FACULTY OF TECHNOLOGY

DEVELOPMENT OF PREDICTIVE MODELS FOR CATALYST DEVELOPMENT

Pekka Uusitalo

DEGREE PROGRAMME IN PROCESS ENGINEERING

Master's thesis

January 2021

ABSTRACT

Development of predictive models for catalyst development

Pekka Uusitalo

University of Oulu, Degree Programme in Process Engineering

Master's thesis 2020, 82 pp. + 1 appendix

Supervisors at the university: Aki Sorsa and Markku Ohenoja

This work was done as a part of the BioSPRINT project, which aims to improve biorefinery operations through process intensification and to replace fossil-based polymers with new bio-based products. The goal was to identify machine learned (ML) models that will accelerate the catalyst identification with high-throughput (HTP) screening methods, identify non-obvious formulations and allow catalyst tuning for different feedstock compositions. Maximum activity for conversion of complex sugar mixtures with optimal selectivity towards the key products of interest is desired.

In the literature part of the thesis, ML was studied in general, where the focus was on different variable selection methods and modeling techniques, more specifically on data-driven modeling. Furthermore, modeling in catalysis was discussed with focus on ML in catalysis. Catalyst screening and selection, descriptor modeling and selection, and predictive modeling in catalysis were studied.

In the experimental part, focus was on developing ML models that predict catalyst performance with relevant descriptors. Dataset for hydrogenation of 5-ethoxymethylfurfural with simple bimetal catalysts, including main metals and promoters, was used as ML model input with the addition of catalyst descriptors found in the literature. Four different responses were used in the experiments: selectivity and conversion with two different solvents. Methods used in the experimental part were discussed in detail, where data collection, preprocessing, variable selection, modeling and model validation were considered. Reference models without variable selection were first identified. Secondly, regularization algorithms were used to identify models. Finally,

models with variable subsets obtained with regularization algorithms were identified. The effect of cross-validation was also studied.

In general, good modeling results were obtained with boosted ensemble tree methods, support vector machine (SVM) methods and Gaussian process regression (GPR) methods. Lasso regression turned out to be the best variable selection method. Good results were obtained with the descriptors found in the literature. It was also shown, that fairly good results can be obtained with only two variables in the studied case. Promoter variables were not considered nearly as important as main metals with variable selection algorithms. Even though the modeling results were good, the variable selection methods were almost purely data-driven, and the actual relevance of the variables cannot be guaranteed.

In the future work, optimization should be studied with the goal of finding catalysts that maximize catalyst performance values based on the model predictions. Also, extrapolation capabilities of the models need to be studied and improved. The studied methods can be easily implemented to other datasets. In the BioSPRINT project, experimental results related to the dehydration reaction of C5 and C6 sugars with simple metal catalysts will be obtained and used with the studied methods.

Keywords: Machine Learning, Heterogeneous Catalysis, Predictive Modeling, Variable Selection

TIIVISTELMÄ

Ennustavien mallien laatiminen katalyytin valmistuksen tehostamiseksi

Pekka Uusitalo

Oulun yliopisto, Prosessitekniikan tutkinto-ohjelma

Diplomityö 2020, 82 s. + 1 liite

Työn ohjaajat yliopistolla: Aki Sorsa ja Markku Ohenoja

Tämä työ tehtiin osana BioSPRINT-projektia, jonka tavoitteena on kehittää biojalostamoiden toimintaa parantamalla niiden prosessitehokkuutta ja korvata fossiilipohjaiset polymeerit uusilla biopohjaisilla tuotteilla. Työn tavoitteena oli muodostaa koneoppimista hyödyntämällä mallit, jotka nopeuttavat optimaalisten katalyyttien löytämistä tehoseulonnan (high-throughput (HTP) screening) avulla, auttavat identifioimaan vaikeasti löydettäviä katalyyttiyhdistelmiä ja mahdollistavat katalyytin valinnan eri lähtöainekoostumuksilla. Tavoitteena on maksimoida monimutkaisten sokeriyhdisteiden konversio ja selektiivisyys halutuiksi tuotteiksi.

Työn kirjallisuusosiossa perehdyttiin koneoppimiseen yleisellä tasolla, missä pääpaino oli muuttujanvalintamenetelmissä ja datapohjaisissa mallinnusmenetelmissä. Lisäksi kirjallisuusosassa tutkittiin mallinnuksen käyttöä katalyysissä, missä pääpaino oli koneoppimisen käytössä. Työssä tarkasteltiin myös katalyyttien seulontaa ja valintaa, laskennallisten muuttujien (deskriptorien) määrittelyä ja valintaa, sekä ennustavan mallinnuksen käyttöä katalyysissä.

Kokeellisessa osiossa painopiste oli koneoppimista hyödyntävien mallien muodostuksessa, jotka ennustavat katalyyttien suorituskykyä oleellisilla deskriptoreilla. Data-aineistona käytettiin 5-etoksimetyylifurfuraalin hydrausreaktion tuloksia yksinkertaisilla kaksikomponenttisilla metallikatalyyteillä, jotka sisältävät päämetallin ja promoottorin. Data-aineistoa täydennettiin kirjallisuudesta löytyvillä katalyyttien deskriptoreilla ja käytettiin koneoppimista hyödyntävien mallien sisääntulona. Tutkimuksissa käytettiin neljää eri vastemuuttujaa: selektiivisyyttä ja konversiota kahdella eri liuottimella. Kokeellisessa osiossa käytetyt menetelmät käytiin läpi

perusteellisesti huomioon ottaen data-aineiston keräämisen, esikäsittelyn, muuttujanvalinnan, mallinnuksen ja mallin validoinnin. Ensin referenssimallit identifioitiin. Tämän jälkeen regularisaatioalgoritmeilla suoritettiin mallinnus. Lopuksi mallinnus suoritettiin käyttämällä muuttujajoukkoja, jotka oli valittu käyttäen regularisaatioalgoritmeja. Myös ristivalidoinnin vaikutusta tutkittiin.

Yleisesti hyvät mallinnustulokset saavutettiin boosted ensemble tree -tekniikalla, tukivektorikoneella ja Gaussian process -regressiolla. Lasso-menetelmä todettiin parhaaksi muuttujanvalinta-algoritmiksi. Hyvät tulokset saavutettiin kirjallisuudesta löytyvien deskriptorien avulla. Tutkimuksissa todettiin myös, että hyvät mallinnustulokset voidaan saavuttaa kyseisessä tutkimustapauksessa jopa vain kahdella muuttujalla. Päämetalleja kuvaavien muuttujien merkitsevyys todettiin paljon suuremmaksi kuin promoottorien vastaavien muuttujien. Saatavia mallinnustuloksia tarkasteltaessa täytyy huomioida, että muuttujanvalinta oli melkein täysin datapohjainen eikä muuttujien varsinaista merkitsevyyttä voida taata.

Jatkossa mallien ennustuksia voidaan hyödyntää optimoinnissa, jossa tavoitteena on etsiä katalyyttiyhdistelmä, joka maksimoi katalyyttien suorituskyvyn. Myös mallin ekstrapolointikykyä täytyy tutkia ja kehittää. Tutkittavat menetelmät ovat helposti sovellettavissa myös muille samantyylisille data-aineistoille. BioSPRINT-projektista saadaan tulevaisuudessa käytettäväksi viisi- ja kuusihiilisten sokerien dehydraatioon perustuva data-aineisto yksinkertaisilla metallikatalyyteillä, jota tullaan käyttämään jatkotutkimuksissa.

Asiasanat: koneoppiminen, heterogeeninen katalyysi, ennustava mallintaminen, muuttujanvalinta

FOREWORD

This thesis is made for BioSPRINT project[1] and as a part of master's degree programme in process engineering in University of Oulu. The objective was to implement predictive modeling methods for catalyst development for dehydration reaction of C5 and C6 sugars.

I want to thank postdoctoral researchers Aki Sorsa and Markku Ohenoja from University of Oulu for supervising the thesis work. I would also like to thank Fernando Russo Abegao from Newcastle University for the guidance and comments on the work.

Oulu, 7.1.2021

Pekka Uusitalo
Pekka Uusitalo

[1] This project has received funding from the Bio Based Industries Joint Undertaking (JU) under grant agreement No 887226. The JU receives support from the European Union's Horizon 2020 research and innovation programme and the Bio Based Industries Consortium.

TABLE OF CONTENTS

LIST OF ABBREVATIONS

1SE	One standard error
5-HMF	5-hydroxymethylfurfural
C5	Five carbon
C6	Six carbon
CV	Cross-validation
DFT	Density Functional Theory
DMSO	Dimethyl sulfoxide
FOM	Figures of Merit
FUR	Furfural
GVL	γ-Valerolactone
HTP	High-throughput
M	Main metal
MAE	Mean absolute error
minMSE	Minimum mean squared error value
ML	Machine Learning
MLP	Machine Learned Interatomic Potentials
MSE	Mean squared error
P	Promoter
PCA	Principal component analysis
PLSR	Partial least squares regression
QM	Quantum Mechanical
RDF	Radial Distribution Function
RMSE	Root mean squared error
RMSLE	Root mean squared logarithmic error
STO	Slater-type orbital
SVD	Singular value decomposition

1 INTRODUCTION

In catalysis, chemical reactions are modified with the use of catalysts. The reaction is accelerated or intensified for commercial purposes. Catalytic reaction is faster than uncatalyzed one, because catalysts decrease the minimal energy needed to start a chemical reaction (activation energies). (Dev et al. 2018) Catalysts can be divided into heterogeneous and homogeneous catalysts. Heterogenous catalysts include elements that are not in the same phase with reactants. In contrast, homogeneous catalysts are in the same phase with reactants, typically in liquid or gas phase. (Luo et al. 2019, p. 16) This topic will be discussed with more detail in the chapter 4.3.

This work is done as a part of the BioSPRINT project, which aims to improve biorefinery operations through process intensification and to replace fossil-based polymers with new bio-based products. A more specific aim in the project is to improve the efficiency of purification processes and conversion of sugars from the hemicelluloses fraction of lignocellulosic biomass. (BioSPRINT 2019) Improvement for conversion is obtained through the design of more efficient catalysts. For the catalyst development, the project's focus is on heterogeneous catalysis and development of super solid acid catalysts with time-efficient high-throughput (HTP) screening methods. The process under study is dehydration reaction of multiple five and six carbon (C5 and C6 respectively) sugars to 5-hydroxymethylfurfural (5-HMF) and furfural (FUR). Maximum activity for conversion of complex sugar mixtures with optimal selectivity towards the key products of interest is desired. (BioSPRINT 2019) This thesis supports the catalyst development in the conversion step by speeding-up the catalyst design in hand with the HTP experiments.

For the studied reaction, hemicelluloses fraction of lignocellulosic biomass is under interest. Lignocellulosic biomass is the most abundant renewable raw material and has great potential as a source for biofuels and biochemicals. It is composed of three main elements: cellulose, hemicellulose, and lignin. (Nebreda 2019, p. 8) Lignocellulosic biomass can be divided into woody and non-woody. Woody lignocellulosic biomass includes hardwoods and softwoods. Non-woody biomass includes agricultural residues, plants, algae and non-wood plant fiber. (Nebreda 2019, p. 24) Lignocellulosic materials can consist hemicellulose (HMC) up to 20-40 % of the total dry weight. This fraction is often not used in industry. (Nebreda 2019, p. 8) Hemicellulose consists of polymers of

hexoses, mainly glucose, mannose, and galactose, and of pentoses, mainly xylose and arabinose, and also sugar acids (Nebreda 2019, p. 28; Zhang 2013, p. 53). It can be degraded into these monomeric components and further converted into biofuels and chemicals (Nebreda 2019, p. 8). HMC does not have crystalline structure like cellulose, which makes it easier to be hydrolyzed by acids, bases, and enzymes (Rinaldi and Schüth 2009, p. 614). This makes polymer conversion to monomers easier.

Figures of merit (FOMs) are needed to estimate and predict catalysts' performance. In catalyst development, the goal is to link catalyst descriptors to FOMs. FOM is a quantitative index describing catalyst's usefulness. FOMs can be, for example, product selectivity, product yield, turnover number, turnover frequency and cost per kg. FOM can also be a combination of several properties. (Maldonado & Rothenberg 2010, p. 1892)

This thesis focuses on developing machine learned (ML) models that predict catalyst performance with relevant descriptors. The goal is to identify models that will accelerate the catalyst identification with HTP screening methods, identify non-obvious formulations and allow catalyst tuning for different feedstock compositions (BioSPRINT 2019). Catalyst library with catalysts' properties and molecular descriptors is formulated and used as ML model inputs. The information in the library is obtained from literature and computations. Performance results from literature and experiments are also used as ML model inputs. Different variable selection methods and modeling techniques are tested with simple bimetal catalysts.

2 MACHINE LEARNING

Machine Learning (ML) methods use data to learn underlying rules of the system studied and to predict system outcomes (Butler et al. 2018, p. 547; Kitchin 2018, p. 230). ML studies focus on developing algorithms that learn from given data without the need of explicit programming. It has been mainly used in computer science and statistical science, but nowadays it is used in a much wider range of applications. (Toyao et al. 2020, p. 2262) ML methods have potential to be easily accessible for molecular and materials modeling without the need of high computational power and specific prior knowledge (Butler et al. 2018, p. 548).

2.1 Data-Driven Modeling

In data-driven modeling, a model that describes the interactions between given inputs and outputs is formed. Specific knowledge about underlying phenomenon or physical behavior of the system is not considered in data-driven modeling. Theory-driven models can be limited by several factors, which can be handled with data-driven modeling. These factors are high computational costs, scalability issues, unclear targets, conditions and parameters, complexity and uncertainty of examined system and incomplete theories of underlying phenomenon. (Toyao et al. 2020, p. 2264)

There are several issues that needs to be considered when using data-driven modeling: First, it is crucial to understand that if the used data is not valid, also the identified model is not valid. Secondly, it can be hard to figure out if the used data and the identified model are valid or not. Thirdly, there can be multiple good model fits to the used data, but no actual fundamental relevance. Therefore, it is important to analyze the used data and its quality carefully. (Toyao et al. 2020, p. 2263-2264)

2.2 Data collection and preprocessing

Wide range of data types can be used with ML. Any data types should be used if they describe the target properties well (Toyao et al. 2020, p. 2264). The way input data is represented affects the learning algorithm’s capability to map the input data to the corresponding output data. Choosing the best representation may not be obvious and may

need specific knowledge from the underlying principles and use of the learning algorithm. (Butler et al. 2018, p. 548)

The quality and quantity of the data are crucial to achieve a good performance with machine learned predictive model. As mentioned in Section 2.1, if the quality of used data is poor, also the quality of identified model will be poor. In addition, increased quantity of data in ML modeling can increase the confidence and reliability of model predictions from a statistical point of view. (Toyao et al. 2020, p. 2264) Usually ML methods require large amount of data to achieve proper outcome. In materials science, the amount of data available is often limited, which makes the modeling challenging. (Butler et al. 2018, p. 553)

Exploratory data analysis should be always done for dataset because it may contain errors, outliers, and noise. If so, dataset requires preprocessing, where missing or incorrect values are handled accordingly. (Butler et al. 2018, p. 547; Toyao et al. 2020, p. 2266) Preprocessing is more often required with large datasets. Normalization and rescaling may be also required for input data and in some cases also for output data. (Toyao et al. 2020, p. 2266)

2.3 Feature Engineering

As mentioned in the previous section, ML algorithm's learning efficiency depends on the used data format. Raw data may be converted into more efficient format via feature engineering. (Butler et al. 2018, p. 548; Guyon & Elisseeff 2003, p. 1170) Feature engineering is often required to obtain optimal input features for the ML algorithm. Furthermore, the acquired ML outputs may be used as new inputs, which are also called as meta features. (Toyao et al. 2020, p. 2266) Feature engineering can have two goals: constructing the best form of the original data or constructing the most efficient features for making predictions. First problem uses unsupervised learning, and second problem is supervised. (Guyon & Elisseeff 2003, p. 1171) In some cases, automated feature engineering can improve performance and make the set of features more compact (Guyon & Elisseeff 2003, p.1179).

There are numerous feature engineering methods including clustering, basic linear transforming of the input variables (e.g. PCA and SVD), more advanced linear transforms

(e.g. Fourier), wavelet transforms and convolutions of kernels, and using simple functions to modify a subset of variables (e.g. products of variables). (Guyon & Elisseeff 2003, p. 1170) Clustering is a popular unsupervised feature engineering method. The idea of clustering is to replace similar variables with a cluster centroid, which will become a feature. Popular algorithms in clustering include K-means and hierarchical clustering. (Guyon & Elisseeff 2003, p. 1171) Singular Value Decomposition (SVD) is another widely used unsupervised feature engineering method. Idea is to form a set of features as linear combinations of the original variables. The best reconstruction of the original data is to be found by optimizing the least squares objective. (Duda et al., 2001; according to Guyon & Elisseeff 2003, p. 1172) Principal component analysis (PCA) is a common model-based evaluation method, that resembles SVD. PCA finds principal components (also called latent variables), which are linear combinations of the original variables without correlations to each other. The objective is to reduce dimensionality while describing as much variance as possible in the original data. (Ras et al. 2014, p. 5966)

2.4 Variable Selection

Variable selection aims to improve prediction accuracy, provide faster and more efficient predictors or models, and provide better understanding of the studied process (Guyon & Elisseeff 2003, p. 1157). Variable and feature selection may also help in data visualization and improve interpretability, reduce measurement and storage requirements, reduce training time, and improve prediction accuracy through dimensionality reduction (Guyon & Elisseeff 2003, p. 1158). Reducing the dimensionality of data by selecting subset from original variables can reduce the expenses of measurement making, storing, and processing (Guyon & Elisseeff 2003, p. 1170).

In variable selection, determining the number of significant variables, choosing the correct hyperparameters, and evaluating the final performance of the system is important. To do this properly, the model performance is needed to be evaluated with out-of-sample data (i.e. data, that has not been used in the training). (Guyon & Elisseeff 2003, p. 1172)

Some variables or features may have individually low importance, but high importance in combination with other variables (Guyon & Elisseeff 2003, p. 1165). Also, by adding presumably redundant variables to a variable subset, better class separation and noise reduction may be achieved (Guyon & Elisseeff 2003, p. 1163). Highly correlated

variables can complement each other. Although, adding perfectly correlated variables does not give any additional information to predictions and are thus redundant. (Guyon & Elisseeff 2003, p. 1164)

Variable selection without response variable (i.e. unsupervised learning) can be done to find the most important variables with respect to certain criterion including saliency, entropy, smoothness, density, and reliability. Salient variable has observations with high variance or large range compared to the other variables. If the distribution of observations is uniform, the variable has high entropy. In a time-dependent situation, variable is smooth if its local curvature is moderate on average. If a variable is highly correlated with many variables, it is in high density area. If the measurement error from repeated measurements is small compared to the variability of the variable values, variable is considered reliable. (Guyon & Elisseeff 2003, p. 1175)

2.4.1 Filter Method

Variable subset selection can be divided into wrappers, filters, and embedded methods (Guyon & Elisseeff 2003, p. 1166). In filter selection, variables are ranked based on some criterion which usually is the correlation coefficient (Guyon & Elisseeff 2003, p. 1158). A subset of variables is selected independently of the chosen predictor as a pre-processing step (Guyon & Elisseeff 2003, p. 1166). Generally, filters are faster than wrapper methods in subset selection. Filtering can also reduce space dimensionality and prevent overfitting. Wrapper or embedded method can be used with a linear predictor as a filter and with the selected variables train a more complex non-linear predictor. Products of input variables can be added to the initial variable set to make linear filters more complex. Although, one may want to reduce complexity of linear filter to reduce the chance of overfitting. (Guyon & Elisseeff 2003, p. 1170)

Variable ranking is a filter method, where preprocessing is made independently from the choose of the predictor. It is often included in variable selection algorithms, because of its simplicity, scalability, and great success rate. A scoring function, where high score means better variable, can be used as ranking criterion to choose the best variable subset. (Guyon & Elisseeff 2003, p. 1159-1160) In variable ranking or nested subset ranking methods a random variable can be involved in data, which is compared with other variables. If a variable has less or equal relevance than the random variable, it will be

discarded. (Guyon & Elisseeff 2003, p. 1173) A nested variable set contains layers of variable subsets by forming a hierarchical structure (i.e. the inner subsets are included in the outer subsets). For example, gradual addition of variables into the model based on the ranking leads to nested subsets.

2.4.2 Wrapper Method

Wrapper methods select subset of variables based on their usefulness to a given predictor (Guyon & Elisseeff 2003, p. 1158). Learning machine of interest is used with different subsets of variables to score them according to their predictive power. Wrapper method is a simple and powerful way for variable selection regardless of the used learning machine. Wrappers are universal and simple when the learning machine is considered as a black box. (Guyon & Elisseeff 2003, p. 1166-1167)

Three factors should be considered, when using wrapper methods: First, a way to search all possible variable subsets should be found. Secondly, one needs to consider how the prediction performance of learning machine can guide the search of optimal variable subset. Thirdly, the used predictor needs to be selected. (Guyon & Elisseeff 2003, p. 1166) Popular predictors include decision trees, Naïve Bayes, least-square linear predictors, and support vector machines (Guyon & Elisseeff 2003, p. 1167).

Wrapper methods can be computationally demanding. This can be overcome with proper search strategies without necessarily needing to sacrifice prediction power. Various methods to search variable subsets can be used, including Genetic Algorithms, which are covered in more details in Section 2.5.6 of this thesis. Coarse search strategies may prevent the overfitting. Greedy strategies are computationally easier and are also good against overfitting. These greedy search strategies never question old decisions when including or excluding variables in the new variable selection decisions. Greedy search strategies can be divided into forward selection and backward elimination. In forward selection, the selection is started with zero variables in the subset and variables are added one by one according to their relevance. In backward elimination, the situation is the opposite. All variables are first included, and variables are eliminated one by one according to their redundancy. Both methods result in nested subsets of variables. (Guyon & Elisseeff 2003, p. 1167) In general, forward selection is computationally more efficient than backward elimination in finding nested subsets of variables. However, forward

selection can find weaker subsets, because the variable importance is estimated without the presence of other variables that are not yet selected. If a variable that is best on its own is wanted to be found, forward selection is more likely to work better, because backward elimination can eliminate that variable early on, when many variables are still included. (Guyon & Elisseeff 2003, p. 1175-1176)

2.4.3 Embedded Methods

Embedded methods resemble wrapper methods, but in addition to optimizing goodness-of-fit-term they also penalize large number of variables (Guyon & Elisseeff 2003, p. 1158). Variable selection is performed during the training phase and are usually specific to the used learning machine (Guyon & Elisseeff 2003, p. 1166). In comparison to wrapper methods, embedded methods can be more efficient: They can make better use of the available data, because splitting the data into training and validation sets are not necessarily needed. Also, solution can be found faster (i.e. algorithm is computationally more efficient) since predictor is not needed to be retrained from scratch again for every variable subset. (Guyon & Elisseeff 2003, p. 1167)

Some embedded methods can guide the search of variable subsets by estimating changes in the objective function value caused by changes in variable subset space. By combining these methods with backward elimination or forward selection, nested subset of variables is formed. (Guyon & Elisseeff 2003, p. 1167) Some learning machines can extract features during the learning process. For example, neural networks extract these features with their internal nodes. (Guyon & Elisseeff 2003, p. 1172) In direct objective optimization, objective function of variable selection is defined, and algorithms are searched to optimize it. Usually, the objective function consists of two terms: the goodness-of-fit to be maximized, and the number variables to be minimized. This method is similar to two-part objective function with regularization term, which shrinks the parameter values by penalizing their great values. (Guyon & Elisseeff 2003, p. 1169)

Advanced wrapper or embedded methods improve the prediction accuracy compared to simple variable ranking methods (e.g. correlation methods), but the improvement is not necessarily significant. With large number of variables, dimensionality can be a problem and with multivariate methods overfitting can happen. (Guyon & Elisseeff 2003, p. 1179)

The overfitting is more likely to happen, when the number of variables is large compared to the number of observations (Guyon & Elisseeff 2003, p. 1165).

