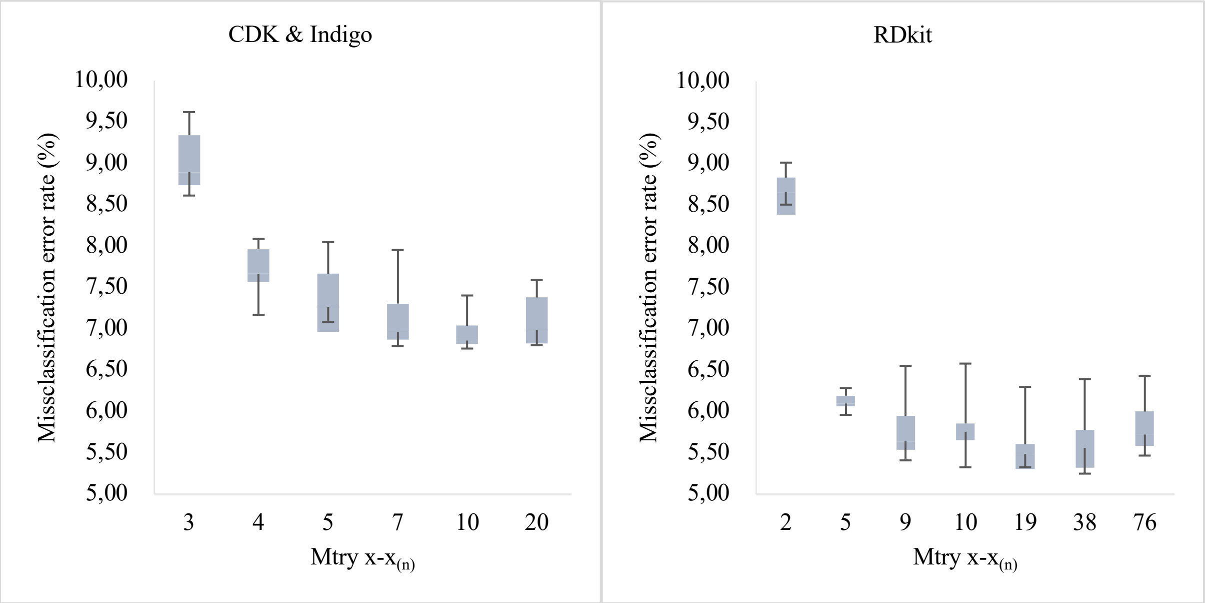


Figure 9: Box plots of number of Mtry after a 10-fold-cross validation for both CDK & INDIGO- and RDKit sets, where the misclassification error in percent are described at y-axis and number of Mtry´s tested at x-axis.

Default settings documented in the theory of the RF algorithm are set to the square root of numbers of features, but as Breiman clarified [21], no optimal set of Mtry can be pre-defined to obtain the "best" model performance. The square root numbers are shown in each plot to 4 respectively 9 in the x-axis of figure 9, showing that the default value is not the best fit in this task. An optimization process was shown to always be sufficient for understanding of each ML algorithm applied in model building [29]. According to the boxplots, was the optimal number of Mtry selected to 10 for CDK & Indigo respectively to 19 for RDKit, acquiring low error rate and would not be that time consuming for calculations.

Next hyperparameter tuned was the number of tree models used in training. The default numbers of trees are selected to 100 and it follows similar agreement defined for Mtry. An optimization study is of need for all hyperparameters obtaining "best" model conditions. For both descriptor sets, 500 numbers of trees were observed to be enough in acquiring good predictability and to not be too demanding for the computational performance (Figure 10). Increasing the number of trees to 1 000 or even 2 000 was verified to influence the execution time and increase the computational cost and demonstrated no significant performance gain while doing that. The final optimized conditions were suggested to 500 number of trees and 10 features of Mtry in CDK & Indigo model versus 500 trees and 19 Mtry for RDKit.

