論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Machine learning-based analysis of solvothermal lignin and lignocellulose liquefaction
著者(和文)	CASTROGARCIA Abraham
Author(English)	Abraham Castro Garcia
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第12794号, 授与年月日:2024年3月26日, 学位の種別:課程博士, 審査員:CROSS JEFFREY SCOTT,大友 順一郎,髙橋 邦夫,石川 敦之,MANZHOS SERGEI,横井 俊之
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第12794号, Conferred date:2024/3/26, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Doctoral Thesis

Machine learning-based analysis of solvothermal lignin and lignocellulose liquefaction

Abraham Castro Garcia

Energy Science and Engineering
Department of Transdisciplinary Science and Engineering
School of Environment and Society
Tokyo Institute of Technology

December 2023

Machine learning-based analysis of solvothermal lignin and lignocellulose conversion

By

Abraham Castro Garcia

Submitted to the Department of Transdisciplinary Science and Engineering in fulfillment of the requirements for the Doctoral Degree of

Philosophy

TOKYO INSTITUTE OF TECHNOLOGY

December 2023

Author……

Abraham Castro Garcia

Supervisor……

Professor: Jeffrey Scott Cross

Abstract

Lignin is an abundant bio-polymer found in all plant matter in significant quantities (8~33% by weight), and its chemical structure is comprised interlinked aromatic units via ether and carbon-carbon chemical bonds. Because of its renewable nature it has been suggested as a possible source of renewable aromatic chemicals that could significantly contribute to the decarbonization of liquid fuel and aromatic chemical production. Among the existing depolymerization methods, solvolysis is seen as one of the most viable candidates due to its ease of application and easy scalability. However, in spite of the large amount of works found in the literature, how these solvolysis processes work remains unclear, with multiple authors claiming that the method they present is the "best" choice and put forward suggestions on what future research should be focused on. This dissertation presents a machine learning-based approach to study and try to understand how different experimental factors influence the outcome of lignin solvolysis experiments, by using data captured from literature and training models that predict experimental performance metrics such as the yield of liquid products and solid residues obtained from solvolysis experiments, focusing first in heterogeneously catalyzed, then homogeneously catalyzed solvolysis and finally studying the possibility of using the measured higher-heating-value (HHV) from bio-oil as an alternative experimental performance metric. The results indicate that catalysts properties (both for heterogeneous and homogeneous) play a comparatively small role in predicting the outcome of solvolysis experiments, and that temperature and reaction time is overwhelmingly more important. The yield of liquid products (bio-oil) was deemed to be an unreliable experiment performance metric at times due to it being influenced by the choice of chemical work-up steps used in a given study, further complicating comparison across studies.

Acknowledgements

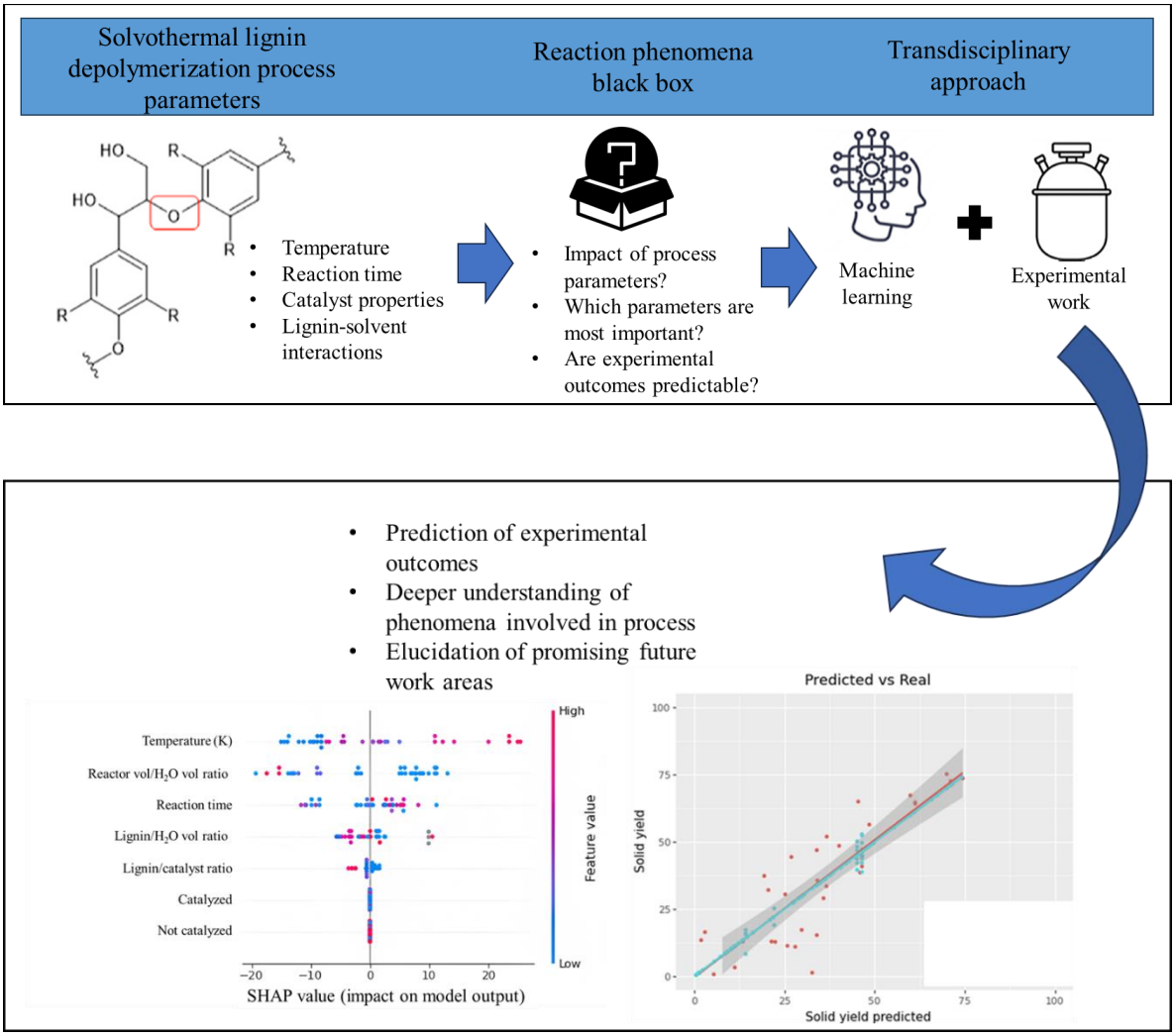
First, I would like to thank my academic advisor, Professor Jeffrey S. Cross, for his continuous support and advice throughout the entirety of my graduate education. His advice helped me grow both as a person and as a researcher.

Special thanks to the Tokyo Tech Academy of Energy and Informatics and the InfoSyEnergy consortium for providing me with a scholarship that has supported me during my doctoral studies. Being part of this program has really enriched my experience as a doctoral student and I could not be more grateful for receiving this opportunity.

I would also like to extend my gratitude to all students of Cross-laboratory whom I have interacted with over the course of the past 5 years. Thanks to May Kristine Carlon, talking with her about research, life and various other random topics over the years was truly enjoyable. Thanks to Muhammad Usman, though his altruism and stubborn optimism clashes with my outlook in life, I realize that he provided me insight and advice that positively impacted my life over the years. Thanks to Keang Kimleng, for listening to my rants and complaints, her presence lightens up my days. To Muhammad Harussani Moklis and Jinesh Mohan, for their interesting conversations and useful advice.

Finally, I would like to thank my family. To my parents and siblings for their constant support and love, making my life brighter and more enjoyable.

Graphical abstract



Contents

List of figures

List of tables

List of publications

1. (Published, chapter 4) Castro Garcia, A., Ching, P. L., So, R. H., Cheng, S., Boonyubol, S., & Cross, J. S. (2023). Prediction of Higher Heating Values in Bio-Oil from Solvothermal Biomass Conversion and Bio-Oil Upgrading Given Discontinuous Experimental Conditions. ACS Omega. https://doi.org/10.1021/acsomega.3c04

2. (Published, chapter 3) Castro Garcia, A., Cheng, S., McGlynn, S. E., & Cross, J. S. (2023). Machine Learning Model Insights into Base-Catalyzed Hydrothermal Lignin Depolymerization. ACS omega, 8(35), 32078-32089. https://doi.org/10.1021/acsomega.3c04168

3. (Published, chapter 2) Castro Garcia, A., Shuo, C., & Cross, J. S. (2022). Machine learning based analysis of reaction phenomena in catalytic lignin depolymerization. Bioresource Technology, 345, 126503. https://doi.org/10.1016/j.biortech.2021.126503

4. (Published, chapter 1) Garcia, A.C.; Cheng, S.; Cross, J.S. Solvolysis of Kraft Lignin to Bio-Oil: A Critical Review. Clean Technol. 2020, 2, 513-528. https://doi.org/10.3390/cleantechnol2040032

Chapter 1:

Introduction to lignin solvolysis

1.1 Introduction

During the course of the last couple of centuries, our easy access to fossil fuels has allowed the development of industry and led to the advanced society we know today. However, in light of our current understanding of the atmospheric greenhouse gas effect and climate change, in addition to the threat of eventual exhaustion of fossil fuel resources, we now desperately need both carbon-free and carbon neutral alternatives that can supplant their role.

While the development of wind power and solar photovoltaic technologies have allowed us to partly fulfill our energy needs, the question remains as to how to replace the role of fossil fuels as raw materials, fuels and chemicals [1]. To overcome this, studies regarding the possibility of converting different kinds of biomass into chemicals and raw materials have gained popularity in recent decades, most prominently the use of lignocellulosic matter as feedstock for processes. Lignocellulosic matter is composed of cellulose, hemicellulose and lignin in different proportions, depending on the plant. This type of biomass has garnered great interest due to its abundance, ease of renewability and potential to be transformed into different kinds of chemicals [2].

Early biomass conversion technologies relied heavily on simple sugars, starches and vegetable oils as raw materials to produce biofuel and led to a debate about the use of food as a source of biofuel synthesis, and its impact on food prices; these were called “first-generation biofuels”. Understanding this problem, attention was paid to non-edible plant matter that could be collected postharvest at relatively low cost; this led to the development of what is called “second-generation biofuels”, which rely entirely on non-edible plant matter, particularly in the production of cellulosic ethanol. Third-generation biofuels rely on algae or bacteria to decompose organic non-edible matter and implement chemical process techniques to turn it into fuels or chemicals [3].

Out of the components of lignocellulosic biomass, lignin accounts for 15 to 40% of its weight on a dry basis [4] and is known to be the most recalcitrant and hard to transform into products of interest, partly due to its chemical resistance [5]. As of 2019 production of lignin from the pulp and paper industry hovers around 50-70 million tons per year, and it is estimated that by 2030 this number will reach 225 million tons due to the implementation of biofuel production mandates [6]. The overall growth in production of chemical wood pulp

over time can be seen in Figure 1-1, which is positively correlated with growth of lignin production.

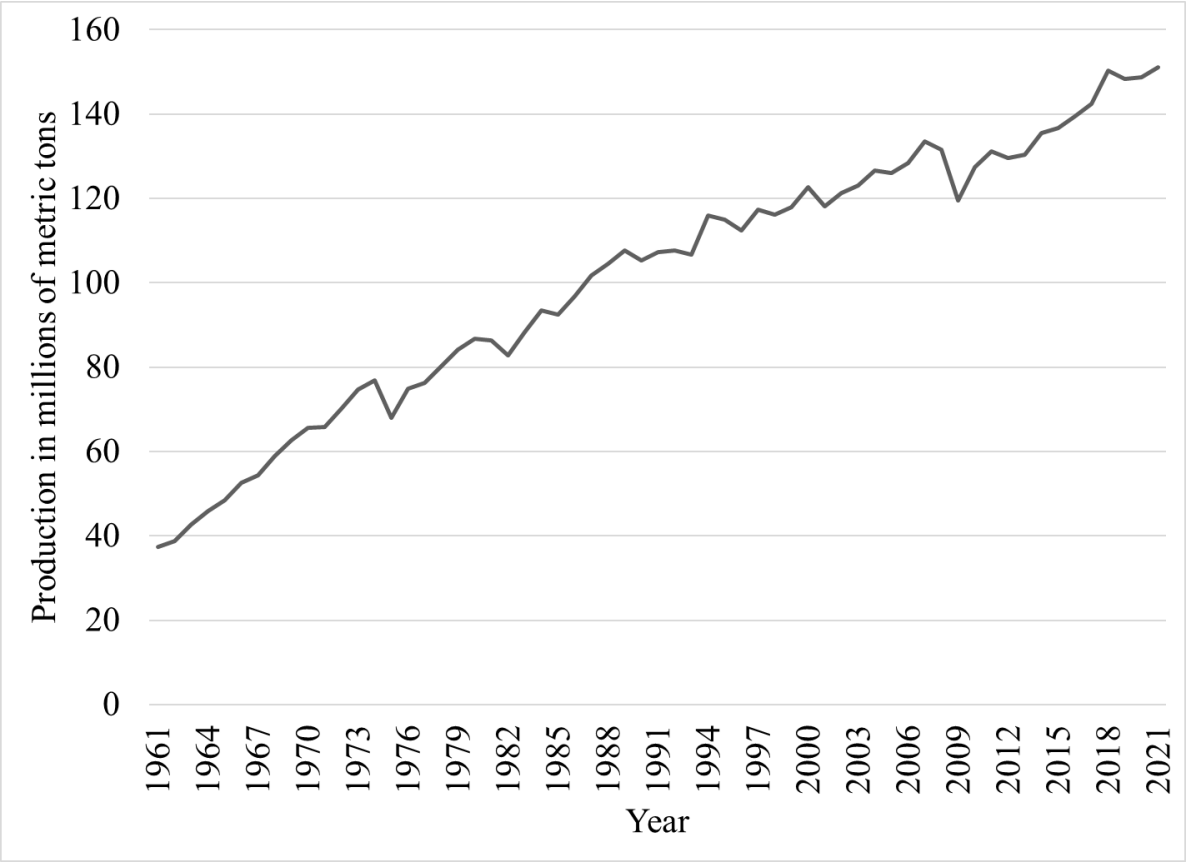


Figure 1-1. Production of chemical wood pulp worldwide from 1961 to 2021.

Lignin's polymeric structure is comprised of monolignols, which are found in variable proportions, depending on the plant source [7] (Figure 1-2).

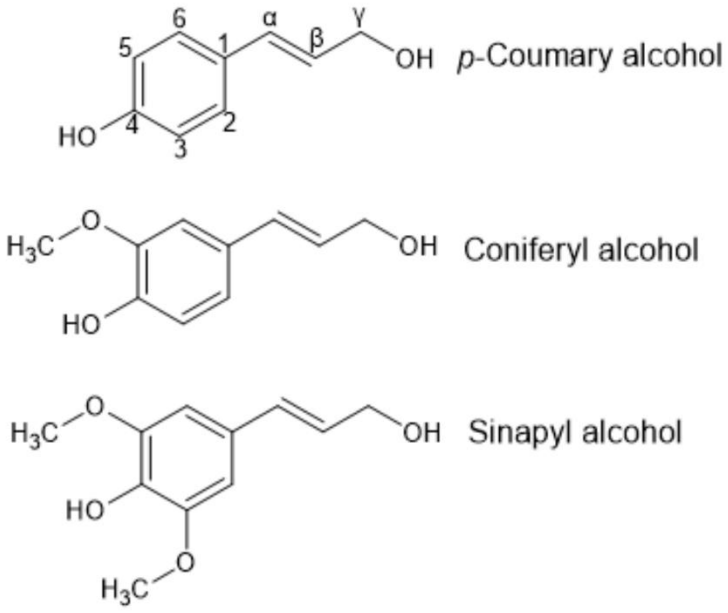


Figure 1-2. Structures of the three occurring monolignols in lignin (from [8]).

Together, these monolignols form the characteristic bonds that make the structure of lignin as shown in Figure 1-3. Depending on the plant from which the lignin originates, and the isolation method used to obtain the lignin, the resulting structure contains different proportions of the chemical bonds, highlighted in red. This can lead to notable differences between lignins in terms of molecular weight, reactivity and solubility.

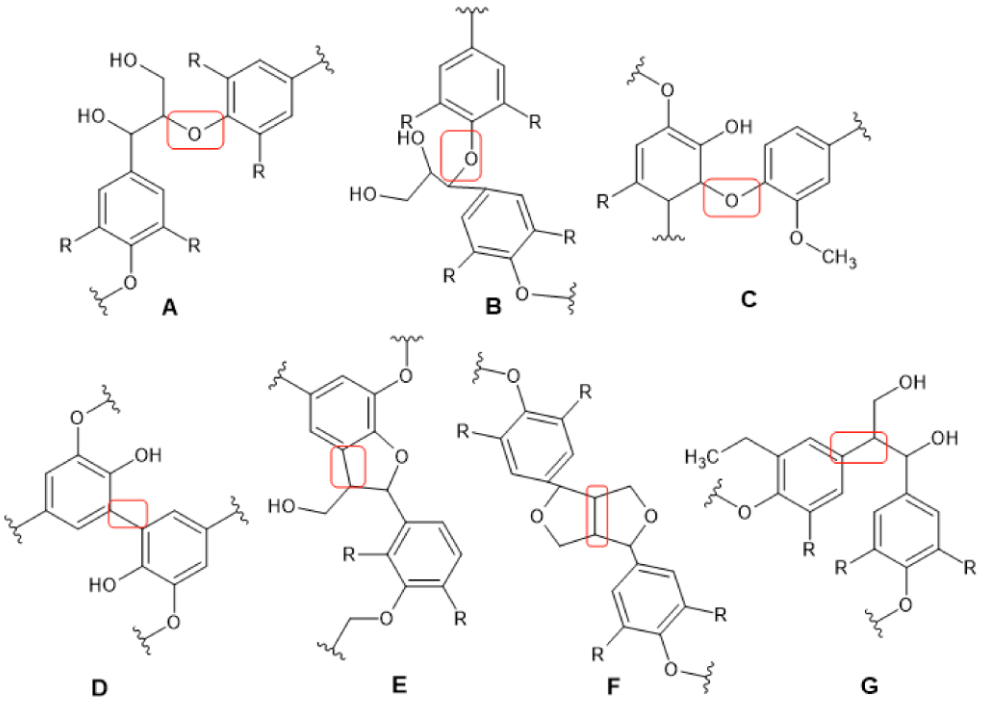


Figure 1-3. Chemical bonds found in lignin, comprised of both ether and carbon–carbon bonds (highlighted in red). (A) β-O-4, (B) α-O-4, (C) 4-O-5, (D) 5-5, (E) β-5, (F) β-β, (G) β-1 (from [8]).

Lignin's structure makes it a potential renewable source of aromatic chemicals that are currently only available from fossil fuel sources. These aromatic chemicals have high economic value as raw materials for diverse industries, as well as being a key component of jet fuel [9].

Diverse thermochemical lignin depolymerization methods exist, such as pyrolysis, solvolysis and gasification [10]. Among these, pyrolysis tends to result in higher rates of solid residue formation and higher oxygen content in the resulting bio-oil, whereas gasification allows high and fast conversion at the expense of solid residue formation and products of lower economic value such as syngas. Solvolysis, by using either a pure solvent or with a solvent mixture, can result in minimal solid residue formation and tends to favor the formation of low-oxygen-containing aromatic monomers [11]. These depolymerization methods, along with their reaction temperature and products, are briefly summarized in Table 1-1. Early studies employing solvolysis to depolymerize lignin suffered from high reaction temperature, long reaction time or high hydrogen pressure [12]; however, catalyst development and better understanding of the reactions involved have led to the development of less severe processes.

Table 1-1. Temperatures and primary products from different depolymerization methods used in lignin depolymerization (from [8].)

Method	Temperature	Primary Products	References
Gasification	700–1000 °C	Syngas	[13]
Pyrolysis	300–600 °C	Gaseous hydrocarbons, bio-oil and solid residue	[10]
Solvolysis	200–350 °C	Bio-oil and solid residue	[10]

1.2. Lignin Solvolysis

Among the various lignin depolymerization methods, studies employing solvolysis make up a very significant portion of the total published studies. The advantages of solvolysis over other lignin depolymerization methods is that by choosing an effective solvent and reaction temperature, mass transfer limitations can be greatly reduced by allowing lignin to properly dissolve, while the temperature distribution inside the reactor (batch or continuous flow) is easier to control.

Water, alcohols, hydrocarbons and other solvents have been used as reaction media with varying degrees of success, usually in conjunction with some sort of homogeneous or heterogeneous catalyst and occasionally molecular hydrogen. This can be seen in the many trends seen in lignin depolymerization studies. While the results reported in these studies are usually centered on the yield and the product distribution in the bio-oil obtained, it is hard to say with certainty which combination of factors is the best performing, with multiple, sometimes very distinct studies reporting bio-oil yields of above 80%. The bio-oil obtained is analyzed through various techniques, most prominently by gas chromatography–mass spectroscopy (GC-MS) and nuclear magnetic resonance (NMR) [14]. GC-MS allows for identification of aromatic monomers contained in bio-oil, but cannot identify larger oligomeric structures, whereas NMR provides insight about the nature of the chemical bonds found in the bio-oil, providing general understanding of the results of the reaction, for example the absence of β-O-4 ether bonds could correlate with good depolymerization results.

As illustrated in Figure 1-4 the performance of a lignin depolymerization reaction through solvolysis for a given type of lignin depends on the interaction between three controlling factors. Concretely, the three pairs of interactions are described as follows: Firstly, the lignin–catalyst interaction is relatively well understood, as catalysts used in studies are designed using analogies on the basis of the functional groups present in lignin, for example promoting hydrogenation, hydrogenolysis or dehydration of hydroxyl or ether functional groups found in lignin moieties. Secondly, this same catalyst can simultaneously react with the reaction media itself, as is the case with alcohols, phenol, water and other solvents, donating hydrogen [15], preventing the formation of solid residue [16], or as alkylating agents [17]; and thirdly,

this same reaction media might display different levels of affinity for lignin, dissolving it to varying degrees at different temperatures, either by itself or in combination with other solvents [18], and may react with the catalyst to decompose and form hydrogen through steam reforming and water gas shift reactions that occur simultaneously for alcohols [19, 20]. The overall depolymerization process and some of the monomers found in bio-oil are shown in Figure 1-5.

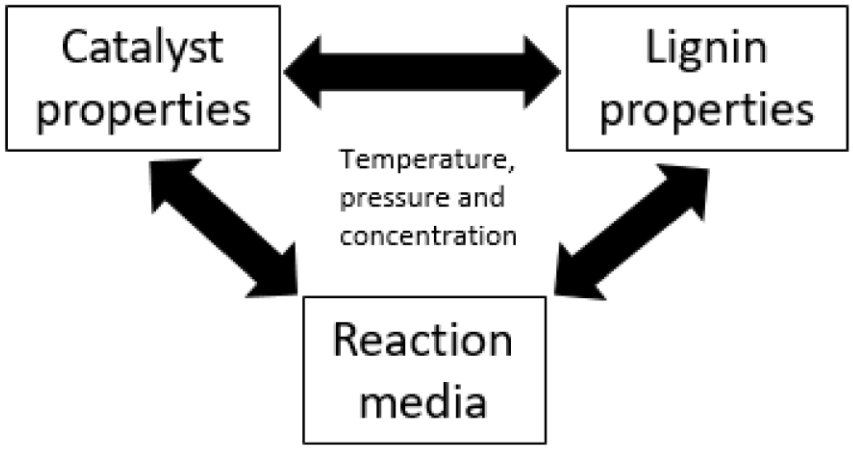


Figure 1-4. Controlling factors of the solvolysis reaction and their interactions, mediated by temperature, pressure and concentration (from [8]).