Regularization algorithms can be interpreted as embedded methods. In this work, lasso, ridge, and elastic net methods will be studied in more detail. Generally, ridge regression shrinks the coefficients towards each other (Friedman 2010, p. 3). It is best used, when large amounts of predictors with non-zero coefficients and with normal distributions are present (Hoerl & Kennard 1970; Friedman 2010; according to Ogutu et al. 2012, p. 2). In case of correlated predictors, their coefficients will shrink equally towards zero (Friedman 2010, p. 3). Ridge regression will not cut predictors completely out from the model like lasso regression does (Ogutu et al. 2012, p. 2-3).

Lasso regression tends to create some coefficients that are equal to zero, which makes the model more interpretable. In addition to interpretability, lasso has generally good stability like ridge regression. (Tibshirani 1996, p. 267) Lasso regression is usually used with large datasets and when efficient and fast algorithms are needed. It does not fit well for situation with highly correlated predictors since it will choose one of the correlated predictors randomly and ignore the rest or with identical predictors algorithm breaks. (Friedman 2010; according to Ogutu et al. 2012, p. 3) Lasso assumes most of the coefficients to be near zero, and only few coefficients to be larger (Ogutu et al. 2012, p. 3). Elastic net combines ridge and lasso regression by combining their penalty terms. The balance between the ridge and lasso penalty terms can be adjusted with alpha value between 0 and 1. Value 0 corresponds to ridge regression and value 1 to lasso regression. (Friedman 2010, p. 3)

2.5 Model identification

ML training can be supervised, semi-supervised or unsupervised, which depends on the type and quantity of available data. Supervised learning is the most used and studied approach in ML in physical science. In supervised learning, both input and output data are given to identify a model, that predicts new outputs with new inputs. Relationships between inputs and outputs can be learned without specific knowledge about the actual system or principles. (Toyao et al. 2020, p. 2263-2264; Butler et al. 2018, p. 548) Supervised learning can be further divided in classification and regression. In classification, discrete (e.g. categories) output values are predicted. In regression,

continuous output values are predicted. In unsupervised modeling, only input data are given, which can be used to identify underlying trends, patterns, or in clustering the data. Semi-supervised learning can be used when there are large amounts of input data with small amounts of output data. (Butler et al. 2018, p. 548)

Normally, when starting modeling, simple models like constant and linear predictors should be used. From there, the complexity of the model can be increased. (Toyao et al. 2020, p. 2266-2267) ML models can often lack interpretability, which can be caused by several aspects: ML algorithms rarely map the interactions in a way that scientists are familiar with. There can also be underlying principles that are not yet known by scientists or they are too complicated to be understood. (Butler et al. 2018, p. 553) Acquiring information about each variables contribution to the model predictions can be useful. There are several methods for this purpose including sensitivity analysis methods. (Toyao et al. 2020, p. 2267)

2.5.1 Linear regression

Linear regression model has one regressor x that is related to response y. In a simple case y is a straight line. Linear regression model can be described by the following Eq. (1),

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{1}$$

where β_0 is the intercept, β_1 is the slope and ε is the error term. β_0 and β_1 are unknown constants (can also be called regression coefficients). (Montgomery et al. 2012, p. 12-13) Usually least squares method, where sum of the squares of the differences between observed and predicted responses is minimized, is used as the objective function to find the model parameters β_0 and β_1 (Montgomery et al. 2012, p. 12-13; Toyao et al. 2020, p. 2265).

2.5.2 Multiple linear regression

In contrast to linear regression model, multiple linear regression (MLR) has two or more regressors related to response y. In Eq. (2), multiple linear regression model with k regressors is given.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon. \quad (2)$$

The model describes a hyperplane in a k-dimensional space. Even though the model is linear, the shape of the surface can be non-linear. Parameters can be identified with the same principle as in linear regression (e.g. least squares). (Montgomery et al. 2012, p. 68-70)

2.5.3 Nonlinear Regression

In nonlinear regression, the coefficients' influence is not linear in contrast to the model output and the coefficients are also not independent. Also, often constrains are needed for variables (e.g. delays must be greater than zero) in nonlinear regression. (Rhinehart 2016, p. 7-8)

Nonlinear regression models can be described by Eq. (3)

$$y = f(\theta, x) + \varepsilon, \ \varepsilon \sim \mathrm{N}(0, \sigma^2), \quad (3)$$

where y is the response, x the regressor (can be multivariate), θ the vector of model parameters describing the relations between x and y through function f, and ε is the residual error term, which is assumed to be normally distributed with mean value of 0 and variance σ^2. (Bates & Watts 1988; according to Baty et al. 2015, p. 4)

2.5.4 Partial Least Squares Regression

When the number of predictors is large compared to observations and collinearity is high, MLR method usually performs poorly. Partial least squares regression (PLSR) is a better option in these situations. There may be only few latent factors in the data that describe most of the variance in response. PLSR searches for these latent variables. After extracting these latent factors, regression step is performed to predict response. (Randall 2016, p. 1-2; Abdi 2003, p. 1-2) Converting correlated features into independent features will cause information loss, which needs to be considered when using PLSR and similar methods (Toyao et al. 2020, p. 2267).

2.5.5 Artificial Neural Networks

Artificial neural networks mimic the operation of human brain (Haykin 2009, p. 1). Although they are not used in this work, it is important to present their basic principles together with the other ML approaches. An example of a multilayer perceptron neural network can be seen in Figure 1. It consists of input layers, hidden layers, and output layers, which all consist of neurons, which operates as processing units. Each neuron is connected to other neurons in previous and/or next layer (depending on the layer position). These connections are weighted with different weight values that can be negative or positive, which can be seen in Figure 2, which represents a model of neuron. In the case of Figure 1, the signals are moving in a feedforward manner. In feedforward networks, all the signals from previous layer neurons move to each neuron in the next layer. These signals are weighted to obtain the neuron inputs as shown in Figure 2. Bias can also be used as the neuron input. The weighted signals and the bias are then summed, and the sum is then used as an input to the activation function (e.g. sigmoid-function or rectified linear unit function). The calculations from the activation function give the activation value of neuron, which usually varies between 0 and 1. This value is then passed to the next layer neurons as a weighted value. (Haykin 2009, p. 11-12, 124-125)

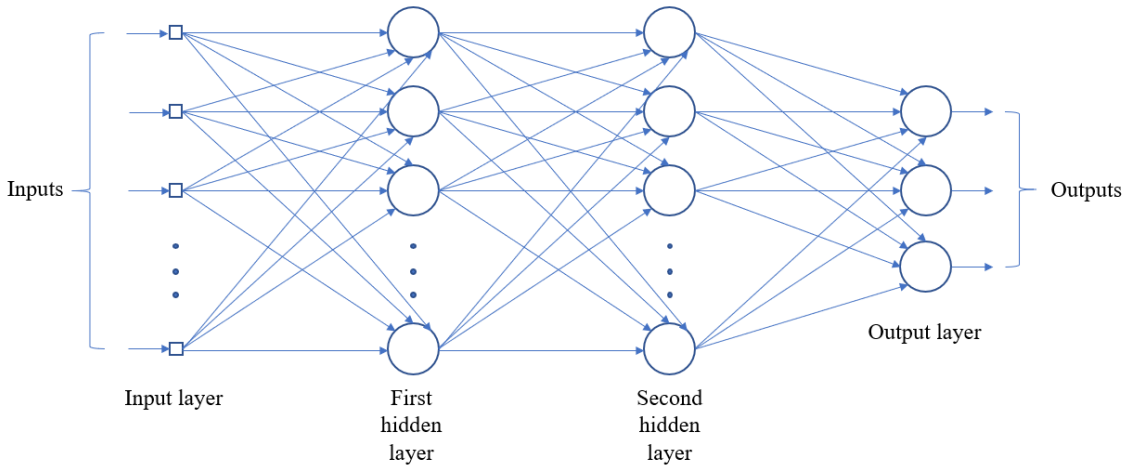

Figure 1. An example of multilayer perceptron neural network with two hidden layers (retell Haykin 2009, p. 124).

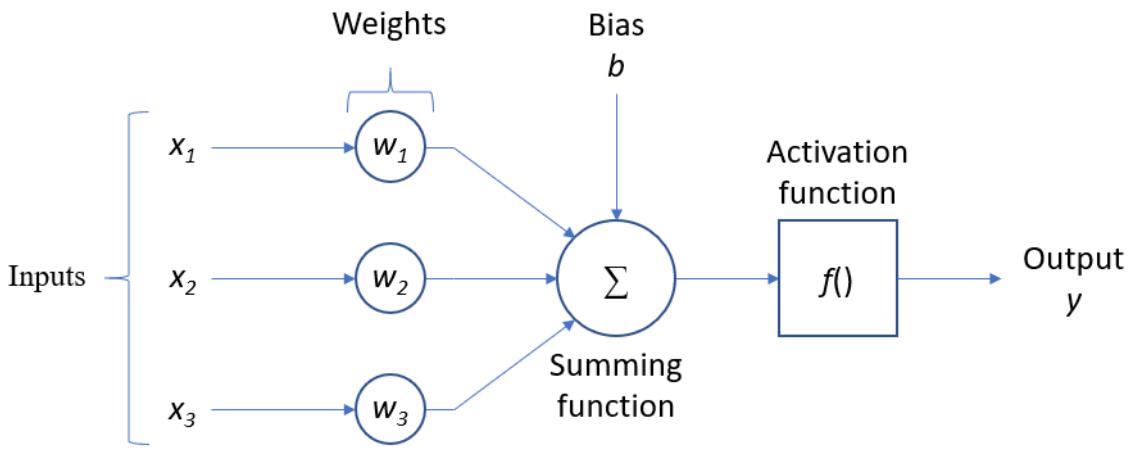

Figure 2. Example model of a neuron with three input signals (retell Haykin 2009, p. 11).

Mean squared error (MSE) can be used as cost function for learning, where mean of sum of squared differences between predicted and measured values are calculated. Then some optimization method can be used to find local or global optimum that minimizes the cost function. A commonly used optimization method is gradient descent, which aims to move towards local minimum by using negative value of gradient and moving towards that direction with predefined step sizes. This process is done iteratively until minimum is found. (Haykin 2009, p. 35-36, 95-96)

A popular method for training a multilayer perceptron neural network is the back-propagation method. First, initial values of the network are calculated. Then, the predicted output values are compared with the target values to gain the error values. By using stochastic gradient descent optimization method, the corrections in weights and biases are calculated. The weights and biases are adjusted accordingly by moving backwards in the network starting from the output layer and moving towards the input layer. The backpropagation learning process can be done in an iterative manner until the performance criterion are met. (Haykin 2009, p. 123-124, 129-141)

2.5.6 Evolutionary Algorithms

Evolutionary algorithms (EAs) mimic the biological evolution (Goldberg 1989; Sarker et al. 2003; according to Sorsa et al. 2010, p. 1, 6). They can be used in modeling, optimization and variable selection. They suit well for large and complex optimization problems (Sarker et al. 2003; according to Sorsa et al. 2010, p. 1). Maybe the most important use of EAs with ML is optimization, but they can also be used in variable selection and modeling. This section covers only genetic algorithms (GAs), which are

one of the most used methods of EAs. Other widely applied EA methods are swarm, differential evolution and genetic programming methods, just to mention a few. Similarly to neural networks, GA or other EAs were not used in the experimental part of this thesis.

Genetic algorithms (GAs) are well-known EAs that suit for various types of optimization problems (Goldberg 1989; according to Sorsa et al. 2010, p. 6). The GA creates new populations until set criterions are met (Sorsa et al. 2010, p. 6). They are based on three basic operations: selection, crossover, and mutation (Sorsa et al. 2010, p. 6). In selection, the principle is to give higher probabilities to more fit chromosomes to act as parents, which will produce offspring (Davis 1991; according to Sorsa et al. 2010, p. 7). When fit chromosomes are mainly chosen as parents, convergence times are lower. If weaker chromosomes are included as parents, convergence times increase but the search space is larger. (Michalewicz 1996; according to Sorsa et al. 2010, p. 7) In crossover, desired properties of chromosomes are combined leading to better chromosomes. Based on probabilities, some parents are placed straight into the new population and some are chosen for crossover. (Goldberg 1989; according to Sorsa et al. 2010, p. 7) In mutation, random changes are made to chromosomes and therefore maintaining diversity of the population. In addition to these three principles, elitism can be used to move predefined number of fittest chromosomes to the new population, thus increasing the convergence rate but decreasing diversity. (Davis 1991; according to Sorsa et al. 2010, p. 9)

2.5.7 Gaussian Process Regression

Gaussian processes use the Bayesian approach. In Bayesian approach, a prior probability distribution is placed over all possible functions and with observed data the prior is updated into posterior probability distribution. Gaussian process specifies a prior directly over function space instead of formulating priors over parameters, which usually simplifies the process. Gaussian Process is defined by a mean function usually with a value zero, and a covariance function $C(x, x')$, which describes the correlation of the function y at observations x and x'. Gaussian process is a stochastic method where variable subset has a joint multivariate Gaussian Distribution. (Williams 2003, p. 466-467) To ensure good performance with kernel methods, including Gaussian processes, the choice of a right kernel function is important. Common kernel function used in practice is the squared-exponential kernel. (Williams 2003, p. 469)

2.5.8 Support Vector Regression

Support Vector Machine (SVM) algorithms can be applied both in classification and regression tasks. It can also be applied for nonlinear problems by using kernel functions. Basic idea is similar with linear regression: a function that predicts training points by minimizing the prediction error is identified. Typically, with SVM the absolute error is minimized instead of the squared error. The difference with SVM algorithms compared to linear regression is that the user specifies a value ε, which defines the width of a tube around the regression function. The tube outlines are called the support vectors. All errors within the formed tube are ignored in the function identification. With linear SVM the tube has shape of a cylinder. If all the training points fit inside the tube (width equal to 2ε), the function is the middle of the flattest tube that enclosed all data. In this case the perceived error is equal to zero. The ε value determines how closely the function is fit to the training data. With large ε values the function is flatter, but the prediction error is bigger. The objective is to simultaneously minimize the prediction error and maximize the flatness of the function, which also reduces the change of overfitting. The tradeoff between the prediction error and flatness of the function can be controlled by setting an upper limit to the absolute values of the coefficients. This value restricts the effect of the support vectors on the shape of the regression function. With larger values, the function can fit the data more accurately and vice versa. (Witten 2011, p. 227-229)

2.5.9 Decision Trees

Decision trees classify observations by sorting them based on different attributes starting from the root and ending in multiple output values from different leaf nodes down the tree. Each node compares the observation with certain attribute, often with a constant value. The node is split into different branches, which correspond to different attribute values. If the attribute value is numeric, node can divide the values in two subsets: values less than the constant value and values greater than the constant. By adding an option where the values are equal to the compared constant, values can be divided into three branches. This is especially useful with integer values. With real-valued attribute, it can be more practical to compare the values against an interval, rather than a single constant. This divides the value in three branches: below, within and above. Numeric attributes are also often tested multiple times with different constant values. (Witten 2011, p. 64-65; Mitchell 1997, p. 52-53) Decision trees are robust to noisy data, to errors in classification of training data and errors in attributes that describe the training data. Decision tree can

also be used with missing values in training data or attributes. (Mitchell 1997, p. 52 & 54)

2.5.10 Ensemble modeling

In ensemble modeling different learned models can be combined. Training data can be divided into subsets, learn a model for each subset and then combine identified models to form an ensemble model. In contrast to individual models, ensemble modeling can make the learning process more effective and increase the predictive power. Popular ensemble modeling methods are bagging, boosting, and stacking, which can be applied to both classification and regression tasks. Problem of ensemble modeling can be loss of interpretability. Hundreds of individual models can be present and important factors can be hard to identify. In ensemble modeling, individual model predictions are converted into a single prediction. Easiest way to do this for classification is to vote (count the individual model predictions for each class and pick the most predicted class) and for regression to calculate average value for predictions. Bagging and boosting follows this principle. Difference is that bagging uses same weights for individual models and boosting uses weights to favor the most successful models. (Witten et al. 2011, p. 351-352) Stacking is usually used with different types of models built with different algorithms rather than with same model types like in bagging and boosting. Stacking uses the so called “meta-learner” to identify the best combination of base learner outputs to form the best predictions. (Witten et al. 2011, p. 369)

Diversity can be created by adding randomization to ensemble modeling. A random forest is created by choosing random decision tree in each iteration and combining the outputs into a single output. Popularly used random forest method identifies a randomized decision tree model in each iteration of the bagging algorithm. (Witten 2011, p. 356-357)

2.6 Hyperparameter tuning

Most learning processes are not fully autonomous, and some guidance is needed. Changes in hyperparameters can affect learning strongly and finding their optimal values can be challenging. Traditionally hyperparameter values have been estimated systematically, with random searches or using heuristics. Autonomous optimization algorithms for parameter tuning have been actively studied recently. (Butler et al. 2018, p. 549)

2.7 Model validation

Best way to evaluate performance of the trained ML model is to test it with completely new data from real-world situations. Although, gaining this new data can be time-consuming and expensive. (Toyao et al. 2020, p. 2267) More feasible option is to divide the existing data into training, validation, and test sets or use cross-validation (CV). Training set is used to train the model, validation set to evaluate hyperparameter tuning and test set to evaluate prediction performance. In cross-validation, data is divided into subparts and each subpart separately into training, test, and validation sets. (Toyao et al. 2020, p. 2266-2267) If the training and validation samples in cross-validation does not represent the whole dataset, unreliable results can be gained. This can be an issue with small datasets and when the model is applied to unseen observations that differ greatly from the observations in the original dataset. (Butler et al. 2018, p. 549)

To estimate the significance of differences in validation errors, statistical tests can be used for validation sets. If large amounts of observations are available, splitting the training data is not necessarily needed. In that case, training errors can be compared with statistical tests. Leave-one-out cross-validation can be used to estimate model performance when data is extremely limited. The leave-one-out cross-validation procedure involves removing one data point from the training set, building model with the remaining training data, and testing with the data point set aside. This procedure is done to all data points and the average of results is used. It should be kept in mind, that often this method leads to overly optimistic results. (Guyon & Elisseeff 2003, p. 1173) Also, random subsampling can be used with small datasets (Toyao et al. 2020, p. 2267).

To evaluate performance of the identified model, evaluation criteria or loss function must be chosen. Typically for regression mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE) or root mean square logarithmic error (RMSLE) is used. (Toyao et al. 2020, p. 2266) Evaluation criteria can be optimized with appropriate tuning algorithm to optimize fit of the ML model to the given data (Toyao et al. 2020, p. 2265).

Three factors should be taken into account, when evaluating the model error: 1) model bias, that describes the mismatch of model to the data, 2) model variance, that describes the sensitivity of the model to small changes in training data, and 3) residual error, that

can be a cause of the noise in the training data, error in measurements or calculations, outliers or missing data. When model bias is high, the model is underfitting and describes the response data poorly. When the model variance is high, the model can be overfitting and is too complex. As the learning advances, fit to training data can improve, but fit to test data declines. (Butler et al. 2018, p. 549) Ultimately, a point where test performance is at highest should be searched.

3 MODELING IN CATALYSIS

Modeling in catalysis can be fundamental, empirical (i.e. data-driven) or their combination (Madaan et al. 2016, p. 129-130). This chapter focuses more on empirical modeling and more specifically in ML models. Catalysis involves many challenges due to its complex nature. First, catalytic reactions are dynamic, they scale from atomic level to large-scale reactors and their time scale ranges must be considered to fully gain understanding of the catalytic reactions. (Kitchin 2018, p. 230; Toyao et al. 2020, p. 2260; Cundari et al. 2001, p. 5475). Second, variety of components and chemical and physical properties affect the outcomes with several catalysts. In addition, relationship between catalyst properties and catalyst performance is often unclear and nonlinear. Thus, optimizing the mixture of components involves multidimensional and multiscale research. (Kitchin 2018, p. 230; Cundari et al. 2001, p. 5475) Finally, heterogeneous catalysts can change in reaction conditions through various phenomena (Kalz et al. 2017) and their active state structures are often complex, which complicates their modeling and further experiments (Goldsmith et al. 2017; according to Goldsmith 2018, p. 2311).

3.1 Machine learning in catalysis

Machine learning has been found to be useful in many applications. In catalysis, Machine Learning (ML) enhances ways to discover catalysts, generates knowledge about catalysis and establishes deeper understanding of relationship between material properties and their catalytic FOMs (Cundari et al. 2001, p. 5475; Kitchin 2018, p. 230; Toyao et al. 2020, p. 2263). With combined computational modeling and/or experiments, catalysts can be rapidly screened, descriptors of catalyst performance can be found, and catalyst synthesis can be enhanced (Goldsmith et al. 2018, p. 2318). Machine learning can also be used in formulating new descriptors used in combination with quantum mechanical (QM) methods and to formulate interatomic potentials (Kitchin 2018, p. 230; Goldsmith et al. 2018, p. 2315).

The use of ML in computational catalysis research and integration with experimental research programs has been increasing (Kitchin 2018, p. 230). Goldsmith et al. (2018, p. 2311) listed several examples, where integration of ML and HTP screening for heterogeneous catalysts were used to predict catalyst FOM for large catalyst spaces.

However, predictions of catalytic FOM are still in their early stages (Goldsmith et al. 2018, p. 9). Synthesis conditions and compositions have been used as model input features for predictive models, which can be seen in (Baumes et al. 2004) and (Baumes et al. 2006). These ML approaches can guide the synthesis towards better catalysts, although data from experiments are often incomplete and can result in poorly generalized models for large chemical spaces (Goldsmith et al. 2018, p. 2312).

According to (Toyao et al. 2020), catalyst preparation-structure-performance relationships have been studied fundamentally, which can accelerate design of new catalysts. Large amounts of data are available from these studies, but not integrated into easily accessible databases. Advances in theoretical and computational chemistry have provided new knowledge to already discovered catalytic processes, but not so much in design of new catalysts. Therefore, the data available is not particularly useful in the design of new catalysts or catalytic reactions. (Toyao et al. 2020, p. 2260-2261)

The early work in applying ML in catalysis has been reviewed by Maldonado & Rothenberg (2010). They highlight, for example, the studies by Artyushkova, Corma, and Serra. Artyushkova et al. (2008) used predictive modeling to find correlations between the structure and electrochemical performance of various non-platinum porphyrin electrocatalysts for oxygen reduction (according to Maldonado & Rothenberg 2010, p. 1892). Genetic algorithms (GAs) and artificial neural networks (ANNs) were combined to predict the performance of virtual catalyst libraries for oxidative dehydrogenation of ethane, where catalyst compositions were used as inputs (Corma et al. 2002; Serra et al. 2007; according to Maldonado & Rothenberg 2010, p. 1892).