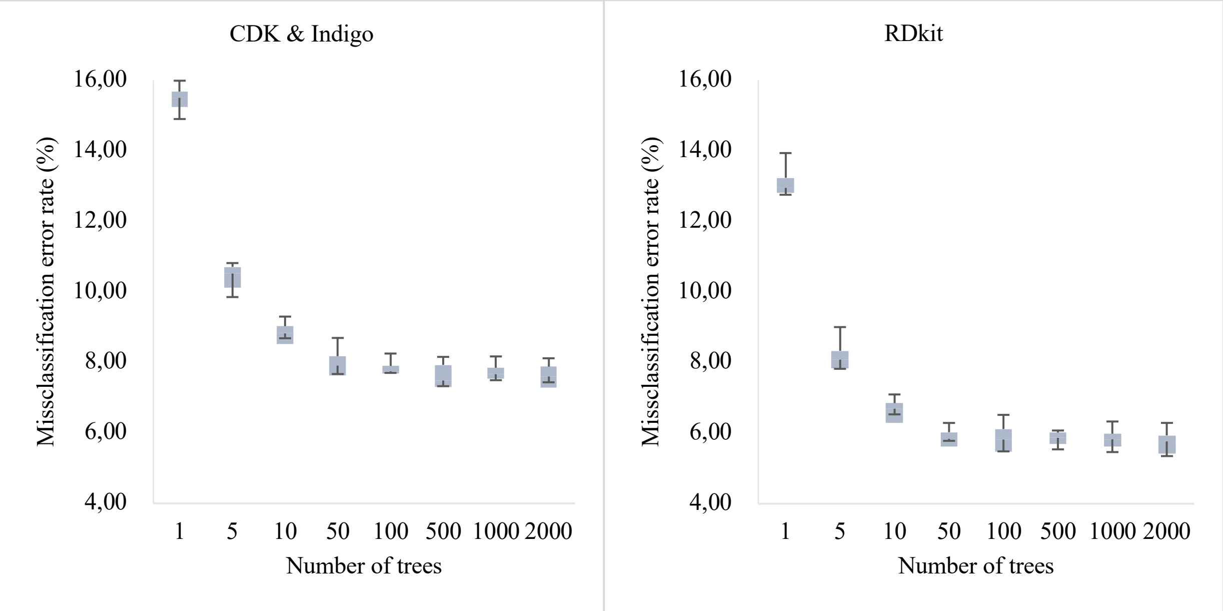


Figure 10: Box plots of number of trees after 10-fold cross validation for both CDK & Indigo- and RDKit sets, where the misclassification error in % are described from 4 -16 % in the y-axis. Number of Mtry´s tested 1-2000 in x-axis.

7.1.3 Final model for prediction of binary classification of yield

The improved "local" models were shown to present comparable results, towards the aim to forecast if a reaction would end up in a specific threshold. That was faster less time consuming than the previous ones. Both models with an additional validation set are shown in Table 2. One concern was if the predictability performance would decrease, hence the dimensionality reduction process. This performance was presented with an AUC of 97.7% respectively 98,5% from the models. The latter one is presenting an increase of 0.5% with the new RDKit set.

Table 2: Random forest performance estimates of both models of different descriptor sets, after a 10-fold cross validation of the binary classification with optimized conditions. Sensitivity, specificity, Cohen´s kappa, accuracies and AUC measures were calculated for evaluation of models as well as for external validation.

Methods	*Number of Mtry*	*Number of trees*	*Descriptor set*	*Set*	*Sensiti vity*	*Specifi city*	*Cohen's kappa*	*Accuracy (%)*	*AUC*
RF algorithm	10	500	CDK & Indigo	Out of bag	0.901	0.943	0.843	92.2	-
				Test	0.899	0.946	0.845	92.2	0.977
External Validation				Validation					0.744
RF algorithm	19	500	RDKit	Out of bag	0.919	0.952	0.871	93.6	-
				Test	0.927	0.950	0.877	93.9	0.985
External Validation				Validation					0.771

To validate the final model, an external validation set with newly documented Suzuki-Miyaura reactions was analysed. Most preferable AUC estimates were received with the RDKit calculated molecular properties, being three percent more accurate than the combined CDK & Indigo ones. A comparison between the independent test set and the external validation set performances presented a difference of 20% in accuracies, which emphasizes the difficulties with this type of task. To evaluate the results, the distinct differences in obtained predictions could depend on several factors. Was the model too biased or was the curation of validation set not performed well enough?

The validation set contained a small size of data in proportion to the size for training and could include reactions where the human factor has had a superior impact on the reaction outcome. For example, a specific reaction could be reported with 20% yield, which then belongs to the reactions defined as "bad", but if another chemist performed the exact reaction and obtained better yield, e.g. of 80%, the reaction would be consider as "good".

The human factor would have a major impact due to the negative obtained results from the external validation. The validation set must be of a greater volume for minimizing the impact of the human factor, to perform a more equitable validation performance. A correlation was shown in the obtained results, which support the hypothesis to predict reaction outcomes based on a simplified representation of molecular properties. For further advances towards the goal of the project, as next task was an attempt to predict specific reactions conditions, e.g. identify best fitted catalyst for a reaction.

7.2 Milestone 2

The obtained results from multiplied combined binary classifications in prediction of several catalysts will be described

7.2.1 Final model for prediction of binary classification of catalysts

Within the full set of 94 000 reactions were the six most used catalysts for this reaction selected. The six models were based on reactions with a representation of only the six different catalysts, resulting in 60 000 reactions. The purpose was to use as many reactions as possible with the desired classification, e.g. as observed in Figure 11. With Pd $(PPh_3)_4$ being set to be the desired catalyst, a final of 59 736 reactions was obtained, with an equal distribution of 50/50 of the two classes, meaning 29 868 for desired catalyst (Class 1) respectively for "all others" (Class 2). In comparison with e.g. XPhosPdG2, that contains only a total of 1 012 reactions after equal sized distribution of the two classes, which is no more than 2% of the entire $Pd(PPh_3)_4$ model.