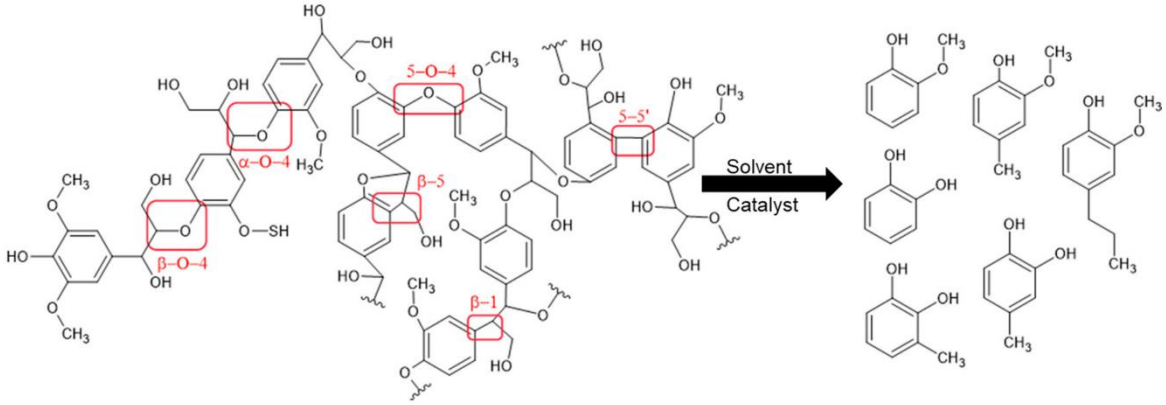


Figure 1-5. Depolymerization of lignin by solvolysis (from [8]).

Together, these interactions, mediated by the temperature, pressure and concentration of the reactants, present determine the yield and the quality of the obtained bio-oil. Most studies published to date have focused exclusively on one of these interactions, while neglecting the others, trying to find the "best" catalyst or the "optimal" reaction conditions for a given reaction media or lignin type. While satisfactory results may be possible by following this approach, it is reasonable to believe that by achieving deeper understanding of all the

interactions that take place simultaneously during the process, we can aim to achieve not only high yields and good quality of bio-oil, but also to do so economically and at industrial scale.

1.3. Purpose of the research

Various statements about which of the experimental factors is more important can be seen in the literature, including but not limited to:

- "The catalyst is the most important aspect"
- "We must minimize H_2 gas consumption"
- "High temperature processes are unfeasible"
- "Water is the only realistic solvent choice"

It is needless to say these statements are only partially true at best, and misleading at worst. In face of the overwhelming number of studies found in the last decade, it would appear that recently researchers are executing studies that can be largely described as "combinatorial", where the purpose or reasoning behind the experiments they propose is limited to using combinations of variables or experimental parameters that had not been tested yet, regardless of the quality of the result.

This, along with the fact that there are already many studies that claim to have "the best" lignin depolymerization method leads to a conundrum: *What is the "best" method of solvothermal lignin (or lignocellulose) depolymerization? Is it possible to clarify to what extent and when certain experimental parameters play a major or minor role in the overall process? What are the metrics we should be using to evaluate such processes?*

While systematic comparison can shed light on these questions, doing it for all kinds of existing lignins and covering all possible experimental parameters amounts to a difficult task, and ultimately this systematic comparison is also influenced by human bias, that is, the belief that certain kinds of studies are better or more valid.

To overcome this problem, using a well-trained machine learning (ML) model could allow us to obtain objective insight as to what experimental parameters are more important, assuming that we can indeed predict the experimental outcomes we are interested in, such as bio-oil yield, solid residue yield or even properties of the bio-oil obtained such as its higher heating value (HHV). Of course, this relies in the assumption that we can indeed predict experimental outcomes using machine learning, and, where this not to be possible, it would bring into question whether we understand solvothermal lignin depolymerization in the first place.

Using the commonly held assumption seen in all fields of science, but especially in data science. If we have data that we know correlates or is causative of a certain phenomenon or process or its output, then, it should be possible to train a ML model that predicts experimental results to a certain extent and possibly provide insight on how different

experimental parameters impact the predicted outcome, in a way that humans might not be able to when dealing with complex data.

1.4. Structure of dissertation

Following this introduction chapter (Chapter 1), this thesis includes two separate studies that focus on predicting the experimental results (bio-oil and solid residue yield) of in heterogeneously catalyzed solvothermal lignin depolymerization (Chapter 2) and hydrothermal, base-catalyzed lignin depolymerization (Chapter 3). Chapter 4 is focused on the prediction of HHV values in bio-oil obtained from solvothermal lignocellulosic conversion processes, as well as understanding the contribution of the different experimental factors involved bo0th in the initial production of the bio-oil and its upgrading to higher quality fuel. Finally, Chapter 5 summarizes the key findings and puts forward possible future lines of work that build on the results presented in this thesis. This is illustrated in Figure 1-6 below.

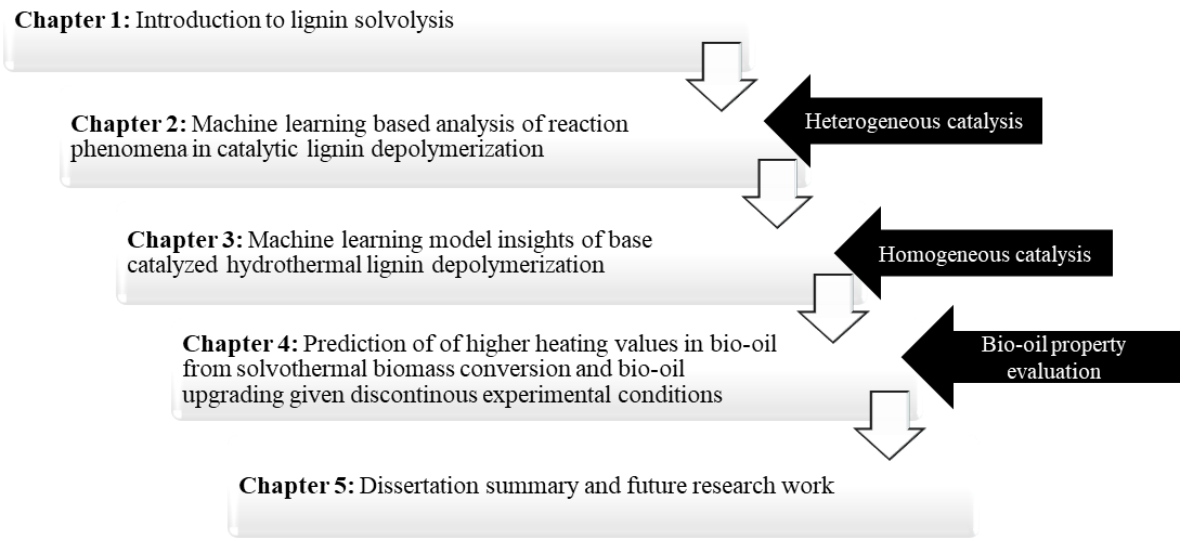


Figure 1-6. General structure of the dissertation.

References

[1] Moriarty, P., &; Honnery, D. (2016). Can renewable energy power the future? Energy Policy, 93, 3–7. https://doi.org/10.1016/j.enpol.2016.02.051
[2] Patel, M., Zhang, X., & Kumar, A. (2016). Techno-Economic and life cycle assessment on lignocellulosic biomass Thermochemical Conversion Technologies: A

Review. Renewable and Sustainable Energy Reviews, 53, 1486–1499. https://doi.org/10.1016/j.rser.2015.09.070
[3] Saladini, F., Patrizi, N., Pulselli, F. M., Marchettini, N., &; Bastianoni, S. (2016). Guidelines for emergy evaluation of First, second and third generation biofuels. Renewable and Sustainable Energy Reviews, 66, 221–227. https://doi.org/10.1016/j.rser.2016.07.073
[4] Novaes, E., Kirst, M., Chiang, V., Winter-Sederoff, H., &; Sederoff, R. (2010). Lignin and biomass: A negative correlation for wood formation and lignin content in trees. Plant Physiology, 154(2), 555–561. https://doi.org/10.1104/pp.110.161281
[5] Fodil Cherif, M., Trache, D., Brosse, N., Benaliouche, F., &; Tarchoun, A. F. (2020). Comparison of the physicochemical properties and thermal stability of organosolv and Kraft lignins from hardwood and softwood biomass for their potential valorization. Waste and Biomass Valorization, 11(12), 6541–6553. https://doi.org/10.1007/s12649-020-00955-0
[6] Bajwa, D. S., Pourhashem, G., Ullah, A. H., & Bajwa, S. G. (2019). A concise review of current lignin production, applications, products and their environmental impact. Industrial Crops and Products, 139, 111526. https://doi.org/10.1016/j.indcrop.2019.111526
[7] Dorrestijn, E., Laarhoven, L. J. J., Arends, I. W. C. E., &; Mulder, P. (2000). The occurrence and reactivity of phenoxyl linkages in lignin and low rank coal. Journal of Analytical and Applied Pyrolysis, 54(1–2), 153–192. https://doi.org/10.1016/s0165-2370(99)00082-0
[8] Garcia, A. C., Cheng, S., & Cross, J. S. (2020). Solvolysis of Kraft lignin to bio-oil: A critical review. Clean Technologies, 2(4), 513–528. https://doi.org/10.3390/cleantechnol2040032
[9] Al-Nuaimi, I. A., Bohra, M., Selam, M., Choudhury, H. A., El-Halwagi, M. M., &; Elbashir, N. O. (2016). Optimization of the aromatic/PARAFFINIC composition of synthetic jet fuels. Chemical Engineering & Technology, 39(12), 2217–2228. https://doi.org/10.1002/ceat.201500513
[10] Pandey, M. P., &; Kim, C. S. (2010). Lignin depolymerization and conversion: A review of Thermochemical Methods. Chemical Engineering &; Technology, 34(1), 29–41. https://doi.org/10.1002/ceat.201000270
[11] Cao, L., Yu, I. K. M., Liu, Y., Ruan, X., Tsang, D. C. W., Hunt, A. J., Ok, Y. S., Song, H., &; Zhang, S. (2018). Lignin valorization for the production of Renewable Chemicals: State-of-the-art review and future prospects. Bioresource Technology, 269, 465–475. https://doi.org/10.1016/j.biortech.2018.08.065
[12] Dorrestijn, E., Kranenburg, M., Poinsot, D., &; Mulder, P. (1999). Lignin depolymerization in hydrogen-donor solvents. Holzforschung, 53(6), 611–616. https://doi.org/10.1515/hf.1999.101

[13] Wang, H., Tucker, M., &; Ji, Y. (2013). Recent development in chemical depolymerization of Lignin: A Review. Journal of Applied Chemistry, 2013, 1–9. https://doi.org/10.1155/2013/838645
[14]. Mattsson, C., Andersson, S.-I., Belkheiri, T., Åmand, L.-E., Olausson, L., Vamling, L., &; Theliander, H. (2016). Using 2d NMR to characterize the structure of the low and high molecular weight fractions of bio-oil obtained from LignoBoost^{TM} Kraft lignin depolymerized in subcritical water. Biomass and Bioenergy, 95, 364–377. https://doi.org/10.1016/j.biombioe.2016.09.004
[15]. Zhang, J. (2018). Catalytic transfer hydrogenolysis as an efficient route in cleavage of lignin and model compounds. Green Energy &; Environment, 3(4), 328–334. https://doi.org/10.1016/j.gee.2018.08.001
[16] Huang, X., Korányi, T. I., Boot, M. D., &; Hensen, E. J. (2015). Ethanol as capping agent and formaldehyde scavenger for efficient depolymerization of lignin to Aromatics. Green Chemistry, 17(11), 4941–4950. https://doi.org/10.1039/c5gc01120e
[17]. Raj, K. J., Malar, E. J. P., &; Vijayaraghavan, V. R. (2006). Shape-selective reactions with AEL and AFI type molecular sieves alkylation of benzene, toluene and ethylbenzene with ethanol, 2-propanol, methanol and T-Butanol. Journal of Molecular Catalysis A: Chemical, 243(1), 99–105. https://doi.org/10.1016/j.molcata.2005.07.040
[18]. Sadeghifar, H., &; Ragauskas, A. (2020). Perspective on technical lignin fractionation. ACS Sustainable Chemistry & Engineering, 8(22), 8086–8101. https://doi.org/10.1021/acssuschemeng.0c01348
[19] Aouad, S., Labaki, M., Ojala, S., Seelam, P., Turpeinen, E., Gennequin, C., Estephane, J., &; Aad, E. A. (2018). A review on the dry reforming processes for hydrogen production: Catalytic materials and technologies. Catalytic Materials for Hydrogen Production and Electro-Oxidation Reactions, 60–128. https://doi.org/10.2174/9781681087580118020007
[20] Pal, D. B., Chand, R., Upadhyay, S. N., & Mishra, P. K. (2018). Performance of Water Gas Shift Reaction Catalysts: A Review. Renewable and Sustainable Energy Reviews, 93, 549–565. https://doi.org/10.1016/j.rser.2018.05.003

Chapter 2:

Machine learning based analysis of reaction phenomena in catalytic lignin depolymerization

2.1 Introduction

In recent decades, the concern for climate change has intensified as the potentially catastrophic consequences it could lead to have become clear. The cause most often blamed is excessive exploitation and usage of fossil fuels that now are vital part of society, as they provide energy in the form of electricity and hydrocarbons that power industry and transportation sectors. Because of this, there has been a strong push towards "decarbonizing" society by relying on non-carbon emitting sources of energy, particularly wind power and solar photovoltaic [1]. However, in spite of the efforts to expand the usage of these technologies, it remains clear that the role currently fulfilled by fossil fuels cannot be so easily replaced, particularly as a source of hydrocarbons that fuels heavy trucks, maritime vessels and airplanes [2]. An alternative often suggested is the conversion of abundant non-edible lignocellulosic biomass into chemicals and biofuels that would replace the role currently served by fossil fuels [3].

Lignocellulosic biomass consists of cellulose, hemicellulose and lignin, out of which cellulose and hemicellulose are used to make paper, with lignin being seen as a waste stream in the papermaking industry, where it usually burned to power the heat-intensive pulping process [4]. Because of lignin's complex three-dimensional aromatic polymer structure, comprised of monolignols [5] it has attracted attention as a potential source of aromatic chemicals, which are exclusively obtained in industrial scale from the processing of hydrocarbons into gasoline. This is further incentivized due to the increase of availability of lignin as a cheap feedstock from the growing cellulosic ethanol industry [6]. However, achieving economical and efficient depolymerization of lignin into valuable aromatic chemicals has seen slow progress in spite of decades of research, with methods such as pyrolysis, solvolysis and catalytic hydrotreatment [7], each of them with their own advantages and challenges. By using these methods, bio-oil, solid residue and gas are obtained. Of these methods, catalytic solvolysis has become the most prominent with a large

number of studies analyzing the impact of different solvent combinations, homogeneous or heterogeneous catalysts made from transition or noble metals, zeolites, activated carbons or alkali salts, possibly aided by the use of hydrogen gas, all of which in conjunction show great activity in the cleavage of the ether bonds found in lignin and guaiacol as model compounds [8].

These studies tend to focus on a single aspect of the depolymerization process, for example, the development of a more active catalyst, or the performance of a specific combination of solvents, ignoring other process parameters that are well known to affect the outcome of the reaction (yield of bio-oil, char and gas). Prior research often fails to address concretely how these experiments advance the understanding of the lignin depolymerization field and contribute to eventually being able to have an economically feasible process. Previous work [8], postulated that in the context of catalytic solvolysis of lignin the outcome of the experiment was largely decided by the interaction between the properties of the catalyst, lignin and reaction media used, mediate by temperature, pressure and concentration of reactants.

These interactions are not meant to be innovative, but rather to highlight that based on existing literature it is clear that these interactions take place and impact the outcome of the experiments. Yet, many published studies are written in a way that would indicate that the one aspect of the experiments they focus on (catalyst or reaction media, usually) is the most important or the only relevant one, neglecting to explain how the other variables impact the result. Some of the most relevant examples are the weight averaged molecular weight (Mw) of lignin which can vary from 1000 g mol-1 to 78400 g mol-1 [9], obtained by using gel-permeation-chromatography (GPC), resulting in lignin, being a large molecule, possibly facing important diffusion limitations if the pores of the given catalyst are small enough, this has been analyzed in previous studies by using ultrafiltration membrane technology [10], obtaining different Mw fractions of lignin containing different chemical bond distribution. In the same vein, surface area, the diameter, volume and geometry of the pores in the catalyst, which is analyzed by the Brunauer–Emmett–Teller (BET) technique is not consistently reported.

This inconsistent reporting of lignin and catalyst properties, in addition to the complex interaction between the experimental variables make it difficult to develop a comprehensive kinetic analysis of lignin depolymerization reactions. One way to tackle this complexity is by using machine learning (ML), a branch of artificial intelligence used to develop models capable of describing complicated data that has had success in studying other biomass related processes, such as solid residue and bio-oil production from pyrolysis of lignocellulosic biomass [11], heavy metal adsorption by using biomass-derived biochars [12] and prediction of lignocellulosic biomass composition [13]. This approach has not yet been attempted to

understand how the lignin depolymerization process parameters impact the bio-oil yield, solid residue formation and reaction time.

Based on prior work, in this study an ML-based approach is put forward to develop a model using the XGBoost regression method that makes use of existing lignin depolymerization data available in the literature to explain the contributions and impact of reaction variables involved in the experiments. Predictive models were created for bio-oil yield, solid residue yield and reaction time for a given yield of bio-oil, with the intent to understand which reaction variables contribute to maximizing bio-oil yield, minimizing solid residue yield and reaction time. XGBoost, being a decision tree-based method was chosen over other methods due to its resistance to overfitting [14] and interpretable variable importance [15]. There is a growing demand for interpretability from ML-developed models [16] to be able to explain how each variable numerically impact the outcome, which is an advantage of using XGBoost in this purpose.

2.2 Materials and methods

2.2.1 Data collection methodology and pre-processing

There are many variables that are known to impact the performance of lignin depolymerization reactions, in the case of this study data that would explain the impact of catalyst textural properties, lignin molecular weight in the yield of bio-oil, solid residues and reaction time for a given bio-oil yield was gathered and analyzed. To gather data from previous work found in the literature a data gathering methodology based on the usage of keywords was designed. In Google scholar the search string lignin depolymerization "BET" "GPC" was used to find relevant studies, finding a total of 488 results consisting of journal publications and theses.

The reasoning behind this search string is as follows: The term “lignin depolymerization” is seen almost without exception in all works related to conversion of lignin, regardless of the method (e.g. pyrolysis, solvolysis, hydrotreatment, etc.), then, the term “BET” stands for Brunauer–Emmett–Teller, an acronym used for the N_2 adsorption analysis technique used to obtain textural properties of catalysts, in the case of this study the focus was in obtaining surface area, pore volume and pore diameter. “GPC” stands for gel permeation chromatography and was included in the search string because it is the most used technique for analyzing the molecular weight of the original lignin used in the experiment, in the case of this study, it was decided to look for the weight averaged molecular weight (Mw). The usage of this search string meant that the studies found invariably involved a catalyst and measurement of molecular weight (either of the lignin, or the products obtained).

The reason for including acronyms in quotations in the Google scholar search string is because acronyms are rarely found in publication titles, thus, largely guaranteeing that the acronyms are found in the body of the text, indicating that the technique was used in the study or is referenced. From the 488 results found using the search string in Google scholar, 95 documents were downloaded, mostly journal publications and 3 doctoral theses.

The number of studies that reported all the key involved variables (surface area, pore volume, pore diameter and number average molecular weight) amounted to 9 [17~25], with the majority of studies reporting only some of the surface properties from BET or in the case of GPC, reporting the molecular weights of the bio-oils obtained, but not the original lignin. Nevertheless, these initial 9 studies provided 102 datasets that would be used in the analysis. The features and labels collected from these studies are shown and described in Table 2-1.

Table 2-1. Names and descriptions of features and labels used.

Feature and label names	Description
Active metal/lignin ratio	Ratio of active metal to lignin in the experiment
Active metal/solvent	Ratio of active metal to solvent in the experiment (mg/mL)
Surface area	Surface area measured for the catalyst (m2/g)
Pore diameter	Average pore diameter measured for the catalyst (nm)
Pore volume	Average pore volume measured for the catalyst (cm3/g)
Mw	Weight averaged molecular weight of the original lignin used
Catalyst/solvent ratio	Ratio of catalyst to solvent in the experiment (mg/mL)
Catalyst/lignin ratio	Ratio of catalyst to lignin in the experiment
Lignin/solvent ratio	Ratio of lignin to solvent in the experiment (mg/mL)
H2 pressure factor	H2 pressure in MPa divided by (reactor volume-solvent volume)
1Reaction time	Reaction time in minutes
Reactor/solvent volume ratio	Ratio of reactor volume to solvent volume used in the experiment
Temperature	Temperature in K
[1]Bio-oil yield	Bio-oil yield from lignin (wt%)
[1]Solid yield	Solid residue yield from lignin (wt%)
*Catalyst zeolite	Catalyst used was a zeolite with no active metals
*Catalyst FePd	Active phase in catalyst was Fe and Pd
*Catalyst Fe	Active phase in catalyst was Fe
*Catalyst Mo	Active phase in catalyst was Mo
*Catalyst MoRu	Active phase in catalyst was Mo and Ru
*Catalyst Ni	Active phase in catalyst was Ni
*Catalyst NiRu	Active phase in catalyst was Ni and Ru
*Catalyst Pd	Active phase in catalyst was Pd
*Catalyst Pt	Active phase in catalyst was Pt
*Catalyst Ru	Active phase in catalyst was Ru
*Solvent E-W	Solvent used was a mixture of ethanol–water (0.5:0.5, volume%)
*Solvent AC	Solvent used was acetone
*Solvent E	Solvent used was ethanol
*Hydrolysis lignin	Hydrolysis lignin was used in the experiment
*Kraft lignin	Kraft lignin was used in the experiments
*Organosolv lignin	Organosolv lignin was used in the experiments
*Sugarcane lignin	Sugarcane lignin was used in the experiments
*Stalk lignin	Stalk lignin was used in the experiments
*[2]Organosolv lignin	Organosolv lignin was used in the experiments

*These features were one-hot encoded due to their categorical nature.

[1]Bio-oil yield and solid yield are labels, reaction time was used as a label in one model.

[2] All lignins used in the studies were one-hot encoded to capture their individual properties, with different two instances of organosolv lignin.

2.2.2 Alternative and null hypotheses

In this study, it was assessed whether average lignin molecular weight (and by extension, kinetic diameter) and surface properties of the catalyst (surface area, average pore diameter and average pore volume) used in the reaction affect the prediction of bio-oil yield, solid residue, and reaction time in heterogeneously catalyzed lignin depolymerization reactions.

In this context, the alternative hypothesis is that differences in lignin Mw and surface properties of the catalyst may significantly impact the prediction accuracy of the developed XGBoost regression model for bio-oil yield, solid residue yield and reaction time for a given bio-oil yield, thus proving that there may be an optimal relation between these properties. The null hypothesis is that these properties do not impact the prediction capability of the developed XGBoost regression models. In Figure 2-1 a Pearson correlation matrix for the features and labels is shown. From this correlation matrix it can be observed that bio-oil yield and solid yield share a negative correlation due to the outcome of the experiments being either bio-oil or solid residue for the most part, with gas comprising a very small part of the results most of the time. Beyond this, the correlation between the features and the yield of bio-oil and solid residue do not seem to be immediately conclusive, with the features that are positively correlated with bio-oil yield being negatively correlated with solid residue yield, which falls in line with their relationship, the exception being the active metal/solvent ratio and catalyst/lignin ratio, which are roughly equal.