More recently, in (Omata 2011), activity of heteropoly acid catalyst supported on active carbon for Friedel-Crafts reaction was predicted with three different ML methods, namely Gaussian process regression (GPR), radial basis function network (RBFN) and support vector machine (SVM). Five main principal components of physicochemical properties of elements of the additives were used as descriptors. Ionic radii, atomic weight and density of the additive element were found to be the most influential variables. Hard Lewis acids were suggested to be related to catalytic activity. Ras et al. (2012) studied the hydrogenation of 5-ethoxymethylfurfural using eight different metals with Al_2O_3 support as catalysts under various conditions and compared the given data with the results from bimetallic supported catalysts. Catalyst descriptors for model were based on Slater

type orbitals. This research proved, that it was possible to describe catalyst performance well with simple models. (Ras et al. 2012)

3.2 Collecting data

At the beginning of modeling, a dataset of catalyst and their experimentally acquired FOMs are needed (Figure 3). Larger datasets with diversity are preferred. HTP screening is the preferred source of the data, which can be integrated with design of experiments. (Maldonado & Rothenberg 2010, p. 1894) Data overload can be a problem with large amount of experiments, which can be dealt with by using databases for storage and combining data. After data construction, visual exploration of data can be done. Statistical experimental design and principal component analysis (PCA) are popular methods, that can be used in catalytic research to gain better understanding of the impacts each variable has on catalytic performance. Also, insights to relationships between various performance indicators can be gained. After initial exploratory data analysis is done, model-based evaluation starts. (Ras et al. 2014, p. 5965-5966)

The data can involve variables and properties such as boiling point, Slater orbitals, and atomic radius. In addition, spectroscopic data may be used. Examples can be found from references (Omata 2011), (Suzuki et al. 2019) and (Toyao et al. 2020). Comprehensive databases, such as "The Open Quantum Materials Database" (OQMD)[2], Catalysis-Hub.org[3] and CatApp[4], also exist, which include for example reaction and activation energy calculations. In Figure 3, three multi-dimensional spaces A, B and C are represented, which contain catalysts, their descriptors and reaction conditions, and figures of merit. The space A contains all the catalyst combinations. The space B contains all the catalyst descriptor values and reaction conditions. Finally, the space C contains the FOM for each catalyst combinations. Examples of variables for each variable space are shown in the figure. The space B is crucial for the identified model performance. To establish a

[2] The Open Quantum Materials Database. Available: http://oqmd.org/.
[3] Catalysis-Hub.org. Available: https://www.catalysis-hub.org/.
[4] CatApp. Available: http://suncat.stanford.edu/theory/it-facilities.

good prediction performance, choosing the right descriptors is of great importance. (Rothenberg 2008, p. 6-7)

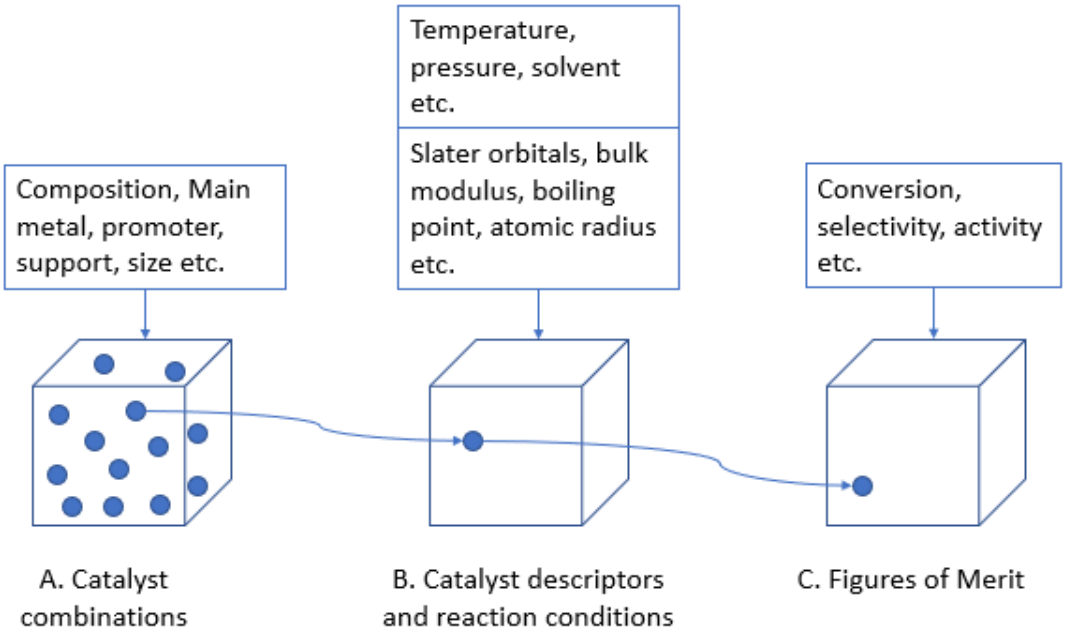

Figure 3. Representation of multi-dimensional spaces A, B and C, which contain catalysts, catalyst descriptors and reaction conditions, and the figures of merit (retell Rothenberg 2008, p. 6).

3.3 Catalyst screening and selection

Catalyst selection can have two different meanings: choosing catalyst formulations for experimental testing or choosing the optimal catalyst to be used in the actual reaction. Machine-based catalyst selection methods, for catalyst synthetizing, are preferred instead of human intuition to avoid bias. The selected catalysts should also be possible to synthesize feasible and consistently. If not, eligible replacement must be considered. (Ras et al. 2014, p. 5971-5972)

Catalyst screening means searching for catalyst formulations, which optimize certain FOM (e.g. conversion and selectivity), by experimentally testing different combinations with different reaction conditions and analyzing the results. Screening for new catalysts and catalytic reactions is difficult and many of these new discoveries have come from trial-and-error approaches. Several multicomponent catalysts that can be used in different conditions has been studied in this empirical approach. (Ishikawa et al. 2018; Sehested 2019; according to Toyao et al. 2020, p. 2260) Variations in catalyst compositions and

experimental conditions are involved in catalysis research. Therefore, a systematic approach is required. (Ras et al. 2014, p. 5963)

Generally, there are three systematic approaches for catalyst screening. First option is to use HTP screening to test everything. Problem is that the search space is enormous. Second approach is using mechanistic theories for catalytic reactions on surfaces and verifying them with kinetic experiments. This approach is expensive and is very laborious. Also, wrong theories can also lead to a dead end. The third option combining HTP screening, data-driven model and mechanistic theories is the most feasible option, as has been shown in numerous studies (see e.g. Maldonado & Rothenberg 2010, Ras et al. 2012).

The third option above combines experimental design, descriptor modeling and experimental validation. Multiple iterations in the workflow are needed to find optimal solutions (Ras et al. 2014, p. 5973). After each iteration, the results should be compared with the set objectives, which typically are related to catalyst performance features (e.g. yield, selectivity, conversion). If the objectives are met, you can move to the next phase. The original objectives may also be unrealistic and must be readjusted during iterations. (Ras et al. 2014, p. 5964)

HTP screening allows calculations of properties of thousands of catalysts in a single study (Butler et al. 2018, p. 547). Before starting HTP screening program, design of efficient catalyst libraries should be considered. Promising areas in catalyst search space, where catalyst performance objectives are closely met, should be highlighted, and outliers discarded. In the start of screening, diversity should be at the highest, whereas in the end, focus should be on the best performing zones. (Baumes et al. 2004, p. 768)

Screening can be assisted with ML to guide the catalyst search in the right direction by finding correlations or by accelerating calculations of the target feature (Goldsmith et al. 2018, p. 2312). Density Functional Theory (DFT) calculations can be used in screening to predict catalyst properties, which are often correlated with adsorbate binding energies (Back et al. 2019, p. 1). EAs suit well for material screening and optimization since they tolerate data with noise and outliers, use population of solids that matches with the synthesis and learn as the calculations proceed, therefore minimizing the iterations by focusing on relevant material spaces (Baumes et al. 2004, p. 767-768).

3.4 Molecular descriptor modeling and selection

In the field of catalysis, descriptors are catalyst's computational features that describe catalyst's properties, and which can be used in predicting catalyst performance (Todeschini & Consonni 2000, p. XI). As mentioned in the beginning of Section 3, understanding catalytic reactions is highly complex problem. Furthermore, formulating appropriate descriptors to describe catalyst's behavior and to predict its performance is a challenging task (Todeschini & Consonni 2000, p. XI). Ensuring that descriptors are relevant to catalyst performance often requires domain knowledge (Kitchin 2018, p. 230; Goldsmith et al. 2018, p. 2314). Also, it is hard to identify descriptors with high predictive power and relevance to the catalysts. Even though the predictive power of a descriptor is low, but it has relevance to the catalysts, it should be considered as a potential candidate to the variable subset. Generally, a descriptor should give more insights to the catalyst properties and/or to be able to participate in predictions of catalyst FOM. (Todeschini & Consonni 2000, p. XI) They should also be able to describe catalytic performance adequately on their own (Ras et al. 2014, p. 5970). Descriptor should be applicable to all catalysts and to be somewhat invariant and predictable. Anyhow, multiple descriptors are needed to fully describe the behavior of catalysts and catalytic reactions. (Todeschini & Consonni 2000, p. XI) Structures of catalysts have found to be strongly relevant to physicochemical, topological, and electronic parameters (Maldonado & Rothenberg 2010, p. 1894). Also, catalytic performance has found to be directly related to distribution of electrons in a metal and the shape of the orbitals (Ras et al. 2012, p. 4).

In ML, it is important to find and select relevant descriptors, which correlate with FOM and have physiochemical relevance. Typically, multiple descriptors are needed to predict catalytic performance, especially when catalyst properties related to good performance are not well known and are needed to be experimentally discovered. Selected descriptors should cover most of the metals used in catalysis. If the knowledge of problem is weak, typically choosing more descriptors is advised. With experimentations, redundant descriptors can be identified and eliminated. (Ras et al. 2014, p. 5970-5971) Computational time should also be considered, when choosing descriptors, since descriptor types and how they are formulated affect in computation times. For example, 2D-descriptors are usually simple and fast while QM calculations are specific but slow to calculate. (Maldonado & Rothenberg 2010, p. 1893).

The fact, that the active site of a heterogeneous complex is often poorly defined limits the development of descriptor-performance relationships for heterogeneous catalysts. In homogeneous catalysis catalyst is well-defined and can be described easily. (Farrusseng et al 2005; Klanner et al. 2004; according to Ras et al. 2014, p. 5968-5969) Descriptor modeling in homogeneous catalysis is far ahead compared to heterogeneous catalysis. In heterogeneous catalysis descriptors are often hard to acquire and usually are obtained from DFT calculations or fundamental characterization of the catalysts, which are time consuming and deep understanding is required. Also, in catalyst characterization, catalysts are needed to be synthesized first before descriptor-performance relationships can be obtained. Thus, screening without experiments is challenging. (Nørskov et al. 2009; according to Ras et al. 2014, p. 5969) Despite the challenges, multiple successful studies involving DFT simulations to describe interactions in reaction network can be found. Also, d-band center has been used in several studies. Simple empirical descriptors though are lacking. (Ras et al. 2014, p. 5969-5970) Large database of chemical properties on transition metal surfaces has found to have great potential in finding new catalytic materials (Mamun et al. 2019, p. 1). Also, scaling the relationship between calculated reaction barriers and binding energies of surface has found to be promising method for data-driven catalyst design (Toyao et al. 2020, p. 2261).

3.4.1 Quantum Mechanical Methods

Quantum Mechanical (QM) calculations are based on probabilities. Using QM based calculations to form predictive models to discover active sites, adsorbate binding strengths and structure-activity relationships is a popular method in catalyst screening. QM computations are computationally expensive, which limits the use of them and are basically limited to small systems. (Ras et al. 2014, p. 5969; Goldsmith et al. 2018, p. 2311, 2315, 2318) QM can create larger datasets than experiments with ease and fill the missing areas in experimental data, from where ML models can be trained (Goldsmith et al. 2018, p. 2312).

Density functional theory (DFT) is a popular computational QM method, which can be used in developing descriptors. Especially, electronic structure calculations with DFT are a strong method to predict a large space of material properties (Jones 2015; according to Mamun et al. 2019, p. 1). Also, the adsorption energies of chemical elements to the surface, from DFT calculations, has found to have a strong relationship with catalytic

activity of surfaces (Mamun et al. 2019, p. 1). In case of hundreds of thousands of calculations, DFT is not suitable considering computational costs (Kitchin 2018, p. 230). Modeling catalysts with DFT is time-consuming even for small number of catalysts and therefore by synthesizing and experimentally testing, the results can be acquired much faster. Creating descriptor-performance relationships can be enhanced by using simple, easily accessible descriptors for metals on the catalyst surface. (Ras et al. 2014, p. 5968-5969)

3.4.2 Surface phase diagrams

Surface phase diagrams can show the expected catalyst composition and surface phase as a function of temperature, pressure, potential or dopant concentration and therefore identify catalyst active sites and reaction mechanisms (Goldsmith et al. 2018, p. 2313).

3.4.3 D-band center

The most used descriptor in heterogeneous catalysis is the energy of the d-band center. D-band center can be related to catalyst activity through linear scaling reactions. (Nørskov et al. 2009; according to Goldsmith et al. 2018, p. 2314) It is commonly predicted using ML (Takikawa et al. 2016; according to Kitchin 2018, p. 230). For example, Kitchin et al. (2018) used the atomic geometry of an active site, with fitted parameters for the size of the d-orbitals of each species, to their model and found out that moderately accurate predictions could be made with only six descriptors.

3.4.4 Machine-learned potentials

Machine-learned interatomic potentials (MLPs) that uses DFT calculations are methods of growing interest (Behler 2015; according to Kitchin 2018, p. 230). Interatomic potentials can be described as mathematical functions that computes the potential energy of an atom system. Interatomic potentials can be formed by using ML with data from QM calculations. These potentials estimate interaction energies compared to QM calculations with increasing efficiency. (Brockherde et al. 2017; according to Goldsmith et al. 2018, p. 2315) Reactive dynamics on catalyst surfaces can be identified, which allows examination of reaction trajectories at realistic temperatures (Kitchin 2018, p. 230). MLPs can accelerate simulations significantly while having moderate accuracy compared to QM calculations (Schütt et al. 2017; according to Goldsmith et al. 2018, p. 2315). Search of

catalyst structures can be accelerated under different operation conditions and longer time and length scales can be simulated (Goldsmith et al. 2018, p. 2316).

3.5 Predictive modeling of catalyst performance

Predictive modeling in catalysis aims to find relevant descriptors to connect experimental data with desired FOMs (Maldonado & Rothenberg 2010, p. 1892, 1895). To build a successful predictive model in catalysis, a proper initial dataset from catalytic experiments and a method to generate and evaluate virtual catalyst libraries are required. Also, identifying a correct model structure and choosing a good validation method is important. (Maldonado & Rothenberg 2010, p. 1894)

Data mining methods can be helpful in building predictive model and finding non-trivial patterns from big data. Predictive features can be extracted, and trends in catalytic reactions and hidden patterns in experimental catalysis can be found, which can be used to guide the search in large catalyst spaces towards better areas. (Goldsmith et al. 2018, p. 2314)

A typical workflow for building predictive models in catalysis starts with creating dataset of catalyst libraries. Each catalyst must be then described by their features, also called fingerprints, which are typically electronic structure properties, physical properties, and atomic properties. The features should explain physicochemical properties of the materials as well as possible and be moderately easy to calculate. After this, ML can be used to find patterns, build models or formulate descriptors, which connects catalyst features with the desired FOMs. (Goldsmith et al. 2018, p. 2312) Usually multiple iterations and experimentations are required for successful application (Maldonado & Rothenberg 2010, p. 1894).

In Figure 4, an example of the catalyst performance prediction procedure can be seen (Madaan et al. 2016, p. 129). First, initial hypothesis for the experiments is formed. Second, the experimental data is gathered by testing or from literature. Third, appropriate descriptors are assigned, and the model is trained by using the chosen descriptors. Fourth, the features of merit are predicted. After that, the results are validated with experiments. If the performance criterion is met, the procedure is finished. Otherwise, descriptors are

assigned again and/or the model is identified again. Also, meta-modelling can be used by using predicted FOM values as inputs to the model to perform virtual catalyst testing.

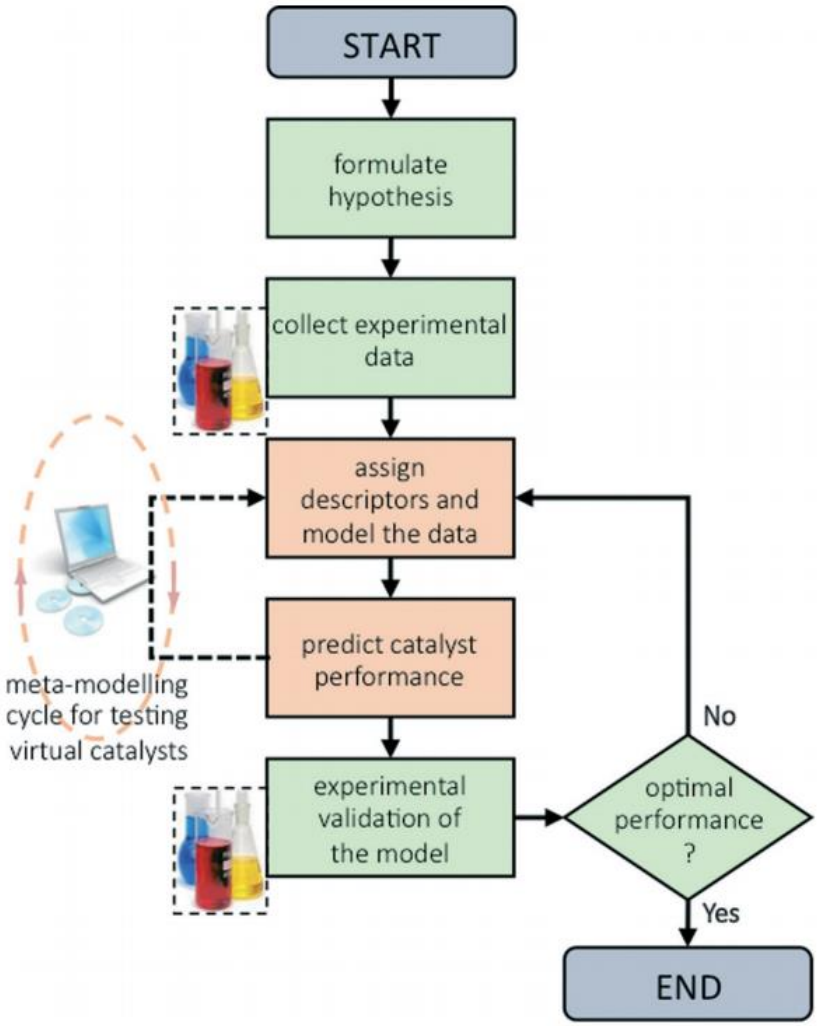

Figure 4. “Schematic flowchart showing the iterative process of hypothesis formulation, data collection, regression modelling, screening of virtual catalysts (meta-modelling), and testing in the lab (experimental validation of the model results).” (Madaan et al. 2016, p. 129) - Published by The Royal Society of Chemistry.

Predictive modeling in catalysis is best used with combining modeling and experimentation teams. Often modeling is carried out by statisticians rather than persons with proper knowledge of catalytic processes like chemists and chemical engineers, which can be problematic. (Ras et al. 2014, p. 5965) Modeling can help all stages in catalyst development. Model complexity should generally start from simple and increase, when the project advances. In the early stages, exploratory data analysis and empirical models should be used and in the later stages more advanced models like kinetic models can be applied. (Ras et al. 2014, p. 5963) Model should be in interpretable form if

possible. Generating flawed models based on bad input feature structure and overfitting the model should naturally be avoided. (Goldsmith et al. 2018, p. 2312)

3.6 Predictive modeling methods

Modeling the catalyst performance can be divided into three categories: fundamental, empirical and their combination. Fundamental approach involves computational chemistry, kinetic modeling and reactor design and focuses on reaction mechanisms and engineering principles. QM models can be involved in fundamental approach. This approach is typically expensive but provides precise information about the system. It is preferably used when mechanistic information and reactor constraints are known. (Madaan et al. 2016, p. 129; Ras et al. 2014, p. 5965) Examples of fundamental approaches in heterogeneous catalysis are listed, for example, in the study of Madaan et al. (2016, p. 129). Empirical approach is based on data and does not make assumptions on reaction mechanisms and reactor configurations. They are based on statistical analysis and are often combined with stochastic optimization methods (e.g. GA). (Beckers et al. 2008; Dragoi et al. 2014; according to Madaan et al. 2016, p. 129; Ras et al. 2014, p. 5965) The lack of universal databases for catalytic FOMs and descriptors hinders the use of data-driven methods in catalysis (Toyao et al. 2020, p. 2261). Empirical approach is fast, but they adapt poorly to new factors and interpretability can be weak. Third approach is combination of fundamental and empirical approaches. Descriptors based on chemical principles are combined with statistical modeling. This approach is preferable for predicting catalyst performance. (Madaan et al. 2016, p. 129; Ras et al. 2014, p. 5965)

Most ML methods for chemical reactions and properties use molecular or atomic descriptors in model identification. Descriptor should be easier to obtain than target properties and low dimensionality should be preferred. (Ghiringhelli et al. 2015; according to Butler et al. 2018, p. 553).

General principle for empirical models is that you can only make reliable predictions for the range of the data used to identify model. For example, if data for yield ranges between 0-70% is captured, you cannot make reliable predictions in yields over 70%. In order to expand prediction range, extrapolation is needed. (Ras et al. 2014, p. 5969)

Predicting performance of solid catalyst is challenging due to sensitive activity changes, as the result of variations in solid surfaces, which are in turn easily resulted by minor changes in the synthesis and pre-treatment conditions (Dacquin et al. 2010; according to Madaan et al. 2016, p. 125). Predictions of solids can be made with combination of QM and classical mechanics. Successful applications have been made in this field, (Nørskov et al. 2009; according to Madaan et al. 2016, p. 125) where performance can be predicted with high cost (Calle-Vallejo et al. 2015; Li et al. 2014; according to Madaan et al. 2016, p. 125). Alternatively, performance can be predicted using data-driven modeling with simple descriptors. These models are less interpretable, but more practical. (Kondratenko et al. 2015; Montemore & Medlin 2014; Harper et al. 2013; according to Madaan et al. 2016, p. 125)

ANNs have been popular also in predictive modeling of catalyst performance. They can be used to form relationship between catalyst properties and performance. After network training, ANNs can be used to identify properties that maximize desired FOMs. ANNs can be also used for classifying catalyst properties in clusters of similar properties. (Cundari et al. 2001, p. 5476)

Physicochemical properties of elements have shown to be great descriptors for ANN to learn catalytic phenomena, which can be seen in the works of Hattori, Kito and Murakami. (1991; 1992; 1994; according to Omata 2011, p. 10948) ANN has been used to form relationship between physicochemical properties and reaction conditions with catalytic performance (Hattori & Kito 1995; Sasaki et al. 1995; according to Goldsmith et al. 2018, p. 2311). Baumes et al. (2004) used ANN to predict performances of catalysts for the Water Gas Shift reaction. A detailed methodology to help identify relevant solid catalyst spaces for HTP testing were also introduced. (Baumes et al. 2004, p. 767) Duvenaud et al. (2015) used ANNs to create new fingerprints for molecules in reactions, which can improve the catalyst synthesis predictions.

4 STUDIED PROCESSES

The target reaction for BioSPRINT project is the simultaneous dehydration of multiple C5 and C6 sugars to produce 5-HMF and FUR. The state of the art of this process will be discussed in Section 4.1. For the project, mainly heterogeneous catalysis will be considered involving solid acid catalysts. (BioSPRINT 2019) Experimental study, and the development of ML approach in this thesis is, however, based on hydrogenation reaction of 5-ethoxymethylfurfural, which will be discussed in Section 4.2. Sections 4.3 and 4.4 introduces relevant catalysts and solvents for the dehydration reaction of C5 and C6 sugars.

4.1 Dehydration of C5 and C6 sugars

Furan derivatives converted from sugars are in great interest in chemistry and catalysis studies, because of their potential in renewable fuel and chemical production (Tong et al. 2010, p. 1 & 11). A key process for biomass derived carbohydrates is dehydration in aqueous phase (Agirrezabal-Telleria et al. 2014, p. 1357).