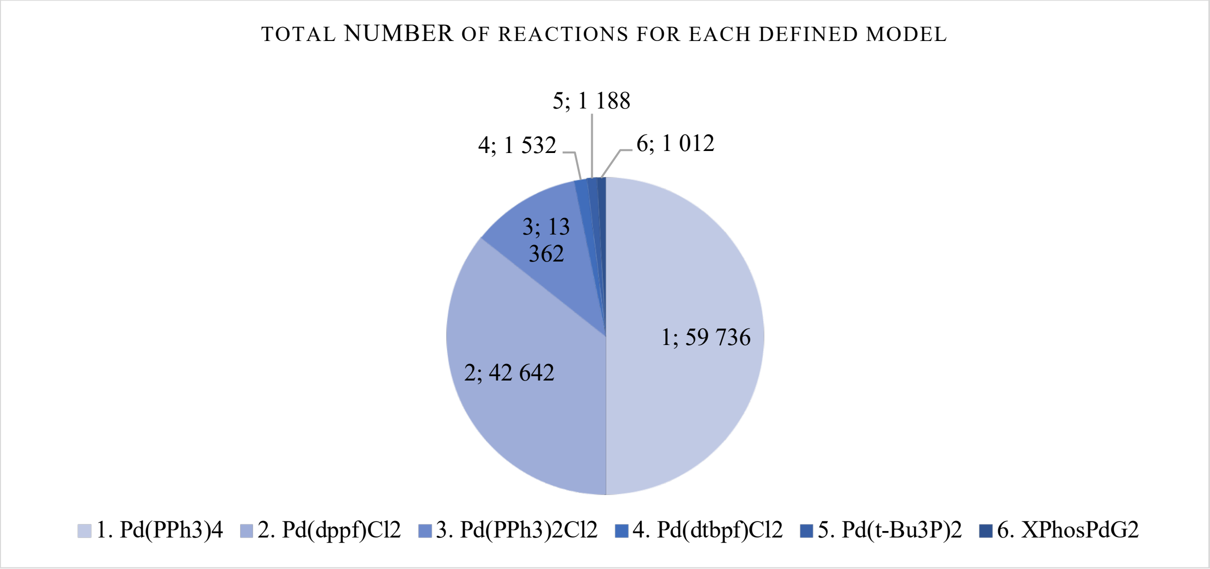


Figure 11: Distribution in volume of total amount of reactions for each defined model, 1) $Pd(PPh_3)_4$ with a representation of 59 736 reactions, 2) $Pd(dppf)Cl_2$ with 42 642 reactions, 3) $Pd(PPh_3)_2Cl_2$ with 13 362 reactions, 4) $Pd(dtbpf)Cl_2$ with 1 532 reactions, 5) $Pd(t\text{-}Bu3P)_2$ with 1 188 reactions and XPhosPdG2 with 1 012 reactions.

The approach was to obtain a comprehensive description of each catalyst and to avoid losing a significant representation that could be essential in anticipating the true description. Due to this, a large variety in the representation (i.e. size) for each constructed model was obtained. The purpose with a large set of different reactions was to minimize the influence of the "human factor" and biases, to obtain a more accurate predictive model.

Due to DR workflow, with the criterion of low variance and too high correlation was different number of descriptors selected for each model. A large representation of different reactions would influence the possibilities of obtaining highly correlated variables, resulting in smaller sets of molecular properties. In comparison of the models consisting of a smaller size of reactions, Model 4 – 6, were a larger representation of descriptors was presented. As performed earlier, is the optimization of model conditions of major impact obtaining best model performances.

It appears that number of Mtry and number of trees increases with smaller table size (Table 3).To reduce the impact of noisy parameters, which leads to a large misclassification error rate, was number of Mtry set to larger than 4 (CDK & Indigo) respectively 9 (RDKit) features, increasing the ability finding better split variables. With the selection of a larger number of trees, was a more exact estimation of the majority votes obtained reducing the uncertainty of the wrong ones. Increasing the algorithm to obtain a more accurate prediction performance.

Table 3: Obtained results after optimization analysis for each model defined. Mtry was tuned together with numbers of trees for achieve a solid predictive performance for each binary classification model.

Model Entry	*Descriptor set*	*Number of Mtry*	*Number of trees*
Model 1	CDK & Indigo	5	500
	RDKit	10	500
Model 2	CDK & Indigo	5	500
	RDKit	10	500
Model 3	CDK & Indigo	5	1000
	RDKit	10	1000
Model 4	CDK & Indigo	5	2000
	RDKit	13	1000
Model 5	CDK & Indigo	5	2000
	RDKit	13	1000
Model 6	CDK & Indigo	5	2000
	RDKit	13	1000

In comparison of the different binary classification performances with defined models, AUC measures and accuracies from both OOB samples and test sets are presented in Figure 12. The accuracy from each respective set, describes how well the ML model identifies each class in an equal size distribution. Larger representation of data indicates to present better or more accurate estimation of identifying the differentiated classes. AUC estimates were received with an average estimate at 95.6%, compared to minor tables with an AUC average of 93.8%. All models presented to possess similar predictability performances, thus not so surprisingly due to each are optimized to perform well under specific conditions.