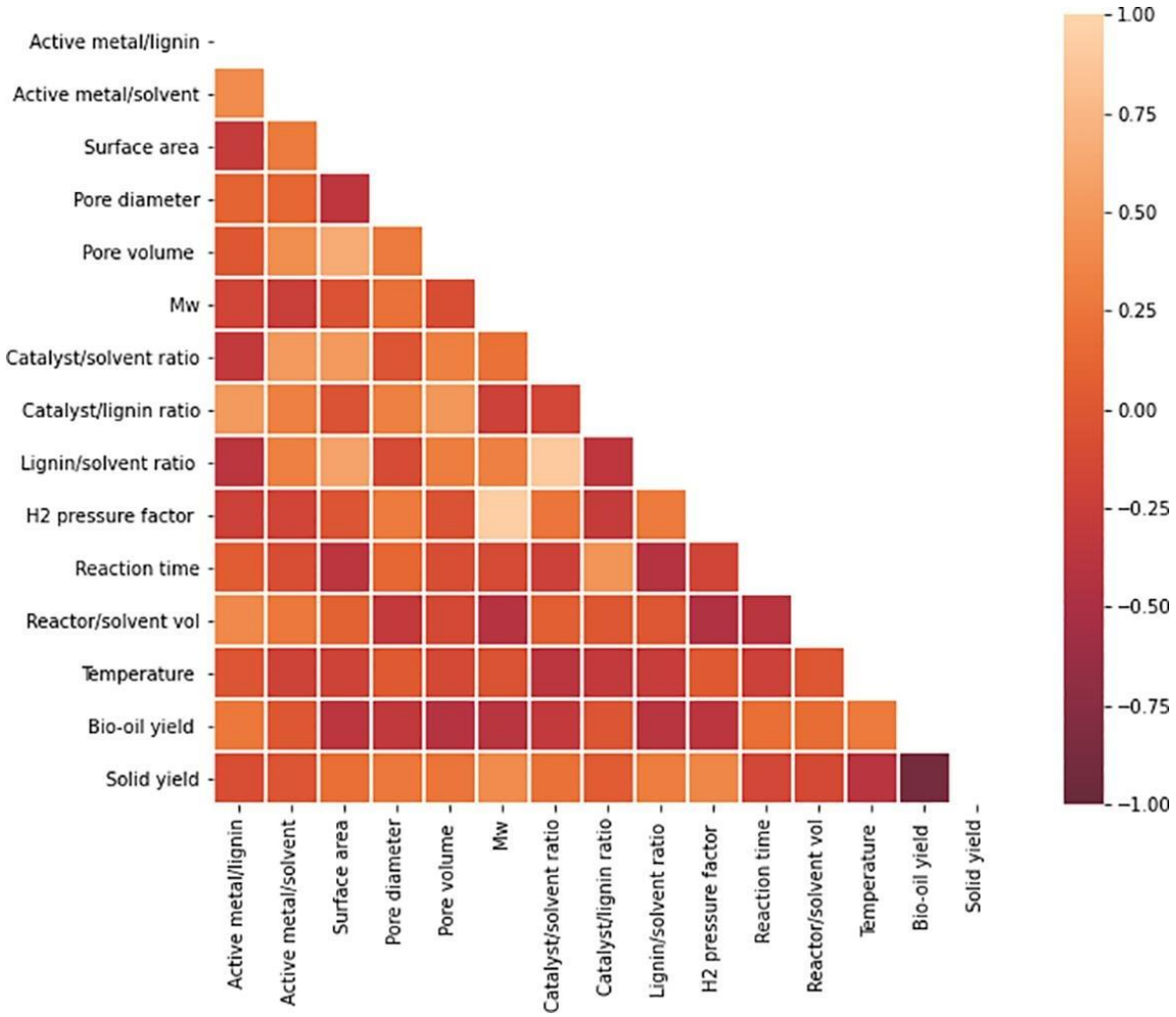


Figure 2-1. Pearson correlation matrix for the features and labels in the dataset.

2.2.3 Machine learning method and evaluation indicators

Scikit-learn free machine learning libraries for python were used to implement Extreme gradient boost regression (XGBoost) [26]. Feature importance was calculated by Scikit-learn's permutation importance (PI) library. Model's hyperparameters were tuned by using gridsearch until optimal performance was found.

The performance of the XGBoost models developed was measured by using Coefficient of determination (R^2) and Root-mean-squared error (RMSE), whose equations are shown below, (1-1) and (1-2).

$$R^2 = 1 - \frac{\sum_i^n \left(Y_i^{exp} - Y_i\right)^2}{\sum_i^n \left(Y_i^{exp} - Y_{avg}^{exp}\right)^2} \qquad (1\text{-}1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_i^n (Y_i^{exp} - Y_i)^2} \qquad (1\text{-}2)$$

Where n represents the number of test samples, Y_i^{exp} denotes the experimental value, and Y_i represents the predicted value. Y_{avg}^{exp} represents the mean value of Y_i^{exp} and Y_i, respectively.

2.3 Results and discussion

2.3.1 Evaluation of XGBOOST model performance for bio-oil yield

In order to evaluate the impact of catalyst surface properties and lignin Mw in the bio-oil yield prediction by using the trained XGBoost model, yield was removed from the dataset and was subsequently tested with all available features, resulting in the predictive performance seen in Figure 2-3, with a training R^2 and RMSE of 0.99 and 0.365, and testing R^2 0.81 and 8.723. The gap between training and testing scores could not be reduced significantly without reducing the test scores, indicating the possibility of limitations in relation to the data used.

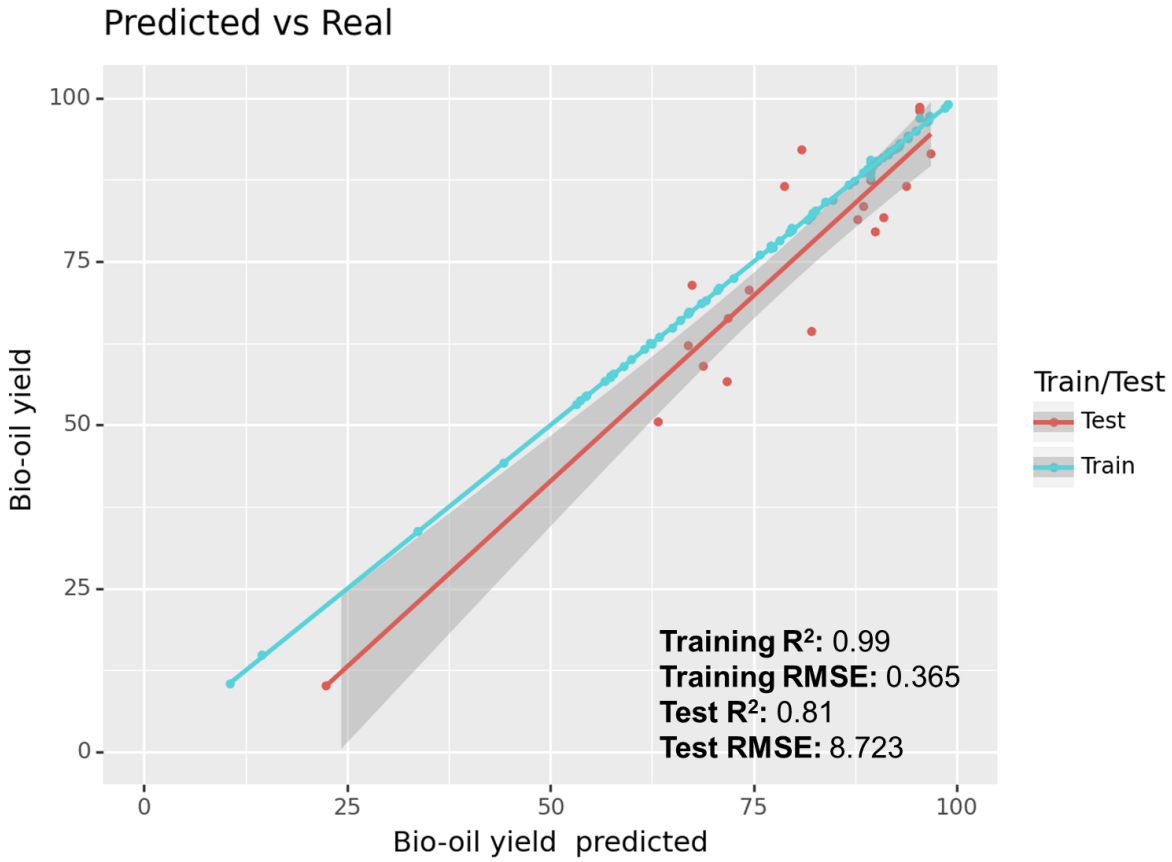


Figure 2-2. Training and test RMSE and R^2 scores for bio-oil yield prediction with XGBoost regression model.

In Table 2-2, permutation importance of the model's features are shown choice of solvent, temperature and reaction time account for roughly 69% of the total feature importance calculated through permutation importance and falls in line with the results reported in [27], where solvolysis of lignin in ethanol was carried out at sub- and super critical temperatures, in this study they found that higher bio-oil yields could be obtained at 200 °C for ethanol, but that higher degree of depolymerization in the resulting bio-oil could be achieved at higher temperatures, at the expense of decreased bio-oil yield. Similarly, usage of ethanol may be correlated to improved bio-oil yield, possibly due to ethanol's tendency to suppress solid residue formation [28]. Other solvolysis-related works have also found indeed that different solvents display different solvolysis performance in lignin depolymerization experiments [29], with ethanol and ethanol-water mixtures being seen in a large number of studies [30].

Table 2-2. Top 5 important features for the XGBoost model for bio-oil yield prediction.

Feature	Importance
Temperature (K)	0.33
Reaction time (min)	0.27
Lignin Mw	0.10
Solvent – Ethanol/Water	0.09
Lignin/Solvent ratio (mg/mL)	0.05

Regarding the catalyst-related features, active metal/solvent and active metal/lignin ratios, along with catalyst/lignin ratio together comprised 10% of the feature importance. Surface and pore properties of the catalyst (surface area, pore diameter and pore volume) contributed about 6.5% of importance and lignin Mw contributed 10%. It is possible that the reason why active metal/solvent ratio displays importance may be because of simultaneous occurrence of ethanol reforming reactions that happen even at low temperatures [31, 32], while these reactions are ideally carried at higher temperatures with catalysts specially tailored towards increasing selectivity towards hydrogen, they are known to happen at lower temperatures and the catalysts used in such experiments show many similarities with the ones found in the dataset, as well as other lignin depolymerization experiments, particularly in the case of works by [33], where porous copper oxides were successfully used in lignin depolymerization precisely because of their activity in reforming methanol. Regarding the feature importance of active metal/lignin and catalyst/lignin ratio it is possible that if indeed lignin fragments face overwhelming mass transfer limitations for a given catalyst, the outer active sites available for reaction become more important, resulting in experiments using catalysts that have higher metal loading and experiments that use more catalyst for a given amount of lignin showing better performance.

2.3.2 Evaluation of XGboost model performance for solid residue yield

Focusing on the features found in the dataset, the XGBoost regression method was used to develop a predictive model the formation of solid residue, again, first using all features available obtaining the results seen Figure 2-4 with training R^2 and RMSE of 0.98 and 1.946, and training scores of R^2 0.77 and 7.933. The gap between training and testing scores could not be reduced significantly without reducing the test scores, indicating the possibility of limitations in relation to the data we are using.

Figure 2-3. RMSE and R^2 score for solid residue yield prediction using XGBoost regression model.

The variable importance from the model in Figure 2-4 is shown in Table 2-3, like in the case of bio-oil yield, features associated with solvolysis of lignin show some importance: Temperature, choice of solvent, reaction time, lignin/solvent ratio and reactor volume/solvent volume ratio comprise together 92% of the feature importance, and because the yields of bio-oil, solid residue and gas share a zero-sum relationship, it is likely that these features share an inverse relationship with the yield of solid residue , this is shown to be true at least from the solvolysis point of view, according to [27] where higher temperatures, lower or no H_2 pressure result in elevated yields of solid residue. Although surface and pore properties showed low importance, surface area may play a role in in-situ generation of H_2 via low-temperature reforming of solvents used in the reaction. This ties into the various transition metals seen in the dataset used , and in fact many of the metals seen in the data set are known to display activity for the reforming reaction, particularly Ni and Pt [32], and would result in the formation of more hydrogen and less solid residues. Catalyst to lignin and solvent ratios partly overlap with the active metal related ratios, but they allow one to account for the properties of the catalyst support used.

Table 2-3. Top 5 important features for the XGBoost model for solid residue yield prediction.

Feature	Importance
Temperature (K)	0.56
Reaction time (mins)	0.18
Solvent – Ethanol/Water	0.09
Lignin/Solvent ratio (mg/mL)	0.05
Catalyst surface area (m^2/gr)	0.04

2.3.4 Underlying chemistry behind models performance and limitations

As indicated in the previous section, the large majority of the feature importance could be explained by reaction conditions and variables commonly associated with lignin solvolysis, with choice of solvent, temperature and reaction time accounting for 62% of the total feature importance for bio-oil yield prediction and in the case of solid residue yield, temperature, H_2 pressure factor, choice of solvent, reaction time, lignin/solvent ratio and reactor volume/solvent volume ratio comprise together 77% of the feature importance,

The alternative hypothesis turned out to be partially true for both solid residue yield and reaction time, but not bio-oil yield. Notably, the surface properties of the catalyst and lignin Mw did not contribute to improving the predictive capability for bio-oil yield. The meaning behind the seeming lack of statistical importance from these features may indicate that for studies similar to the ones used to obtain the data, lignin fragments may face overwhelming mass transfer limitations for most micro and lower mesoporous catalysts, thus not taking advantage of the large surface area available and active metals found within them. In the same vein, this could raise concerns about pore shape in the catalysts used, an aspect that has not been studied for lignin depolymerization reactions yet.

Additionally, this would also explain the large gap in reaction rates for a given bio-oil yield seen in homogeneously catalyzed lignin depolymerization reactions, where alkali salts are used to great effect to quickly depolymerize lignin. Homogeneous catalysts may not be necessarily more active than some of the heterogeneous ones, but do not face any mass transfer limitations. Considering the results found from the creation of these models, the relationship between reaction factors postulated in [8] can be re-interpreted as shown below in Figure 2-6 and indeed the interactions between each factor can be individually tested, such as in the usage of model compounds to test ease of bond cleavage for a given catalyst [34], the usage of different alcohols to prevent lignin condensation (shown in Figure 2-7, with ethanol as the alcohol) [35] and in-situ conversion reforming of the solvent to produce H_2 [36], which can be represented by (1-3), all of them contributing towards the outcome of the lignin depolymerization reaction, taking into account these interactions could lead to the development of more meaningful studies could lead to an efficient and economical lignin depolymerization process.

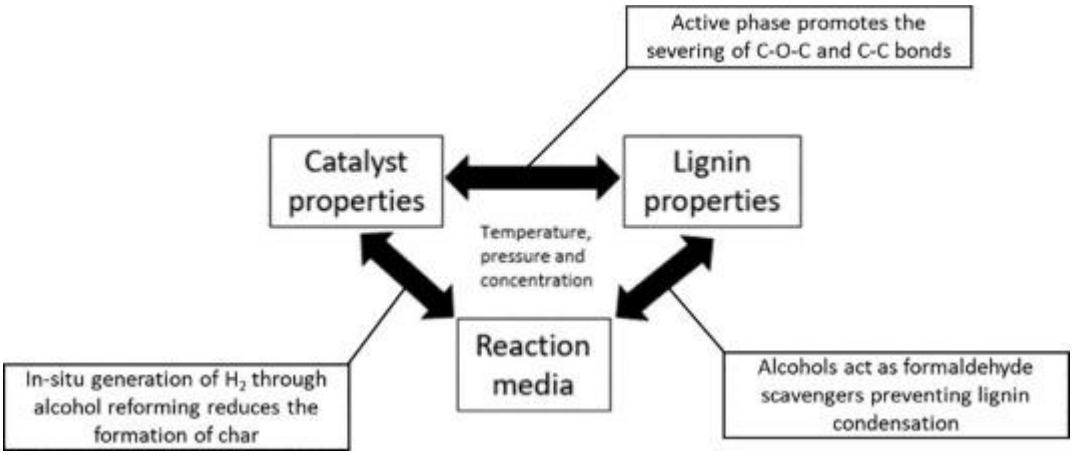


Figure 2-4. Re-interpretation of reaction factor interaction based on results of the ML models.

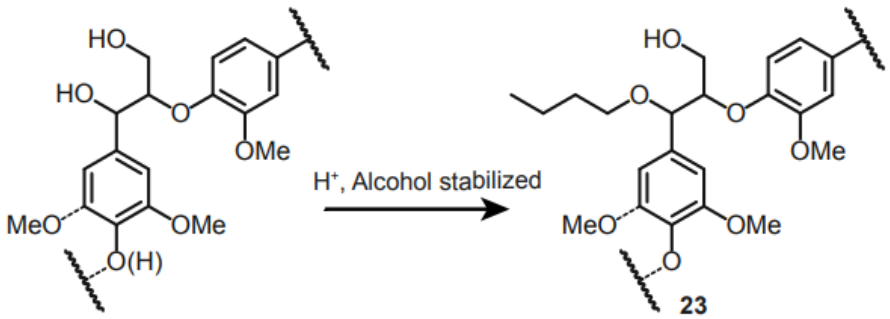


Figure 2-5. Stabilization of β–O–4 Linkages by α-OH etherification (adapted from [35]).

$$C_2H_5OH \rightarrow CH_4 + CO + H_2 \qquad (1\text{-}3)$$

The limitations associated with the data used come from the lack of studies that do comprehensive analysis of both the catalyst and lignin used in their experiments, thus, the results for this dataset may not extrapolate well for studies that involve other variables, such as the usage of formic acid as H_2 donor [37], simultaneous usage of various catalysts [38] and oxidative gas media [39]. However, the resulting variable importance from the models obtained can be tied to studies that analyze specific behaviors found in lignin depolymerization reactions, allowing to go beyond using machine learning just as a "black box" that can accurately predict, but ultimately not explain why the model works the way it does. The possibility of following and studying a particular reaction mechanism in relation to lignin depolymerization is an attractive idea. Doing this would probably be easier by focusing on studies with model compounds such as guaiacol or biphenyl (for C-O-C and C-C bonds, respectively), then, focusing on a single aspect of the experiment, such as the composition or synthesis of the catalyst.

The possibility of using ML to detect patterns and principles that would escape human perception is a topic of interest in recent publications that touch on the concept of Fermi energy and chemical potential [40 and 41].

2.4. Conclusions

In this study, the XGBoost regression method was used to develop models that predict bio-oil yield and solid residue yield in heterogeneously catalyzed lignin depolymerization reactions using data available in the literature. The predictive capability of the resulting models varied, with the prediction bio-oil yield and solid residue yield achieving test R^2 scores of 0.81 and 0.77. The variable importance found in these models indicates that the

ultimate outcome of lignin depolymerization reactions are mostly dependent on variables associated with solvolysis and not necessarily surface properties of the catalyst.

It must be emphasized that the data used for training this model represents a small fraction of the possible experimental space, thus, it may not be representative of the whole. However, considering the distribution of the data that used, it would be reasonable to claim that differences in catalyst performance may only be appreciable at when comparing experiments that share the same temperature, reaction time and solvent, only then the impact of the catalyst properties would be easy to assess.

Because this study focused only on the modeling of experimental data involving heterogeneous catalysts, the next chapter was dedicated to predicting the yields of bio-oil and solid residue for homogeneously catalyzed-hydrothermal lignin depolymerization studies, where, due to the comparably low variety of experimental parameters the interpretation of the model should reveal useful insight regarding lignin solvolysis in general.

References:

[1] Heide, D., von Bremen, L., Greiner, M., Hoffmann, C., Speckmann, M., & Bofinger, S. (2010). Seasonal optimal mix of wind and solar power in a future, highly renewable Europe. Renewable Energy, 35(11), 2483–2489. https://doi.org/10.1016/j.renene.2010.03.012

[2] Moriarty, P., & Honnery, D. (2016). Can renewable energy power the future? Energy Policy, 93, 3–7. https://doi.org/10.1016/j.enpol.2016.02.051

[3] Won Seo, M., Hoon Lee, S., Nam, H., Lee, D., Tokmurzin, D., Wang, S., & Kwon Park, Y. (2021). Recent advances of Thermochemical Conversieon processes for Biorefinery. Bioresource Technology, 126109. https://doi.org/10.1016/j.biortech.2021.126109

[4] Ekielski, A., & Mishra, P. K. (2020). Lignin for Bioeconomy: The present and future role of technical lignin. International Journal of Molecular Sciences, 22(1), 63. https://doi.org/10.3390/ijms22010063

[5] Vanholme, R., Demedts, B., Morreel, K., Ralph, J., & Boerjan, W. (2010). Lignin biosynthesis and structure. Plant Physiology, 153(3), 895–905. https://doi.org/10.1104/pp.110.155119

[6] Sanchez, A., & Gomez, D. (2014). Analysis of historical total production costs of cellulosic ethanol and forecasting for the 2020-Decade. Fuel, 130, 100–104. https://doi.org/10.1016/j.fuel.2014.04.037

[7] Pandey, M. P., & Kim, C. S. (2010). Lignin depolymerization and conversion: A review of Thermochemical Methods. Chemical Engineering & Technology, 34(1), 29–41. https://doi.org/10.1002/ceat.201000270

[8] Garcia, A. C., Cheng, S., & Cross, J. S. (2020). Solvolysis of Kraft lignin to bio-oil: A critical review. Clean Technologies, 2(4), 513–528. https://doi.org/10.3390/cleantechnol2040032

[9] Tolbert, A., Akinosho, H., Khunsupat, R., Naskar, A. K., & Ragauskas, A. J. (2014). Characterization and analysis of the molecular weight of lignin for Biorefining Studies. Biofuels, Bioproducts and Biorefining, 8(6), 836–856. https://doi.org/10.1002/bbb.1500

[10] Chen, C., Jin, D., Ouyang, X., Zhao, L., Qiu, X., & Wang, F. (2018). Effect of structural characteristics on the depolymerization of lignin into phenolic monomers. Fuel, 223, 366–372. https://doi.org/10.1016/j.fuel.2018.03.041

[11] Ullah, Z., khan, M., Raza Naqvi, S., Farooq, W., Yang, H., Wang, S., & Vo, D.-V. N. (2021). A comparative study of machine learning methods for bio-oil yield prediction – a genetic algorithm-based features selection. Bioresource Technology, 335, 125292. https://doi.org/10.1016/j.biortech.2021.125292

[12] Zhu, X., Wang, X., & Ok, Y. S. (2019). The application of machine learning methods for prediction of metal sorption onto biochars. Journal of Hazardous Materials, 378, 120727. https://doi.org/10.1016/j.jhazmat.2019.06.004

[13] Xing, J., Luo, K., Wang, H., & Fan, J. (2019). Estimating biomass major chemical constituents from ultimate analysis using a random forest model. Bioresource Technology, 288, 121541. https://doi.org/10.1016/j.biortech.2019.121541

[14] Shahri, N. H., Lai, S. B., Mohamad, M. B., Rahman, H. A., & Rambli, A. B. (2021). Comparing the performance of AdaBoost, XGBoost, and logistic regression for Imbalanced Data. Mathematics and Statistics, 9(3), 379–385. https://doi.org/10.13189/ms.2021.090320

[15] Wen, H.-T., Wu, H.-Y., & Liao, K.-C. (2022). Using XGBoost regression to analyze the importance of input features applied to an artificial intelligence model for the biomass Gasification System. Inventions, 7(4), 126. https://doi.org/10.3390/inventions7040126

[16] Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. Harvard Data Science Review, 1(2). https://doi.org/10.1162/99608f92.5a8a3a3d

[17] Chen, P., Zhang, Q., Shu, R., Xu, Y., Ma, L., & Wang, T. (2017). Catalytic depolymerization of the hydrolyzed lignin over mesoporous catalysts. Bioresource Technology, 226, 125–131. https://doi.org/10.1016/j.biortech.2016.12.030

[18] Zeng, Z., Xie, J., Guo, Y., Rao, R., Chen, B., Cheng, L., Xie, Y., & Ouyang, X. (2021). Hydrogenolysis of lignin to produce aromatic monomers over FePd bimetallic catalyst supported on HZSM-5. Fuel Processing Technology, 213, 106713. https://doi.org/10.1016/j.fuproc.2020.106713

[19] Du, B., Chen, C., Sun, Y., Yu, M., Yang, M., Wang, X., & Zhou, J. (2020). Catalytic conversion of lignin to bio-oil over PTA/MCM-41 catalyst assisted by ultrasound acoustic cavitation. Fuel Processing Technology, 206, 106479. https://doi.org/10.1016/j.fuproc.2020.106479

[20] Feng, L., Li, X., Wang, Z., & Liu, B. (2021). Catalytic hydrothermal liquefaction of lignin for production of aromatic hydrocarbon over metal supported mesoporous catalyst. Bioresource Technology, 323, 124569. https://doi.org/10.1016/j.biortech.2020.124569

[21] Jae-Young Kim. Depolymerization features of lignin under supercritical ethanol state in the presence of metal catalysts. Doctoral thesis, 2017. Seoul National University. https://s-space.snu.ac.kr/handle/10371/121088

[22] Cheng, S., Yuan, Z., Leitch, M., Anderson, M., & Xu, C. (C. (2013). Highly efficient de-polymerization of organosolv lignin using a catalytic hydrothermal process and production of phenolic resins/adhesives with the depolymerized lignin as a substitute for phenol at a high substitution ratio. Industrial Crops and Products, 44, 315–322. https://doi.org/10.1016/j.indcrop.2012.10.033

[23] Cheng, S., Wilks, C., Yuan, Z., Leitch, M., & Xu, C. (C. (2012). Hydrothermal degradation of alkali lignin to bio-phenolic compounds in sub/supercritical ethanol and water–ethanol co-solvent. Polymer Degradation and Stability, 97(6), 839–848. https://doi.org/10.1016/j.polymdegradstab.2012.03.044