Conversion of biomass is challenging due to biomass's over-functionalized composition, where alcohols, ethers, esters, and carboxylic acids are present (Ras et al. 2009, p. 3175). Also, typically poorly characterized components are involved and large number of minerals in various concentrations are present, which also depends on the source and the type of biomass (Vassilev et al. 2012; according to Zhang 2013, p. 68). These minerals can function as a catalyst, passivate catalysts or be inert (Zhang 2013, p. 68).

Improvement of conversion and selectivity are needed to produce furan derivatives commercially from sugars. Multifunctional catalysts formed from combination of transition metals and solid acid/base catalysts can potentially allow several reaction steps to take place in a single reactor and remove the need for expensive intermediate separation processes. Catalyst recyclability and efficient product separation are also interesting topics in catalysis studies. (Tong et al. 2010, p. 11)

4.1.1 Production of 5-HMF

5-Hydroxymethylfurfural (5-HMF) with valuable chemical properties has great potential as a renewable intermediate chemical to substitute non-renewable raw materials for various polymers and chemicals including renewable fuels. 5-HMF can be produced from carbohydrate hexoses such as fructose and glucose with dehydration. (Tong et al. 2010, p. 1; Zhang 2013, p. 68; Rosatella et al. 2011, p. 754)

5-HMF is mainly produced via dehydration of monosaccharides. Disaccharides and polysaccharides can also be used as starting materials, where additional hydrolysis reaction step is normally used for depolymerization. (Rosatella et al. 2011, p. 757) Production of 5-HMF from fructose (C6 sugar) via acid-catalyzed dehydration in aqueous phase is the most studied production route. More recently, fructose via glucose (C5 sugar) isomerization and direct cellulose as feedstock has also been studied. (Rosatella et al. 2011, p. 754; Zhang 2013, p. 68) Glucose is a more abundant feedstock than fructose, but generally fructose has had better reaction rates and selectivities than glucose (Kuster 1990; according to Chheda et al. 2007, p. 343). Due to the abundance of glucose, development of 5-HMF production with glucose as starting material is of great interest (Tong et al. 2010, p. 4).

The production of 5-HMF industrially has been limited by the high production cost (Zhao et al. 2007, p. 1597). Mainly, due to challenging dehydration reaction with possible side reactions (Rosatella et al. 2011, p. 757). These side reactions are increased with the use of acid catalysts leading to increased purification costs (Zhao et al. 2007, p. 1597). In order to increase yield of 5-HMF and avoid side reactions, there are two approaches (Tong et al. 2010, p. 3). First option is to use multifunctional catalysts that promote 5-HMF formation, while demoting the side reactions and the second option is to use continuous removal of products. Also, biphase reaction systems allow better 5-HMF recovery and lowers the degradation of 5-HMF during the reaction phase, leading to better performance (Zhang 2013, p. 69).

4.1.2 Production of furfural

Much like 5-HMF, furfural has potential to be transformed into fuels and useful chemicals from hemicellulose (Mariscal et al. 2016; according to Luo et al. 2019, p. 15).

C5 sugar monomers, mainly xylose, of lignocellulosic HMC in raw biomass can be used to produce FUR (Werpy et al. 2004; Lange et al. 2012; according to Luo et al. 2019, p. 17). Production of FUR from xylose has been studied by several researchers, especially with acid catalysts (Chheda et al. 2007, p. 343). FUR has been produced industrially from pentosan-rich biomasses via xylose cyclodehydration with low yields of circa 50 % (Mamman et al. 2008; according to Agirrezabal-Telleria et al. 2014, p. 1357). Producing FUR from biomass involves several complications: Homogeneous H_2SO_4 type catalysts increase corrosion, harden separation and have poor selectivities. Separations based on steam require high amounts of energy and also have high distillation costs. (Zeitsch 2000; according to Agirrezabal-Telleria et al. 2014, p. 1357) More feasible reaction routes and production technologies to produce FUR from HMC are needed (Agirrezabal-Telleria et al. 2014, p.1357). These technologies (e.g. recyclable solvents) should minimize solvent-FUR separation and purification costs (Agirrezabal-Telleria et al. 2014, p. 1358). Reactions in aqueous phase with high temperature and use of stripping agents (e.g. N_2) have shown great FUR yield increases (Jing & Lü 2007; Agirrezabal-Telleria et al. 2012; according to Agirrezabal-Telleria et al. 2014, p. 1358). In these conditions, stripping of FUR from the reaction medium can be achieved at high selectivity, separation of stripping-agent is easier and further FUR purification stages can be reduced (Agirrezabal-Telleria et al. 2014, p. 1358).

4.1.3 Combined production of 5-HMF and furfural

Various compositions of catalysts and solvents in the simultaneous production of FUR and 5-HMF from multiple sugar monomers (mainly glucose, fructose, and xylose) has been experimented. Yet, feasible option for production is still lacking. (Chheda et al. 2007, p. 343) Biphasic reaction systems are promising for feasible production of 5-HMF and FUR simultaneously from cheap and abundant renewable feedstocks (Chheda et al. 2007, p. 349). Chheda et al. (2007) studied simultaneous production of 5-HMF and FUR in a biphasic reactor with multiple sugar monomer feeds (glucose, fructose, and xylose). Optimal reaction conditions were identified, and those conditions were used with polysaccharides corresponding to sugar monomers. Good selectivities and conversions were achieved with both sugar monomer feeds and polysaccharide feeds. This study showed that production of furan derivatives straight from polysaccharide feeds are potential routes, where good results were achieved without the need for additional hydrolysis reaction step for depolymerization of polysaccharides. (Chheda et al. 2007, p.

342) Process variables that were influential to the yields were solution pH, initial sugar concentration, acid properties, dimethyl sulfoxide (DMSO) and extraction solvent content (Chheda et al. 2007, p. 349). In the work of Agirrezabal-Telleria et al. (2014), with combination of xylose and glucose feeds, high FUR and 5-HMF selectivities were achieved. N_2-stripping allowed high product purity and selectivity. (Agirrezabal-Telleria et al. 2014, p. 1367)

4.2 Hydrogenation of 5-ethoxymethylfurfural

Experimental part is based on Ras et al. papers (2009, 2012). The experimental data was obtained from Ras et al. (2009) paper and values for Slater-type orbitals were obtained from Ras et al. (2012) paper. The process for the experimental part is hydrogenation of 5-ethoxymethylfurfural with alumina-supported heterogeneous catalysts. The desired product for the reaction is 5-ethoxymethylfurfuryl alcohol, which is a potential additive for diesel fuel. Several by-products are involved in the reaction (more detail in the paper of Ras et al. 2009). This reaction resembles the dehydration reaction of C5 and C6 sugars. A 16-parallel trickle-flow reactor was used in the experiment of 48 different catalyst combinations with three different temperatures (80, 100 and 120 °C) with diethyl carbonate and 1,4-dioxane solvents. 5-ethoxymethylfurfural can be easily obtained via acid-catalyzed dehydration of C6 sugars with ethanol solvent. (Ras et al. 2009, p. 3175) The experiments and the gathered dataset are explained more in detail in the Section 5.1.

4.3 Catalysts

Catalysts are typically divided into homogeneous and heterogeneous. Homogeneous catalysts are dissolved in the reaction medium, which makes the interaction with solid biomass easier. (Luo et al. 2019, p. 16) On the other hand, homogeneous reactions catalyzed by acids tend to have challenges in product and catalyst separations, increased corrosion, pollution by acidic wastewater and large amounts of catalyst typically going to waste (Al-Mubaddel et al. 2006, p. 17). In general, heterogeneous catalysts are easier to separate and recover (Rinaldi and Schüth 2009, p. 612). Heterogeneous catalysts often have difficulties with deactivation in aqueous solutions or with biphasic systems, where salts are involved (Sahu & Dhepe 2012; Bernal et al. 2014; according to Luo et al. 2019, p. 16).

Catalysts for dehydration of carbohydrates can be classified into organic acids, inorganic acids, salts, Lewis acids and others (Cottier & Descotes 1991; according to Rosatella et al. 2011, p. 757). In addition, catalysts for 5-HMF synthesis has been classified as mineral acids, organic acids, solid acids and metal-containing acids (Tong et al. 2010, p. 3). The most common acid catalysts for 5-HMF production have been mineral acids, H_2SO_4, H_3PO_4 and HCl and Lewis acids (Harris & Feather 1974; Kuster & van der Baan 1977; Moye & Goldsack 1966; according to Zhang 2013, p. 56).

In addition, so called super solid acids can be used to promote the dehydration reactions. "Any acid with an acidity stronger than that of 100% H_2SO_4 (Hammett acidity function: $H_0 =< -12$) is called a superacid." (Gillespie 1968; according to Al-Mubaddel et al. 2006, p. 4) The most common solid super acids in catalysis according to Al-Mubaddel et al. (2006, p. 6) are: liquid acids mounted on suitable supports, combined acids (a combination of inorganic acids), sulfate ion promoted metal oxides (e.g., SO_4^{2-}/ZrO_2), metal promoted superacid and nafion-H.

Hydrolysis of HMC with inorganic acids, mainly HCl and H_2SO_4, and solid acids has been used to produce FUR (Zhu et al. 2016; Fusaro et al. 2015; Danon et al. 2014; Dhepe & Sahu 2010; according to Hui et al. 2019, p. 49). Problem of these acid catalysts have been poor selectivity, poor separation of product and catalyst and deactivation of solid acids (Hui et al. 2019, p. 49). In the degradation of HMC, acid strength of catalyst has found to have great importance (Henon et al. 2003; according to Hui et al. 2019, p. 50). In the production of 5-HMF from fructose, the conversion of fructose and the selectivity of 5-HMF have been found to be related to the acid type, acid structural properties and acid pore volume distribution (Moreau et al. 1996, p. 1).

Active and stable catalysts are needed in catalysis (Madaan et al. 2016, p. 125). They should also be highly selective towards desired products to maximize cost-effectiveness of the process (Rinaldi and Schüth 2009, p. 612). Development of these kind of catalysts for biomass conversion is challenging due to several difficulties to be considered: thermal stability, deactivation, low conversion and multifunctional catalyst interactions. (Rinaldi & Schüth 2009, p. 616, 618-619, 622-623).

Lewis/Brønsted (L/B) surface properties have strong importance when developing catalysts for FUR production. Acid properties and their nature can be further correlated

to the catalytic carbohydrate conversion activity and reaction mechanisms. (Agirrezabal-Telleria et al. 2014, p. 1358) Bifunctional catalysts have been studied in xylose dehydration, where the reaction kinetics has been correlated with the Lewis/Brønsted acid ratio (Choudhary et al. 2011; Weingarten et al. 2011; according to Agirrezabal-Telleria et al. 2014, p. 1362). Dehydration and degradation reactions could be controlled by changing the L/B acid ratios to optimize FUR production. However, glucose in xylose feed will reduce FUR yield due to increasing side reactions, which should be taken into account with multiple sugar feeds. (Agirrezabal-Telleria et al. 2014, p. 1358, 1362)

In contrast to liquid acid catalysts in the production of 5-HMF, solid acid catalyst have several advantages: easier product separation, better catalyst recyclability, capability to operate in high temperatures with shorter reaction times and promotion of 5-HMF formation instead of degradation in longer period of time and improved 5-HMF selectivity. (Tong et al. 2010, p. 4) Common solid acid catalysts found in the literature in the dehydration of carbohydrates are H-form zeolites, ion-exchange resins, vanadyl phosphate and ZrO_2 (Tong et al. 2010, p. 4). Zhang (2013, p. 56) listed several solid acid catalysts reported in the production of 5-HMF: Amberlyst-15, Nafion NR50, SO_4/ZrO_2, $Nb_2O_5–nH_2O$, H-ZSM5, H-Beta and H-Mordenite. Takagaki et al. (2009, p. 6276) used Amberlyst-15 as solid acid catalyst for dehydration of fructose, where catalyst was able to be used three times without loss of activity. Moreau et al. (1996) used dealuminated H-form mordenites for dehydration of fructose to 5-HMF. Compared to mineral acids and ion-exchange resins these H-form mordenites can operate at higher temperatures and can be easily regenerated and used for several times (Moreau et al. 1996, p. 223).

4.4 Solvents

Lignocellulosic biomaterials are typically solids with poor solubility in most solvents, which leads to the need of dispersants. Poor solubility can make the reactions seen as heterogeneous concerning the substrate. In these cases, substrate surface processes affect reaction rates. (NIST 2018; according to Rinaldi and Schüth 2009, p. 616)

Rosatella et al. (2011) listed several articles, where different organic extraction solvents were used. These organic solvents have been proven to be efficient in reducing the side-reactions of 5-HMF (Rosatella et al. 2011, p. 757). Dimethyl sulfoxide (DMSO) is one of these commonly used organic solvents. Unfortunately, separation of DMSO from 5-HMF

and water and formation of toxic side-products via degradation of DMSO limit the use of it. (Tong et al. 2010, p. 5) Luo et al. (2019) listed multiple research, where γ-Valerolactone (GVL) were proven as successful solvent in the production of FUR (Luo et al. 2019, p. 17).

5 MATERIALS AND METHODS

5.1 The dataset used

The goal of this study was to accelerate catalyst development and optimization with scalable catalysts for simultaneous dehydration reaction of C5 and C6 sugars. Because of the lack of relevant datasets with simple catalysts, where electronic, atomic, and structural properties can be easily related to predicted responses, dataset for hydrogenation of 5-Ethoxymethylfurfural were used to implement and compare the variable selection and modeling techniques. This reaction resembles the dehydration reaction of 5-HMF and FUR.

Dataset from Ras et al. (2009) consists of conversion and selectivity as responses in three different temperatures: 80, 100 and 120 °C. In the study, 8 different main metals (Au, Cu, Ir, Ni, Pd, Pt, Rh, Ru) and 6 promoters (Bi, Cr, Fe, Na, Sn, W) were used as catalysts. Al_2O_3 support was used in all observations. Main metal had a loading of 1 wt% and promoter 10 mol% related to main metal. Feedstock composition was not changed in the studied dataset. Each main metal was tested with each promoter in three different temperatures leading in 144 observations. Two solvents were used in the study: diethyl carbonate solvent and 1,4-dioxane solvent. Conversion and selectivity with both solvents were kept as different responses leading in four response variables: conversion with diethyl carbonate solvent (C), selectivity with diethyl carbonate solvent (S), conversion with 1,4-dioxane solvent (C1) and selectivity with 1,4-dioxane solvent (S1). Because of this, the solvent was not considered as input feature. Ir/W observations were not available for the first solvent and because of that, Ir/W observations were removed in the preprocessing stage.

Database was collected into an Excel spreadsheet, where also the calculations of Slater orbitals' interaction terms and quadratic terms were performed. The data was then read to MATLAB®. In Figure 5, exploratory data analysis is given in form of scatter plots for each response dataset pairs and histograms for each response dataset in the diagonal can be seen. It is notable, that with 1,4-dioxane solvent, the dataset structures are poor for modeling; with conversion most of the observations are near zero, and either 100 or 0 with selectivity. This leads to response variables with low entropy (i.e. with non-uniform distribution).

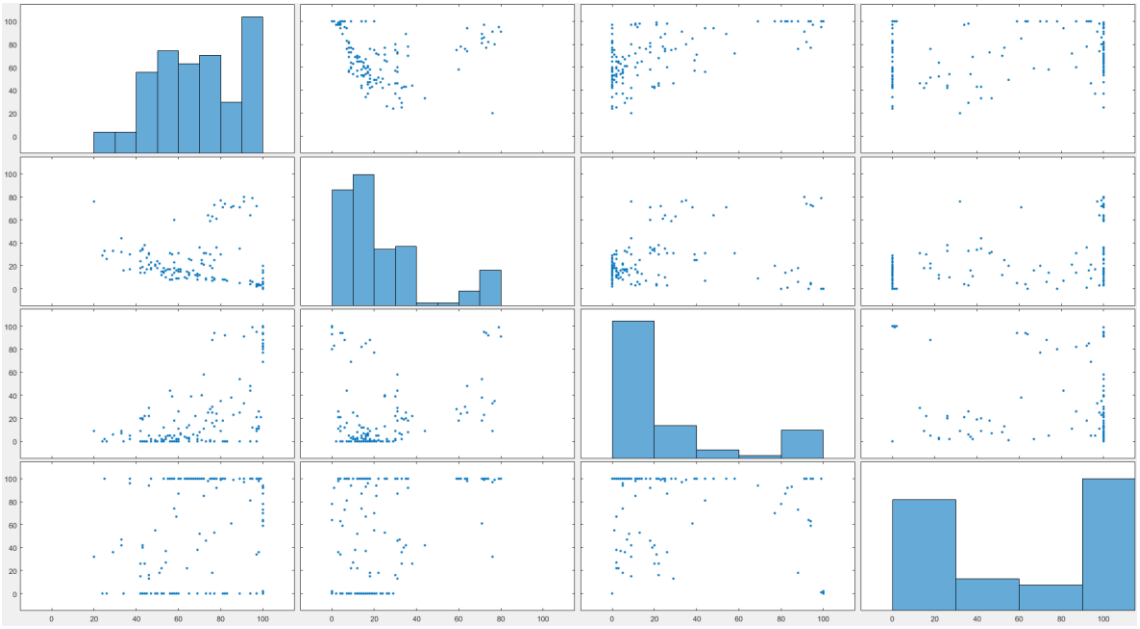

Figure 5. Scatter plot for each response dataset pairs and histograms in the diagonal for each response dataset. First row and column shows the scatter plots and histogram for conversion with diethyl carbonate solvent (C), second row and column for selectivity with diethyl carbonate solvent (S), third row and column for conversion with 1,4-dioxane solvent (C1), and fourth for selectivity with 1,4-dioxane solvent (S1).

5.2 MATLAB® tools

Preprocessing, variable selection, modeling, model validation, and statistical analysis was performed with MATLAB®. Statistics and Machine Learning Toolbox in MATLAB® was used in the study, which includes Regression Learner App. Regression Learner App was used to perform modeling and model validations. In the app, all models were used except for stepwise linear regression, because of the long calculation times. In MATLAB® automated hyperparameter optimization is available for several model structures and was used with ensemble tree model.

5.3 Inputs or descriptors

For the used elements (8 main metals + 6 promoters), input variables were added based on the periodic table (Gray 2020). Slater orbitals were also used with their interaction terms and quadratic terms. Also, some d-band center values were gathered, but because of the missing values, these variables were eliminated. The initial gathered database consisted of 65 variables and 4 responses. Categorical variables were converted into dummy variables, which lead to increase of variables. After the preprocessing in

MATLAB, there were 121 variables in dataset 1 and 141 variables in dataset 2 (dataset 1 does not consist Slater orbital interaction terms and quadratic terms).

5.3.1 Slater orbitals

Slater orbitals (STO) were calculated according to paper of Ras et al. (2012). Calculations resulted in 4 variables: r_{APEX}, $R(r)_{APEX}$, FWHH and SKEW. Ras et al. (2012) described these variables the following way: r_{APEX} is defined as "Distance of maximum probability of encountering a valence electron", $R(r)_{APEX}$ is the "maximum value of the probability distribution", FWHH is the "width of the probability distribution at the half height" and SKEW is the "measure for the asymmetry of the distribution".

5.4 Data preprocessing

Data were preprocessed in MATLAB. Variables and observations with missing values were removed. Categorical variables were converted into dummy variables. For each categorical variable, dummy variables equal to categories present were created. For example, for categorical variable "element" 14 dummy variables were created, since 8 main metals and 6 promoters are present. Each dummy variable defines each category with binary values 0 or 1: value 1 means true, and value 0 false (e.g. if dummy variable for Au is 1, main metal used is Au). In the preprocessing stage, also input data was normalized and variable names were added accordingly. In addition, data was formulated in a form, from where it was easily used in Regression Learner App and in MATLAB workspace.

5.5 Variable selection

Lasso and ridge regression, and their combination, elastic net, were used mainly as variable selection methods. Variables were also chosen based on "expert knowledge" i.e. data and variables were studied and based on the studies certain variables were removed or added. Cross-validation with K-fold value 10 was used with all regularization methods.

For ridge regression and elastic net, threshold value for coefficient values were identified to choose which variables to select to the subset. More relevant variables have bigger coefficient values and the variables with higher coefficient values than the threshold value

were kept in the subset. Threshold values were adjusted manually so that the gained variable subsets were kept in reasonable size. For lasso regularization, the variable selection is easier because the algorithm automatically makes the most of the coefficient values zero. Variables with coefficient values 0 were eliminated. Although, as mentioned in Section 2.4.3, with correlated variables (which are present in the used datasets), lasso algorithm can perform poorly, because it picks one of the correlated variables and ignores the rest.

`fitrlinear` function in MATLAB was used to perform variable selection with ridge and lasso regularization. Vector for different lambda values were created and the lambda value that minimizes MSE was chosen. `fitrlinear` function with cross-validation trains a model for each fold. Therefore, chosen variables were determined for each model (in this case 10). After variable subsets for each fold were determined, variables that occurred at least in half of the folds (in this case 5 or more) were chosen in the final subset. Variable selection with `fitrlinear` function and ridge regression were performed with Stochastic Gradient Descent solver. With lasso in contrast, Sparse Reconstruction by Separable Approximation (SpaRSA) was used. Least squares learner was used for both methods. MATLAB-algorithms are described more in detail in the MATLAB documentation.

`lasso` function was also used in MATLAB to perform lasso variable selection and elastic net variable selection. Elastic net variable selection was performed with alpha value 0.5. Alpha value determines the weight of lasso versus ridge optimization. Alpha value 1 represents lasso regression and alpha value close to zero represents ridge regression. For both methods, variable selection was performed with two different lambda values: lambda value that gives minimum mean squared error value (minMSE) and lambda value that is the largest lambda value one standard error away from minMSE lambda value (1SE). Thus, 1SE lambda value gives smaller variable subset with larger MSE value.

5.6 Modeling

Regression Learner App in MATLAB was mostly used to identify models. All models (except for stepwise linear regression) were used in the app including: linear models, decision tree models, ensemble models, SVM models, and GPR models. In MATLAB workspace ensemble tree model with hyperparameter optimization was used. Also, linear

models, PLSR and regularization algorithms were used to identify reference models. Finally, the modeling was done with subsets selected by variable selection algorithm, and also at some extent with manually chosen variables.

5.7 Model validation

Cross-validation were used as model validation method. The K-fold number of 10 was used in both modeling and variable selection. The effect of K-fold number was also studied. With higher K-fold values, more optimistic results are gained and with smaller K-fold values vice versa. The right K-fold value should be chosen somewhere in between. RMSE value and R-squared value for both training data and test data were used to estimate model performance.

6 RESULTS AND DISCUSSION

6.1 Modeling without variable selection

First, initial reference models were identified without variable selection. Some values in the results were not able to be calculated and are thus missing. In Table 1, the results of reference models are given for conversion C. It can be seen that good results (R^2 value 0.96 for training set and 0.86 test set) can be achieved with Quadratic SVM model without variable selection. With other methods, R^2 value with the training set is approximately 0.7 and 0.6 for the test data.

Table 1. Results with reference models for response C.

Method	RMSE (Train)	R^2 (Train)	RMSE (Test)	R^2 (Test)	Database	Input variables
Linear model without CV (`fitlm`)	11.68	0.72	-	-	1	121
Linear model with CV (`fitrlinear`)	11.44	0.71	13.34	0.60	1	121
Quadratic SVM with CV (best model with Regression Learner App)	4.46	0.96	7.97	0.86	1	121
PLSR with CV	11.09	0.72	12.28	-	1	121
Linear model without CV (`fitlm`)	11.68	0.72	-	-	2	141
Linear model with CV (`fitrlinear`)	11.57	0.70	13.54	0.59	2	141
Quadratic SVM with CV (best model with Regression Learner App)	4.45	0.96	8.11	0.86	2	141
PLSR with CV	11.09	0.72	12.39	-	2	141

The same results for selectivity S can be seen in Table 2. Good results were gained with fine tree model with R^2 values 0.95 for the training set and 0.91 for the test set. Also, good results were gained with boosted ensemble tree model. With other methods, the results vary between 0.52-0.64 with the training set and approximately between 0.41-0.55 with the test set.