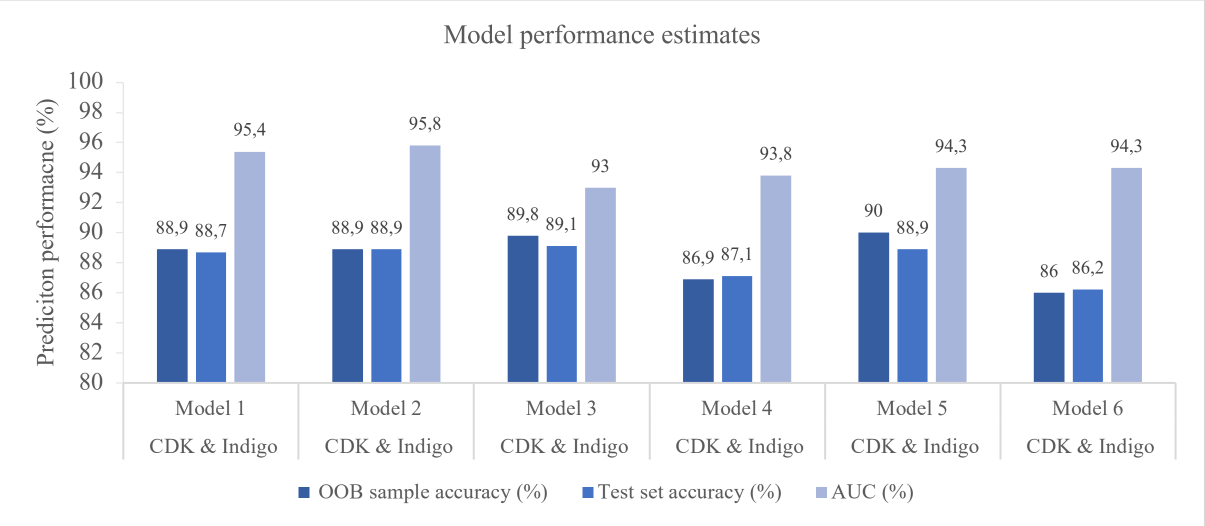


Figure 12: Estimated predictive performance from each model, OOB sample accuracy, test set accuracy and AUC measures from ROC-curve obtained from test set is shown, where each model is represented with CDK & Indigo descriptors.

From results in Figure 13 was it possible to notice a slight better prediction performance using the RDKit models, an overall increase of accuracy and AUC estimates was observed to an average of 96.8% compared to an averaged of 94.4% for CDK & Indigo.
To minimize the number of incorrect estimates in further analysis and obtain the "best" model for ranking appropriate catalyst, was the RDKit models selected as candidate for the final ranking procedure. An external validation (small set of reactions with documented catalysts) was performed on four of the six models. The effect of the human factor was already known to have a large impact, it was further of interest to analyse if models with a larger representation of reactions would show a more coveted trend in predictability performance.

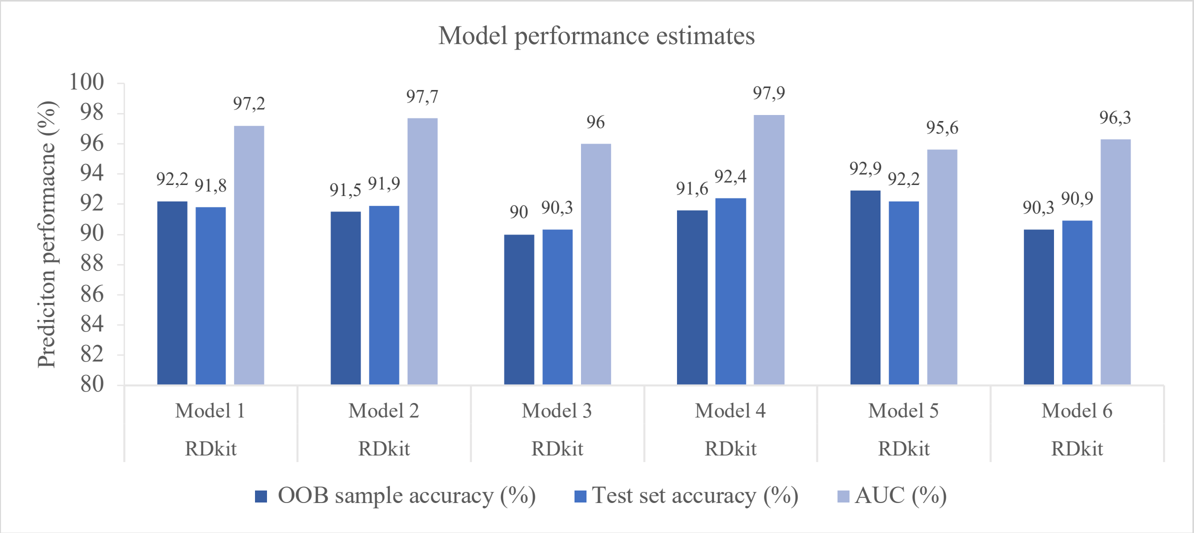


Figure 13: Estimated predictive performance from each model, OOB sample accuracy, test set accuracy and AUC measures from ROC-curve obtained from test set is shown, where each model is represented with RDKit descriptors.