[24] Park, J., Oh, S., Kim, J.-Y., Park, S. Y., Song, I. K., & Choi, J. W. (2016). Comparison of degradation features of lignin to phenols over PT catalysts prepared with various forms of carbon supports. RSC Advances, 6(21), 16917–16924. https://doi.org/10.1039/c5ra21875f

[25] Tymchyshyn, M., Rezayan, A., Yuan, Z., Zhang, Y., & Xu, C. C. (2020). Reductive hydroprocessing of hydrolysis lignin over efficient bimetallic catalyst MoRu/AC. Industrial & Engineering Chemistry Research, 59(39), 17239–17249. https://doi.org/10.1021/acs.iecr.0c01151

[26] Chen, T., & He, T. (2020, June 11). xgboost: eXtreme Gradient Boosting. Retrieved January 14, 2022, from https://mran.microsoft.com/snapshot/2020-07-15/web/packages/xgboost/vignettes/xgboost.pdf

[27] Kim, J.-Y., Oh, S., Hwang, H., Cho, T.-su, Choi, I.-G., & Choi, J. W. (2013). Effects of various reaction parameters on solvolytical depolymerization of lignin in sub- and supercritical ethanol. Chemosphere, 93(9), 1755–1764. https://doi.org/10.1016/j.chemosphere.2013.06.003

[28] Huang, X., Korányi, T. I., Boot, M. D., & Hensen, E. J. (2015). Ethanol as capping agent and formaldehyde scavenger for efficient depolymerization of lignin to Aromatics. Green Chemistry, 17(11), 4941–4950. https://doi.org/10.1039/c5gc01120e

[29] Kouris, P. D., van Osch, D. J., Cremers, G. J., Boot, M. D., & Hensen, E. J. (2020). Mild thermolytic solvolysis of technical lignins in polar organic solvents to a crude lignin oil. Sustainable Energy & Fuels, 4(12), 6212–6226. https://doi.org/10.1039/d0se01016b

[30] Sun, Z., Fridrich, B., de Santi, A., Elangovan, S., & Barta, K. (2018). Bright side of lignin depolymerization: Toward new platform chemicals. Chemical Reviews, 118(2), 614–678. https://doi.org/10.1021/acs.chemrev.7b00588

[31] Morgenstern, D. A., & Fornango, J. P. (2005). Low-temperature reforming of ethanol over copper-plated Raney nickel: a new route to sustainable hydrogen for transportation. Energy & Fuels, 19(4), 1708–1716. https://doi.org/10.1021/ef049692t

[32] Palma, V., Ruocco, C., Meloni, E., Gallucci, F., & Ricca, A. (2018). Enhancing Pt-Ni/CeO2 performances for ethanol reforming by catalyst supporting on high surface silica. Catalysis Today, 307, 175–188. https://doi.org/10.1016/j.cattod.2017.05.034

[33] Barta, K., Matson, T. D., Fettig, M. L., Scott, S. L., Iretskii, A. V., & Ford, P. C. (2010). Catalytic disassembly of an organosolv lignin via hydrogen transfer from supercritical methanol. Green Chemistry, 12(9), 1640. https://doi.org/10.1039/c0gc00181c

[34] Guadix-Montero, S., & Sankar, M. (2018). Review on catalytic cleavage of C–C inter-unit linkages in lignin model compounds: Towards lignin Depolymerisation. Topics in Catalysis, 61(3-4), 183–198. https://doi.org/10.1007/s11244-018-0909-2

[35] Lan, W., & Luterbacher, J. S. (2019). Preventing lignin condensation to facilitate aromatic monomer production. CHIMIA International Journal for Chemistry, 73(7), 591–598. https://doi.org/10.2533/chimia.2019.591

[36] Elias, K. F. M., Bednarczuk, L., Assaf, E. M., Ramírez de la Piscina, P., & Homs, N. (2019). Study of Ni/CeO2–ZnO catalysts in the production of H2 from acetone steam reforming. International Journal of Hydrogen Energy, 44(25), 12628–12635. https://doi.org/10.1016/j.ijhydene.2018.10.191

[37] Ouyang, X., Huang, X., Zhu, Y., & Qiu, X. (2015). Ethanol-enhanced liquefaction of lignin with formic acid as an in situ hydrogen donor. Energy & Fuels, 29(9), 5835–5840. https://doi.org/10.1021/acs.energyfuels.5b01127

[38] Limarta, S. O., Ha, J.-M., Park, Y.-K., Lee, H., Suh, D. J., & Jae, J. (2018). Efficient depolymerization of lignin in supercritical ethanol by a combination of metal and base catalysts. Journal of Industrial and Engineering Chemistry, 57, 45–54. https://doi.org/10.1016/j.jiec.2017.08.006

[39] Liu, C., Wu, S., Zhang, H., & Xiao, R. (2019). Catalytic oxidation of lignin to valuable biomass-based platform chemicals: A Review. Fuel Processing Technology, 191, 181–201. https://doi.org/10.1016/j.fuproc.2019.04.007

[40] Link, M., Gao, K., Kell, A., Breyer, M., Eberz, D., Rauf, B., & Köhl, M. (2023, May 16). Machine learning the phase diagram of a strongly interacting Fermi Gas. Physical Review Letters. https://doi.org/10.1103/PhysRevLett.130.203401

[41] Hyttinen, N., Pihlajamäki, A., & Häkkinen, H. (2022). Machine learning for predicting chemical potentials of multifunctional organic compounds in atmospherically relevant solutions. The Journal of Physical Chemistry Letters, 13(42), 9928–9933. https://doi.org/10.1021/acs.jpclett.2c02612

Chapter 3:

Machine learning model insights of base catalyzed hydrothermal lignin depolymerization

3.1 Introduction

Biomass conversion has been suggested as an alternative, CO_2-neutral source of hydrocarbons and has seen large success over the past decades in the conversion of edible biomass such as sugars to produce ethanol [1] and oils to produce biodiesel [2]. However, the conversion of edible biomass has been judged as unsustainable due to its impact on food prices, and in turn, the usage of non-edible biomass as a feedstock has been suggested instead [3]. Among these, the conversion of lignocellulose and in particular cellulose has seen the most progress with the emergence of the cellulosic ethanol industry [4]. Lignin usage, on the other hand, has lagged despite its abundance, accounting for roughly 15 to 25 % of the total lignocellulosic biomass available worldwide [5] and possessing a polymeric aromatic structure that could provide aromatic chemicals currently only available from fossil fuels in any meaningful quantities [6]. These aromatic chemicals are foreseen to remain a staple of the chemical industry due to their large number of applications in the making of plastics and as precursors for the chemical synthesis of drugs and materials [7].

Depolymerizing lignin is not an easy task however, as seen from the large body of research in the literature, with numerous studies employing different methods, thermochemical, catalytic, or biological [8]. Among these, thermochemical and catalytic methods have obtained the largest success, usually obtaining a mixture of gas, solid residue, and bio-oil as products from the reaction, in different proportions depending on the severity of the process, reaction media, and presence or absence of a catalyst [9]. While the merits of using transition metal-containing catalysts are well understood, resulting in a higher yield of bio-oil containing less oxygen and lower formation of solid residue [10], their application at an

industrial scale leaves much to be desired due to their relatively fast poisoning and low cost-benefit ratio when compared to simple strong alkali salts, such as NaOH or KOH in homogeneous reaction media [11].

Due to the great number of possible combinations of catalysts, reaction media and qualities of lignin with different properties, lignin depolymerization research across studies can be perceived as complex or confusing, with many studies reporting very different results that provide only partial or no justification for the choice of metals in the catalyst they use, or the reaction media chosen. In previous work [12], machine learning (ML) algorithms were used to develop predictive models for yield of bio-oil and solid residue and interestingly, in spite of the large variety of experimental variables and properties, it was revealed that in most cases the predicted yield of bio-oil and solid residue could be attributed to specific process variables, such as temperature, the ratio of lignin to solvent used and the usage of certain solvent combinations. However, the literature data used in this work represented only a very small fraction of the published data available in the reaction space, due to it being gathered only from certain studies that fit the search criteria used in that study. Another recent study [13] focuses more broadly in predicting and optimizing the results of hydrothermal biomass liquefaction highlighted the potential in modelling seemingly simple processes to better understand how the process variables impact the experimental results obtained.

Among the existing lignin depolymerization methods, base catalyzed depolymerization in water as reaction media has attracted attention due to its economic feasibility [14], short reaction times [15], and reaction performance comparable to that seen in transition metal-catalyzed processes, in terms of bio-oil yield [16], particularly for pulping-derived lignins. Comparatively speaking, base catalyzed hydrothermal depolymerization is simpler than transition metal catalyzed depolymerization in organic solvents, due to the lower number of possible interactions among the reactants. This makes base catalyzed depolymerization seemingly simpler to understand and interpret.

Inspired by the recent developments in usage of ML in study thermochemical and hydrothermal conversion of biomass [17, 18]. In this study, base catalyzed and non-catalyzed hydrothermal depolymerization of lignin was investigated, for the first time, by combining ML modeling to predict the yield of bio-oil and solid residue and experimental work to test the validity of the models. Explainable variable importance for the models was obtained through two different methodologies in order to obtain insight on how the process variables in lignin depolymerization impact the results of the experiments. The results revealed that prediction of solid residues is consistent with the experimental results, but reliably predicting bio-oil yield may be difficult in absence of clearer and more comprehensive characterization of the lignin used.

3.2 Materials and methods

3.2.1 Data collection and pre-processing

A methodology was developed to manually gather data from the existing literature related to hydrothermal and homogeneously catalyzed lignin depolymerization reactions, by utilizing methods published in a previous study [12]. First, an extensive literature search in Scopus was carried out by using the search string base catalyzed lignin, finding a total of 146 documents that were then sorted. Of these, 57 documents were downloaded and 12 were selected to obtain the data [19-30], resulting in 143 experimental data points captured, out of which 60 are base catalyzed and 83 are not. In this search targeted documents contained the variables seen in Table 3-1; these variables were deemed the minimum necessary reported information required to predict the outcome of a base catalyzed lignin depolymerization experiment based on existing knowledge in the literature. Of these, reactor volume to solvent volume ratio serves as a rough estimate for the pressure since all studies involved used water as a solvent, and the relative fraction of lignin to water is very small. This study's leading hypothesis is that the parameters outlined in Table 3-1 should reliably predict the yield of bio-oil and solid residue from base catalyzed or not catalyzed hydrothermal lignin depolymerization. Although many studies report the variables outlined in Table 3-1, many studies contain methodological or experimental variables that deviate from the ones seen in the final dataset, notably the works by Miller, et al. [31, 32] that are considered foundational studies are not included due to them involving the usage of organic solvents. Another notable study by Schutyser, et al. [33] was considered but ultimately not used as part of the data set due to their usage of high pressure O_2. All experiments in the final dataset are limited to those involving water reaction media and not involving the use of pressurized air or oxygen.

"Solid residues" was chosen instead of "char" due to the ambiguity related to the terminology used; in many studies, they use the words char, coke, and repolymerized lignin without clear consistency, thus "solid residues" represents all of the solids obtained post-reaction. All data was manually extracted from the chosen studies either directly from tables or graphs by using graph digitizing software.

Originally, gas yield was also intended to be captured, however, gas yield was inconsistently reported in papers thus deemed not fit to be used in this study.

Table 3-1. Machine learning features and label names, along with their descriptions.

Feature and label names	Description
Lignin/H_2O ratio	Ratio of lignin to solvent in the experiment (mg/mL)
Lignin/catalyst ratio	Ratio of lignin to the homogeneous catalyst used in the experiment (wt/wt)
*Catalyzed or uncatalyzed	Whether the experiment was catalyzed or not
Reactor volume/H_2O ratio	Ratio of the volume of the reactor used to the volume of solvent in the experiment (vol/vol)
Reaction time	Reaction time in seconds
Temperature	Temperature in K
[1]Bio-oil yield	Bio-oil yield from lignin (wt%)
[1]Solid residue yield	Solid residue yield from lignin (wt%)

* This variable was one-hot encoded due to its categorical nature.
[1]Bio-oil and solid residue yield are labels.

In order to further clarify the distribution of the data captured from literature, violin plots were elaborated for all features and labels, illustrated in Figure 3-1.

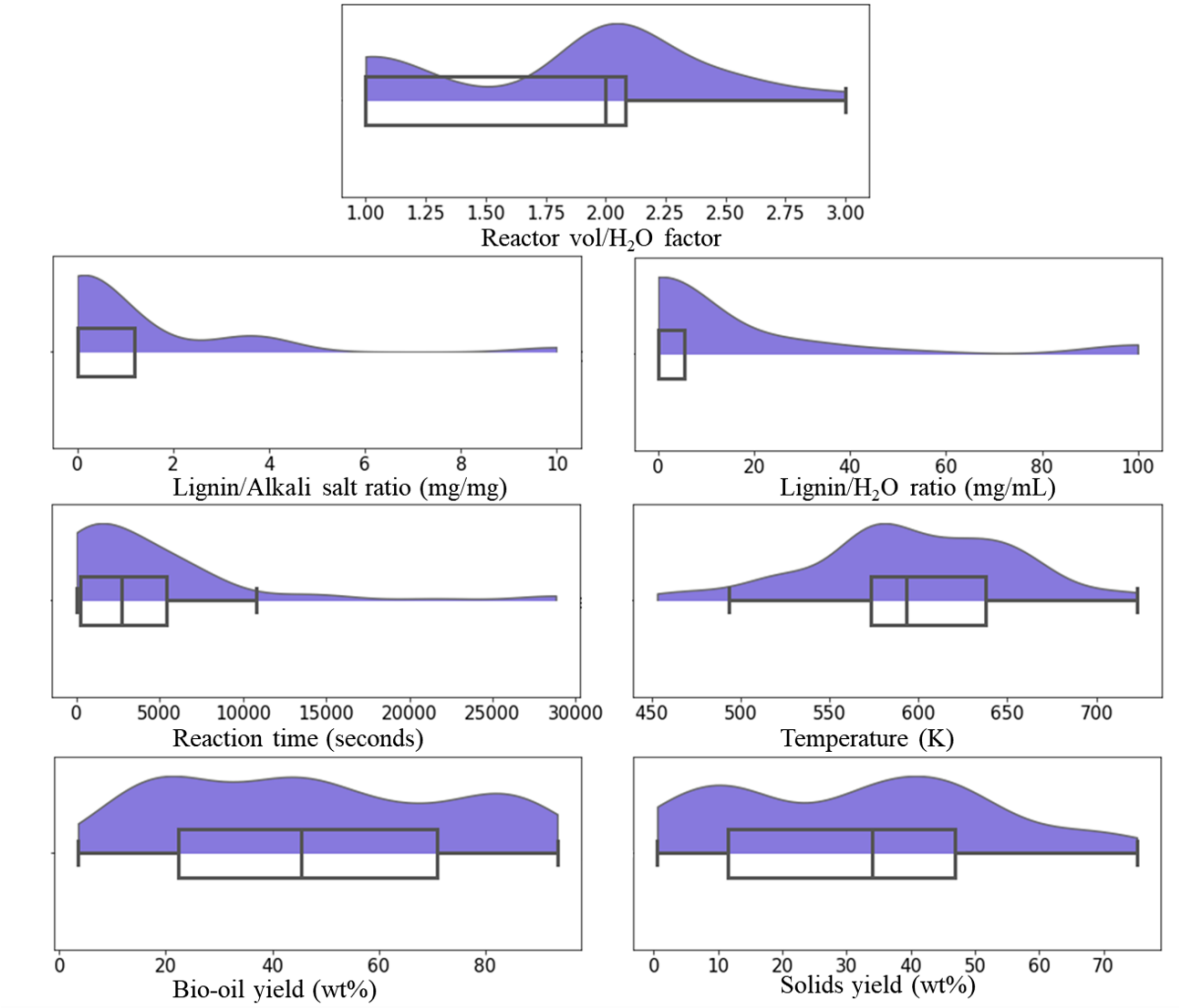


Figure 3-1. Compiled violin plots for features and labels captured from literature.

It is clear from Figure 3-1 that the distribution of the data capture is most often not normal, and in the case of reaction time, lignin/catalyst ratio and lignin/H_2O ratio they heavily skewed towards low values. While it is worth noting that most ML methods do not rely on assumptions of normal distribution in the data used to work properly, it does demonstrate that the experimental parameters found in literature are less diverse than one could imagine. This is most likely due to the natural tendency of researchers to further test experimental conditions that yielded good results for others, though it must be stated that it is not representative of the entire possible experimental space. Due to the lack of normality in the data captured, Spearman's rank correlation was used to analyze the data and is presented in Figure 3-2.

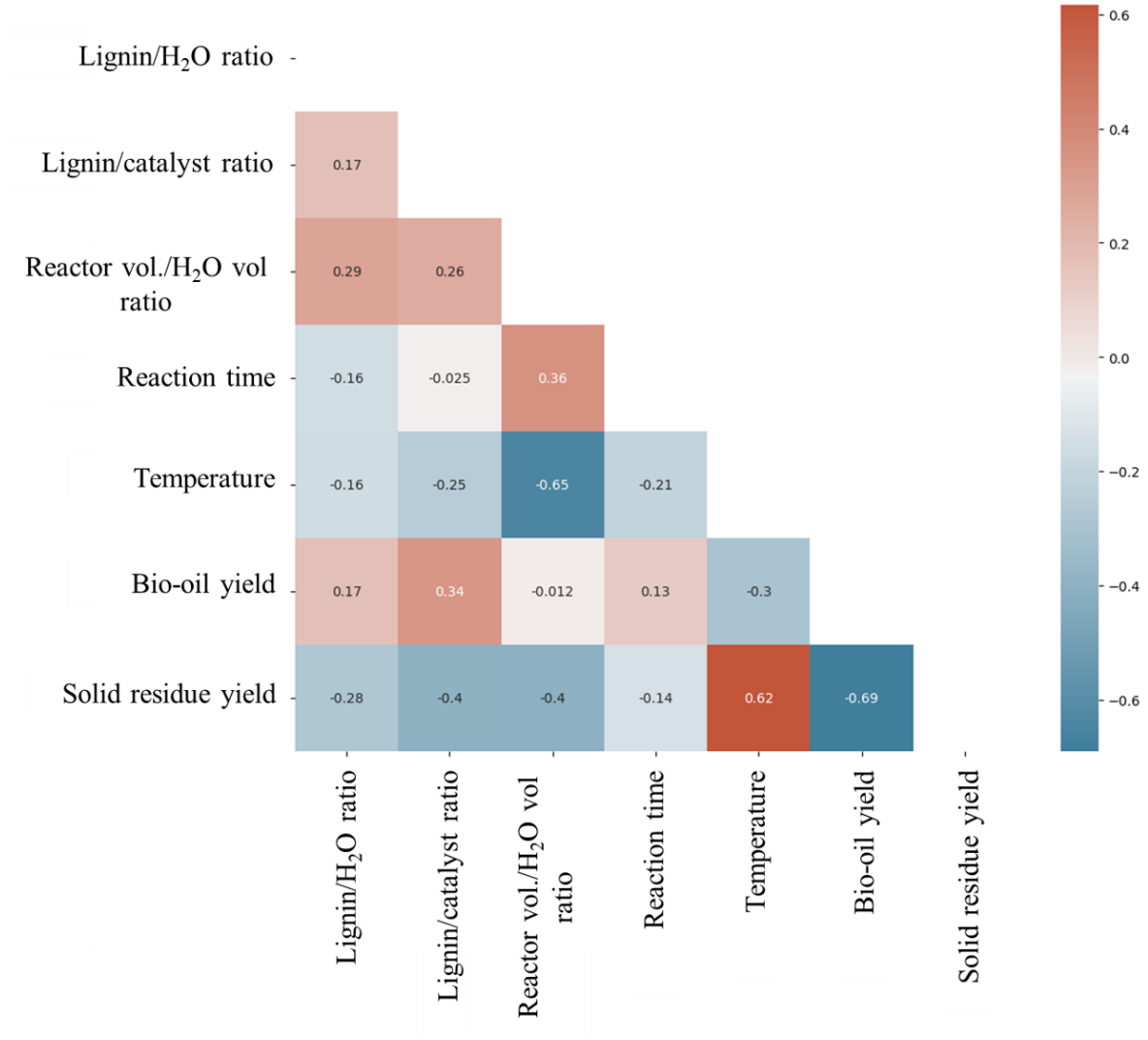


Figure 3-2. Spearman's rank correlation heatmap for data captured from literature.

While a clear correlation can be observed between the yield of bio-oil and solid residue due to their mutually exclusive nature, most other correlations are weak. Temperature, however, shows a very strong positive correlation with solid residue yield, which makes sense based on previous trends in literature.

During the data gathering phase two assumptions were made, firstly that the catalytic activity of different alkali salts is roughly the same, with NaOH, KOH and Na_2CO_3 being featured in this dataset. It is foreseen that this simplification would be a source of inaccuracy in the model, as it is known that there are measurable differences amongst different alkali salts with different lignins [34]. The second assumption is that heating rate did not play a significant role in the outcome of the experiment, it was chosen to not extract this data as it

was often not reported in the selected studies or could not be inferred reliably in the case of continuous flow experiments.

3.2.2 Machine learning methods, evaluation indicators and feature importance calculation

Scikit-learn free machine learning libraries for python were used to implement Extreme gradient boost regression (XGBoost) [35], other decision tree-based models (Random Forest, Gradient Boosting Regression, AdaBoost) were tested but XGBoost showed marginally higher performance. XGBoost is an ensemble algorithm that employs decision trees as weak learners and was chosen due to its ease of understanding, lack of necessity for data to follow a normal distribution, and capacity to handle continuous and categorical data. Most importantly the algorithm robust against overfitting, outliers, and noise [36]. In Figure 3-3 a diagram of how XGBoost model works is shown.

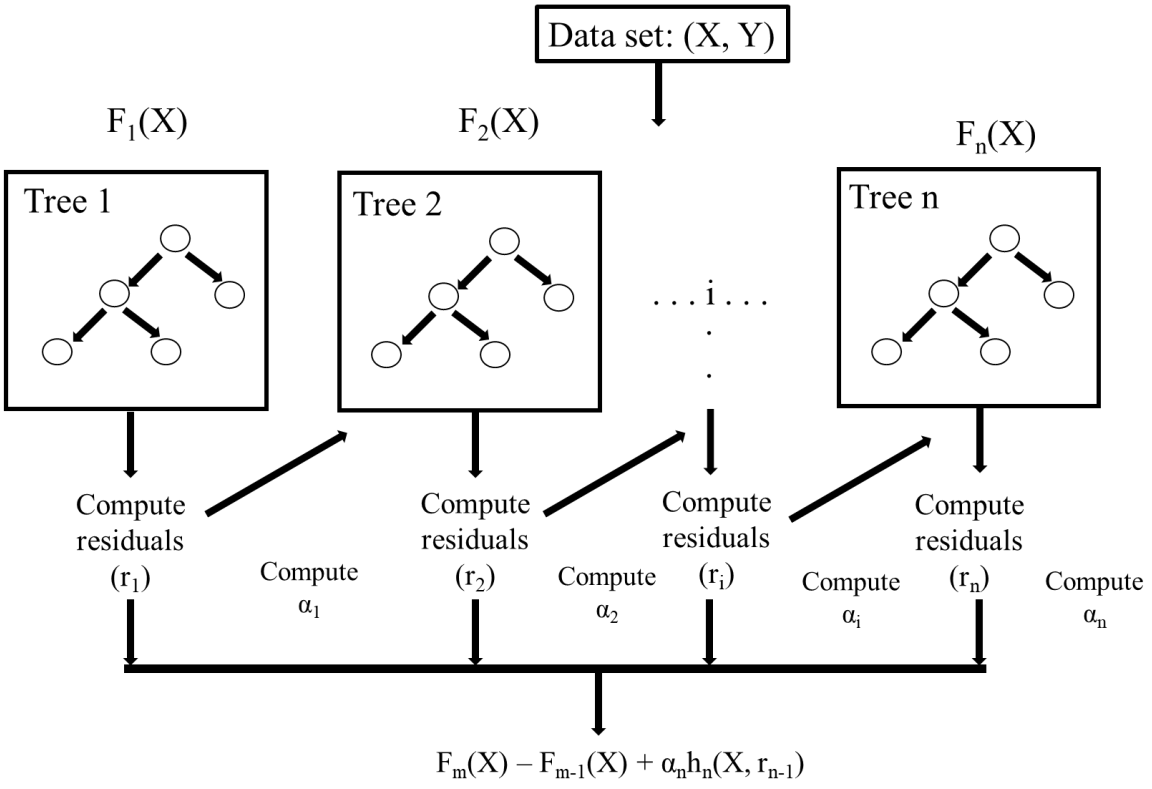


Figure 3-3. Representation of how XGBoost model works.