Table 2. Results with reference models for response S.

Method	RMSE (Train)	R^2 (Train)	RMSE (Test)	R^2 (Test)	Database	Input variables
Linear model without CV (`fitlm`)	13.00	0.64	-	-	1	121
Linear model with CV (`fitrlinear`)	14.01	0.54	15.49	0.43	1	121
Fine tree with CV (best model with Regression Learner App)	4.37	0.95	6.28	0.91	1	121
PLSR with CV	12.34	0.64	14.08	-	1	121
Linear model without CV (`fitlm`)	13.00	0.64	-	-	2	141
Linear model with CV (`fitrlinear`)	14.21	0.52	15.75	0.41	2	141
Fine tree with CV (best model with Regression Learner App)	4.37	0.95	6.07	0.91	2	141
PLSR with CV	12.34	0.64	14.29	-	2	141

The results with reference models for response C1 are presented in Table 3. It can be seen, that the RMSE values are high compared to high R^2 values, which can be explained with challenging dataset structure. The best results were gained with quadratic SVM model with R^2 value of 0.97 for the training set and 0.92 for the test set. Good results were also gained with cubic SVM, fine tree and boosted ensemble tree models. With other methods, the R^2 value is between 0.79-0.84 for the training set and approximately 0.75-0.80 for the test set.

Table 3. Results with reference models for response C1.

Method	RMSE (Train)	R^2 (Train)	RMSE (Test)	R^2 (Test)	Database	Input variables
Linear model without CV (`fitlm`)	13.11	0.84	-	-	1	121
Linear model with CV (`fitrlinear`)	13.92	0.80	15.70	0.75	1	121
Quadratic SVM with CV (best model with Regression Learner App)	5.34	0.97	8.88	0.92	1	121
PLSR with CV	12.44	0.84	14.27	-	1	121
Linear model without CV (`fitlm`)	13.11	0.84	-	-	2	141
Linear model with CV (`fitrlinear`)	14.13	0.79	15.40	0.76	2	141
Quadratic SVM with CV (best model with Regression Learner App)	5.36	0.97	9.43	0.91	2	141
PLSR with CV	12.44	0.84	13.76	-	2	141

The results with reference models for response S1 can be seen in the Table 4. The best results are gained with boosted ensemble tree model with R^2 value 0.79 for the training set and 0.52 for the test set. With other methods, R^2 values vary between 0.33-0.41 for the training set and approximately 0.20-0.25 for the test set. It can be noticed, that worst results are gained for S1 when compared with other modeled response variables.

Table 4. Results with reference models for response S1.

Method	RMSE (Train)	R^2 (Train)	RMSE (Test)	R^2 (Test)	Database	Input variables
Linear model without CV (`fitlm`)	35.23	0.41	-	-	1	121
Linear model with CV (`fitrlinear`)	35.21	0.35	38.71	0.21	1	121
Boosted ensemble tree with CV (best model with Regression Learner App)	19.93	0.79	30.63	0.52	1	121
PLSR with CV	33.43	0.41	36.31	-	1	121
Linear model without CV (`fitlm`)	35.23	0.41	-	-	2	141
Linear model with CV (`fitrlinear`)	35.64	0.33	39.03	0.20	2	141
Boosted ensemble tree with CV (best model with Regression Learner App)	19.93	0.79	31.61	0.48	2	141
PLSR with CV	33.43	0.41	37.50	-	2	141

6.2 Regularization methods

Three different regularization methods were tested: lasso, ridge and elastic net. Models were identified without performing variable selection beforehand. In contrast to models in Section 6.1, models were identified with lesser number of variables by excluding the variables with coefficient values equal or close to zero. The results for conversion C can be seen in Table 5. The R^2 value with the training data is approximately 0.7 with all the used regularization algorithms and the test results are slightly worse. The same results for selectivity S can be seen in the Table 6. The R^2 value for the training set varies between 0.55-0.64. The test results are slightly worse. It can be noticed that ridge regression with `fitrlinear` function gives worse results than lasso and elastic net with `lasso` function. The same conclusion can be made for responses C1 and S1 in Tables 7 and 8. For conversion C1, the R^2 values are relatively high compared to the RMSE values, due to challenging dataset structure. The R^2 values vary between 0.81-0.84 for the training data and are slightly worse for the test data. As can be seen in Table 8, the results for selectivity S1 are the worst from the four responses, which can be explained with poor data structure.

The R^2 values vary between 0.34-0.41 for the training set and approximately 0.23-0.29 for the test set.

Table 5. Modeling results with regularization algorithms for response C.

Method	RMSE (Train)	R^2 (Train)	RMSE (Test)	R^2 (Test)	Database	Input variables
Lasso with CV	11.25	0.72	12.14	-	1	121
Ridge with CV (`fitrlinear`)	11.39	0.71	12.65	0.64	1	121
Elastic net with CV	11.15	0.72	12.37	-	1	121
Lasso with CV	11.30	0.71	12.25	-	2	141
Ridge with CV (`fitrlinear`)	11.41	0.71	12.41	0.65	2	141
Elastic net with CV	11.17	0.72	12.64	-	2	141

Table 6. Modeling results with regularization algorithms for response S.

Method	RMSE (Train)	R^2 (Train)	RMSE (Test)	R^2 (Test)	Database	Input variables
Lasso with CV	12.35	0.64	13.68	-	1	121
Ridge with CV (`fitrlinear`)	13.74	0.55	14.36	0.51	1	121
Elastic net with CV	12.35	0.64	13.49	-	1	121
Lasso with CV	12.86	0.61	13.95	-	2	141
Ridge with CV (`fitrlinear`)	13.83	0.55	14.56	0.50	2	141
Elastic net with CV	12.36	0.64	13.71	-	2	141

Table 7. Modeling results with regularization algorithms for response C1.

Method	RMSE (Train)	R^2 (Train)	RMSE (Test)	R^2 (Test)	Database	Input variables
Lasso with CV	12.49	0.84	13.51	-	1	121
Ridge with CV (`fitrlinear`)	13.62	0.81	14.42	0.79	1	121
Elastic net with CV	12.48	0.84	13.73	-	1	121
Lasso with CV	12.52	0.84	13.83	-	2	141
Ridge with CV (`fitrlinear`)	13.46	0.81	14.00	0.80	2	141
Elastic net with CV	12.46	0.84	14.05	-	2	141

Table 8. Modeling results with regularization algorithms for response S1.

Method	RMSE (Train)	R^2 (Train)	RMSE (Test)	R^2 (Test)	Database	Input variables
Lasso with CV	33.59	0.41	37.56	-	1	121
Ridge with CV (`fitrlinear`)	35.54	0.37	38.31	0.23	1	121
Elastic net with CV	33.64	0.41	37.18	-	1	121
Lasso with CV	33.57	0.41	37.06	-	2	141
Ridge with CV (`fitrlinear`)	35.41	0.34	37.67	0.26	2	141
Elastic net with CV	33.55	0.41	37.33	-	2	141

6.3 Modeling with variable selection

Finally, modeling was performed with variable selection. The best results for modeling of conversion C can be seen in Figure 6 and Figure 7. The best results were obtained with ensemble tree and cubic SVM models. Relatively good results were gained with R^2 value 0.93 for the training set and 0.86 for the test set with seven variables: temperature, Brinell hardness (M), electron affinity (M), covalent radius (M), neutron mass absorption (M), SKEW (M) and second lattice angle (P) (M refers to main metal and P refers to promoter).

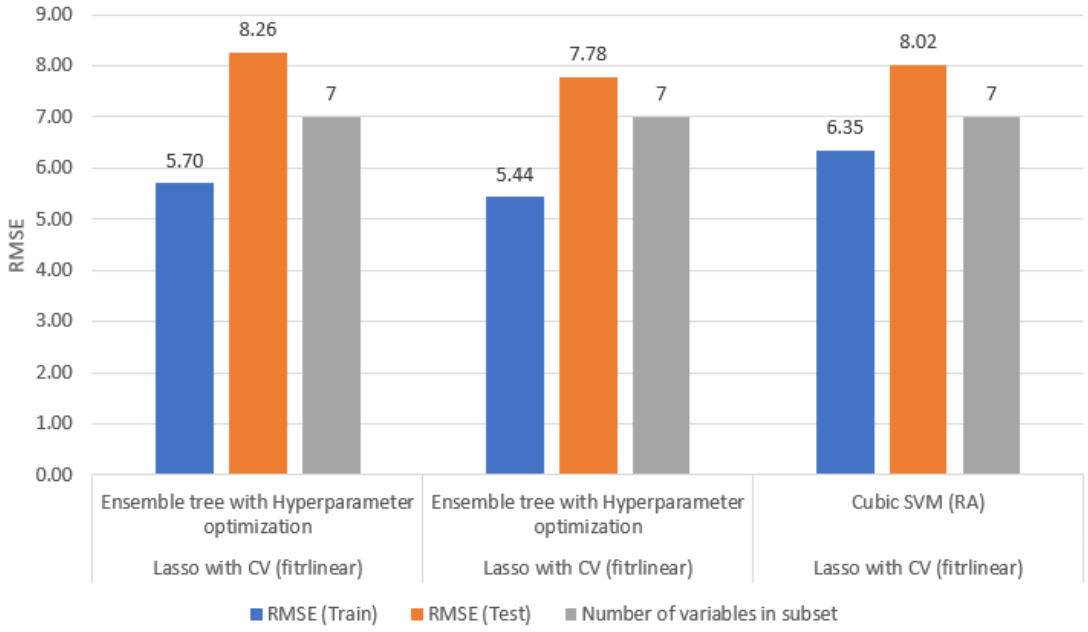

Figure 6. The RMSE values and the number of variables used for the best methods for conversion (C) predictions.

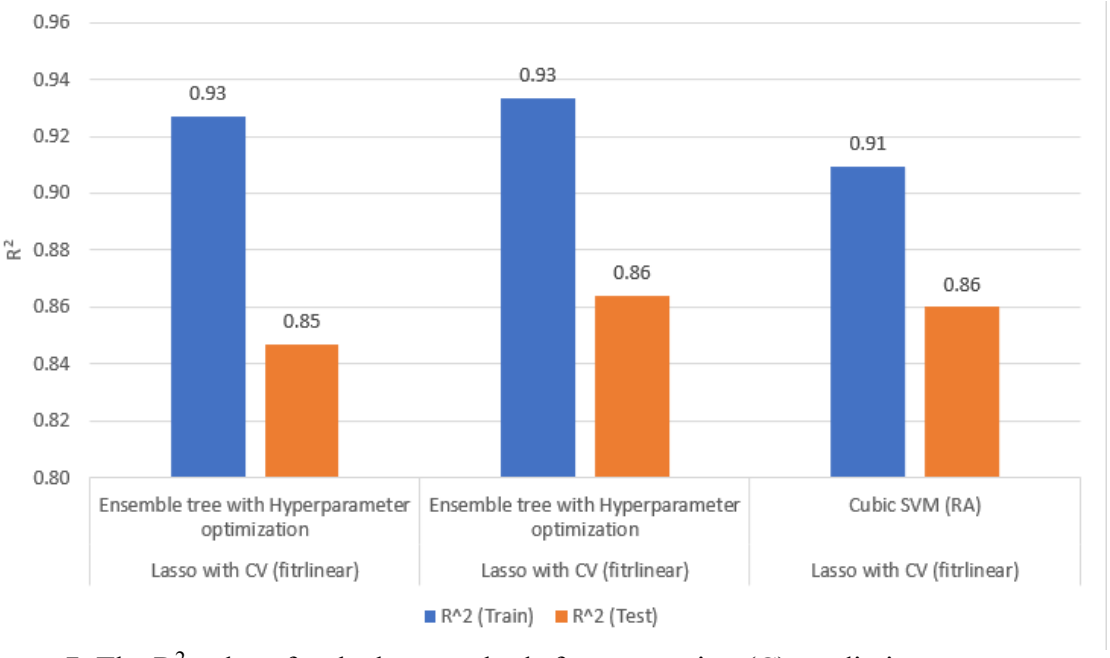

Figure 7. The R^2 values for the best methods for conversion (C) predictions.

The best results for modeling of selectivity S can be seen in Figure 8 and Figure 9. The best results were gained with fine tree, ensemble tree, SVM and GPR models. The best R^2 values in these cases were 0.94 for the training set and 0.92 for the test set. The models had only four variables: temperature, boiling point (M), bulk modulus (M) and electron affinity (M).

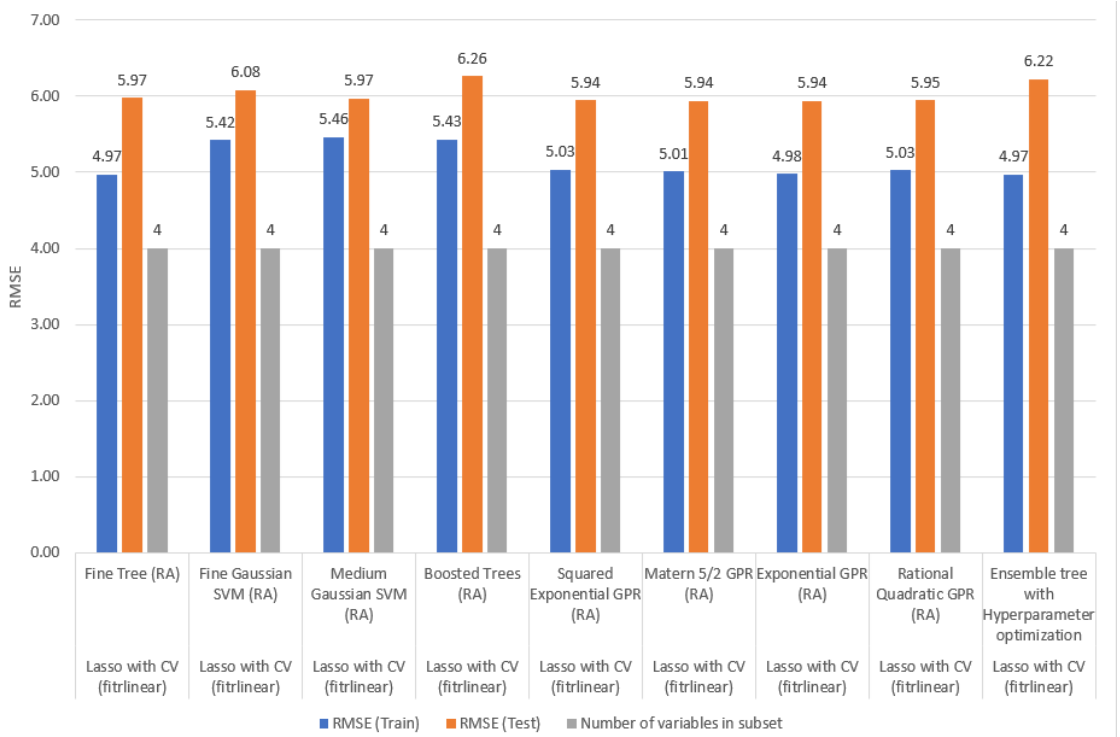

Figure 8. The RMSE values and the number of variables used for the best methods for selectivity (S) predictions.

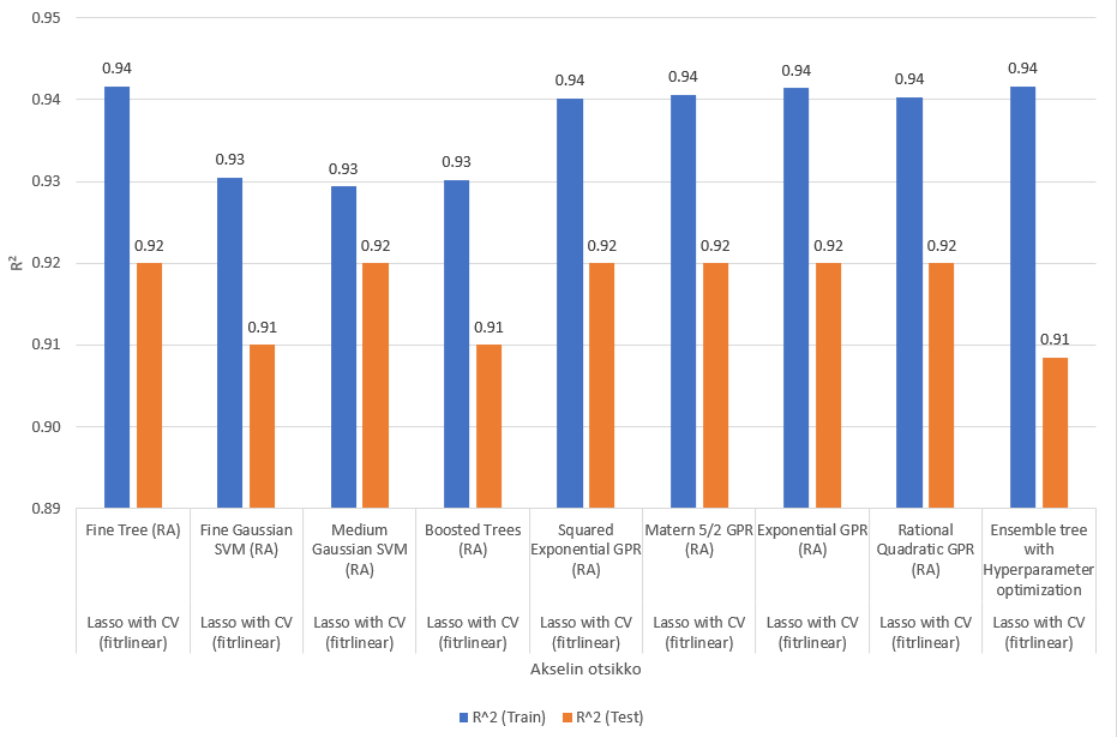

Figure 9. The R^2 values for the best methods for selectivity (S) predictions.

The best results for modeling of conversion C1 can be seen in Figure 10 and Figure 11. The same trend as with the reference models continues in these results; the RMSE value is relatively high compared to R^2 values, because of the challenging data structure. The best results were gained with ensemble tree, SVM and GPR models. Good results were gained with R^2 value 0.93 for both the training and test sets with only five variables:

dummy variable for Ir (M), dummy variable for Pd (M), temperature, Brinell hardness (M) and rAPEX (M).

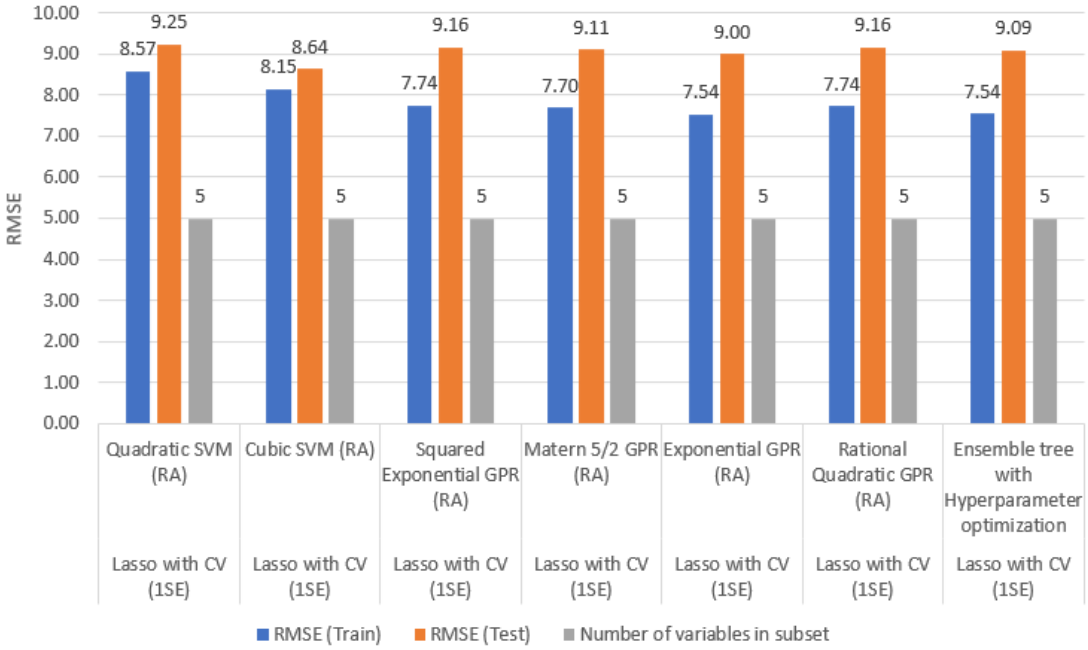

Figure 10. The RMSE values and the number of variables used for the best methods for conversion (C1) predictions.

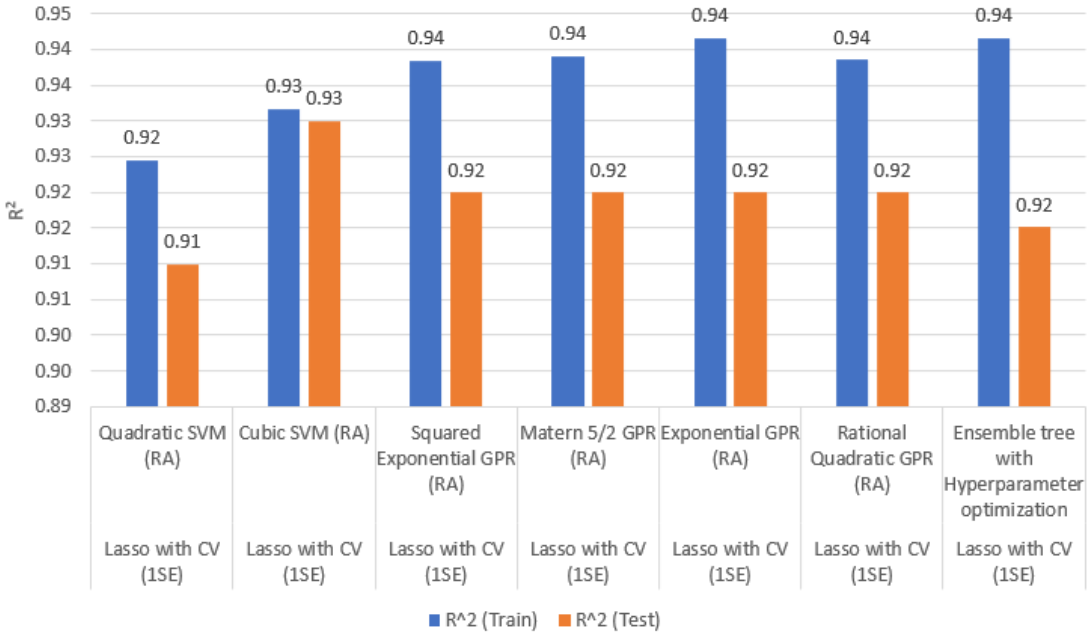

Figure 11. The R^2 values for the best methods for conversion (C1) predictions.

The best results for modeling of selectivity S1 can be seen in Figure 12 and Figure 13. The best results were gained with fine tree, ensemble tree, and GPR models. Poor results compared with other response variables were gained with R^2 value 0.67 for the training set and 0.55 for the test set. Six variables: temperature, density (M), boiling point (M), Brinell hardness (M), electrical conductivity (M) and neutron cross section (M), were

selected to these models. As mentioned earlier, poor results can be explained with bad data structure. Data is not evenly spread since the data values are mostly either 100 or 0.