AUC measures were obtained from the selected models of catalysts with calculated RDKit properties on the validation set (Table 4). Models with a smaller representation of data (Model 4 and 5) presented an AUC average of only 40.9% compared to an average of 77.8% (Model 1 and 2) a difference of 36.9%. This result is not unexpected and is and has been of general interest for further understanding in this project, as the volume of representation proves to be of a major impact.

Table 4: Obtained AUC measures after external validation, by implementation of validation set in model 1,2,4 respectively 5. Model 1 and 2 represents the AUC estimation with a larger representation of data, compared to Model 4 and 5 which contributes to reduced representation.

Model Entry	*Catalyst*	*AUC (%)*
Model 1	$Pd(PPh_3)_4$	75.1
Model 2	$Pd(dppf)Cl_2$	80.6
Model 4	$Pd(t\text{-}Bu3P)_2$	45.1
Model 5	$Pd(dtbpf)Cl_2$	63.6

There was of further interest to obtain a multiplied binary classification model to rank most preferable catalysts later implemented for practical use and experimental validation.

7.3 Milestone 3

A method to define a final model for ranking of the most preferable catalysts for unknown candidate reactions will be described, as well as the experimental validation of the obtained predictive performance.

7.3.1 Ranking candidate catalyst

Selected candidate reactions were implemented into each defined model to obtain predictive performance scores, meaning which catalyst would be best suited to obtain good reaction outcome, i.e. highest yield. Prediction scores are presented according to a desirable yield between 80-100% in Table 5.

Table 5: Final prediction scores for models 1 - 6, were highest percentages corresponding to the probability of obtaining high yield for respective candidate reactions A and B.

Model Entry	*Catalyst*	*Prediction scores (=>80% yield) for A*	*Prediction scores (=>80% yield) for B*
Model 1	$Pd(PPh_3)_4$	77.3%	71.8%
Model 2	$Pd(dppf)Cl_2$	28.0%	28.6%
Model 3	$Pd(PPh_3)_2Cl_2$	31.9%	46.1%
Model 4	$Pd(dtbpf)Cl_2$	34.7%	34.4%
Model 5	$Pd(t\text{-}Bu3P)_2$	48.8%	51.1%
Model 6	XPhosPdG2	23.6%	31.4%

The obtained predictions for A versus B were almost identical. Both reactions are similar in structure, with identical boron components. The catalyst showing most preferable prediction scores was $Pd(PPh_3)_2$, one of the most documented catalyst that has been proven to work well in Suzuki-Miyaura cross-coupling reactions. The catalyst obtaining lowest prediction scores was $Pd(dppf)Cl_2$ and XPhosPdG2.

7.3.2 Experimental analysis on candidate reactions

To confirm the hypothesis, each reaction was run in duplicates to obtain a minimum of reproducible results to validate the model performances. Candidate reaction A (Figure 14) was performed with Butanol/H_2O as solvent system together with potassium carbonate. This resulted in a conversion of only 67% product with $Pd(PPh_3)_4$ and 84% with the least preferable catalysts, $Pd(dppf)Cl_2$ and is in contradiction to the hypothesis. The presence of the competitive (fast) side reaction, beta-hydride elimination, during these types of reactions could be one explanation of the presented outcome. It seems to be considerably higher with the sufficiently more reactive catalyst, $Pd(PPh_3)_4$, compared to $Pd(dppf)Cl_2$ and occurs under highly polar conditions using n-Butanol as solvent. Another explanation could be the ML models giving inaccurate predictions. For further understanding different conditions for reaction A were performed using as solvent the less polar 1,4-dioxane. This, with a conversion of and 92% for catalyst 1, 93% for catalyst 2 and 94% for catalyst 3.