Where α_i and r_i are the regularization parameters and residuals computed with the i^{th} tree respectively, and h_i is a function that is trained to predict the residuals r_i using X for the i^{th}

tree. To compute α_i the residuals computed r_i are used and the following is computed, as shown in (1):

$$\arg \min_{\alpha} = \sum_{i=1}^{m} L(Y_i, F_{i-1}(X_i) + \alpha h_i(X_i, r_{i-1})) \qquad (1)$$

Where *L(Y, F(X))* is a differentiable loss function.

The reason why only decision tree-based models were tested is due to the heterogeneous nature of the data collected from multiple studies that may include implicit or explicit differences in terms of experimental practices. Purity of reactants and lignin used as well as differences in reactor design and workup were accounted for as much as possible in the data gathered, however, there may be differences or characteristics not explicitly stated in the studies where the data comes from. Additionally, XGBoost has the benefit of acknowledging sparse data distributions when assigning scores, which is relevant due to the nature of the data used in this study, as seen in Figure 3-1, where some of the distributions are heavily skewed in one direction due to the large number of instances of "0".

The data available was separated into training and testing data, accounting for 75% and 25% of the total respectively, 5-fold cross-validation was carried out in each instance to observe the bias and variance in each case. The performance of the models was evaluated by using coefficient of determination (R^2) and root-mean-squared-error (RMSE), whose equations are shown below, (2-1) and (2-2).

$$R^2 = 1 - \frac{\sum_i^n \left(Y_i^{exp} - Y_i\right)^2}{\sum_i^n \left(Y_i^{exp} - Y_{avg}^{exp}\right)^2} \qquad (2\text{-}1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_i^n (Y_i^{exp} - Y_i)^2} \qquad (2\text{-}2)$$

Where n represents the number of test samples, Y_i^{exp} denotes the experimental value, and Yi represents the predicted value. Y_{avg}^{exp} represents the mean value of Y_i^{exp} and Y_i, respectively.

Feature importance for the models was obtained both by permutation importance (PI) and SHAP (Shapley Additive explanations) values. PI calculates the total reduction in loss or impurity contributed by all splits for a given feature. This method is computationally very efficient and has been widely used in a variety of applications [37]. SHAP values is a method of feature importance estimation based on cooperative game theory where each feature is interpreted as a "player" and is used to increase transparency and interpretability of machine learning models [38]. Partial dependency plots were used estimate the impact of using different temperatures and reaction time in the yields of bio-oil and solid residues.

3.2.3 Materials

Dealkaline lignin (DL) was bought from Asahi Kasei chemicals, Japan. DL was completely soluble in aqueous alkali solutions of pH > 10. Other chemicals include NaOH (98%), acetone (98%), dichloromethane (99.8%) (DCM) and hydrochloric acid, all bought from Wako chemicals, Japan.

3.2.4 Base catalyzed lignin depolymerization experiments

Depolymerization experiments were carried out in 50 mL Taiatsu (Japan) TPR-5 reactors equipped with a pressure gauge, thermocouple and gas line for experiments at 250 °C, for experiments at 350 °C self-made autoclave reactors were used. Across all experiments 125-250 mg of DL was loaded in the reactors, then a varying amount (10 to 30 g) of water and 125 to 250 mg of NaOH was added. The reactor was then sealed and heated using heating jackets at an average of 10 °C/min until reaching the target temperature. Reaction time includes the heating ramp. During the course of the experiment pressure increased from atmospheric pressure to 3~5 MPa mainly due to water vapor pressure, as the amount of gas produced by the lignin was comparatively smaller. After the reaction time had elapsed the reactors were quickly cooled down by using an electric fan until the temperature dropped below 50 °C. The reactors were then opened to release the gas products and the liquid and solid products were then dumped into a beaker. The reactors were thoroughly washed and scrubbed with distilled water to remove any particles from the walls of the reactors.

The liquid and solid products were then acidified by using 2 M HCl until the pH reached 1~2 to precipitate lignin oligomers. Subsequently, the products were filtered with a pre-weighted 110mm filter paper and the aqueous fraction was combined with 30 mL DCM and thoroughly mixed to extract any present water-soluble aromatics. The aromatics in the DCM solution were later separated by using a rotary evaporator. Precipitated lignin was washed with acetone to separate acetone soluble products (ASP) which are reported as part of the bio-oil yield. Yields of bio-oil and solid residue were defined as follows:

$$\text{Bio} - \text{oil yield (wt\%)} = \left(\frac{\text{weight of ASP+DCM soluble organics}}{\text{weight of initial lignin}}\right) \text{x } 100 \qquad (4)$$

$$\text{Solid residue yield (wt\%)} = \left(\frac{\text{weight of solid residue}}{\text{weight of initial lignin}}\right) \text{x } 100 \qquad (5)$$

The gas fraction produced was not captured nor analyzed in these experiments.

3.3 Results and discussion

3.3.1 Evaluation of machine learning model performance for bio-oil yield prediction

In order to check for overfitting, XGBoost hyperparameters were tuned by using gridsearch, in addition, the models were compared with a regression algorithm (Gaussian process). The results shown in Table 3-2 indicate that although minor differences in R^2 and RMSE scores were found, the overall conclusions obtained using the XGBoost with the default hyperparameters would not change dramatically. Hence, further analysis was done based on the XGBoost model with default hyperparameters and hereinafter XGBoost with default hyperparameters is referred to as XGBoost.

Table 3-2. Accuracy/error measures for BO yield and solid residue yield prediction.

Model		**Training**		**Test**	
	Output	**R^2**	**RMSE (MJ/kg)**	**R^2**	**RMSE (MJ/kg)**
XGBoost (default parameters])	BO yield	0.99	1.85	0.83	10.52
	Solid yield	0.99	1.73	0.76	11.21
XGBoost (gridsearch optimized parameters)	BO yield	0.92	6.90	0.83	10.60
	Solid yield	0.907	5.92	0.813	9.98
Gaussian Process regression	BO yield	0.99	0.28	0.86	9.60
	Solid yield	0.99	0.26	0.86	10.42

Using the literature data, a prediction model for bio-oil yield was developed by first removing the yield of solid residues from the datasets, as bio-oil, solid residue and gas share a zero-sum relationship as lignin is the only reactant in these experiments. Shown in Figure 3-4, the prediction performance scores for the model developed can be observed with RMSE of 10.522 and R^2 Score of 0.836. The gray band indicates the 95% confidence interval.

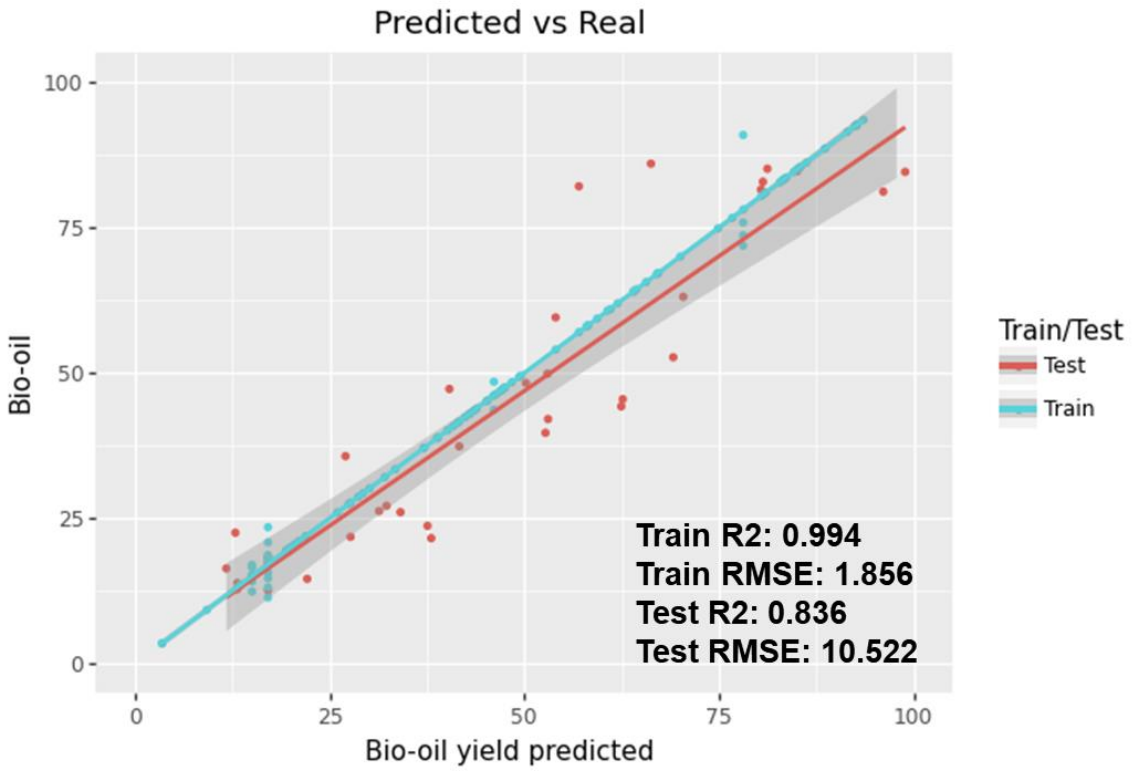


Figure 3-4. Prediction performance of XGBoost model for bio-oil yield and its associated RMSE and R^2 scores.

The model displayed better prediction capability at low and middle bio-oil yields (20~60 wt%), with predictions on the higher end accounting for most of the RMSE.

Based on this, the PI variable importance is also shown in Table 3-3, where reactor volume to H_2O volume ratio holds the highest importance followed by lignin to H_2O ratio, lignin to catalyst ratio, temperature and reaction time, in that order.

Table 3-3. PI feature importance for prediction on bio-oil from the XGBoost model.

Feature	**Importance**
Reactor vol/H_2O vol ratio	0.600
Lignin/H_2O ratio (mg/mL)	0.189
Lignin/Catalyst ratio	0.079
Temperature	0.079
Reaction time	0.050

These variables are known to have an impact in the outcome of the experiment based on the existing knowledge in the literature, however, the magnitude at which they do is not clear when comparing one variable against another. Though the PI importance values shown reflect "how important" one variable is compared to other, it must be kept in mind that this variable importance only pertains to a limited part of the possible experimental space.

It is well understood that temperatures close to the critical temperature (647.14 K) of water can offer higher yields of bio-oil [24]. Simultaneously, when the temperature exceeds the critical temperature of the water, a higher likelihood of gas-forming reactions taking place is also a possibility, both in the presence or absence of catalysts [39].

Regarding the importance of the ratio of lignin to solvent, it stands to reason that re-polymerization behavior that results in the formation of solid residues is intensified when the concentration of lignin is high, and because of the zero-sum relationship between bio-oil, solid residue, and gas, it makes sense that it shows high importance, it is also a behavior seen in lignin solvolysis studies [40].

The ratio of lignin to catalyst on the other hand is a more complicated issue, with various reports on what is the optimal concentration of catalyst in their experiments [41] from 2 to 4 wt% based on the quantity of lignin, these differences could be attributed to the nature of the lignin used in the experiments, as it is known that depending on the method used to isolate the lignin, its solubility in water or other solvents varies [42], thus impacting the results. It is understood that the presence of strong alkali salts in hydrothermal lignin depolymerization increases the yield of liquid products by incentivizing alkaline hydrolysis and dehydration reactions, the first of which is directly responsible for the cleave of β-O-4 bonds, and the second being indirectly aiding the depolymerization process by removing hydroxyl groups attached to both the aromatic ring itself, as well as the side chains attached to them.

The importance of reaction time is difficult to assess, as the published literature data used in this study included experiments with reaction times as short as 0.5 seconds [23], however, due to the homogeneous nature of these experiments it is reasonable to assume that there are no mass transfer limitations with the catalyst, and the severing of C-O-C bonds is known to be a fast reaction when using model compounds [43], it is likely that at least partial depolymerization into bio-oil-like mixture of monomers and oligomers can be achieved even with short reaction times.

The ratio of reactor volume to H_2O volume holds most of the importance, this ratio correlates to the operational pressure at a given temperature, however, because of this, its impact is also dependent on the temperature of the experiment.

In contrast to Table 3-3, in Figure 3-5 a beeswarm representation of the SHAP values obtained for the same model that also represent the importance of the features in the model

are shown. In this plot, features are ordered from top to bottom in order of magnitude of importance in the vertical axis, in the horizontal axis the magnitude of positive or negative contribution towards the predicted bio-oil yield can be seen, mediated by the color of the points, where red means high and blue means low. It is immediately clear that the order of the features in Figure 3-5 is not the same as that shown in Table 3-3, due to the way SHAP and PI are calculated.

The beeswarm plot in Figure 3-5 allows for interesting observations regarding temperature, as it indicates that very high values negatively impact the yield of bio-oil, which in the case of this study would mean 663~723 K, approaching gasification temperatures, while purple to blue values can be positively correlated with high bio-oil yield. Similarly, in terms of the ratio of lignin to alkali salt, a low to middle value appears to have either no impact or negative impact on the bio-oil yield, while a high loading of alkali salt increases bio-oil yield, presumably by increasing lignin solubility and guaranteeing interaction with the catalytically active OH^- ion the reaction media. The interpretation of the reactor volume to water volume ratio indicates that a low volume of water within the reactor with regards to the total volume available is negatively correlated with bio-oil yield. This ratio is meant to serves as a proxy for pressure inside the reactor under the assumption that most of the pressure during the process can be attributed to the presence of water at high temperatures, and not necessarily due to gas products. The interpretation of lignin to solvent ratio and reaction time is complicated as both high and low values can be seen over the horizontal axis. However, it is important to note that this means there are interactions between variables that allow for this to be true, particularly for absence/presence of catalyst, for example that a short reaction time in the presence of high alkali salt concentration could still result in high bio-oil yield, or that repolymerization into solid residues is minimized in the presence of enough alkali salts, thus leading to higher bio-oil yield as a co-consequence.

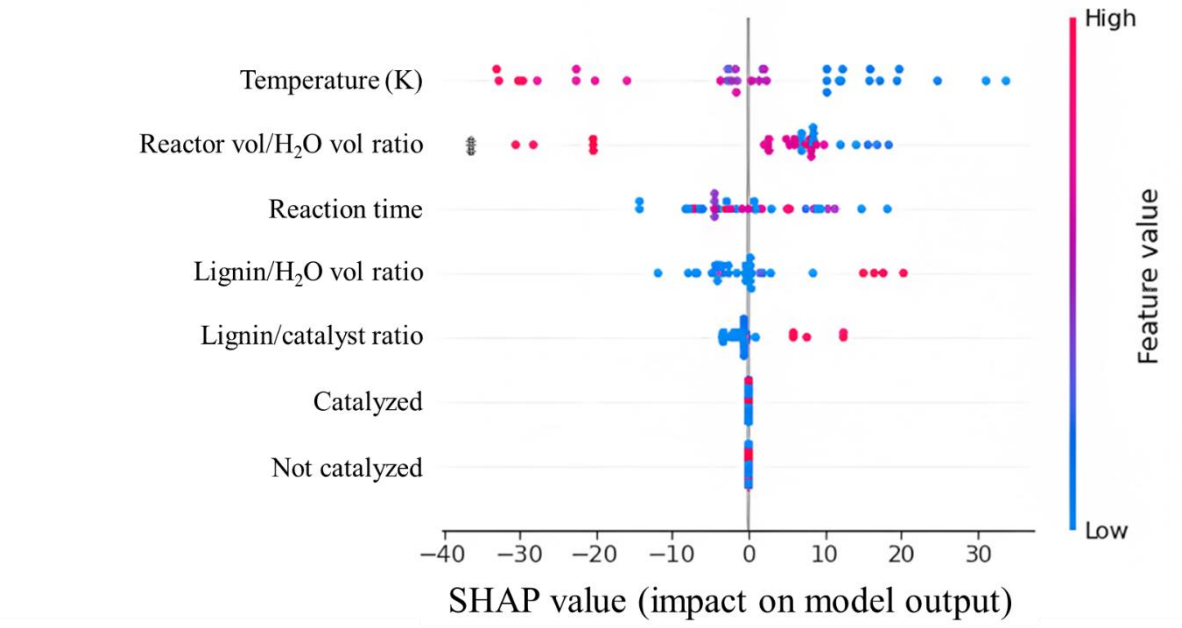


Figure 3-5. SHAP values beeswarm plot for XGBoost bio-oil prediction model.

Of the controllable variables in a given lignin depolymerization experiment, process temperature and reaction time are thought of as the most critical and straightforward directly controllable process parameters. In Figure 3-6, partial dependency plots that illustrates the impact of temperature and reaction time on bio-oil yield is shown, where boxplots show the possible range bio-oil yield for the various experiments in the dataset under the assumption that only temperature is changed to the values seen in the 10^{th}, 50^{th} and 90^{th} percentile of process temperature found in the dataset. For temperature, it a clear tendency can be observed for predicted bio-oil yield to decrease as temperature increases, which makes sense given that the higher temperature values in the dataset approach gasification temperatures, which results in the production of non-condensable gases at the expense of bio-oil yield and solid residue yield. For reaction time, a clear trend cannot be observed and with predicted bio-oil yield values fluctuating between slightly lower and then higher as the reaction time moves from the 50^{th} percentile to the 90^{th}. It is possible that the strongly skewed, non-normal distribution of reaction time in the dataset may negatively affect the interpretability of this variable.

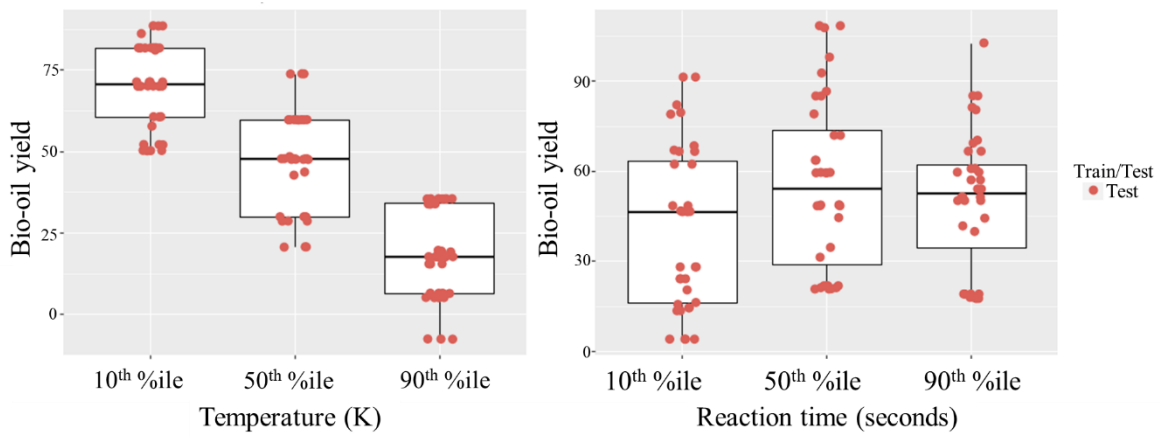


Figure 3-6. Partial dependency plots for temperature and reaction time impact on bio-oil yield.

3.3.2 Evaluation of machine learning model performance for solid residue yield prediction

The ML prediction models for solid residue were developed by first removing the yield of bio-oil from the datasets, as bio-oil, solid residue, and gas share a zero-sum relationship. In Figure 3-7, where the prediction performance of the model is displayed as a R^2 score and RMSE value; XGBoost model can be seen with a R^2 score of 0.764 and an RMSE of 11.218. The gray band in the graph represents the 95% confidence interval.

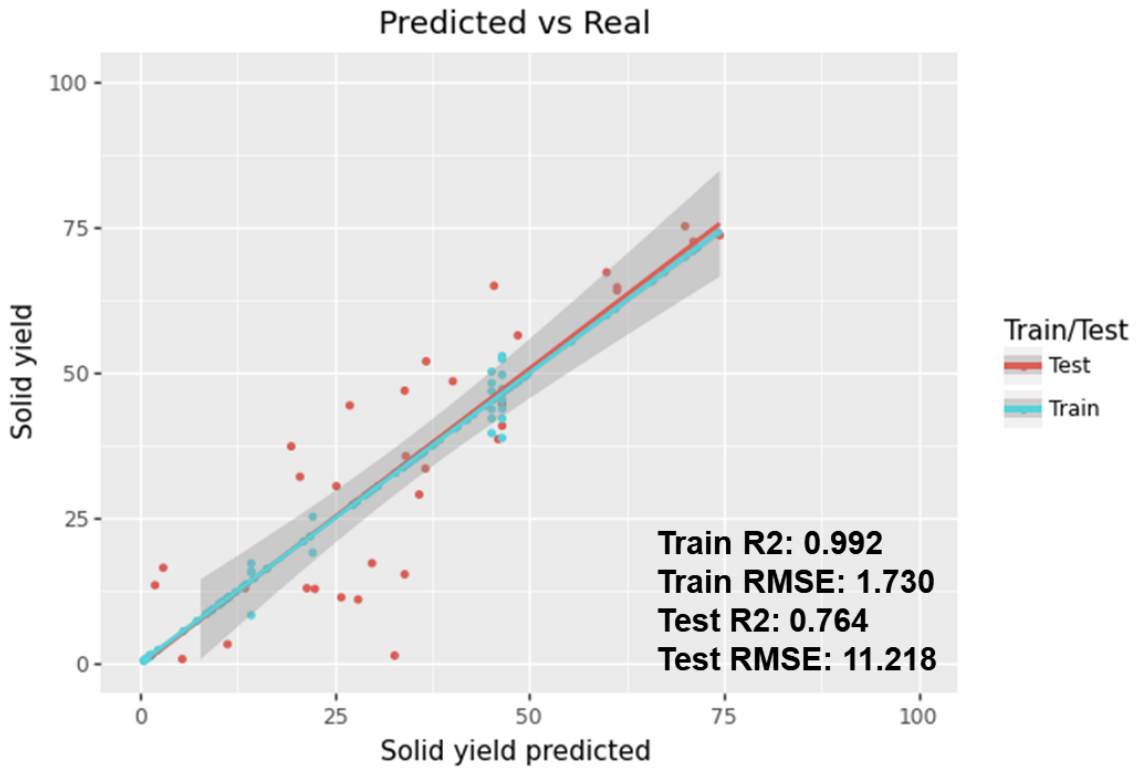


Figure 3-7. Prediction performance of XGBoost model for solid residue yield and its associated RMSE and R^2 scores.

The PI variable importance for the prediction of solid residues is shown in Table 3-4, where the magnitude and order of the features seen is different from that observed for bio-oil prediction. While it is not possible to extrapolate from these values which temperature is optimal for minimizing the yield of solid residue, within the context of the study there are data from experiments where low temperatures resulted in a higher likelihood of re-polymerization of lignin [24], note that in the data gathering, repolymerized lignin, char or coke were not distinguished, as different authors had used these terms disregarding their definition.

Additionally, it is also possible that traces of cellulose may contribute to the formation of char [44], and various studies included in the dataset use lignins that may contain such traces [19, 27]. In a similar manner to bio-oil yield, a high concentration of lignin in the reaction media may result in a higher likelihood that the species responsible for the formation of repolymerized lignin and char react to form a more solid residue. It is known in the literature that to allow the maximum possible lignin concentration in pilot-scale processes, formaldehyde-reacting chemical species are co-fed to the reactor, such as phenol [41]. While the presence of strong alkali salts is known to increase the yield of liquid products, it is also understood that their presence can also lead to the formation of phenolate ions, which can

then react with formaldehyde-like species resulting in condensation of fragments that results in more solid residues, as well as phenolate-ketone aldol condensation, this is illustrated in Figure 3-8.

Table 3-4. PI feature importance for prediction on solid residue yield from XGBoost model.