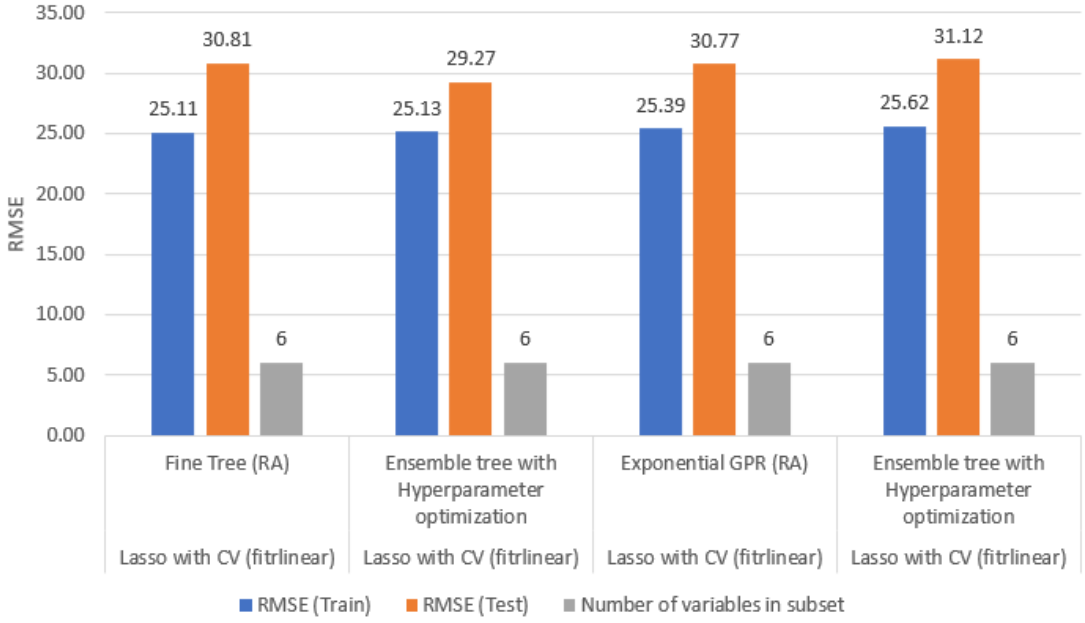

Figure 12. The RMSE values and the number of variables used for the best methods for selectivity (S1) predictions.

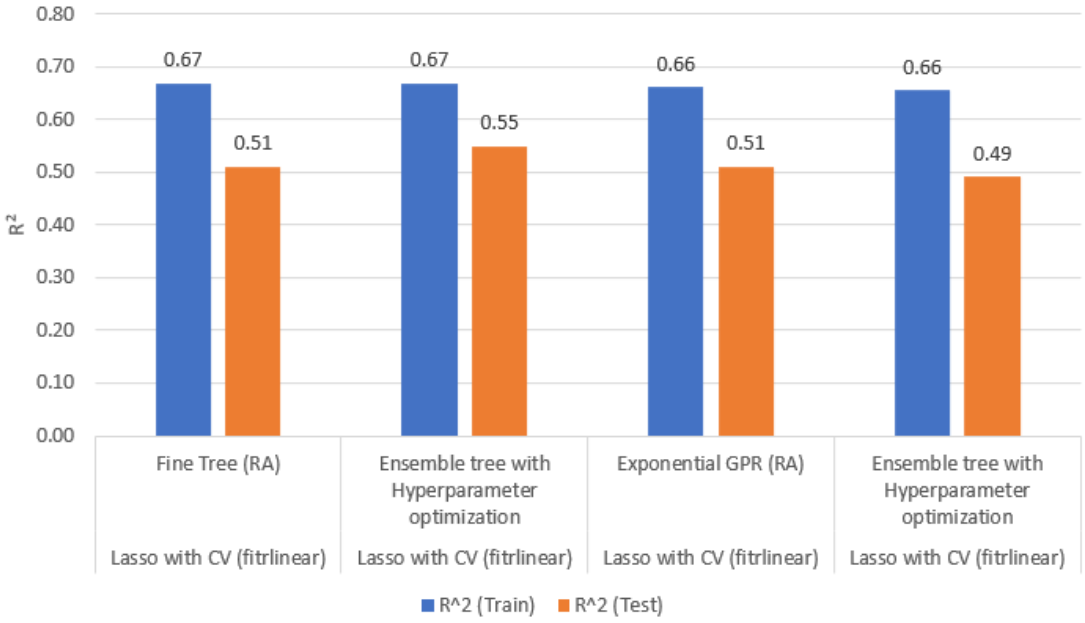

Figure 13. The R^2 values for the best methods for selectivity (S1) predictions.

6.4 Summary of modeling results

Summary of the best results for each response can be seen in Table 9. In general, ensemble tree with hyperparameter optimization and with lasso variable selection was the best combination. High R^2 values were gained for selectivity S and conversion C1 with small

variable subsets. C1's RMSE values are relatively high, since the prediction error for a small number of observations is considerable. Also, relatively high R^2 values were gained for conversion C with larger variable subset. The results for selectivity S1 were poor. Removing some observations with extreme value from the dataset before modeling could improve the results for S1.

Table 9. Summary of best results for each response.

Response	Number of inputs	RMSE (test)	R^2 (test)
C	7	7.78	0.86
S	4	5.94	0.92
C1	5	9.00	0.93
S1	6	29.27	0.55

6.5 Model validation

The effect of the number of folds in K-fold cross-validation was studied (see Section 5.7), and the results can be seen in Figure 14. The Regression Learner App was used to identify SVM cubic model with nine variables chosen with Lasso variable selection. K-fold values from 2 to 20 were used. It can be seen from the figure, that the performance stabilizes with small changes in R^2 values after K-fold value 5. A drop in R^2 value can be seen with K-fold value 13, which is possibly explained with poor dataset division. Poor R^2 values are gained with K-fold values from 2 to 4. Based on these results, the number of folds was set to 10 to be used in both variable selection phase and modeling phase.

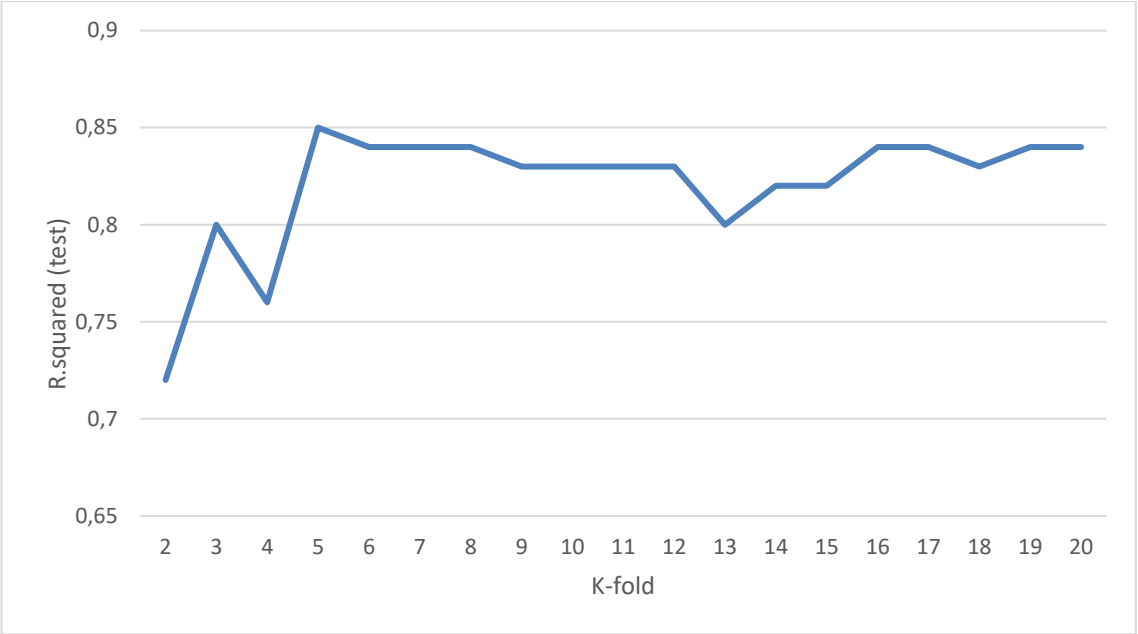

Figure 14. Results of K-fold testing with SVM cubic model.

6.6 Variable occurrences

The abundance of each variable with different methods were evaluated with four different regularization methods:

- elastic net (EN) with `lasso` function with alpha value 0.5 and minMSE lambda value,
- lasso with `lasso` function with minMSE lambda value (L1),
- lasso with `fitrlinear` function (L2) and
- ridge with `fitrlinear` function (R).

Each algorithm was executed 20 times resulting in total number of 80 iterations. For the elastic net variable selection, the following threshold values (see Section 5.5) were used: 2.3 for conversion C, 1.2 for selectivity S, 5 for conversion C1 and 3.5 for selectivity S1. Same values were used for both datasets. For ridge regression, the following threshold values were used: 2 for conversion C, 1.5 for selectivity S, 5 for conversion C1 and 5 for selectivity S1.

In Appendix 1, a list of variables used in the variable selection can be seen. The variables are assigned by numbers, which are used in the following result tables. Also, units and definitions are included in the tables. A total number of 141 variables are present. In Table

10 variable occurrences for conversion C with dataset 1 with four different regularization methods and their sum are represented. Eight most often occurring variables are shown in the table. It can be seen that temperature, Brinell hardness (M) and SKEW (M) occur in every chosen subset. It is also notable that no promoter variables are chosen. In Table 11 the same results are shown with dataset 2 (including Slater interactions and quadratic terms) for seven most occurring variables. It can be seen that the results differ from dataset 1 results. A strong Slater interaction term (for RAPEX and FWHH (M)) and a quadratic term (for SKEW (M)) are present. Also, one promoter variable (second lattice angle (P)) is present, which mainly describes the presence of promoter Bi, because the value is constant with every other promoter.

Table 10. Variable occurrences for conversion (C) with dataset 1 with four different regularization method and their sum.

	45.	57.	83.	4.	5.	71.	1.	65.
EN	20	20	20	20	20	0	20	0
L1	20	20	20	20	20	0	0	0
L2	20	20	20	0	0	20	0	7
R	20	20	20	20	0	20	10	20
SUM	80	80	80	60	40	40	30	27

45. Temperature, 57. Brinell hardness (M), 83. SKEW (M), 4. Dummy variable for Ni (M), 5. Dummy variable for Pd (M), 71. Covalent radius (M), 1. Dummy variable for Au (M), 65. Electron affinity (M).

Table 11. Variable occurrences for conversion (C) with dataset 2 with four different regularization method and their sum.

	45.	57.	91.	59.	6.	121.	87.
EN	20	20	0	0	20	0	0
L1	20	20	20	20	18	15	0
L2	20	20	20	20	0	20	8
R	20	20	20	0	0	0	20
SUM	80	80	60	40	38	35	28

45. Temperature, 57. Brinell hardness (M), 91. Interaction term for RAPEX and FWHH (M), 59. Bulk modulus (M), 6. Dummy variable for Pt (M), 121. Second lattice angle (P), 87. Quadratic term for SKEW (M)

In Table 12 variable occurrences for selectivity S with dataset 1 with four different regularization methods and their sum can be seen. 10 most often occurring variables are listed. It can be noticed that three promoter variables (dummy variable for Fe (P), dummy variable for group 8 in periodic table (P) and resistivity (P)) are present. In Table 13 the

same results are shown for dataset 2 for 10 most often occurring variables. The results are basically the same and no Slater interaction or quadratic terms are present.

Table 12. Variable occurrences for selectivity (S) with dataset 1 with four different regularization method and their sum.

	45.	59.	65.	50.	5.	6.	11.	17.	32.	117.
EN	20	0	20	0	20	20	20	20	20	0
L1	20	20	0	20	20	20	20	19	13	12
L2	20	20	20	20	0	0	0	0	0	10
R	20	20	20	19	0	0	0	0	0	0
SUM	80	60	60	59	40	40	40	39	33	22

45. Temperature, 59. Bulk modulus (M), 65. Electron affinity (M), 50. Boiling point (M), 5. Dummy variable for Pd (M), 6. Dummy variable for Pt (M), 11. Dummy variable for Fe (P), 17. Dummy variable for group 9 in periodic table (M), 32. Dummy variable for group 8 in periodic table (P), 117. Resistivity (P)

Table 13. Variable occurrences for selectivity (S) with dataset 2 with four different regularization method and their sum.

	45.	59.	65.	50.	5.	6.	11.	17.	32.	117.
EN	20	0	20	0	20	20	20	20	20	0
L1	20	20	0	20	20	20	20	19	16	14
L2	20	20	20	20	0	0	0	0	0	6
R	20	20	20	18	0	0	0	0	0	0
SUM	80	60	60	58	40	40	40	39	36	20

45. Temperature, 59. Bulk modulus (M), 65. Electron affinity (M), 50. Boiling point (M), 5. Dummy variable for Pd (M), 6. Dummy variable for Pt (M), 11. Dummy variable for Fe (P), 17. Dummy variable for group 9 in periodic table (M), 32. Dummy variable for group 8 in periodic table (P), 117. Resistivity (P)

In Table 14 variable occurrences for conversion C1 with dataset 1 with four different regularization methods and their sum is shown. 13 most often occurring variables are present. Two promoter variables are present (SKEW (P) and Speed of sound (P)). In Table 15 the same results are shown for 11 most often occurring variables for dataset 2. Some difference can be seen from dataset 1 results. One Slater interaction term (for RAPEX and FWHH (M)) and quadratic term (for rAPEX (M)) can be seen in the table.

Table 14. Variable occurrences for conversion (C1) with dataset 1 with four different regularization method and their sum.

	45.	5.	65.	83.	1.	3.	57.	66.	78.	80.	82.	131.	110.
EN	20	20	20	0	20	20	0	0	0	0	0	0	0
L1	20	20	0	20	20	20	20	0	0	20	0	20	20
L2	20	0	20	20	0	0	20	20	20	20	20	18	9
R	20	20	20	19	0	0	0	20	20	0	19	0	0
SUM	80	60	60	59	40	40	40	40	40	40	39	38	29

45. Temperature, 5. Dummy variable for Pd (M), 65. Electron affinity (M), 83. SKEW (M), 1. Dummy variable for Au (M), 3. Dummy variable for Ir (M), 57. Brinell hardness (M), 66. First ionization energy, 78. Neutron cross section (M), 80. rAPEX (M), 82. FWHH (M), 131. SKEW (P), 110. Speed of sound (P)

Table 15. Variable occurrences for conversion (C1) with dataset 2 with four different regularization method and their sum.

	45.	65.	1.	3.	5.	57.	78.	84.	91.	66.	110.
EN	20	20	20	20	20	0	0	0	0	0	0
L1	20	0	20	20	20	20	0	20	20	0	20
L2	20	20	0	0	0	20	20	20	20	20	10
R	20	20	0	0	0	0	20	0	0	11	0
SUM	80	60	40	40	40	40	40	40	40	31	30

45. Temperature, 65. Electron affinity, 1. Dummy variable for Au (M), 3. Dummy variable for Ir (M), 5. Dummy variable for Pd (M), 57. Brinell hardness (M), 78. Neutron cross section (M), 84. Quadratic term for rAPEX (M), 91. Interaction term for RAPEX and FWHH (M), 66. First ionization energy (M), 110. Speed of sound (P)

In Table 16 variable occurrences for selectivity S1 with dataset 1 with four different regularization methods and their sum is shown. 14 most often occurring variables are shown. Four promoter variables are chosen (dummy variable for Cr (P), first ionization energy (P), density (P) and Electronegativity (P)). In Table 17 the same results are shown for 14 variables with dataset 2. The results are nearly the same, when compared to the results for dataset 1. Slater interaction or quadratic terms are not seen.

Table 16. Variable occurrences for selectivity (S1) with dataset 1 with four different regularization method and their sum.

	45.	57.	1.	2.	3.	4.	6.	8.	10.	16.	17.	114.	81.	96.	112.
EN	20	19	20	20	20	20	20	20	20	20	20	0	0	0	0
L1	20	0	20	20	20	20	20	20	20	20	20	20	0	15	20
L2	20	20	0	0	0	0	0	0	0	0	0	20	19	16	11
R	20	20	0	0	0	0	0	0	0	0	0	0	14	0	0
SUM	80	59	40	40	40	40	40	40	40	40	40	40	33	31	31

45. Temperature, 57. Brinell hardness (M), 1. Dummy variable for Au (M), 2. Dummy variable for Cu (M), 3. Dummy variable for Ir (M), 4. Dummy variable for Ni (M), 6. Dummy variable for Pt (M), 8. Dummy variable for Rt (M), 10. Dummy variable for Cr (P), 16. Dummy variable for group 8 in the periodic table (M), 17. Dummy variable for group 9 in the periodic table (M), 114. First ionization energy (P), 81. RAPEX (M), 96. Density (P), 112. Electronegativity (P)

Table 17. Variable occurrences for selectivity (S1) with dataset 2 with four different regularization method and their sum.

	45.	57.	1.	2.	3.	4.	6.	8.	16.	17.	114.	10.	64.	96.
EN	20	10	20	20	20	20	20	20	20	20	0	19	0	0
L1	20	0	20	20	20	20	20	20	20	20	20	20	0	16
L2	20	20	0	0	0	0	0	0	0	0	20	0	20	17
R	20	20	0	0	0	0	0	0	0	0	0	0	13	0
SUM	80	50	40	40	40	40	40	40	40	40	40	39	33	33

45. Temperature, 57. Brinell hardness (M), 1. Dummy variable for Au (M), 2. Dummy variable for Cu (M), 3. Dummy variable for Ir (M), 4. Dummy variable for Ni (M), 6. Dummy variable for Pt (M), 8. Dummy variable for Rt (M), 10. Dummy variable for Cr (P), 16. Dummy variable for group 8 in the periodic table (M), 17. Dummy variable for group 9 in the periodic table (M), 114. First ionization energy (P), 10. Dummy variable for Cr (P), 64. Electronegativity (M), 96. Density (P)

6.7 Conclusions from variable occurrences

From the results above, the following conclusions can be made: (1) Promoter variables seem to be more relevant in the prediction of selectivity than conversion. (2) Slater variables and their interaction and quadratic terms seem to be more relevant for predicting conversion than selectivity. (3) Temperature was chosen for every variable subset. Also, Brinell hardness for main metal was found to be a strong variable. In general, main metal variables are much more relevant than promoter variables. It is also notable that `fitrlinear` function rarely chooses dummy variables in the variable subset. Therefore, the chosen variable subsets for different methods differ quite a bit. For comparison, the modeling was also performed with the two most important variables, namely the temperature and Brinell hardness (M), and the results can be seen in the Table 18. The results look surprising; The RMSE and R^2 values are only slightly worse than with the

variable subsets chosen by variable selection algorithms. In general for all responses, the best results were gained with GPR, ensemble tree and SVM models.

Table 18. The RMSE and R^2 values for the best models with two variables: Temperature and Brinell hardness (M).

Response	Number of inputs	RMSE (test)	R^2 (test)
C	2	8.43	0.84
S	2	5.95	0.92
C1	2	8.57	0.93
S1	2	29.39	0.55

6.8 Other issues

The datasets were scaled to a range from 0 to 1 and with the gained data, variable selection algorithms were tested. With `fitrlinear` function, significant difference between the scaled and non-scaled data could be seen. This could be, because `fitrlinear` function uses derivates to solve the model parameter identification problem. With `lasso` function there were no difference with scaled data. Modeling was also tested with non-scaled and scaled data and no difference could be seen.

For temperature, °C was used as the unit and for other temperature related variables K was used.

6.9 Recommendations for ML assisted catalyst development and future work

Good results were obtained with descriptors found in the literature. It was also shown, that fairly good results can be obtained with only two variables in this case. However, the performance with non-seen data should be further studied. Brinell hardness for main metal was found to have high predictive power with the used dataset. Promoter variables were not considered important with variable selection algorithms, which can be problematic in deriving optimal catalyst formulations where both the main metal and promoter need to be selected. With the current methods, the most relevant promoter variables need to be manually added into the variable subset. In general, best results were gained with ensemble tree, GPR and SVM models. From the studied variable selection

algorithms, lasso algorithm was the best. The studied methods can be easily implemented with actual datasets related to the dehydration reaction of C5 and C6 sugars.

Even though the modeling results were good, the variable selection methods were almost purely data-driven and the actual relevance of the variables cannot be guaranteed. Also, some of the values used in the dataset are computational and are likely to consist some amount of error (such as uncertainties in atomic radius, lattice angles etc.). Lasso algorithm was used with dataset consisting of highly correlated variables, which can lead to lasso algorithm picking one variable and ignoring the rest resulting in loss of possibly significant variables (see Section 2.4.3).

In the BioSPRINT project, experimental results related to the dehydration reaction of C5 and C6 sugars with simple metal catalysts will be obtained. The dataset should also include different feedstock and catalyst compositions, which possibly leads to new numerical variables. Therefore, some of the binary valued dummy variables can be excluded (e.g. dummy variables for main metal and promoter). In the future work, optimization should be studied with the goal of finding catalysts that give the maximum FOM values based on the model predictions. Extrapolation capabilities of the models needs to be studied and improved by excluding some temperature value or some catalyst composition from the used training set. Also, possibility of predicting multiple responses with single model needs to be considered. Other relevant descriptors should be found and added to the used dataset (e.g. d-band center). In addition, occurrences of variable combinations can also be studied to find variables that are strong together in the predictions. Instead of appointing threshold values for ridge regression and elastic net variable selection methods, lambda value could have been increased to penalize the number of variables in the subset more efficiently. This should be considered in the following works. In addition, different lambda values could have been tested to gain sufficient modeling results with minimal number of variables in the chosen subset. Elastic net was only used with alpha value 0.5 and could be tested with different alpha values as well. Hyperparameter optimization was only used with ensemble tree model, which could be applied also to other methods.

7 SUMMARY

In Section 2, ML was studied in general, where the focus was on different variable selection methods and modeling techniques, more specifically on data-driven modeling. Also, data collection, preprocessing, feature engineering, hyperparameter tuning and model validation were discussed. Section 3 covered modeling in catalysis focusing on ML in catalysis. Catalyst screening and selection, descriptor modeling and selection, and predictive modeling in catalysis were studied. The state of the art for dehydration of C5 and C6 sugars to produce 5-HMF and FUR was given in Section 4. Catalysts were also discussed with the focus on heterogeneous super solid acid catalysts. In addition, relevant solvents for the reaction were briefly studied. Dataset for hydrogenation of 5-ethoxymethylfurfural were used in the experimental part with the addition of catalyst descriptors found in the literature.

The process for the experimental part was presented in Section 4.2 and Section 5.1. Methods used in the experimental part were discussed in detail in Section 5, where data collection, preprocessing, variable selection, modeling and model validation were considered. The obtained dataset was introduced in Sections 5.1 and 5.3. The used descriptors were listed in Appendix 1 and discussed in Section 5.3. Variable selection methods readily available in MATLAB®, including regularization algorithms, were mainly used in the experiments. Reference models without variable selection were first identified. Secondly, regularization algorithms were used to identify models. Finally, models with variable subsets obtained with regularization algorithms were identified. The effect of cross-validation was studied in Section 6.5. Variable occurrences in variable selection were listed in Section 6.6, where Brinell hardness for main metal was found to have high predictive power. In Section 6.9, recommendations for ML assisted catalyst development and future work were discussed.

In general, good modeling results were gained with boosted ensemble tree methods, SVM methods and GPR methods. Lasso regression turned out to be the best variable selection method. Good results were gained with the descriptors found in the literature. It was also shown, that fairly good results can be gained with only two variables in the studied case. Promoter variables were not considered nearly as important as main metals with variable

selection algorithms. The studied methods can be easily implemented with actual datasets related to the dehydration reaction of C5 and C6 sugars.