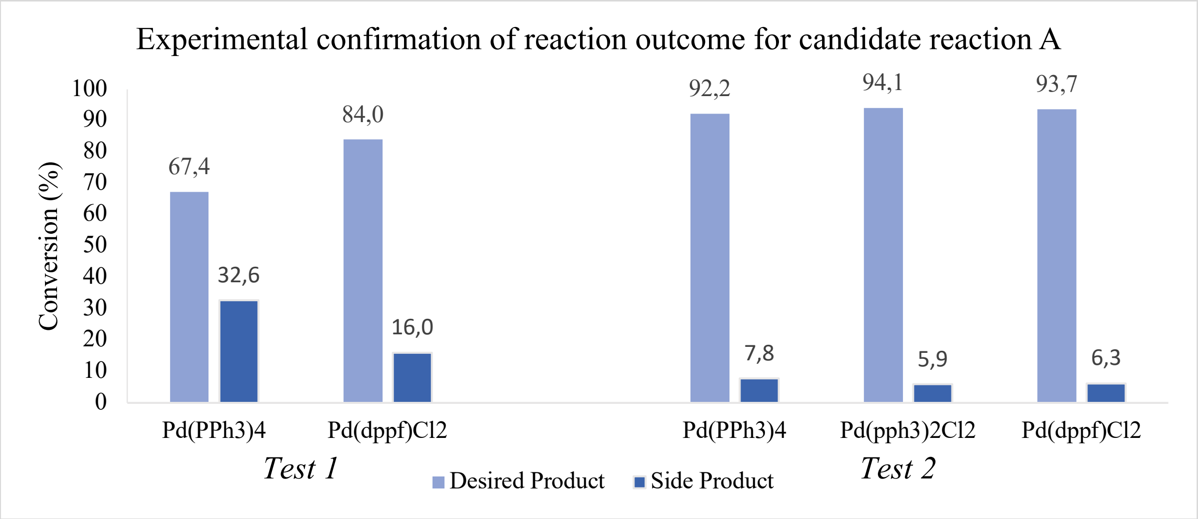


Figure 14: Obtained experimental outcomes for candidate reaction A. Test 1) was performed with n-Butanol and potassium carbonate, at 80°C for both Pd(PPh$_3$)$_4$ (ranked as number 1) and Pd(dppf)Cl$_2$ (ranked as number 5). Conversion of product was 67.4% respectively 84.0%. Test 2) was performed to obtain better results, which are more preferable for Pd(PPh$_3$)$_4$ and an improvement was shown with 1,4-dioxane as solvent for all of the three catalysts. A negative trend is still observed for Pd(PPh$_3$)$_4$.

The least optimal catalyst (Pd(dppf)Cl$_2$) was observed to perform remarkably well in both tests. This could indicate that the defined approach of ranking most preferable catalysts is questionable. The combination of solvent and base were selected to work in such a manner to perform well for different types of reactions. The importance of solvent in this case was due to the attendance of high polarity components, i.e. boronic acid. The combination of 1,4-dioxane as solvent and potassium carbonate as base, were selected to be a tremendous fit for this type of reaction and catalysts. Pd(dppf)Cl$_2$ was presented to be of a good fit with a probability of 28.0% to perform within the defined threshold for the desired outcome. One consideration was if the selected condition system for testing would have been the "optimal" in synthesis of product with Pd(dppf)Cl$_2$ as catalyst for candidate reaction A, thereby the impressive results.

The resulting conversions for each reaction outcome of reaction B, presented less impressive results (Figure 15) and a negative trend was observed. A similar trend was observed earlier in the first test for candidate A. The catalysts presenting highest probability of obtaining a conversion of product above 80.0% turned out to be least optimal ones, i.e. the opposite was observed versus what was predicted. Only 53.8% conversion was obtained with Pd(PPh$_3$)$_4$ and the least considerable catalysts to 89.3% for desired product. The selected conditions were therefore changed to achieve a better understanding of the outcome. Bromobenzonitrile showed to be more sensitive to the selected base, and to improve the reactivity during trans-metalation was Caesium fluoride selected. Generally, it was observed that all reactions improved, Pd(PPh$_3$)$_4$ went from 53.8% to 78.6% , Pd(PPh$_3$)$_2$Cl$_2$ from 82.7% to 94.7% and Pd(dppf)Cl$_2$ from 89.3% to 94.0%. All results deemed to disprove the defined hypothesis.