Feature	Importance
Lignin/H_2O ratio (mg/mL)	0.740
Reactor vol/H_2O vol ratio	0.150
Temperature	0.050
Reaction time	0.039
Lignin/Catalyst	0.029

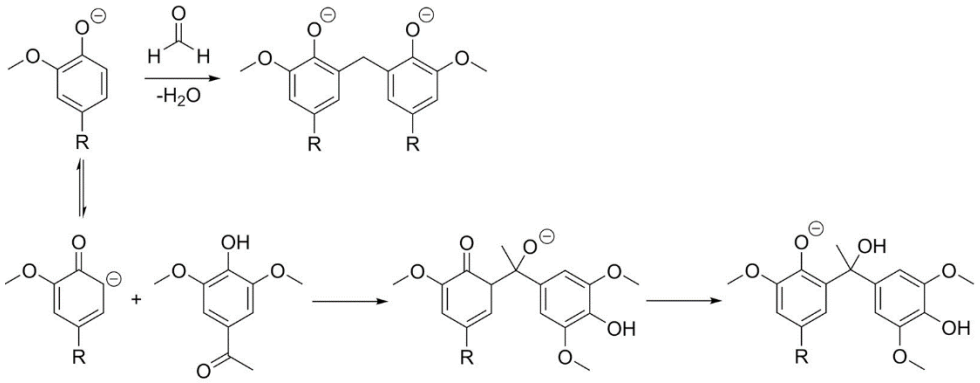


Figure 3-8. Phenolate-formaldehyde condensation (top), phenolate-ketone aldol reaction (bottom) (adapted from [24]).

In contrast to the values shown in Table 3-4, the beeswarm plot for SHAP values shown in Figure 3-9 tells a much more detailed and nuanced perspective on feature importance than that obtained from PI in Table 3-4. Here it can be observed that much like in the case of bio-oil yield, especially high temperatures are associated with higher yields of solid residue, being consistent with the zero-sum nature of bio-oil, solid residue and gas yield. High reaction times tend to be associated with higher solid residue yields, though not always, as there are some instances of high reaction times that did not particularly affect the yield of solid residue in either direction, again, indicating in those cases that an interaction between variables happen that allows for this to be possible. It is worth noting that the ratio of reactor volume to water volume in the experiment does not seem to play a large role in solid residue prediction, like it did for bio-oil yield, and that both lignin to catalyst and lignin to solvent ratio show a pattern opposite of that seen in beeswarm plot for bio-oil yield prediction in Figure 3-5.

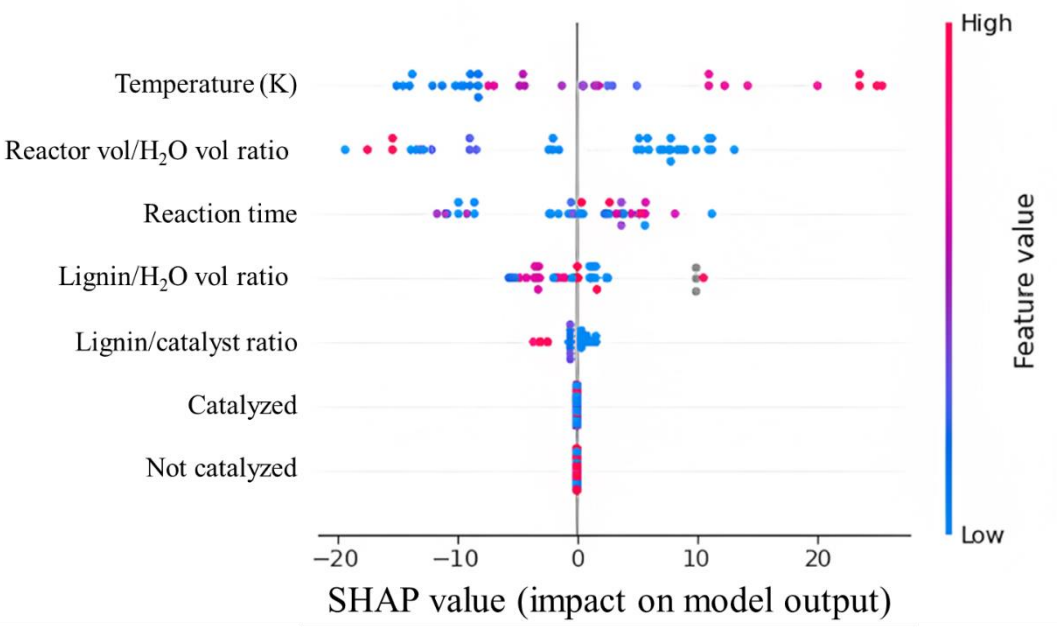


Figure 3-9. SHAP values bee swarm plot for XGBoost solid residue prediction model.

In Figure 3-10, a partial dependency plot displaying the effect of process temperature and reaction time on solid residue yield is shown. Higher temperature seems to be clearly associates with higher yield of solid residues, however, at the temperature found in the 50th percentile, a very erratic box-plot can be seen, which includes a few outliers (black dots). It is unclear why the predicted values are like this. Given that the temperature distribution is relatively normal-shaped and data is most abundant around the 50th percentile, it most likely not due to data sparsity. Reaction time on the other hand shows a decreasing trend as reaction time increases. As previously mentioned, this may be due to lignin's tendency to repolymerize as reaction time increases in particular circumstances. There may also be an interaction effect between temperature and reaction time with regards to solid residue yield, where higher temperatures result in higher solid residue yield if reaction time is extended.

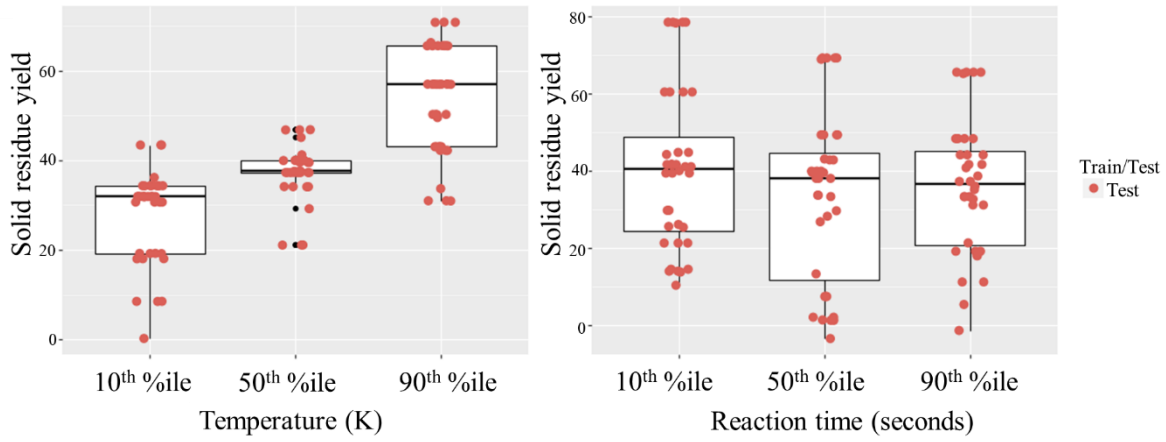


Figure 3-10. Partial dependency plots for temperature and reaction time impact on solid residue yield.

3.3.3 Experimental validation of predictive models

In order to further validate the predictive models made and obtain insight on the experiments they are based on; a series of experiments were carried out as shown in Table 3-5, with the experimental results reported being the average of triplicate runs. All experiments were carried out with NaOH as catalyst due to it being the most commonly seen in the published literature. Temperatures of 250 and 350 °C were chosen for the experiments due to them being the most representative within the dataset and also because of limitations on the temperature and pressure limits of the reactors used. As seen in Table 3-5, the experimental results can deviate strongly from the predictions made with the models, particularly in regard to bio-oil yield, with prediction of solid residues falling within the expected prediction error range of the trained model.

Considering the results obtained from the experimental work, literature was consulted to find experimental reasons as to why the bio-oil yield could vary across studies despite using the same or very similar reaction conditions. Two main sources of variation can be pointed at, first, the difference of solubility of different lignins in NaOH solutions [42] partly due to different phenol hydroxyl group content which impact the way the reaction occurs. Second, upon comparison of the various methodologies involved in the studies used as a source of data for the modeling, it is clear that the choices made in the post-reaction handling of the lignin products plays a large role in the resulting calculated yield of bio-oil and solid residue. The recovery of the soluble lignin phase by using an organic solvent and whether that is considered part of the bio-oil yield or not is questionable. From the data science-perspective, the dataset used contains multiple experimental instances where the results fluctuated heavily

for bio-oil yield at a given temperature and time, which may hinder the predictions made by the model.

Table 3-5. Experimental validation of predictive models for bio-oil yield and solid residue yield

T (°C)	Reaction time (hr)	Lignin-solvent ratio	Lignin-catalyst ratio	Reactor volume-solvent volume ratio	Experimental bio-oil yield (wt%) (E. bio-oil yield)	Predicted bio-oil yield (wt%)	Prediction gap for bio-oil[1]	Experimental solid residue yield (wt%)	Predicted solid residue yield (wt%)	Prediction gap for solid residue[1]
250	1	16.66	1	3.33	28.9	57.93	29.03	45.38	39.57	5.81
250	3	16.66	1	3.33	19.70	71.43	51.73	41.25	43.51	2.26
250	1	16.66	0.5	3.33	35.15	55.68	20.53	30.75	39.37	8.62
250	3	16.66	0.5	3.33	24.74	69.07	44.33	37.01	43.54	6.53
250	1	16.66	0.5	1.66	24.20	91.17	66.97	39.75	20.46	19.29
250	3	16.66	0.5	1.66	20.08	84.9	64.82	37.34	25.18	12.16
350	1	12.5	1	2	24.8	58.98	34.18	44.8	27.90	16.9
350	3	12.5	1	2	19.6	61.38	41.78	24.8	27.28	2.48
350	1	25	2	2	16.53	58.92	42.39	30.8	29.40	1.4
350	3	25	2	2	14.8	55.37	40.57	18.66	33.83	15.17
350	1	25	1	2	32.8	58.92	26.12	24.13	29.40	5.27
350	3	25	1	2	15.53	54.46	38.93	23.00	33.83	10.83

[1] "Prediction gap" is defined as the absolute difference between the model prediction and the experimental result.

3.3.4 Recommendations based on current and previous work

This study is the first attempt to use machine learning to predict experimental outcomes in base catalyzed or uncatalyzed hydrothermal lignin depolymerization.

The results reveal that significant predictive performance can be achieved for solid residues from the experiments, but bio-oil yield shows a large prediction gap when compared to the experiments done to test the model.

Previous work on predicting lignin depolymerization performance in heterogeneously catalyzed reactions [12] included more reaction parameters yet achieved higher prediction performance. In Table 3-6, a comparative table is shown, that includes previous work in predicting bio-oil yield in heterogeneously catalyzed lignin depolymerization and hydrothermal liquefaction of various feedstocks.

Table 3-6. Comparison of model bio-oil yield predicting performance with previous and related recent literature.

Feedstock and conversion method	Model used	Data used	Input features	Test R^2	Test RMSE	Ref.
Lignin; base catalyzed/non-catalyzed hydrothermal liquefaction	Extreme gradient boosting	143	6 features; Temperature, residence time, lignin/H_2O ratio, lignin/catalyst ratio reactor vol/H_2O ratio, based catalyzed (Catalyzed/uncatalyzed)	0.83	10.52	This study
Lignin; heterogeneously catalyzed liquefaction	Random forest	102	6 features; Temperature, residence time, solvent choice, active metal/lignin ratio, catalyst/solvent ratio and active metal/solvent ratio	0.90	6.03	[12]
Lignocellulosic waste; hydrothermal liquefaction	Random forest	117	10 features; Elemental composition (C, H, N, O and S), atomic ratio (H/C, O/C and N/C), temperature and residence time.	0.85	5.83	[45]
Algae; hydrothermal liquefaction	Gradient boost regression	310	15 features; Elemental composition (C, H, N, O and S), atomic ratio (H/C, O/C and N/C), biological composition (content of protein, lipids or carbohydrate), temperature and residence time and ash content	0.90	4.69	[46]
Wet biomass and wastes; hydrothermal liquefaction	Extreme gradient boosting	325	17 features; Elemental composition (C, H, O and N), atomic ratio (H/C, O/C and N/C), biological composition (content of protein, lipid or carbohydrate), ash content, residence time, temperature, initial pressure, reactor size, biomass loading, water and water to biomass ratio	0.87	5.41	[47]
Various types of biomass waste; hydrothermal liquefaction	Gaussian process regression	652	10 features; Elemental composition (C, H, O, N and S), ash content, operating dry matter, temperature, residence time and pressure	0.95	0.038	[13]

From Table 3-6, it can be appreciated that the prediction performance for bio-oil in this study is lower than those seen in other recent studies. It must be kept in mind that the other studies save for [12] focus on hydrothermal liquefaction of non-lignin feedstocks and do not involve the use of catalysts. However, what can be appreciated across these other studies is that comprehensive characterization of the feedstock used is associated with better prediction performance, for the most part. Many lignin depolymerization studies (including the ones used for data gathering in this study), do not characterize the lignin used, which could partly explain the gap in performance. From the perspective of SHAP values, this study's practical interpretability is high due to its focus on purely operation conditions, instead of intrinsic properties of the feedstock such as that seen in

[45] where marginally higher R^2 score was obtained but relies largely on feedstock's properties to predict the yield of bio-oil.

This study and the prior studies both fail to account for the fact that while bio-oil yield may be predictable, the more important target of lignin depolymerization is the yield of aromatic monomers obtained from lignin. Predicting aromatic monomer yield is a more difficult task, as very few publications do comprehensive quantification of the monomers obtained and often use different methodologies to do so, for example only analyzing the monomers in the aqueous phase of the experiment, the light oil phase, or all the phases, sometimes not explicitly stating what they did too. Regardless, the possibility of predicting the yield of aromatic monomers from a given lignin depolymerization experiment will depend not only on the reaction parameters used but also on the inherent chemical structure of the lignin used, as the upper limit of aromatic monomers obtainable depends on the magnitude of C-O-C bonds present in the lignin, as C-C bonds are often impossible to break during most thermochemical processes that do not involve a noble metal catalyst [48] or use temperatures that border on gasification. These chemical bonds can be quantified by nuclear magnetic resonance (NMR) [49].

Modeling the impact of homogeneous base catalysts such as the ones involved in this study represents a different challenge from that seen in other studies that attempt to describe material properties through the use of various intrinsic-measurable properties such as surface area and pore properties, or those based on chemistry principles such as adsorption energies on transition metal surfaces [50]. In contrast to this, the description of catalytically active chemical bases (NaOH for example) could be simpler as the role that the Na^+ and the OH^- ions play is well understood [51] and would require fewer descriptors.

Additional data is always desired when creating these kinds of predictive models, however, since each data point comes from a single experiment, it is understandably difficult to gather data in the magnitudes seen in other areas where ML modeling has been applied, such as sales, weather or classification of pictures. An emergent approach that could resolve this issue is using at least partially simulated data from experiments as part of the data used, however, complete simulation of lignin depolymerization processes has not been found in the literature, with the exception of studies that focus purely on theorical interactions of specific chemical bonds with catalysts [52], but not the complete depolymerization process itself perhaps due to the complexity of simulating heterogeneous polymers interacting with a solid catalyst.

Although this study does not deal with the lignin-first biorefinery concept, a recent work by [53] outlines extensive guidelines for the analyzing of data from lignin-first approaches, including feedstock analysis and process parameters, with the ambition of uniting the lignin-first research community around a common set of reportable metrics. These guidelines comprise standards and best practices or minimum requirements for feedstock analysis, stressing reporting of the fractionation efficiency, product yields, solvent mass balances, catalyst efficiency, and the requirements for additional reagents such as reducing, oxidizing, or capping agents.

These guidelines and minimum reporting requirements when publishing a paper can potentially allow for easier usage of data from literature in future ML-lignin depolymerization related studies, by guaranteeing that data across studies is compatible with each other. The formation of the guidelines needs further discussion in academic societies or perhaps in this journal in order to advance the lignin depolymerization machine learning field. Whether the guidelines described in [53] would be partially or entirely compatible with homogeneously catalyzed lignin depolymerization studies is an issue that needs to be analyzed critically.

3.4 Conclusions and future directions

Herein, XGBoost ML method was used to develop predictive models for bio-oil yield and solid residue from base catalyzed lignin depolymerization reactions, achieving R^2 scores of 0.80 and 0.87, respectively for the best model in each case. Results indicate that the relation between reaction parameters such as temperature and ratio of lignin to catalyst or solvent does not follow a linear relationship and are different for bio-oil yield and solid residue yield prediction. Yields of bio-oil and solid residue may be poor metrics for reaction performance evaluation, therefore, alternative metrics such as target chemical concentration or chemical bond concentration were suggested. Based on the contrast between the modelling and experimental work done, recommendations on how to report experimental results from lignin depolymerization were suggested, including proper characterization of lignin properties and experimental techniques, although more in-depth discussion and analysis is necessary for defining useful guidelines. Future research that regarding hydrothermal lignin liquefaction (catalyzed or not) should ideally aim to take advantage of data science and machine learning tools to address fundamental phenomenological issues such as under what conditions can C-O-C and C-C bonds be broken and what role (beneficial or not) do certain characteristics of lignin play, such as molecular weight distribution and ash content.

Having modelled and analyzed the two categories of studies that the lignin solvolysis literature can be divided into, it is possible to suggest that prediction of bio-oil yield and solid residues may not be the best possible performance metric for this kind of studies, and instead, metrics that are less influenced by the chemistry work-up used during the experiments should be chosen, ideally ones that are easy to standardize.

In the next chapter focus is shifted towards predicting and understanding how experimental parameters impact the HHV of the resulting bio-oil from lignin (and lignocellulosic) biomass solvolysis, as a possible alternative experimental performance metric.

References:

[1] *Maps and data - global ethanol production by country or region*. Alternative Fuels Data Center: Maps and Data - Global Ethanol Production by Country or Region. (2021, June). Retrieved January 8, 2022, from https://afdc.energy.gov/data/10331

[2] Hill, J., Nelson, E., Tilman, D., Polasky, S., & Tiffany, D. (2006). Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. *Proceedings of the National Academy of Sciences*, *103*(30), 11206–11210. https://doi.org/10.1073/pnas.0604600103

[3] Atabani, A. E., Mahlia, T. M. I., Anjum Badruddin, I., Masjuki, H. H., Chong, W. T., & Lee, K. T. (2013). Investigation of physical and chemical properties of potential edible and non-edible feedstocks for biodiesel production, a comparative analysis. *Renewable and Sustainable Energy Reviews*, *21*, 749–755. https://doi.org/10.1016/j.rser.2013.01.027

[4] Sanchez, A., & Gomez, D. (2014). Analysis of historical total production costs of cellulosic ethanol and forecasting for the 2020-Decade. *Fuel*, *130*, 100–104. https://doi.org/10.1016/j.fuel.2014.04.037

[5] Watkins, D., Nuruddin, M., Hosur, M., Tcherbi-Narteh, A., & Jeelani, S. (2015). Extraction and characterization of lignin from different biomass resources. *Journal of Materials Research and Technology*, *4*(1), 26–32. https://doi.org/10.1016/j.jmrt.2014.10.009

[6] Stijepovic, M. Z., Vojvodic-Ostojic, A., Milenkovic, I., & Linke, P. (2009). Development of a kinetic model for catalytic reforming of naphtha and parameter estimation using Industrial Plant Data. *Energy & Fuels*, *23*(2), 979–983. https://doi.org/10.1021/ef800771x

[7] Schutyser, W., Renders, T., Van den Bosch, S., Koelewijn, S.-F., Beckham, G. T., & Sels, B. F. (2018). Chemicals from lignin: An interplay of lignocellulose fractionation, depolymerisation, and upgrading. *Chemical Society Reviews*, *47*(3), 852–908. https://doi.org/10.1039/c7cs00566k

[8] Chio, C., Sain, M., & Qin, W. (2019). Lignin utilization: A review of lignin depolymerization from various aspects. *Renewable and Sustainable Energy Reviews*, *107*, 232–249. https://doi.org/10.1016/j.rser.2019.03.008

[9] Pandey, M. P., & Kim, C. S. (2010). Lignin depolymerization and conversion: A review of Thermochemical Methods. *Chemical Engineering & Technology*, *34*(1), 29–41. https://doi.org/10.1002/ceat.201000270

[10] Castro Garcia A., Cheng, S., & Cross, J. S. (2020). Solvolysis of Kraft lignin to bio-oil: A critical review. Clean Technologies, 2(4), 513–528. https://doi.org/10.3390/cleantechnol2040032

[11] Otromke, M., White, R. J., & Sauer, J. (2019). Hydrothermal base catalyzed depolymerization and conversion of technical lignin – An introductory review. *Carbon Resources Conversion*, *2*(1), 59–71. https://doi.org/10.1016/j.crcon.2019.01.002

[12] Castro Garcia, A., Shuo, C., & Cross, J. S. (2022). Machine learning based analysis of reaction phenomena in catalytic lignin depolymerization. *Bioresource Technology*, *345*, 126503. https://doi.org/10.1016/j.biortech.2021.126503

[13] Shafizadeh, A., Shahbeig, H., Nadian, M. H., Mobli, H., Dowlati, M., Gupta, V. K., Peng, W., Lam, S. S., Tabatabaei, M., & Aghbashlo, M. (2022). Machine learning predicts and optimizes hydrothermal liquefaction of biomass. Chemical Engineering Journal, 445, 136579. https://doi.org/10.1016/j.cej.2022.136579

[14] Funkenbusch, L. L. T., Mullins, M. E., Vamling, L., Belkhieri, T., Srettiwat, N., Winjobi, O., Shonnard, D. R., & Rogers, T. N. (2018). Technoeconomic assessment of hydrothermal liquefaction oil from lignin with catalytic upgrading for Renewable Fuel and chemical production. *WIREs Energy and Environment*, *8*(1). https://doi.org/10.1002/wene.319

[15] Yong, T. L.-K., & Matsumura, Y. (2012). Reaction kinetics of the lignin conversion in Supercritical Water. *Industrial & Engineering Chemistry Research*, *51*(37), 11975–11988. https://doi.org/10.1021/ie300921d

[16] Nguyen, T. D., Maschietti, M., Belkheiri, T., Åmand, L.-E., Theliander, H., Vamling, L., Olausson, L., & Andersson, S.-I. (2014). Catalytic depolymerization and conversion of Kraft lignin into liquid products using near-critical water. *The Journal of Supercritical Fluids*, *86*, 67–75. https://doi.org/10.1016/j.supflu.2013.11.022

[17] Li, H., Chen, J., Zhang, W., Zhan, H., He, C., Yang, Z., Peng, H., Leng, L. (2023). Machine-learning-aided thermochemical treatment of biomass: A Review. Biofuel Research Journal, 10(1), 1786–1809. https://doi.org/10.18331/brj2023.10.1.4

[18] Zhang, W., Chen, Q., Chen, J., Xu, D., Zhan, H., Peng, H., Pan, J., Vlaskin, M., Leng, L., Li, H. (2023). Machine learning for hydrothermal treatment of biomass: A Review. Bioresource Technology, 370, 128547. https://doi.org/10.1016/j.biortech.2022.128547

[19] Toledano, A., Serrano, L., & Labidi, J. (2012). Process for olive tree pruning lignin revalorisation. *Chemical Engineering Journal*, *193-194*, 396–403. https://doi.org/10.1016/j.cej.2012.04.068

[20] Erdocia, X., Prado, R., Corcuera, M. Ã., & Labidi, J. (2014). Influence of reaction conditions on lignin hydrothermal treatment. *Frontiers in Energy Research*, *2*. https://doi.org/10.3389/fenrg.2014.00013

[21] Fernández-Rodríguez, J., Erdocia, X., Sánchez, C., González Alriols, M., & Labidi, J. (2017). Lignin depolymerization for phenolic monomers production by sustainable processes. *Journal of Energy Chemistry*, *26*(4), 622–631. https://doi.org/10.1016/j.jechem.2017.02.007

[22] Mahmood, N., Yuan, Z., Schmidt, J., & (Charles) Xu, C. (2013). Production of polyols via direct hydrolysis of Kraft lignin: Effect of process parameters. *Bioresource Technology*, *139*, 13–20. https://doi.org/10.1016/j.biortech.2013.03.199

[23] Yong, T. L.-K., & Matsumura, Y. (2013). Kinetic analysis of lignin hydrothermal conversion in sub- and supercritical water. *Industrial & Engineering Chemistry Research*, *52*(16), 5626–5639. https://doi.org/10.1021/ie400600x