REFERENCE LIST

Abdi, H., 2003. Partial Least Squares (PLS) Regression. In: Lewis-Beck M., Bryman A., Futing T. (Eds.) Encyclopedia of Social Sciences Research Methods. 1st Edition. Thousand Oaks: Sage. ISBN 978-0761923633. Available: https://personal.utdallas.edu/~herve/node4.html [Accessed 17.12.2020].

Agirrezabal-Telleria, I., Requies, J., Güemez, M.B., Arias, P.L., 2012. Furfural production from xylose + glucose feedings and simultaneous N2-stripping [online document]. Green Chem., 14 (11), p. 3132–3140. Available: https://doi.org/10.1039/C2GC36092F [Accessed 15.12.2020].

Agirrezabal-Telleria, I., Guo, Y., Hemmann, F., Arias, P.L., Kemnitz, E., 2014. Dehydration of xylose and glucose to furan derivatives using bifunctional partially hydroxylated MgF_2 catalysts and N_2-stripping [online document]. Catal. Sci. Technol. 4 (5), p. 1357–1368. Available: https://doi.org/10.1039/C4CY00129J [Accessed 15.12.2020].

Al-Mubaddel, F., Al-Zeghayer, Y.S., Al-Masry, W.A., Jibril, B.Y., 2006. Prospects of Using Solid Superacids as Catalysts in Petrochemical Industry [online document]. ChemInform, 37 (42). Available: https://doi.org/10.1002/chin.200642271 [Accessed 15.12.2020].

Artyushkova K., Pylypenko S., Olson TS., Fulghum JE., Atanassov P., 2008. Predictive Modeling of Electrocatalyst Structure Based on Structure-to-Property Correlations of X-ray Photoelectron Spectroscopic and Electrochemical Measurements [online document]. Langmuir 24 (16), p. 9082-9088. Available: https://doi.org/10.1021/la801089m [Accessed 14.12.2020].

Back, S., Yoon, J., Tian, N., Zhong, W., Tran, K., Ulissi, Z.W., 2019. Convolutional Neural Network of Atomic Surface Structures To Predict Binding Energies for High-Throughput Screening of Catalysts [online document]. J. Phys. Chem. Lett. 10 (15), p. 4401–4408. Available: https://doi.org/10.1021/acs.jpclett.9b01428 [Accessed 17.12.2020].

Bates, D.M., Watts, D.G., 1988. Nonlinear regression analysis: Its applications. New York: Wiley. ISBN 0471-816434

Baty F., Ritz C., Charles S., Brutsche M., Flandrois J.-P., Delignette-Muller M.-L., 2015. A Toolbox for Nonlinear Regression in R: The Package nlstools [online document]. In: J. Stat. Soft., 66 (5), p. 1-21. Available: https://doi.org/10.18637/jss.v066.i05 [Accessed 1.12.2020].

Baumes, L., Farrusseng, D., Lengliz, M., Mirodatos, C., 2004. Using Artificial Neural Networks to Boost High-throughput Discovery in Heterogeneous Catalysis [online document]. QSAR Comb. Sci. 23 (9), p. 767–778. Available: https://doi.org/10.1002/qsar.200430900 [Accessed 17.12.2020].

Baumes, L., Serra, J., Serna, P., Corma, A., 2006. Support vector machines for predictive modeling in heterogeneous catalysis: a comprehensive introduction and overfitting investigation based on two real applications [online document]. J. Comb. Chem., 8 (4), p. 583-596. Available: https://doi.org/10.1021/cc050093m [Accessed 14.12.2020].

Beckers, J., Clerc, F., Blank, J.H., Rothenberg, G., 2008. Selective Hydrogen Oxidation Catalysts via Genetic Algorithms [online document]. Adv. Synth. Catal., 350 (14-15), p. 2237-2249. Available: https://doi.org/10.1002/adsc.200800374 [Accessed 14.12.2020].

Behler, J., 2015. Constructing high-dimensional neural network potentials: A tutorial review [online document]. International Journal of Quantum Chemistry, 115 (16), p. 1032–1050. Available: https://doi.org/10.1002/qua.24890 [Accessed 14.12.2020].

Bernal, H.G., Bernazzani, L., Galletti, A.M.R., 2014. Furfural from corn stover hemicelluloses. A mineral acid-free approach [online document]. Green Chem., 16 (8), p. 3734–3740. Available: https://doi.org/10.1039/C4GC00450G [Accessed 15.12.2020].

BioSPRINT, 2019. BioSPRINT - Biorefining of sugars via Process Intensification, Sealed proposal. Unpublished.

Brockherde, F., Vogt, L., Li, L., Tuckerman, M.E., Burke, K., Müller, K.R., 2017. Bypassing the Kohn-Sham equations with machine learning. Nat Commun., 8 (1), p. 1-10. Available: https://doi.org/10.1038/s41467-017-00839-3 [Accessed 14.12.2020].

Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., Walsh, A., 2018. Machine learning for molecular and materials science [online document]. Nature, 559 (7715), p. 547–555. Available: https://doi.org/10.1038/s41586-018-0337-2 [Accessed 1.12.2020].

Calle-Vallejo, F., Loffreda, D., Koper, M. T., Sautet, P., 2015. Introducing structural sensitivity into adsorption-energy scaling relations by means of coordination numbers [online document]. Nature chemistry, 7 (5), p. 403–410. Available: https://doi.org/10.1038/nchem.2226 [Accessed 15.12.2020].

Chheda, J.N., Roman-Leshkov, Y., Dumesic, J.A., 2007. Production of 5-hydroxymethylfurfural and furfural by dehydration of biomass-derived mono- and poly-saccharides [online document]. Green Chem., 9 (4), p. 342-350. Available: https://doi.org/10.1039/B611568C [Accessed 15.12.2020].

Choudhary, V., Pinar, A.B., Sandler, S.I., Vlachos, D.G., Lobo, R.F., 2011. Xylose Isomerization to Xylulose and its Dehydration to Furfural in Aqueous Media [online document]. ACS Catal., 1 (12), p. 1724–1728. Available: https://doi.org/10.1021/cs200461t [Accessed 17.12.2020].

Corma, A., Serra, J.M., Argente, E., Botti, V., Valero, S., 2002. Application of Artificial Neural Networks to Combinatorial Catalysis: Modeling and Predicting ODHE Catalysts [online document]. ChemPhysChem, 3 (11), p. 939-945. Available: https://doi.org/10.1002/1439-7641(20021115)3:11%3C939::AID-CPHC939%3E3.0.CO;2-E [Accessed 14.12.2020].

Cottier, L., Descotes, G., 1991. 5-Hydroxymethylfurfural syntheses and chemical transformations [online document]. Trends in Heterocyclic Chemistry, 2, p. 233-248. Available: http://www.researchtrends.net/tia/abstract.asp?in=0&vn=2&tid=3&aid=4635&pub=1991&type=3 [Accessed 15.12.2020].

Cundari, T.R., Deng, J., Zhao, Y., 2001. Design of a Propane Ammoxidation Catalyst Using Artificial Neural Networks and Genetic Algorithms [online document]. Ind. Eng. Chem. Res. 40 (23), p. 5475–5480. Available: https://doi.org/10.1021/ie010316v [Accessed 1.12.2020].

Dacquin, J.-P., Cross, H.E., Brown, D.R., Duren, T., Williams, J.J., Lee, A.F., Wilson, K., 2010. Interdependent lateral interactions, hydrophobicity and acid strength and their influence on the catalytic activity of nanoporous sulfonic acid silicas [online document]. Green Chem., 12 (8), p. 1383-1391. Available: https://doi.org/10.1039/C0GC00045K [Accessed 15.12.2020].

Danon, B., Marcotullio, G., de Jong, W., 2014. Mechanistic and kinetic aspects of pentose dehydration towards furfural in aqueous media employing homogeneous catalysis [online document]. Green Chem., 16 (1), p. 39–54. Available: https://doi.org/10.1039/C3GC41351A [Accessed 17.12.2020].

Davis, L., 1991. Handbook of genetic algorithms. New York: Van Nostrand Reinhold, 385 p. ISBN 0-442-00173-8.

Dev, A., Srivastava, A.K., Karmakar, S., 2018. Chapter 12 - New Generation Hybrid Nanobiocatalysts: The Catalysis Redefined. In: Mustansar Hussain, C. (Ed.), Handbook of Nanomaterials for Industrial Applications. Elsevier, p. 217–231. ISBN 978-0-12-813351-4

Dhepe, P.L., Sahu, R., 2010. A solid-acid-based process for the conversion of hemicellulose [online document]. Green Chem., 12 (12), p. 2153–2156. Available: https://doi.org/10.1039/C004128A [Accessed 17.12.2020]

Dragoi, E.N., Horoba, C.A., Mamaliga, I., Curteanu, S., 2014. Grey and black-box modelling based on neural networks and artificial immune systems applied to solid dissolution by rotating disc method [online document]. Chemical Engineering and Processing: Process Intensification, 82, p. 173–184. Available: https://doi.org/10.1016/j.cep.2014.06.005 [Accessed 14.12.2020].

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern Classification. 2nd edition. USA: John Wiley & Sons. ISBN 978-0-471-05669-0

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., Adams, R.P., 2015. Convolutional Networks on Graphs for

Learning Molecular Fingerprints [online document]. Available: https://arxiv.org/abs/1509.09292 [Accessed 15.12.2020].

Farrusseng, D., Klanner, C., Baumes, L., Lengliz, M., Mirodatos, C., Schüth, F., 2005. Design of Discovery Libraries for Solids Based on QSAR Models [online document]. QSAR Comb. Sci., 24, p. 78-93. Available: https://doi.org/10.1002/qsar.200420066 [Accessed 14.12.2020].

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent [online document]. J. Stat. Soft., 33 (1), p. 1-22. Available: https://doi.org/10.18637/jss.v033.i01 [Accessed 1.12.2020].

Fusaro, M.B., Chagnault, V., Postel, D., 2015. Reactivity of d-fructose and d-xylose in acidic media in homogeneous phases [online document]. Carbohydrate Research, 409, p. 9–19. Available: https://doi.org/10.1016/j.carres.2015.03.012 [Accessed 17.12.2020].

Ghiringhelli, L.M., Vybiral, J., Levchenko, S.V., Draxl, C., Scheffler, M., 2015. Big Data of Materials Science: Critical Role of the Descriptor [online document]. Phys. Rev. Lett., 114 (10), p. 105503-1-105503-5. Available: https://doi.org/10.1103/PhysRevLett.114.105503 [Accessed 14.12.2020].

Gillespie, R.J., 1968. Fluorosulfuric acid and related superacid media [online document]. Acc. Chem. Res. 1 (7), p. 202–209. Available: https://doi.org/10.1021/ar50007a002 [Accessed 18.12.2020].

Goldberg, D.E., 1989. Genetic algorithms in search, optimization, and machine learning. USA: Addison-Wesley, 412 p. ISBN 0-201-15767-5

Goldsmith, B.R., Peters, B., Johnson, J.K., Gates, B.C., Scott, S.L., 2017. Beyond Ordered Materials: Understanding Catalytic Sites on Amorphous Solids. ACS Catal. 7 (11), p. 7543–7557. Available: https://doi.org/10.1021/acscatal.7b01767 [Accessed 18.12.2020].

Goldsmith, B.R., Esterhuizen, J., Liu, J., Bartel, C.J., Sutton, C., 2018. Machine learning for heterogeneous catalyst design and discovery [online document]. AIChE J., 64 (7), p. 2311–2323. Available: https://doi.org/10.1002/aic.16198 [Accessed 1.12.2020].

Gray, T., 2020. Periodic table [online document]. Available from: https://periodictable.com/ [Accessed 1.12.2020].

Guyon, I., Elisseeff, A., 2003. An Introduction to Variable and Feature Selection [online document]. J. Mach. Learn. Res., 3, p. 1157-1182. Available: https://dl.acm.org/doi/10.5555/944919.944968 [Accessed 18.12.2020].

Harper, K.C., Vilardi, S.C., Sigman, M.S., 2013. Prediction of Catalyst and Substrate Performance in the Enantioselective Propargylation of Aliphatic Ketones by a Multidimensional Model of Steric Effects [online document]. J. Am. Chem. Soc., 135 (7), p. 2482–2485. Available: https://doi.org/10.1021/ja4001807 [Accessed 15.12.2020].

Harris, J.F., Feather, M.S., 1974. Intramolecular C-2 -> C-1 Hydrogen Transfer Reactions during the Conversion of Aldoses to 2-Furaldehydes [online document]. J. Org. Chem., 39 (5), p. 724-725. Available: https://pubs.acs.org/doi/pdf/10.1021/jo00919a036?casa_token=EiwKpSuqK0oAAAAA:VW8DoGnc9whVTgMwZD8FOny0Xr-xyol5aiXTE1Fp3Y65am-FXh9IxTXqHR5WIXbhWIiTkgIqHmHInzg [Accessed 18.12.2020].

Hattori, T., Kito, S., 1995. Neural network as a tool for catalyst development [online document]. Catalysis Today, 23 (4), p. 347–355. Available: https://doi.org/10.1016/0920-5861(94)00148-U [Accessed 15.12.2020].

Haykin, S., 2009. Neural networks and Learning Machines. 3rd Edition. New York: Prentice Hall, 906 p. ISBN 978-0-13-147139-9

Henon, E., Bohr, F., Sokolowski-Gomez, N., Caralp, F., 2003. Degradation of three oxygenated alkoxy radicals of atmospheric interest: $HOCH_2O^{\cdot}$, $CH_3OCH_2O^{\cdot}$, $CH_3OCH_2OCH_2O^{\cdot}$. RRKM theoretical study of the β-C–H bond scission and the 1,6-isomerisation kinetics [online document]. Phys. Chem. Chem. Phys., 5 (24), p. 5431–5437. Available: https://doi.org/10.1039/B313251J [Accessed 17.12.2020].

Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems [online document]. Technometrics 12 (1), p. 55–67. Available: https://doi.org/10.1080/00401706.1970.10488634 [Accessed 1.12.2020].

Hui, W., Zhou, Y., Dong, Y., Cao, Z.-J., He, F.-Q., Cai, M.-Z., Tao, D.-J., 2019. Efficient hydrolysis of hemicellulose to furfural by novel superacid SO_4H-functionalized ionic liquids [online document]. Green Energy & Environment, 4 (1), p. 49–55. Available: https://doi.org/10.1016/j.gee.2018.06.002 [Accessed 15.12.2020].

Ishikawa, S., Zhang, Z., Ueda, W., 2018. Unit Synthesis Approach for Creating High Dimensionally Structured Complex Metal Oxides as Catalysts for Selective Oxidations [online document]. ACS Catal. 8 (4), p. 2935–2943. Available: https://doi.org/10.1021/acscatal.7b02244 [Accessed 14.12.2020].

Jing, Q., Lü, X., 2007. Kinetics of Non-catalyzed Decomposition of D-xylose in High Temperature Liquid [online document]. Chinese Journal of Chemical Engineering, 15 (5), p. 666–669. Available: https://doi.org/10.1016/S1004-9541(07)60143-8 [Accessed 15.12.2020].

Jones, R.O., 2015. Density functional theory: Its origins, rise to prominence, and future [online document]. Rev. Mod. Phys., 87 (3), p. 897. Available: https://doi.org/10.1103/RevModPhys.87.897 [Accessed 14.12.2020].

Kalz, K.F., Kraehnert, R., Dvoyashkin, M., Dittmeyer, R., Gläser, R., Krewer, U., Reuter, K., Grunwaldt, J.-D., 2017. Future Challenges in Heterogeneous Catalysis: Understanding Catalysts under Dynamic Reaction Conditions [online document]. ChemCatChem, 9 (1), p. 17–29. Available: https://doi.org/10.1002/cctc.201600996 [Accessed 14.12.2020].

Kitchin, J.R., 2018. Machine learning in catalysis [online document]. Nat. Catal., 1, p. 230–232. Available: https://doi.org/10.1038/s41929-018-0056-y [Accessed 1.12.2020].

Kito, S., Hattori, T., Murakami, Y., 1991. DETERMINATION OF SYNERGISTICALLY GENERATED ACID STRENGTH BY NEURAL NETWORK COMBINED WITH EXPERIMENT [online document]. Analytical Sciences, 7 (Supple), p. 761–764. Available: https://doi.org/10.2116/analsci.7.Supple_761 [Accessed 15.12.2020].

Kito, S., Hattori, T., Murakami, Y., 1992. Estimation of the acid strength of mixed oxides by a neural network [online document]. Ind. Eng. Chem. Res. 31 (3), p. 979–981. Available: https://doi.org/10.1021/ie00003a046 [Accessed 15.12.2020].

Kito, S., Hattori, T., Murakami, Y., 1994. Estimation of catalytic performance by neural network — product distribution in oxidative dehydrogenation of ethylbenzene [online document]. Applied Catalysis A: General, 114 (2), p. 173–178. Available: https://doi.org/10.1016/0926-860X(94)80169-X [Accessed 15.12.2020].

Klanner, C., Farrusseng, D., Baumes, L., Lengliz, M., Mirodatos, C., Schüth, F., 2004. The Development of Descriptors for Solids: Teaching "Catalytic Intuition" to a Computer. Angewandte Chemie International Edition, 43 (40), p. 5347-5349. Available: https://doi.org/10.1002/anie.200460731 [Accessed 14.12.2020].

Kondratenko, E.V., Schlüter, M., Baerns, M., Linke, D., Holena, M., 2015. Developing catalytic materials for the oxidative coupling of methane through statistical analysis of literature data [online document]. Catal. Sci. Technol., 5 (3), p. 1668-1677. Available: https://doi.org/10.1039/C4CY01443J [Accessed 15.12.2020].

Kuster, B.F.M., van der Baan, H.S., 1977. The influence of the initial and catalyst concentrations on the dehydration of D-fructose [online document]. Carbohydr. Res., 54 (2), p. 165-176. Available: https://doi.org/10.1016/S0008-6215(00)84806-5 [Accessed 18.12.2020].

Kuster, B.F.M., 1990. 5-Hydroxymethylfurfural (HMF). A Review Focussing on its Manufacture [online document]. Starch/Stärke, 42 (8), p. 314–321. Available: https://doi.org/10.1002/star.19900420808 [Accessed 15.12.2020].

Lange, J.-P., van der Heide, E., van Buijtenen, J., Price, R., 2012. Furfural—A Promising Platform for Lignocellulosic Biofuels [online document]. ChemSusChem, 5 (1), p. 150–166. Available: https://doi.org/10.1002/cssc.201100648 [Accessed 15.12.2020].

Li, H., Brothers, E.N., Hall, M.B., 2014. Computational Exploration of Alternative Catalysts for Olefin Purification: Cobalt and Copper Analogues Inspired by Nickel

Bis(dithiolene) Electrocatalysis [online document]. Inorg. Chem., 53 (18), p. 9679–9691. Available: https://doi.org/10.1021/ic5011538 [Accessed 15.12.2020].

Luo, Y., Li, Z., Li, X., Liu, X., Fan, J., Clark, J.H., Hu, C., 2019. The production of furfural directly from hemicellulose in lignocellulosic biomass: A review [online document]. Catalysis Today, 319, p. 14–24. Available: https://doi.org/10.1016/j.cattod.2018.06.042 [Accessed 15.12.2020].

Madaan, N., Shiju, N.R., Rothenberg, G., 2016. Predicting the performance of oxidation catalysts using descriptor models [online document]. Catal. Sci. Technol., 6 (1), p. 125–133. Available: https://doi.org/10.1039/C5CY00932D [Accessed 1.12.2020].

Maldonado, A.G., Rothenberg, G., 2010. Predictive modeling in homogeneous catalysis: a tutorial [online document]. Chem. Soc. Rev. 39 (6), p. 1891. Available: https://doi.org/10.1039/b921393g [Accessed 1.12.2020].

Mamman, A.S., Lee, J.-M., Kim, Y.-C., Hwang, I.T., Park, N.-J., Hwang, Y.K., Chang, J.-S., Hwang, J.-S., 2008. Furfural: Hemicellulose/xylose-derived biochemical [online document]. Biofuels, Bioproducts and Biorefining, 2 (5), p. 438–454. Available: https://doi.org/10.1002/bbb.95 [Accessed 15.12.2020].

Mariscal, R., Maireles-Torres, P., Ojeda, M., Sádaba, I., López Granados, M., 2016. Furfural: a renewable and versatile platform molecule for the synthesis of chemicals and fuels [online document]. Energy Environ. Sci., 9 (4), p. 1144–1189. Available: https://doi.org/10.1039/C5EE02666K [Accessed 15.12.2020].

Michalewicz, Z., 1996. Genetic algorithms + data structures = evolution programs. 3rd edition. Berlin: Springer, 387 p. ISBN 978-3-662-03315-9

Mitchell, T.M., 1997. Machine Learning [online document]. New York: McGraw-Hill, 414 p. ISBN 0-07-042807-7

Montemore, M.M., Medlin, J.W., 2014. Scaling relations between adsorption energies for computational screening and design of catalysts [online document]. Catal. Sci. Technol., 4 (11), p. 3748–3761. Available: https://doi.org/10.1039/C4CY00335G [Accessed 15.12.2020].

Montgomery, D.C., Peck, E.A., Vining, G.G., 2012. Introduction to Linear Regression Analysis, 5th Edition. 672 p. ISBN: 978-0-470-54281-1

Moreau, C., Durand, R., Razigade, S., Duhamet, J., Faugeras, P., Rivalier, P., Ros, P., Avignon, G., 1996. Dehydration of fructose to 5-hydroxymethylfurfural over H-mordenites [online document]. Applied Catalysis A: General, 145 (1-2), p. 211–224. Available: https://doi.org/10.1016/0926-860X(96)00136-6 [Accessed 17.12.2020].

Moye, C.J., Goldsack, R.J., 1966. Reaction of ketohexoses with acid in certain non-aqueous sugar solvents [online document]. Journal of Applied Chemistry, 16 (7), p. 206–208. Available: https://doi.org/10.1002/jctb.5010160703 [Accessed 18.12.2020].

Nebreda, A.P., 2019. Valuable monomers and oligomers from hemicelluloses [online document]. Thesis (PhD). Åbo Academi University. Available: http://urn.fi/URN:ISBN:978-952-12-3794-2 [Accessed 21.12.2020].

NIST, 2018. NIST Chemistry WebBook [online document]. Available: http://webbook.nist.gov/chemistry/ [Accessed 17.12.2020].

Nørskov, J.K., Bligaard, T., Rossmeisl, J., Christensen, C.H., 2009. Towards the computational design of solid catalysts [online document]. Nat. Chem., 1 (1), p. 37-46. Available: https://pubmed.ncbi.nlm.nih.gov/21378799 [Accessed 14.12.2020].

Ogutu, J.O., Schulz-Streeck, T., Piepho, H.-P., 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions [online document]. BMC Proceedings, 6 (S2):S10, p. 1-6. Available: https://doi.org/10.1186/1753-6561-6-S2-S10 [Accessed 18.12.2020].

Omata, K., 2011. Screening of New Additives of Active-Carbon-Supported Heteropoly Acid Catalyst for Friedel–Crafts Reaction by Gaussian Process Regression [online document]. Ind. Eng. Chem. Res., 50 (19), p. 10948–10954. Available: https://doi.org/10.1021/ie102477y [Accessed 18.12.2020].

Randall, T.D., 2016. An Introduction to Partial Least Squares Regression [online document]. Available: https://stats.idre.ucla.edu/wp-content/uploads/2016/02/pls.pdf [Accessed 1.12.2020].

Ras, E.-J., Louwerse, M.J., Rothenberg, G., 2012. New tricks by very old dogs: predicting the catalytic hydrogenation of HMF derivatives using Slater-type orbitals [online document]. Catal. Sci. Technol., 2 (12), p. 2456-2464. Available: https://doi.org/10.1039/c2cy20193c [Accessed 18.12.2020].