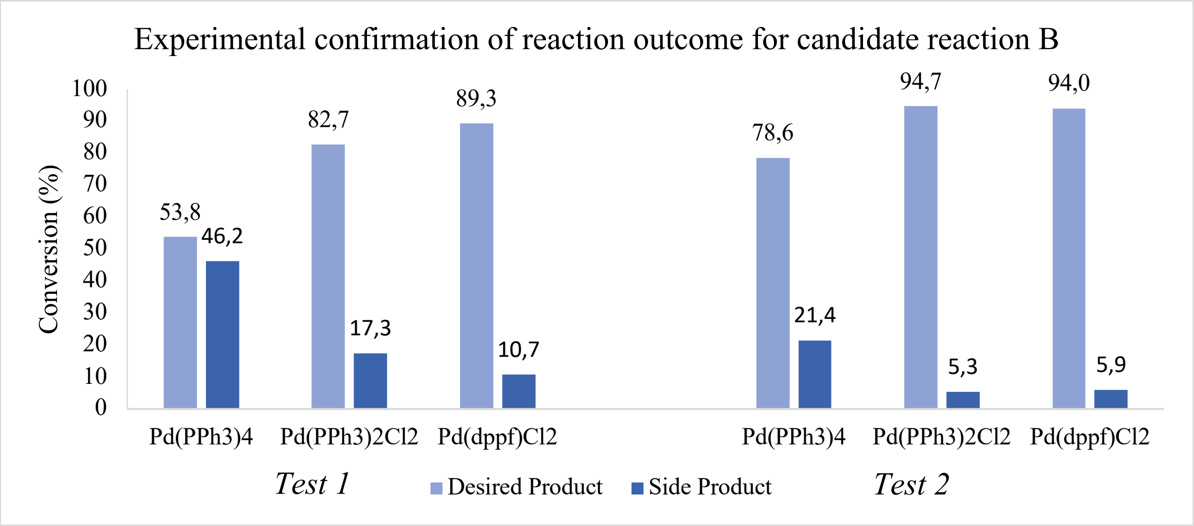


Figure 15: Obtained experimental outcomes for candidate reaction B. Test 1) was performed with 1,4-Dioxane and potassium carbonate, at 80°C for both Pd(PPh$_3$)$_4$ (ranked as number 1), Pd(PPh$_3$)$_2$Cl$_2$ (ranked as 3) and Pd(dppf)Cl$_2$ (ranked as number 6). Conversion of product were 53,8%, 82,7 and 83,9%. Test 2) was performed for obtaining better results, which are more preferable for Pd(PPh$_3$)$_4$ and an improvement was shown with caesium fluoride as base for all of three catalysts. A negative trend is still shown for Pd(PPh$_3$)$_4$.

The method strongly suggests that Pd(PPh$_3$)$_4$ would be the most considerable catalysts with a probability of 77.0% but shows inadequate results in laboratorial performance. The final ranking model would be considered to be revisited, since solvent, base and temperature was not part of the original prediction. These conditions have been proved to be of a significant impact in reaction outcomes and it may not be surprising to have obtained a bad fit of theory versus experimental results. The approach for validation can be questioned, if the scope of different solvent systems, bases and temperatures were selected to get an estimation of how well the model was ranking the different catalysts it could be improved. The obtained information from the experimental analysis was enough to conclude there are improvements needed to obtain a better model for prediction of catalysts.

8 Conclusion

The applications of ML have been of large interest since the big data era began and it continues to develop. The difficulties of finding drug candidates or optimizing time-consuming processes could be solved by use of machine learning and data mining. The aim of this study was to optimize a tedious time-consuming process by implementing a machine learning tool to rank the best catalysts for a specific reaction. The initial focus was the widely applied Suzuki-Miyaura reaction in drug development investigations.

The concluding remarks in this study refer to the impressive performance of the Random Forest algorithm in predicting the outcomes for a distinct type of reaction. The improved binary classification model was shown to perform at a low computational cost with an AUC of 97.0% when a Suzuki-Miyaura reaction would end up in a certain threshold, i.e. with ≥60% and ≤40%, yields. Why this particular task was of interest for this study was explained by questionable results from previous studies [4]. To avoid the presented complications (in predicting reaction outcomes), another approach towards the task was applied, incorporating the use of a "local" model.

The complications with "global" models include an insufficient coverage of steric, electronic effects, etc. The "global" models are shown to represent a large set of reactions (millions) with distinct groups of type reactions. These global models do not include more than a hundred reactions for a unique type and will not be enough to describe the intrinsic complexity of a reaction. To present a more comprehensive analysis, an approach to limit the scope to only one type reaction was used to predict possible reaction outcomes and was shown to increase the predictability with 32.0%.

The difficulty for a scientist in optimizing a Suzuki Miyaura reaction has been known to depend on the reactivity and selectivity of one catalyst. Furthermore, six diverse Random Forest-based models were defined to predict a catalyst as proof of concept towards the aim of this study and all models presented astonishing results with an average of 90.0% accuracy. With the experimental outcome, being opposite to the expected prediction, a new hypothesis for future approach will be required. The intrinsic complexity of designing a Random Forest model towards selecting specific conditions, was the lack of sufficient molecular representation from currently available descriptors. As concluded, the representation of a category has a significant impact on the performance of the model. The difference in the number of catalysts per model is caused by the size of reactions, which resulted in an insufficient coverage of a specific catalyst.

To summarise, an improved binary classification model was designed to predict two yield classes. The models for ranking best catalyst selections for a specific reaction showed promising correlations. However, the model requires further investigation, mainly regarding the catalyst representation but also other factors, including solvent systems, base, etc. A human factor in the experimental laboratory work cannot be ruled out, as may be a potential issue with some of the literature data. At first glance, the Random Forest algorithm was shown to have high and stable predictions. However, models with a minor representation of reactions have a major drawback in practical use, which was discovered through the experimental validation in this thesis.

9 Future work

In principal the correct future work is already present, an interest is to investigate the scope of development with the Random Forest model (in predicting yield classes) towards new opportunities. A study to obtain prediction results with other metal-catalysed cross-coupling reactions (e.g. Heck, Negishi etc.), would be interesting due to being a commonly used transformation in developing new medicinal candidates. It would also be interesting to investigate if it is possible to obtain a multiple binary classification model, based on a combined dataset with different cross-coupling reactions.

There is also an interest to explore the accuracy of the experimental validation, to address the doubtfulness from earlier performance. A new strategical laboratorial approach (factorial designs) will be of interest as validation procedure in future work.

Furthermore, to design an approach to identify a wider range of more commercially available catalysts as well as obtain narrower information on different bases, temperatures, and solvent combinations. Together with an investigation in how to create a new descriptor value for obtain a better representation of a catalyst, will be of priority. Ultimately, if any future performed experiment is started based on the results from predicting the outcome would be close to ideal.