[24] Jensen, M. M., Madsen, R. B., Becker, J., Iversen, B. B., & Glasius, M. (2017). Products of hydrothermal treatment of lignin and the importance of ortho-directed repolymerization reactions. Journal of Analytical and Applied Pyrolysis, 126, 371–379. http`s://doi.org/10.1016/j.jaap.2017.05.009

[25] Hidajat, M. J., Riaz, A., Park, J., Insyani, R., Verma, D., & Kim, J. (2017). Depolymerization of concentrated sulfuric acid hydrolysis lignin to high-yield aromatic monomers in basic sub- and supercritical fluids. *Chemical Engineering Journal*, *317*, 9–19. https://doi.org/10.1016/j.cej.2017.02.045

[26] Rana, M., Taki, G., Islam, M. N., Agarwal, A., Jo, Y.-T., & Park, J.-H. (2019). Effects of temperature and salt catalysts on depolymerization of Kraft lignin to aromatic phenolic compounds. *Energy & Fuels*, *33*(7), 6390–6404. https://doi.org/10.1021/acs.energyfuels.9b00808

[27] Dell'Orco, S., Miliotti, E., Lotti, G., Rizzo, A. M., Rosi, L., & Chiaramonti, D. (2020). Hydrothermal depolymerization of Biorefinery lignin-rich streams: Influence of reaction conditions and catalytic additives on the organic monomers yields in biocrude and aqueous phase. *Energies*, *13*(5), 1241. https://doi.org/10.3390/en13051241

[28] Long, J., Xu, Y., Wang, T., Shu, R., Zhang, Q., Zhang, X., Fu, J., & Ma, L. (2014). Hydrothermal depolymerization of lignin: Understanding the structural evolution. *BioResources*, *9*(4). https://doi.org/10.15376/biores.9.4.7162-7175

[29] Abad-Fernández, N., Pérez, E., & Cocero, M. J. (2019). Aromatics from lignin through ultrafast reactions in water. *Green Chemistry*, *21*(6), 1351–1360. https://doi.org/10.1039/c8gc03989e

[30] Jiang, W., Lyu, G., Wu, S., & Lucia, L. A. (2016). Near-critical water hydrothermal transformation of industrial lignins to high value phenolics. *Journal of Analytical and Applied Pyrolysis*, *120*, 297–303. https://doi.org/10.1016/j.jaap.2016.05.017

[31] Miller, J. E., Evans, L., Littlewolf, A., Trudell, D. E. (1999). Batch microreactor studies of lignin and lignin model compound depolymerization by bases in alcohol solvents. Fuel, 78(11), 1363–1366. https://doi.org/10.1016/s0016-2361(99)00072-1

[32] Miller, J. E., Evans, L., Littlewolf, A., Trudell, D. E. (2002). Batch microreactor studies of lignin depolymerization by bases. 1. alcohol solvents. https://doi.org/10.2172/800959

[33] Schutyser, W., Kruger, J. S., Robinson, A. M., Katahira, R., Brandner, D. G., Cleveland, N. S., Mittal, A., Peterson, D. J., Meilan, R., Román-Leshkov, Y., Beckham, G. T. (2018). Revisiting alkaline aerobic lignin oxidation. Green Chemistry, 20(16), 3828–3844. https://doi.org/10.1039/c8gc00502h

[34] Belkheiri, T., Andersson, S.-I., Mattsson, C., Olausson, L., Theliander, H., & Vamling, L. (2018). Hydrothermal liquefaction of Kraft lignin in sub-critical water: The influence of the sodium and potassium fraction. Biomass Conversion and Biorefinery, 8(3), 585–595. https://doi.org/10.1007/s13399-018-0307-9

[35] Chen, T., & He, T. (2020, June 11). *xgboost: eXtreme Gradient Boosting*. Retrieved January 14, 2022, from https://mran.microsoft.com/snapshot/2020-07-15/web/packages/xgboost/vignettes/xgboost.pdf

[36] Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, 1–15. https://doi.org/10.1007/3-540-45014-9_1

[37] Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. Bioinformatics, 26(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

[38] Molnar, Christoph. "Interpretable machine learning: A guide for making black box models explainable." (2018). https://christophm.github.io/interpretable-ml-book/

[39] Furusawa, T., Sato, T., Saito, M., Ishiyama, Y., Sato, M., Itoh, N., & Suzuki, N. (2007). The evaluation of the stability of Ni/MgO catalysts for the gasification of lignin in Supercritical Water. *Applied Catalysis A: General*, *327*(2), 300–310. https://doi.org/10.1016/j.apcata.2007.05.036

[40] Kim, J.-Y., Oh, S., Hwang, H., Cho, T.-su, Choi, I.-G., & Choi, J. W. (2013). Effects of various reaction parameters on solvolytical depolymerization of lignin in sub- and supercritical ethanol. *Chemosphere*, *93*(9), 1755–1764. https://doi.org/10.1016/j.chemosphere.2013.06.003

[41] Katahira, R., Mittal, A., McKinney, K., Chen, X., Tucker, M. P., Johnson, D. K., & Beckham, G. T. (2016). Base-catalyzed depolymerization of Biorefinery Lignins. *ACS Sustainable Chemistry & Engineering*, *4*(3), 1474–1486. https://doi.org/10.1021/acssuschemeng.5b01451

[42] Sameni, J., Krigstin, S., & Sain, M. (2017). Solubility of lignin and acetylated lignin in organic solvents. *BioResources*, *12*(1). https://doi.org/10.15376/biores.12.1.1548-1565

[43] Wu, X., Fu, J., & Lu, X. (2013). Kinetics and mechanism of hydrothermal decomposition of lignin model compounds. *Industrial & Engineering Chemistry Research*, *52*(14), 5016–5022. https://doi.org/10.1021/ie302898q

[44] Scheirs, J., Camino, G., & Tumiatti, W. (2001). Overview of water evolution during the thermal degradation of cellulose. *European Polymer Journal*, *37*(5), 933–942. https://doi.org/10.1016/s0014-3057(00)00211-1

[45] Leng, L., Zhang, W., Chen, Q., Zhou, J., Peng, H., Zhan, H., Li, H. (2022). Machine learning prediction of nitrogen heterocycles in bio-oil produced from hydrothermal liquefaction of biomass. Bioresource Technology, 362, 127791. https://doi.org/10.1016/j.biortech.2022.127791

[46] Zhang, W., Li, J., Liu, T., Leng, S., Yang, L., Peng, H., Jiang, S., Zhou, W., Leng, L., Li, H. (2021). Machine learning prediction and optimization of bio-oil production from hydrothermal liquefaction of algae. Bioresource Technology, 342, 126011. https://doi.org/10.1016/j.biortech.2021.126011

[47] Katongtung, T., Onsree, T., Tippayawong, N. (2022). Machine learning prediction of biocrude yields and higher heating values from hydrothermal liquefaction of wet biomass and wastes. Bioresource Technology, 344, 126278. https://doi.org/10.1016/j.biortech.2021.126278

[48] Hernández-Ramos, F., Fernández-Rodríguez, J., Alriols, M. G., Labidi, J., & Erdocia, X. (2020). Study of a renewable capping agent addition in lignin base catalyzed depolymerization process. *Fuel*, *280*, 118524. https://doi.org/10.1016/j.fuel.2020.118524

[49] Evstigneev, E. I. (2011). Factors affecting lignin solubility. Russian Journal of Applied Chemistry, 84(6), 1040–1045. https://doi.org/10.1134/s1070427211060243

[50] Capanema, E. A., Balakshin, M. Y., Kadla, J. F. (2004). A comprehensive approach for quantitative lignin characterization by NMR spectroscopy. Journal of Agricultural and Food Chemistry, 52(7), 1850–1860. https://doi.org/10.1021/jf035282b

[51] Roberts, V. M., Stein, V., Reiner, T., Lemonidou, A., Li, X., & Lercher, J. A. (2011). Towards quantitative catalytic lignin depolymerization. Chemistry - A European Journal, 17(21), 5939–5948. https://doi.org/10.1002/chem.201002438

[52] Monti, S., Srifa, P., Kumaniaev, I., Samec, J. S. (2018). ReaxFF simulations of lignin fragmentation on a palladium-based heterogeneous catalyst in methanol–water solution. The Journal of Physical Chemistry Letters, 9(18), 5233–5239. https://doi.org/10.1021/acs.jpclett.8b02275

[53] Abu-Omar, M. M., Barta, K., Beckham, G. T., Luterbacher, J. S., Ralph, J., Rinaldi, R., Román-Leshkov, Y., Samec, J. S., Sels, B. F., Wang, F. (2021). Guidelines for performing lignin-first biorefining. Energy & Environmental Science, 14(1), 262–292. https://doi.org/10.1039/d0ee02870c

Chapter 4:

Prediction of higher heating values in bio-oil from solvothermal biomass conversion and bio-oil upgrading given discontinuous experimental conditions

4.1 Introduction

It is foreseen that liquid hydrocarbons will still be part of our lives for the foreseeable future, due to several key technologies that cannot be electrified or powered by alternative sources, such as fuel for aviation, heavy trucks, and maritime vessels [1].

To overcome this, extensive research has been devoted to the conversion of renewable biomass resources that can potentially be transformed into molecules that can fulfill the role that fossil fuels currently serve [2]. Great success has been found in the conversion of edible biomasses such as vegetable oils and simple carbohydrates to biodiesel [3] and ethanol [4], respectively. However, these have attracted criticism for their potential impact on food prices, and thus, the conversion of non-edible biomass has been promoted [5].

Amongst non-edible biomass, lignocellulosic biomass is the most abundant. However, due to its recalcitrance, high-temperature thermochemical conversion methods are used to convert it to more useful forms, such as bio-oil or syngas [6]. The production of bio-oil from lignocellulosic feedstocks has seen large progress especially through solvothermal conversion methods that allow

the use of moderate temperatures, and often results in bio-oil with better higher heating value (HHV), which is key for fuel purposes.

This bio-oil, nevertheless, still requires to be upgraded to improve its HHV and other fuel properties such as viscosity and corrosiveness, by reducing its oxygen content by using organic solvents and hydrogen gas, usually in the presence of a catalyst [7]. The reactions associated with the upgrading of bio-oil fall under the umbrella term of hydrodeoxygenation (HDO), wherein oxygen is removed by the action of hydrogen through hydrogenation or hydrogenolysis [8]. However, compared to the hydrocarbon mixtures found in crude oil, the biomass and bio-oil contain a higher diversity of molecules and structures and it is not easy to keep track of which reactions are happening. The distribution of chemicals found in the bio-oil being dependent on the properties of the lignocellulosic biomass feedstock used to produce it, with lignin-heavy feedstocks resulting in higher concentrations of aromatic chemicals and cellulose being converted into ketones, furans, and sugars [9]. This variation in lignocellulosic biomass composition in addition to the great diversity of experimental choices such as solvents, catalysts, presence or absence of hydrogen gas, and reaction conditions results in a large number of possible experiments. This also resulted in extensive research that has become more popular in recent decades, as seen in figure 4-1, where the number of results for the search string bio-oil upgrading in Web of Science shows a sudden increase at the start of 2010's.

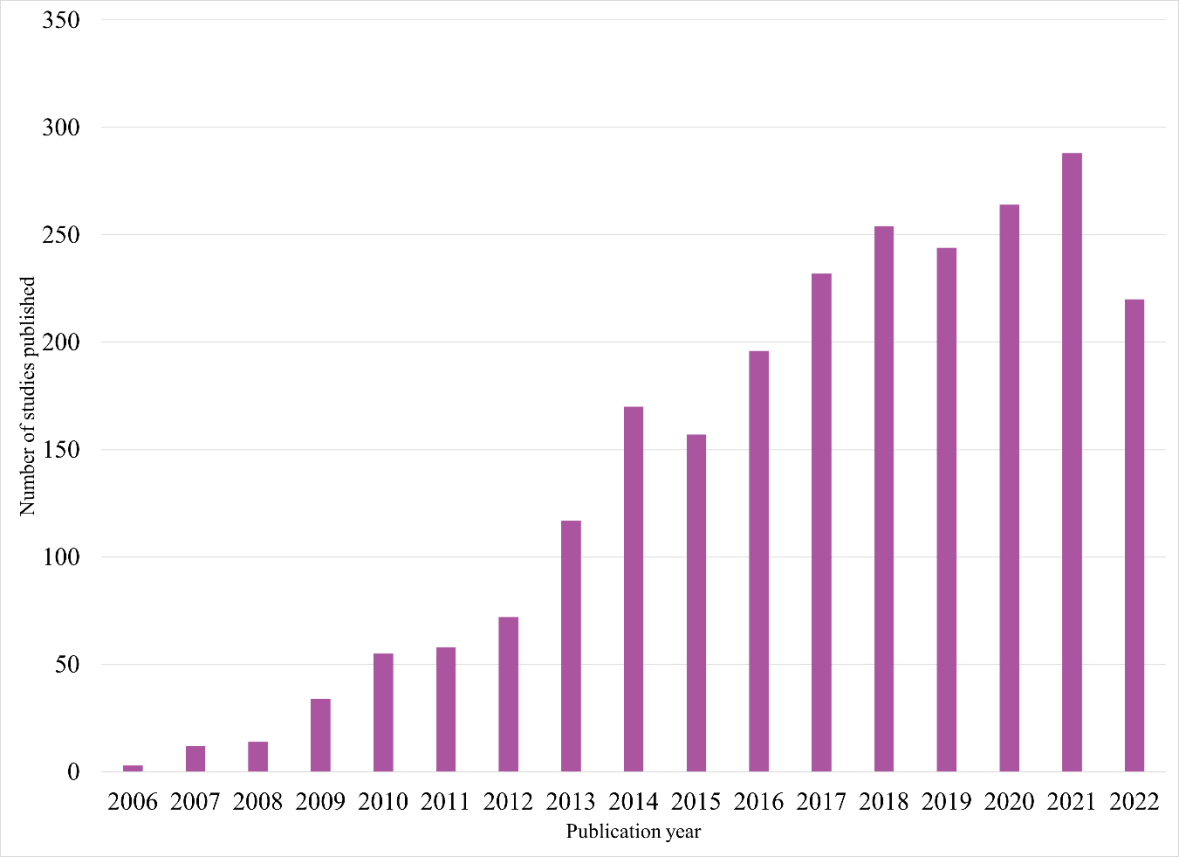


Figure 4-1. Publication trend of studies related to bio-oil upgrading found in Web of Science using bio-oil upgrading as the search string.

Many studies on the topic of "biomass to bio-oil" and "bio-oil upgrading" have very different methodologies that are many times not necessarily justified in terms of why they choose certain reactants, process conditions, or strategies in their experiments.

Typical types of studies in solvothermal biomass conversion to bio-oil or bio-oil upgrading include:

- Non-catalyzed, solvolysis/upgrading in water or mixtures of organic solvents and water [10]
- Solvolysis catalyzed by metallic heterogeneous catalysts in the presence of H_2 gas.
- Processing with non-standard heating methods, such as microwaves [11].
- Bio-oil or solid feedstock (SF) being either untreated woody biomass [12] or isolated lignin [13].
- Usage of extremely low (~200 °C) [14] or really high temperatures (350 °C~) [15].

In turn, this makes the studies hard to compare and results in slow progress in this research field, as extrapolations from very different studies do not seem superficially compatible.

Of the metrics evaluated for the production of bio-oil and its upgrading, HHV is often pointed as the most important, with viscosity, corrosiveness, and other fuel properties largely correlating in a positive way with the increase of HHV [16]. In addition to this, higher HHV bio-oil is more economically valuable as it can be used on higher-standard combustion engines of different kinds.

Recent advances in the usage of machine learning have allowed for the development of models that can predict HHV in raw biomass [17] using data from proximate analysis, torrefied biomass as a function of the treatment conditions [18] and bio-oil derived from hydrothermal liquefaction of wet biomass and wastes [19], providing not only good prediction performance, with R^2 scores ranging from 0.83 to 0.93, in spite of relying on small, human-curated datasets originating from literature. In addition to the prediction performance, interpretation of the models made was also possible through the use of partial dependency plots [19] or SHAP values [18], obtaining insight into how the features in the model impact the phenomena or properties responsible for the values obtained.

Inspired by these studies, in this work, knowledge of chemical engineering and machine learning were combined. Specifically, I used knowledge on the biomass liquefaction processes to construct a dataset based on studies that share common experimental conditions and variables. Then, machine learning was used to bridge the differences in feedstock, solvent choice, catalysts active media and experimental variables, which cannot be accomplished by simple linear or curve

fitting. The resulting models can predict the change in HHV in biomass-to-bio-oil processes and the upgrading of the bio-oil by relying on data extracted from the literature, obtaining simultaneously variable importance that provides insight into the mechanisms involved in the processes as well as useful observations for future research on bio-oil production and upgrading. The results in this study demonstrate that a few processing conditions across studies have the biggest impact in the resulting HHV of the bio-oil produced or upgraded. To date, this is the first study to predict the increase in HHV of bio-oil as a function of its upgrading process conditions.

4.2 Materials and methods

4.2.1 Data collection and pre-processing

Many process variables are known to affect biomass-to-bio-oil and bio-oil upgrading**.** Based on the extensive consultation on the literature and previous work [20] the studies that report the experimental data outlined in Table 4-1 were gathered, including HHV which was used as the target of the predictive model. These features were chosen based on comprehensive analysis of the different measurable process parameters described across studies on this topic. The change in HHV from the original biomass to bio-oil, as well as the change in bio-oil before and after upgrading was denoted as "ΔHHV". After this, an exhaustive search of the literature was carried out to look for studies that focused on solvothermal upgrading of biomass to bio-oil and bio-oil upgrading. This was done by using the following search strings in Web of Science in July of 2022:

- Solvolysis biomass
- Bio-oil upgrading

This resulted in an initial number of 172 and 2,392 documents, respectively, that were then screened to make sure they reported as many of the process variables noted in Table 4-1. After the screening process, a total of 15 and 29 papers were selected that fit the criteria, resulting in a total of 175 and 211 data points. The references to these papers are available in Appendix A supplementary data file. Studies that were not selected were deemed unfit due to non-compatible methodologies or underreporting of results and process conditions. This search for data and processing centers exclusively on changes to HHV or resulting HHV in bio-oil from experiments, but not directly on the chemical species found in the bio-oil, which are beyond the scope of this work. All methodologies from selected papers were carefully analyzed for compatibility of results amongst each other. The HHV values used for dataset were either directly measured by the researchers or calculated through Dulong's formula using the BO's elemental composition.

Table 4-1. Machine learning features and label names, along with their descriptions.

Feature and label names	Description
Elemental composition (wt%)	Concentration of C, H, O, and N elements in the feedstock
BO/SF original HHV (MJ/kg)	Original higher heating value of the feedstock, measured or calculated from the elemental composition
*Catalyst name	Active phase of the catalyst used in the experiment
*Solvent name	Solvent name
Solid feedstock name	Feedstock name, all wood and grass varieties were grouped together
Reaction time (min)	Reaction time in minutes
Temperature (°C)	Temperature in °C
Active metal/solvent ratio (mg/mL)	Ratio of active metal in catalyst to solvent
Active metal/feedstock ratio (mg/mg)	Ratio of active metal in catalyst to feedstock
Catalyst metal/solvent ratio (mg/mL)	Ratio of catalyst to solvent
Catalyst metal/feedstock ratio (mg/mg)	Ratio of catalyst to feedstock
Feedstock/solvent ratio (mg/mL)	Ratio of feedstock to solvent
H^+ ion added (mol)	Moles of H^+ added as strong acid equivalent
H_2 pressure factor (MPa H_2*mL)	Estimation of H_2 gas used, defined as the product of the pressure of H_2 and the "difference between reactor volume and solvent volume"
Reactor volume - solvent volume (mL)	Defined as the reactor volume minus the volume of solvent used
[1]Final HHV (MJ/kg)	Final HHV value of the bio-oil, measured or calculated
[1]ΔHHV (MJ/kg)	Change of HHV value of the bio-oil, measured or calculated

* This variable was one-hot encoded due to its categorical nature.
[1]Final HHV and ΔHHV were the labels in this study.

From the data captured, the distribution of values for elemental concentration, reaction time, reaction temperature, original HHV value and change in HHV value after processing is shown as violin and box-plots in Figure 4-2 a) for bio-oil and Figure 4-2 b) for SF. These features were represented as violin plots due to their direct impact in the outcome of a given experiment.

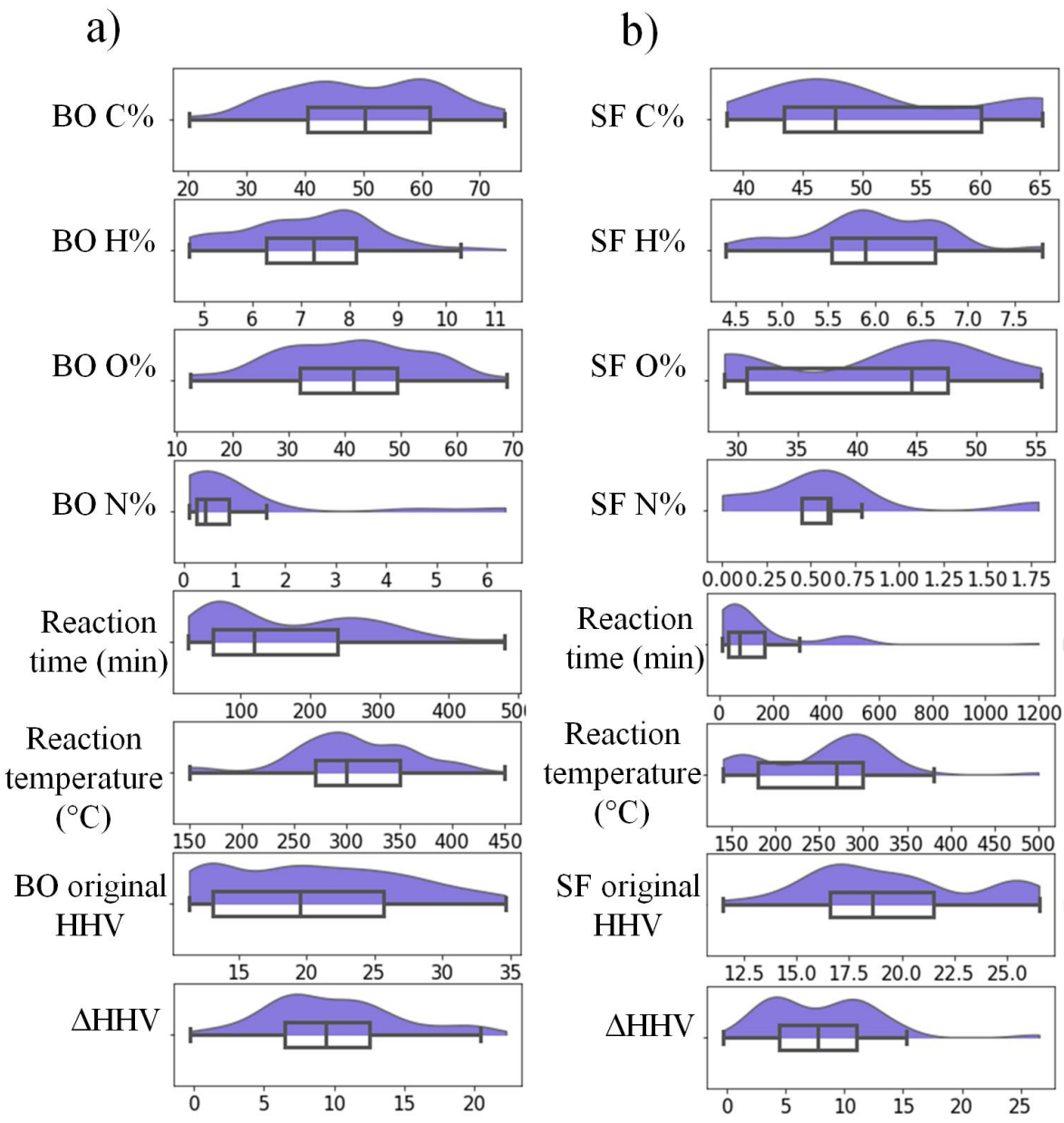


Figure 4-2. Violin distribution and box-plots for elemental composition, reaction time, reaction temperature, original HHV and change in HHV after processing for a) Bio oil (BO) and b) SF

Here, it can be observed that distributions of the features and labels are most often not normally distributed. Although most machine learning methods do not rely on assumptions of data normality to work properly, it is clear that the measures of correlation such as Pearson's correlation matrix would not work properly (as it relies on data normality).