Ras, E.-J., Maisuls, S., Haesakkers, P., Gruter, G.-J., Rothenberg, G., 2009. Selective Hydrogenation of 5-Ethoxymethylfurfural over Alumina-Supported Heterogeneous Catalysts [online document]. Adv. Synth. Catal., 351 (18), p. 3175-3185. Available: https://doi.org/10.1002/adsc.200900526 [Accessed 15.12.2020].

Ras, E.-J., Rothenberg, G., 2014. Heterogeneous catalyst discovery using 21st century tools: a tutorial [online document]. RSC Adv., 4 (12), p. 5963-5974. Available: https://doi.org/10.1039/c3ra45852k [Accessed 1.12.2020].

Rhinehart, R.R., 2016. Nonlinear Regression Modeling for Engineering Applications: Modeling, Model Validation, and Enabling Design of Experiments. Chichester, UK; Hoboken, NJ: John Wiley & Sons, 400 p. ISBN 978-1-118-59796-5

Rinaldi, R., Schüth, F., 2009. Design of solid catalysts for the conversion of biomass [online document]. Energy Environ. Sci., 2 (6), p. 610-626. Available: https://doi.org/10.1039/B902668A [Accessed 18.12.2020].

Rosatella, A.A., Simeonov, S.P., Frade, R.F.M., Afonso, C.A.M., 2011. 5-Hydroxymethylfurfural (HMF) as a building block platform: Biological properties, synthesis and synthetic applications [online document]. Green Chem., 13 (4), p. 754-793. Available: https://doi.org/10.1039/C0GC00401D [Accessed 15.12.2020].

Rothenberg, G., 2008. Data mining in catalysis: Separating knowledge from garbage [online document]. Catalysis Today, 137 (1), p. 2–10. Available: https://doi.org/10.1016/j.cattod.2008.02.014 [Accessed 18.12.2020].

Sahu, R., Dhepe, P.L., 2012. A One-Pot Method for the Selective Conversion of Hemicellulose from Crop Waste into C5 Sugars and Furfural by Using Solid Acid Catalysts [online document]. ChemSusChem, 5 (4), p. 751–761. Available: https://doi.org/10.1002/cssc.201100448 [Accessed 15.12.2020].

Sarker, R., Kamruzzaman, J., Newton, C., 2003. EVOLUTIONARY OPTIMIZATION (EvOpt): A BRIEF REVIEW AND ANALYSIS [online document]. Int. J. Comp. Intel. Appl., 03 (04), p. 311–330. Available: https://doi.org/10.1142/S1469026803001051 [Accessed 17.12.2020].

Sasaki, M., Hamada, H., Kintaichi, Y., Ito, T., 1995. Application of a neural network to the analysis of catalytic reactions Analysis of NO decomposition over Cu/ZSM-5 zeolite [online document]. Applied Catalysis A: General, 132 (2), p. 261–270. Available: https://doi.org/10.1016/0926-860X(95)00171-9 [Accessed 15.12.2020].

Schütt, K., Arbabzadah, F., Chmiela, S., Müller, K.R., Tkatchenko, A., 2017. Quantum-chemical insights from deep tensor neural networks [online document]. Nat. Commun., 8 (1), p. 1-8. Available: https://doi.org/10.1038/ncomms13890 [Accessed 14.12. 2020].

Sehested, J., 2019. Industrial and Scientific Directions of Methanol Catalyst Development [online document]. J. Catal., 371, p. 368−375. Available: https://doi.org/10.1016/j.jcat.2019.02.002 [Accessed 14.12.2020].

Serra, J.M., Corma, A., Valero, S., Argente, E., Botti, V., 2007. Soft Computing Techniques Applied to Combinatorial Catalysis: A New Approach for the Discovery and Optimization of Catalytic Materials [online document]. QSAR Comb. Sci., 26 (1), p. 11-26. Available: https://doi.org/10.1002/qsar.200420051 [Accessed 14.12.2020].

Suzuki, K., Toyao, T., Maeno, Z., Takakusagi, S., Shimizu, K., Takigawa, I., 2019. Statistical Analysis and Discovery of Heterogeneous Catalysts Based on Machine Learning from Diverse Published Data [online document]. ChemCatChem 11 (18), p. 4537–4547. Available: https://doi.org/10.1002/cctc.201900971 [Accessed 1.12.2020].

Takagaki, A., Ohara, M., Nishimura, S., Ebitani, K., 2009. A one-pot reaction for biorefinery: combination of solid acid and base catalysts for direct production of 5-hydroxymethylfurfural from saccharides [online document]. Chem. Commun., Issue 41, p. 6276–6278. Available: https://doi.org/10.1039/B914087E [Accessed 17.12.2020].

Takigawa, I., Shimizu, K., Tsuda, K., Takakusagi, S., 2016. Machine-learning prediction of the d-band center for metals and bimetals [online document]. RSC Adv., 6 (58), p. 52587–52595. Available: https://doi.org/10.1039/C6RA04345C [Accessed 18.12.2020].

Tibshirani, R., 1996. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58 (1), p. 267–288. Available: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x [Accessed 1.12.2020].

Todeschini, R., Consonni, V., 2000. Handbook of Molecular Descriptors. Weinheim: Wiley-VCH, 667 p. ISBN 9783527613106

Tong, X., Ma, Y., Li, Y., 2010. Biomass into chemicals: Conversion of sugars to furan derivatives by catalytic processes [online document]. Applied Catalysis A: General, 385 (1-2), p. 1–13. Available: https://doi.org/10.1016/j.apcata.2010.06.049 [Accessed 15.12.2020].

Toyao, T., Maeno, Z., Takakusagi, S., Kamachi, T., Takigawa, I., Shimizu, K., 2020. Machine Learning for Catalysis Informatics: Recent Applications and Prospects [online document]. ACS Catal., 10 (3), p. 2260–2297. Available: https://doi.org/10.1021/acscatal.9b04186 [Accessed 18.12.2020].

Vassilev, S.V., Baxter, D., Andersen, L.K., Vassileva, C.G., Morgan, T.J., 2012. An overview of the organic and inorganic phase composition of biomass [online document]. Fuel, 94, p. 1–33. Available: https://doi.org/10.1016/j.fuel.2011.09.030 [Accessed 15.12.2020].

Weingarten, R., Tompsett, G.A., Conner, W.C., Huber, G.W., 2011. Design of solid acid catalysts for aqueous-phase dehydration of carbohydrates: The role of Lewis and Brønsted acid sites [online document]. Journal of Catalysis, 279 (1), p. 174–182. Available: https://doi.org/10.1016/j.jcat.2011.01.013 [Accessed 17.12.2020].

Werpy, T., Petersen, G., 2004. Top Value Added Chemicals from Biomass: Volume I -- Results of Screening for Potential Candidates from Sugars and Synthesis Gas [online document]. United States: US Department of Energy. DOE/GO-102004-1992. Available: https://doi.org/10.2172/15008859 [Accessed 18.12.2020].

Williams, C. K. I., 2003. Gaussian Processes. In: Arbib, M. A. (ed.) The Handbook of Brain Theory and Neural Networks. 2nd Edition. Cambridge, Mass: MIT Press, p. 466–470. ISBN 978-0-262-01197-6

Witten, I. H, Frank, E., Hall M. A., 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition. Burlington, MA: Morgan Kaufmann, 629 p. ISBN 978-0-12-374856-0

Zeitsch, K.J., 2000. The Chemistry and Technology of Furfural and Its Many By-Products. 1st edition. The Netherlands: Elsevier, 376 p. ISBN 9780444503510

Zhang, Z.C., 2013. Emerging Catalysis for 5-HMF Formation from Cellulosic Carbohydrates. In: Suib, S.L., (ed.). New and Future Developments in Catalysis. Amsterdam: Elsevier, p. 53–71. ISBN 978-0-444-53878-9.

Zhao, H., Holladay, J.E., Brown, H., Zhang, Z.C., 2007. Metal Chlorides in Ionic Liquid Solvents Convert Sugars to 5-Hydroxymethylfurfural [online document]. Science, 316 (5831), p. 1597-1600. Available: https://doi.org/10.1126/science.1141199 [Accessed 15.12.2020].

Zhu, S., Xue, Y., Guo, J., Cen, Y., Wang, J., Fan, W., 2016. Integrated Conversion of Hemicellulose and Furfural into γ-Valerolactone over Au/ZrO 2 Catalyst Combined with ZSM-5 [online document]. ACS Catal., 6 (3), p. 2035–2042. Available: https://doi.org/10.1021/acscatal.5b02882 [Accessed 17.12.2020].

Appendix 1. Variables used in the variable selection.

Table 1. Variables used in the variable selection, part 1. M refers to main metal and P to promoter.

Number	Variable	Unit	Definition
1.	Dummy variable for Au (M)	N/A	Explains the presence of Au as main metal.
2.	Dummy variable for Cu (M)	N/A	Explains the presence of Cu as main metal.
3.	Dummy variable for Ir (M)	N/A	Explains the presence of Ir as main metal.
4.	Dummy variable for Ni (M)	N/A	Explains the presence of Ni as main metal.
5.	Dummy variable for Pd (M)	N/A	Explains the presence of Pd as main metal.
6.	Dummy variable for Pt (M)	N/A	Explains the presence of Pt as main metal.
7.	Dummy variable for Rh (M)	N/A	Explains the presence of Rh as main metal.
8.	Dummy variable for Rt (M)	N/A	Explains the presence of Rt as main metal.
9.	Dummy variable for Bi (P)	N/A	Explains the presence of Bi as promoter.
10.	Dummy variable for Cr (P)	N/A	Explains the presence of Cr as promoter.
11.	Dummy variable for Fe (P)	N/A	Explains the presence of Fe as promoter.
12.	Dummy variable for Na (P)	N/A	Explains the presence of Na as promoter.
13.	Dummy variable for Sn (P)	N/A	Explains the presence of Sn as promoter.
14.	Dummy variable for W (P)	N/A	Explains the presence of W as promoter.
15.	Dummy variable for d-block in the periodic table (M)	N/A	Explains, if the main metal belongs to d-block in the periodic table.
16.	Dummy variable for group 8 in the periodic table (M)	N/A	Explains, if the main metal belongs to group 8 in the periodic table.
17.	Dummy variable for group 9 in the periodic table (M)	N/A	Explains, if the main metal belongs to group 9 in the periodic table.
18.	Dummy variable for group 10 in the periodic table (M)	N/A	Explains, if the main metal belongs to group 10 in the periodic table.
19.	Dummy variable for group 11 in the periodic table (M)	N/A	Explains, if the main metal belongs to group 11 in the periodic table.
20.	Dummy variable for period 4 in the periodic table (M)	N/A	Explains, if the main metal belongs to period 4 in the periodic table.
21.	Dummy variable for period 5 in the periodic table (M)	N/A	Explains, if the main metal belongs to period 5 in the periodic table.

Table 2. Variables used in the variable selection, part 2. M refers to main metal and P to promoter.

Number	Variable	Unit	Definition
22.	Dummy variable for period 6 in the periodic table (M)	N/A	Explains, if the main metal belongs to period 6 in the periodic table.
23.	Dummy variable for face-centered cubic crystalline structure (M)	N/A	Explains, if the main metal has face-centered cubic crystalline structure.
24.	Dummy variable for simple hexagonal crystalline structure (M)	N/A	Explains, if the main metal has simple hexagonal crystalline structure.
25.	Dummy variable for space group number 194 (M)	N/A	Explains, if the main metal has space group number 194.
26.	Dummy variable for space group number 225 (M)	N/A	Explains, if the main metal has space group number 194.
27.	Dummy variable for d-block in the periodic table (P)	N/A	Explains, if the promoter belongs to d-block in the periodic table.
28.	Dummy variable for p-block in the periodic table (P)	N/A	Explains, if the promoter belongs to p-block in the periodic table.
29.	Dummy variable for s-block in the periodic table (P)	N/A	Explains, if the promoter belongs to s-block in the periodic table.
30.	Dummy variable for group 1 in the periodic table (P)	N/A	Explains, if the promoter belongs to group 1 in the periodic table.
31.	Dummy variable for group 6 in the periodic table (P)	N/A	Explains, if the promoter belongs to group 6 in the periodic table.
32.	Dummy variable for group 8 in the periodic table (P)	N/A	Explains, if the promoter belongs to group 8 in the periodic table.
33.	Dummy variable for group 14 in the periodic table (P)	N/A	Explains, if the promoter belongs to group 14 in the periodic table.
34.	Dummy variable for group 15 in the periodic table (P)	N/A	Explains, if the promoter belongs to group 15 in the periodic table.
35.	Dummy variable for period 3 in the periodic table (P)	N/A	Explains, if the promoter belongs to period 3 in the periodic table.
36.	Dummy variable for period 4 in the periodic table (P)	N/A	Explains, if the promoter belongs to period 4 in the periodic table.
37.	Dummy variable for period 5 in the periodic table (P)	N/A	Explains, if the promoter belongs to period 5 in the periodic table.
38.	Dummy variable for period 6 in the periodic table (P)	N/A	Explains, if the promoter belongs to period 6 in the periodic table.
39.	Dummy variable for base-centered monoclinic crystalline structure (P)	N/A	Explains, if the promoter has base-centered monoclinic crystalline structure.
40.	Dummy variable for base-centered cubic crystalline structure (P)	N/A	Explains, if the promoter has base-centered cubic crystalline structure.
41.	Dummy variable for centered tetragonal crystalline structure (P)	N/A	Explains, if the promoter has centered tetragonal crystalline structure.
42.	Dummy variable for space group number 12 (P)	N/A	Explains, if the promoter has space group number 12.

Table 3. Variables used in the variable selection, part 3. M refers to main metal and P to promoter.

Number	Variable	Unit	Definition
43.	Dummy variable for space group number 141 (P)	N/A	Explains, if the promoter has space group number 141.
44.	Dummy variable for space group number 229 (P)	N/A	Explains, if the promoter has space group number 229.
45.	Temperature	°C	Temperature of experiment.
46.	Atomic number (M)	N/A	Atomic number of the element in periodic table.
47.	Atomic weight (M)	g/mol	Defines the weight of an atom.
48.	Density (M)	g/cm^3	Defines materials mass per unit volume.
49.	Melting point (M)	K	Defines the temperature value at which the element changes its phase from solid to liquid.
50.	Boiling point (M)	K	Defines the temperature value at which the element changes its phase from liquid to gas.
51.	Heat of fusion (M)	kJ/mol	The quantity of heat necessary to change a solid to a liquid without temperature change.
52.	Heat of vaporization (M)	kJ/mol	The quantity of heat necessary to change a liquid to a solid without temperature change.
53.	Specific heat capacity (M)	J/(kg*K)	The quantity of heat necessary for a given mass to produce a unit change in its temperature.
54.	Thermal conductivity (M)	W/(m*K)	A measure of materials ability to conduct heat.
55.	Thermal expansion (M)	K^{-1}	Defines materials ability to change its shape, area, volume, and density to a temperature change.
56.	Molar volume (M)	m^3/mol	Volume occupied by one mole of the substance at the given temperature and pressure.
57.	Brinell hardness (M)	MPa	Definition of materials hardness tested by applying pressure with indenter on the material.
58.	Mohs hardness (M)	N/A	Defines materials scratch resistance.
59.	Bulk modulus (M)	GPa	Defines materials resistance to compression.
60.	Shear modulus (M)	GPa	Describe materials response to shear stress.
61.	Young modulus (M)	GPa	Defines materials resistance to elastic changes.
62.	Speed of sound (M)	m/s	Defines how fast speed will travel in the material.

Table 4. Variables used in the variable selection, part 4. M refers to main metal and P to promoter.

Number	Variable	Unit	Definition
63.	Valence of ion (M)	N/A	Defines the number of electrons in the materials valence orbital.
64.	Electronegativity (M)	N/A	Defines atoms ability to attract a shared pair of electrons with another.
65.	Electron affinity (M)	kJ/mol	Defines the change in energy of a neutral atom, when an electron is added to the atom to form a negative ion.
66.	First ionization energy (M)	kJ/mol	The amount of energy needed to remove one electron from an atom.
67.	Second ionization energy (M)	kJ/mol	The amount of energy needed to remove two electrons from an atom.
68.	Electrical conductivity (M)	S/m	Defines materials ability to conduct electric current.
69.	Resistivity (M)	m*Ω	Defines materials ability to resist electric current.
70.	Atomic radius (M)	pm	Measure of the size of atoms in element.
71.	Covalent radius (M)	pm	Measure of the size of atom that forms part of one covalent bond.
72.	First lattice angle (M)	N/A	Defines first dimension's angle in unit cell that describes the crystal structure.
73.	Second lattice angle (M)	N/A	Defines second dimension's angle in unit cell that describes the crystal structure.
74.	Third lattice angle (M)	N/A	Defines third dimension's angle in unit cell that describes the crystal structure.
75.	First lattice constant (M)	pm	Defines first dimension's length in unit cell that describes the crystal structure.
76.	Second lattice constant (M)	pm	Defines second dimension's length in unit cell that describes the crystal structure.
77.	Third lattice constant (M)	pm	Defines third dimension's length in unit cell that describes the crystal structure.
78.	Neutron cross section (M)	b	Defines the likelihood of interaction between an incident neutron and a target nucleus.
79.	Neutron mass absorption (M)	m^2/kg	Thermal neutron mass absorption coefficient.
80.	STO variable rAPEX (M)	N/A	Distance of maximum probability of encountering a valence electron.

Table 5. Variables used in the variable selection, part 5. M refers to main metal and P to promoter.

Number	Variable	Unit	Definition
81.	STO variable RAPEX (M)	N/A	Maximum value of the probability distribution (i.e. STOs).
82.	STO variable FWHH (M)	N/A	Width of the probability distribution (i.e. STOs) at half height (half of the maximum).
83.	STO variable SKEW (M)	N/A	Measure for the asymmetry of the probability distribution (i.e. STOs).
84.	Quadratic term for rAPEX (M)	N/A	Quadratic term for rAPEX.
85.	Quadratic term for RAPEX (M)	N/A	Quadratic term for RAPEX.
86.	Quadratic term for FWHH (M)	N/A	Quadratic term for FWHH.
87.	Quadratic term for SKEW (M)	N/A	Quadratic term for SKEW.
88.	Interaction term for rAPEX and RAPEX (M)	N/A	Interaction term for rAPEX and RAPEX.
89.	Interaction term for rAPEX and FWHH (M)	N/A	Interaction term for rAPEX and FWHH.
90.	Interaction term for rAPEX and SKEW (M)	N/A	Interaction term for rAPEX and SKEW.
91.	Interaction term for RAPEX and FWHH (M)	N/A	Interaction term for RAPEX and FWHH.
92.	Interaction term for RAPEX and SKEW (M)	N/A	Interaction term for RAPEX and SKEW.
93.	Interaction term for FWHH and SKEW (M)	N/A	Interaction term for FWHH and SKEW.
94.	Atomic number (P)	N/A	Atomic number of the element in periodic table.
95.	Atomic weight (P)	g/mol	Defines the weight of an atom.
96.	Density (P)	g/cm^3	Defines materials mass per unit volume.
97.	Melting point (P)	K	Defines the temperature value at which the element changes its phase from solid to liquid.
98.	Boiling point (P)	K	Defines the temperature value at which the element changes its phase from liquid to gas.
99.	Heat of fusion (P)	kJ/mol	The quantity of heat necessary to change a solid to a liquid without temperature change.
100.	Heat of vaporization (P)	kJ/mol	The quantity of heat necessary to change a liquid to a solid without temperature change.
101.	Specific heat capacity (P)	J/(kg*K)	The quantity of heat necessary for a given mass to produce a unit change in its temperature.
102.	Thermal conductivity (P)	W/(m*K)	A measure of materials ability to conduct heat.

Table 6. Variables used in the variable selection, part 6. M refers to main metal and P to promoter.

Number	Variable	Unit	Definition
103.	Thermal expansion (P)	K^{-1}	Defines materials ability to change its shape, area, volume, and density to a temperature change.
104.	Molar volume (P)	m^3/mol	Volume occupied by one mole of the substance at the given temperature and pressure.
105.	Brinell hardness (P)	MPa	Definition of materials hardness tested by applying pressure with indenter on the material.
106.	Mohs hardness (P)	N/A	Defines materials scratch resistance.
107.	Bulk modulus (P)	GPa	Defines materials resistance to compression.
108.	Shear modulus (P)	GPa	Describe materials response to shear stress.
109.	Young modulus (P)	GPa	Defines materials resistance to elastic changes.
110.	Speed of sound (P)	m/s	Defines how fast speed will travel in the material.
111.	Valence of ion (P)	N/A	Defines the number of electrons in the materials valence orbital.
112.	Electronegativity (P)	N/A	Defines atoms ability to attract a shared pair of electrons with another.
113.	Electron affinity (P)	kJ/mol	Defines the change in energy of a neutral atom, when an electron is added to the atom to form a negative ion.
114.	First ionization energy (P)	kJ/mol	The amount of energy needed to remove one electron from an atom.
115.	Second ionization energy (P)	kJ/mol	The amount of energy needed to remove two electrons from an atom.
116.	Electrical conductivity (P)	S/m	Defines materials ability to conduct electric current.
117.	Resistivity (P)	m*Ω	Defines materials ability to resist electric current.
118.	Atomic radius (P)	pm	Measure of the size of atoms in element.
119.	Covalent radius (P)	pm	Measure of the size of atom that forms part of one covalent bond.
120.	First lattice angle (P)	N/A	Defines first dimension's angle in unit cell that describes the crystal structure.
121.	Second lattice angle (P)	N/A	Defines second dimension's angle in unit cell that describes the crystal structure.

Table 7. Variables used in the variable selection, part 7. M refers to main metal and P to promoter.

Number	Variable	Unit	Definition
122.	Third lattice angle (P)	N/A	Defines third dimension's angle in unit cell that describes the crystal structure.
123.	First lattice constant (P)	pm	Defines first dimension's length in unit cell that describes the crystal structure.
124.	Second lattice constant (P)	pm	Defines second dimension's length in unit cell that describes the crystal structure.
125.	Third lattice constant (P)	pm	Defines third dimension's length in unit cell that describes the crystal structure.
126.	Neutron cross section (P)	b	Defines the likelihood of interaction between an incident neutron and a target nucleus.
127.	Neutron mass absorption (P)	m^2/kg	Thermal neutron mass absorption coefficient.
128.	Slater variable rAPEX (P)	N/A	Distance of maximum probability of encountering a valence electron.
129.	Slater variable RAPEX (P)	N/A	Maximum value of the probability distribution (i.e. STOs).
130.	Slater variable FWHH (P)	N/A	Width of the probability distribution (i.e. STOs) at half height (half of the maximum).
131.	Slater variable SKEW (P)	N/A	Measure for the asymmetry of the probability distribution (i.e. STOs).
132.	Quadratic term for rAPEX (P)	N/A	Quadratic term for rAPEX.
133.	Quadratic term for RAPEX (P)	N/A	Quadratic term for RAPEX.
134.	Quadratic term for FWHH (P)	N/A	Quadratic term for FWHH.
135.	Quadratic term for SKEW (P)	N/A	Quadratic term for SKEW.
136.	Interaction term for rAPEX and RAPEX (P)	N/A	Interaction term for rAPEX and RAPEX.
137.	Interaction term for rAPEX and FWHH (P)	N/A	Interaction term for rAPEX and FWHH.
138.	Interaction term for rAPEX and SKEW (P)	N/A	Interaction term for rAPEX and SKEW.
139.	Interaction term for RAPEX and FWHH (P)	N/A	Interaction term for RAPEX and FWHH.
140.	Interaction term for RAPEX and SKEW (P)	N/A	Interaction term for RAPEX and SKEW.
141.	Interaction term for FWHH and SKEW (P)	N/A	Interaction term for FWHH and SKEW.