10 Acknowledgements

Thank to my supervisors Fernando Huerta, Ulf Tedebark and Alexander Minidis at RISE Research Institutes of Sweden in Södertälje for giving me the opportunity to working on this project. For always being available for questions and for many supportive suggestions towards proceeding to the final goal of this project.

11 References

1. Aggarwal VK, Staubitz AC, Owen M (2006) Optimization of the Mizoroki-Heck reaction using Design of Experiment (DoE). Org Process Res Dev 10:64–69

2. Bullock KM, Mitchell MB, Toczko JF (2008) Optimization and scale-up of a suzuki#miyaura coupling reaction: Development of an efficient palladium removal technique. Org Process Res Dev 12:896–899

3. Fonseca MH, List B (2004) Combinatorial chemistry and high-throughput screening for the discovery of organocatalysts. Curr Opin Chem Biol 8:319–326

4. Skoraczyñski G, DIttwald P, Miasojedow B, Szymkuc S, Gajewska EP, Grzybowski BA, Gambin A (2017) Predicting the outcomes of organic reactions via machine learning: Are current descriptors sufficient? Sci Rep 7:1–9

5. Smith JS, Roitberg AE, Isayev O (2018) Transforming Computational Drug Discovery with Machine Learning and AI. ACS Med Chem Lett 9:1065–1069

6. Hochreiter S, Klambauer G, Rarey M (2018) Machine Learning in Drug Discovery. J Chem Inf Model 58:1723–1724

7. Segler MHS, Waller MP (2017) Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. Chem - A Eur J 23:5966–5971

8. Schneider N, Lowe DM, Sayle RA, Landrum GA (2015) Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. J Chem Inf Model 55:39–53

9. Derek T. Ahneman, Jesús G. Estrada, Shishi Lin SDD and AGD (2018) Predicting reaction performance in C−N cross-coupling using machine learning. 190:Coursera vs Udacity for Machine Learning

10. Culberson JC, Feuston BP, Svetnik V, Tong C, Liaw A, Sheridan RP (2003) Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J Chem Inf Comput Sci 43:1947–1958

11. Aires de Sousa J, Marcou G, Horvath D, Latino DARS, de Luca A, Rietsch V, Varnek A (2015) Expert System for Predicting Reaction Conditions: The Michael Reaction Case. J Chem Inf Model 55:239–250

12. Engkvist O, Norrby PO, Selmi N, Lam Y hong, Peng Z, Sherer EC, Amberg W, Erhard T, Smyth LA (2018) Computational prediction of chemical reactions: current status and outlook. Drug Discov Today 23:1203–1218

13. Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF (2017) Prediction of Organic Reaction Outcomes Using Machine Learning. ACS Cent Sci 3:434–443

14. Gao H, Struble TJ, Coley CW, Wang Y, Green WH, Jensen KF (2018) Using Machine Learning to Predict Suitable Conditions for Organic Reactions. ACS Cent Sci 4:1465–

1476

15. Genuer R, Poggi JM, Tuleau-Malot C, Villa-Vialaneix N (2017) Random Forests for Big Data. Big Data Res 9:28–46

16. Banerjee P, Siramshetty VB, Drwal MN, Preissner R (2016) Computational methods for prediction of in vitro effects of new chemical structures. J Cheminform 8:1–11

17. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Thiel K, Wiswedel B (2009) KNIME - the Konstanz information miner. ACM SIGKDD Explor Newsl 11:26

18. Murphy KP (2012) Machine learning: a probabilistic perspective. In: Chance Encount. Probab. Educ. The MIT Press, London, pp 1–1098

19. Mitchell T (1997) Machine Learning. McGraw-Hill Science/Engineering/Math; (March 1, 1997)

20. The Elements of Statistical Learning:Data Mining, Inference and P (2009) The Elements of Statistical Learning:Data Mining, Inference, and Prediction. Math Intell.

21. Breiman L (1999) Random Forests. 5–32

22. Andrew PB (1997) The use of the area under the {ROC} curve in the evaluation of machine learning algorithms. Pattern Recognit 30:1145–1159

23. Majnik M, Bosnić Z (2013) ROC analysis of classifiers in machine learning: A survey. Intell Data Anal 17:531–558

24. Reaxys. https://www.reaxys.com/#/login. Accessed 30 Mar 2019

25. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. J Chem Inf Comput Sci 43:493–500

26. RDKit. http://www.rdkit.org/. Accessed 30 Mar 2019

27. Sambasivarao K, Kakali L, Dhurke K (2002) Recent applications of the Suzuki–Miyaura cross-coupling reaction in organic synthesis. Tetrahedron 58:9633–9695

28. Maluenda I, Navarro O (2015) Recent developments in the Suzuki-Miyaura reaction: 2010-2014. Molecules 20:7528–7557

29. Probst P, Wright MN, Boulesteix AL (2019) Hyperparameters and tuning strategies for random forest. Wiley Interdiscip Rev Data Min Knowl Discov 1–19

30. Nguyen KT, Blum LC, Van Deursen R, Reymond JL (2009) Classification of organic molecules by molecular quantum numbers. ChemMedChem 4:1803–1805