4.2.2 Machine learning method used and evaluation indicators

The processes in this study were modeled using an Extreme Gradient Boosting (XGBoost) machine. It must be noted that other ML methods were also tested, however, XGBoost showed marginally higher performance. XGBoost has been extensively applied in modeling-related processes. This work opts to highlight two main attributes of XGBoost as a modeling approach which makes it a strong candidate for the intended application. First, XGBoost is an ensemble of regression trees (visually summarized in Figure 4-3). Being part of the family of ensemble models means that the XGBoost model is typically composed of hundreds of regression trees, which each make a partial estimate of the variable being predicted, i.e., ΔHHV or final HHV in this case.

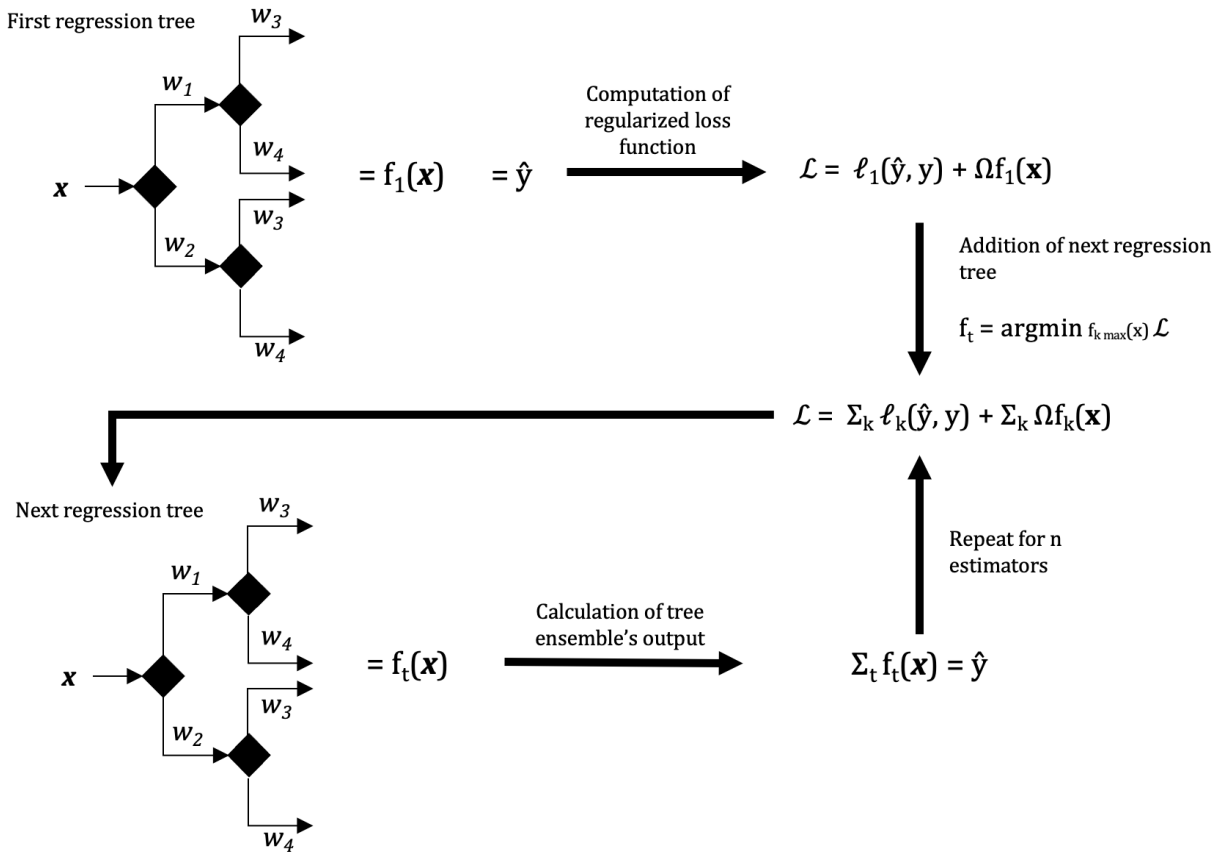


Figure 4-3. Visual representation of the learning process of XGBoost.

Regression trees are composed of layered 'branches' and scored 'leaves' (leaf weight, *wk*) as shown in Figure 4-3. At each branch, a data point is assigned to a leaf or branch in the next layer according to the value of a certain feature. Features that are relatively important to the predicted variable will assume this role in many branches. The last layer of branches is assigned to a leaf with the continuous score assigned to this leaf serving as the contribution of that regression tree to the predicted variable. Individually, the regression trees are 'weak learners', being oversimplified models and having a tendency for overfitting because of their structure. However, when the output

of these models is aggregated, gross errors and noise can be averaged out, while the consistent inferences across many regression trees are highlighted.

In particular, a regression tree that assigns conditional scores in the manner described is well-suited to the data in this study. This study uses potentially heterogeneous data collected from multiple studies which may have implicit and explicit differences. The examples of implicit differences include undocumented details about the experimental methodology (e.g., purity of reactants used, number of effective catalytic active sites in the catalyst, or differences in work-up during experiments) while the examples of explicit differences include differences in feedstock and catalyst, which are clearly identifiable. Most likely, these differences would result in a skewed, multimodal, or otherwise, non-normal distribution, which makes them unsuitable for most of the models (e.g., fitted lines or curves). Regression trees are not under any assumption of a probability distribution and are mostly deemed to be appropriate for these applications.

The second reason why XGBoost was selected for this study is that its assignment of scores acknowledges the potential of sparse datasets. In general, sparse datasets are those with many 0 elements. This is a common phenomenon in machine learning. In the context of application in this research, some data points may be missing one or more features as the documented factors and parameters differ from study to study. In addition, categorical variables such as the solvent type or the catalyst type need to be one-hot encoded, which means an integer 0 and 1 is assigned to indicate if a data point uses a particular solvent. The presence of 0 values in the dataset tends to be a problem for most models, which must consider the zeroes as a continuous variable. In that sense, the high frequency of zeroes could make the mean magnitude of certain parameters seem lower. Parametric methods which are reliant on these statistics would thus be skewed in response to the presence of the 0s. On the other hand, XGBoost is sparse-aware in the sense that, in its assignment of a next branch or leaf, a specific assignment can be made for the 0 value, thus allowing it to accommodate missing and one-hot encoded data.

The performance of the model was evaluated by using the coefficient of determination (R^2) and root-mean-squared-error (RMSE), whose equations (3-1) and (3-2) are shown below, respectively.

$$R^2 = 1 - \frac{\sum_i^n \left(Y_i^{exp} - Y_i\right)^2}{\sum_i^n \left(Y_i^{exp} - Y_{avg}^{exp}\right)^2} \qquad (3\text{-}1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_i^n (Y_i^{exp} - Y_i)^2} \qquad (3\text{-}2)$$

Where n represents the number of test samples, Y_i^{exp} denotes the experimental value, and Y_i represents the predicted value. Y_{avg}^{exp} represents the mean value of Y_i^{exp} and Y_i, respectively.

Data for all models was split for 80% training and 20% testing, all models used n_estimators = 5000, with all other hyperparameters being set to were taken from defaults in the Scikit-learn libraries. K-fold validation was used to evaluate model performance.

4.2.3 Feature importance calculation

Feature importance for the models was obtained by using the SHapley Additive exPlanations (SHAP) values [21]. Each feature has a corresponding calculated SHAP value representing the contribution of that feature toward the prediction. These values are based on the marginal contribution of the feature, or the difference between the values predicted by a model including that feature and one without it. The difference may either be positive or negative, indicating whether or not that feature makes a positive or negative contribution to the prediction. A multivariate model such as this research can be decomposed into many models using different combinations of features as inputs. As such, SHAP values are the weighted sum of the marginal contributions of each feature to the prediction. This is illustrated by equations (3)~(5)

$$mc_{x1,\{x1\}} = \hat{y}_{\{x1\}} - \hat{y}_{\{\emptyset\}} \quad (3\text{-}3)$$

$$mc_{x1,\{x1,x2\}} = \hat{y}_{\{x1,x2\}} - \hat{y}_{\{x2\}} \quad (3\text{-}4)$$

$$SHAP_{x1} = w_1 * mc_{x1,\{x1\}} + w_2 * mc_{x1,\{x1,x2\}} + w_3 * mc_{x1,\{x1,x3\}} + \cdots \quad (5)$$

$$where \sum_i w_i = 1$$

Each data point has a corresponding predicted value, which corresponds to its SHAP value. The SHAP values for the dataset can be interpreted collectively to understand the general behavior of the model for different inputs. This can be used to confirm the logic of the model, i.e., whether or not it follows the known or hypothesized effect of certain parameters on HHV. It can also confirm that less important features have zero contribution instead of adding noise to the prediction. As the calculated SHAP values follow a continuous scale, the SHAP values of one-hot-encoded categorical features cannot be reliably interpreted and were excluded from the analysis.

4.3 Results and discussion

4.3.1 Evaluating prediction accuracy

In a similar manner to the previous chapter, in order to check for overfitting, XGBoost hyperparameters were tuned by using gridsearch. The results shown in Table 4-2 indicate that although minor differences in R^2 and RMSE scores were found, the overall conclusions obtained using the XGBoost with the default hyperparameters would not change dramatically. Hence, further analysis was done based on the XGBoost model with default hyperparameters and hereinafter XGBoost with default hyperparameters is referred to as XGBoost.

This study consists of four models, given two predicted values and two processes (biomass solvolysis to produce bio-oil and bio-oil upgrading). The accuracy of each model was evaluated based on common error measures (i.e., R^2, RMSE) and compared the linear plot of predicted and reported values. The error measures indicate good fitting of the XGBoost model for final HHV and ΔHHV prediction in both the training and test sets. The R^2 values range from 0.96-0.99 (training) and 0.77-0.86 (test) across the four models. The RMSE, which corresponds to the average difference in real units of HHV, ranges from 0.42-0.89 MJ/kg (training) and 1.78-2.16 MJ/kg (test). A summary of the error measures is shown in Table 4-2.

Table 4-2. Accuracy/error measures for final HHV and ΔHHV prediction of models trained.

Model		**Training**		**Test**	
XGBoost (default parameters])		R^2	RMSE (MJ/kg)	R^2	RMSE (MJ/kg)
	Solvolysis of lignocellulosic SF				
	Final HHV prediction	0.99	0.42	0.83	1.94
	ΔHHV prediction	0.99	0.42	0.79	1.78
	Bio-oil upgrading				
	Final HHV prediction	0.97	0.89	0.86	2.12
	ΔHHV prediction	0.96	0.89	0.77	2.16
XGBoost (gridsearch optimized parameters)	**Solvolysis of lignocellulosic SF**				
	Final HHV prediction	0.93	1.18	0.83	1.91
	ΔHHV prediction	0.94	1.14	0.64	2.36
	Bio-oil upgrading				
	Final HHV prediction	0.94	1.11	0.86	2.06
	ΔHHV prediction	0.95	0.97	0.73	2.40

In figure 4-4, the linear plots of the predicted and reported HHV values are used to provide more insight into the sources of error. For solvolysis, there are a few predicted values with a large deviation from the real reported value, which skews the entire RMSE. Specifically, some data points for the final HHV are significantly overpredicted on the lower end of the regression line, indicating insufficient data on the lowest HHV values. Conversely, using ΔHHV shows a wider variance in error values, yet a lower error magnitude as a whole. As the final HHV can be derived from ΔHHV and initial HHV, the latter model may be more reliable for making estimations on HHV. On the other hand, the regression lines for bio-oil upgrading indicate a tendency to underpredict lower values and overpredict higher values. This is more evident for the ΔHHV, while all data points for the final HHV adhere to the regression line, except for a few outliers.

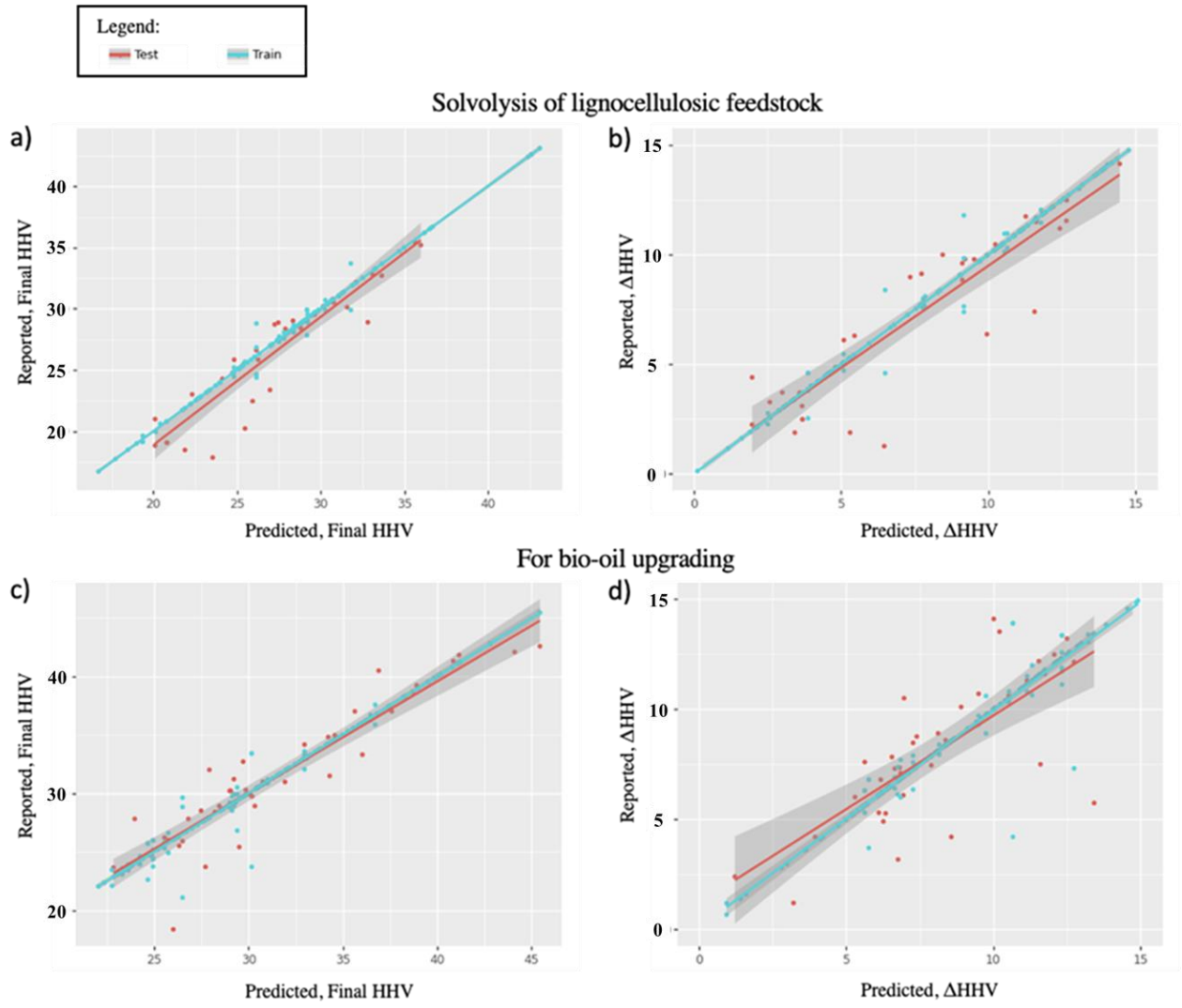


Figure 4-4. Model performance for final HHV and ΔHHV for lignocellulosic SF conversion to bio-oil through solvolysis and bio-oil upgrading. a) Final HHV for solvolysis bio-oil, b) ΔHHV for solvolysis bio-oil, c) Final HHV for bio-oil upgrading, and d) ΔHHV for bio-oil upgrading.

4.3.2 Evaluating the model's logic and interpretability

SHAP values were used to understand the model's logic for prediction. In figure 4-5, the beeswarm plots of the most important variables for the models are displayed for both final HHV and ΔHHV of solvothermal conversion to bio-oil of lignocellulosic SF to bio-oil in a) and b), and c) and d) for bio-oil upgrading.

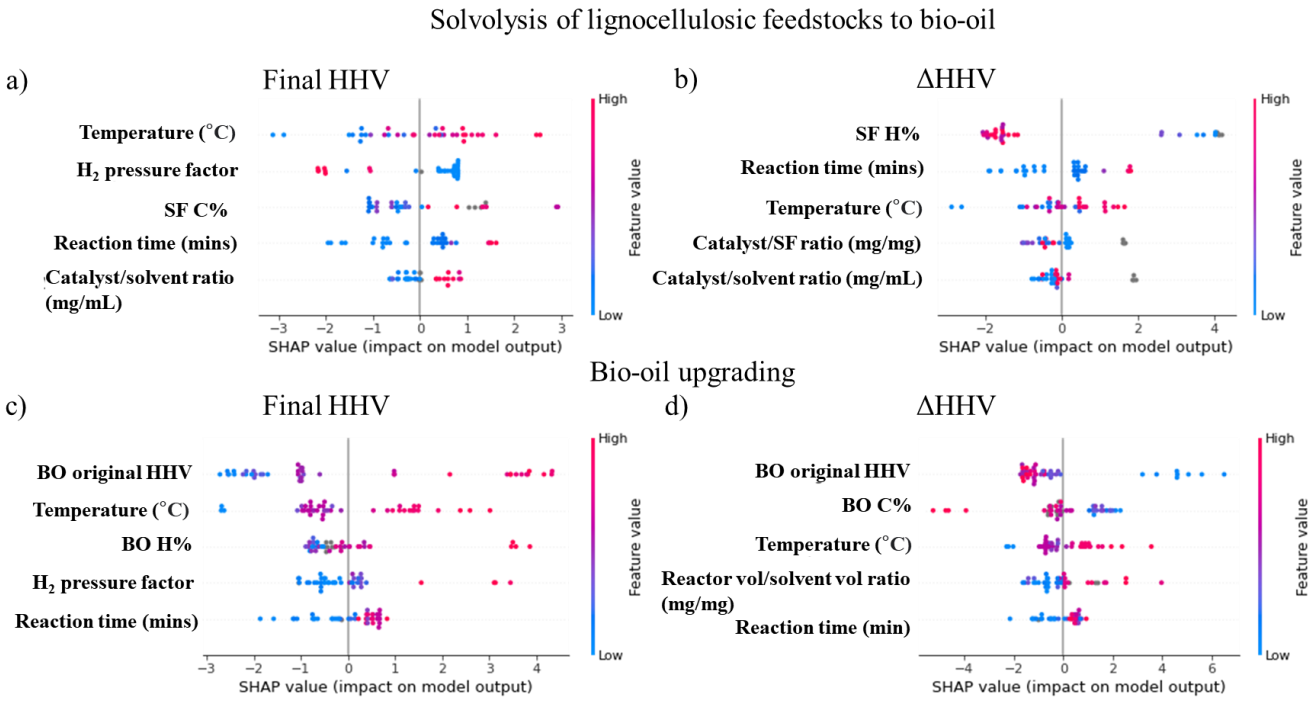


Figure 4-5. SHAP values for (a) final HHV and (b) ΔHHV from solvolysis of lignocellulosic SF; and SHAP values for (c) final HHV and (d) ΔHHV from bio-oil upgrading.

In figure 4-5 a), we can observe a clear tendency for final HHV. The value of the final HHV increases with the increasing temperature and the decrease in temperature is associated with the lower final HHV in the produced bio-oil. In the context of the data used to train the model, this could be explained by the reactions associated with the removal of oxygen, such as hydrogenolysis and hydrogenation, which are more prevalent in the temperature range of 250 - 350 °C [22].

The concentration of elemental carbon in the original biomass is shown to be important for predicting the final value of HHV of the produced bio-oil. Its importance is largely a co-consequence of the zero-sum relationship in elemental composition, meaning that higher carbon content is associated with higher hydrogen content and lower oxygen content. It is important to note that a high starting oxygen content in the feedstock does not necessarily mean that the final HHV of the resulting bio-oil will be low, but rather that the conditions of the solvothermal process would have to be tailored to remove more oxygen. Longer reaction times are associated with higher final HHV of the bio-oil, which stands to reason, as if oxygen-removing reactions continue to take

place for a longer time, it should result in lower final oxygen content, therefore, higher final HHV. However, at the right reaction conditions, short reaction times can still result in a high final HHV of the resulting bio-oil, depending on the combination of reactants and process conditions. The extended reaction time may be detrimental to the yield of bio-oil in cases where lignin-derived compounds can re-polymerize into higher molecular weight fragments that form char [23]. This can be prevented by the usage of short-chain alcohols such as methanol and ethanol as solvents in the reaction, due to their ability to react with lignin-derived reaction intermediaries and prevent the formation of char [24] examples of which represent a large part of the used dataset.

The role of catalyst to solvent ratio can be an issue of reaction optimization depending on the choice of heterogeneous catalyst, or the absence of it. Many of the transition metal species found in the experiments that compose the dataset used can react with the short-chain alcohols used (methanol or ethanol) to generate hydrogen [25] or promote alkylation reactions [26] that ultimately result in higher final HHV of the bio-oil produced. It must be pointed out that there are a significant number of non-catalyzed experiments in the dataset used, in which case, the ratio of catalyst to solvent was defined as 0, thus the above observations would not apply in these cases. The distribution of H_2 pressure factor indicates that the lower the number (less moles of hydrogen) the higher the resulting final HHV should be. This goes against the common understanding that more H_2 gas should result in higher HHV, due to the removal of oxygen in the bio-oil. This unexpected pattern may be due to the sparse distribution of values for this variable found in our dataset, which can be observed in the violin plot for the H_2 factor found in the supplemental file.

Figure 4-5 b) shares many parallels with figure 4-5 (a) such as elemental starting elemental composition, reaction time, and temperature. However, the interpretation of the catalyst/solvent ratio is more difficult in this case as the values are closely clustered. The ratio of catalyst/feedstock shows an ambiguous trend, with both low and high values sometimes being associated with lower ΔHHV. This is probably due to the presence of multiple optimal ratios of catalyst/feedstock in the dataset from different studies.

Figure 4-5 c) shows that the starting HHV of the bio-oil holds the highest importance in the prediction of the final HHV. High starting HHV in bio-oil is strongly associated with low final HHV from the upgrading process. This is due to the strong correlation between oxygen content and HHV, where high starting HHV values go in hand with low oxygen content. In a similar vein, the carbon content in the bio-oil is a strong predictor for the final HHV of the bio-oil, due to the relation between elemental composition and HHV. Regarding the temperature, a clear correlation between higher temperature and a higher resulting final HHV can be observed. The temperatures in the high quartiles are positively associated with higher resulting final HHV, which falls in the temperature range of 250 - 350 °C previously mentioned. The distribution of H_2 pressure factor also shows a clear relation to the final HHV of the bio-oil, where low values (a smaller number of moles of hydrogen used) result in low final HHV. With regards to the reaction time, it is clear that longer reaction times are associated with higher final HHV.

For figure 4-5 d), the bio-oil original HHV plays an important role in the resulting ΔHHV. This can be observed from the large cluster of SHAP values on the left side of figure 4-5 d). From the perspective of the chemical components of the bio-oil, it makes sense that if a given bio-oil sample already contains little remaining oxygen, the resulting possible increase in HHV depends on how much oxygen remains to be removed. This can also be seen that the high carbon content in the original bio-oil was also strongly associated with lower final HHV. Temperature, on the other hand, displays a different spread of SHAP values which can be seen in the final HHV case for bio-oil upgrading (shown in figure 4-5 c), still ultimately follows a similar trend where higher temperature results in higher ΔHHV. The relation between the ratio of solvent volume to reactor volume was intended to be used as an approximation of the process pressure, on the assumption that the resulting pressure at high temperature is mostly a consequence of the solvent at high temperature and not because of gas being generated. Based on this assessment, it appears that lower process pressure can be associated with higher ΔHHV. However, a significant number of points also cluster around the SHAP value of 0, indicating that there are circumstances where the feature has no impact. Lastly, the reaction time displays a very similar pattern as shown in figure 4-5 c).

Based on the performance of the model and the distribution of variable importance found, the simulation of hypothetical process conditions can be executed. Amongst the directly controllable process variables, temperature and reaction time are considered the most impactful ones since both variables are in the top five highest importance in the prediction of ΔHHV in figure 4-5. With this in mind, figure 4-6 shows partial dependency plots for ΔHHV values using different percentiles of temperature and reaction time, in a) and b) for solvolysis of biomass to bio-oil and c) and d) for bio-oil upgrading.

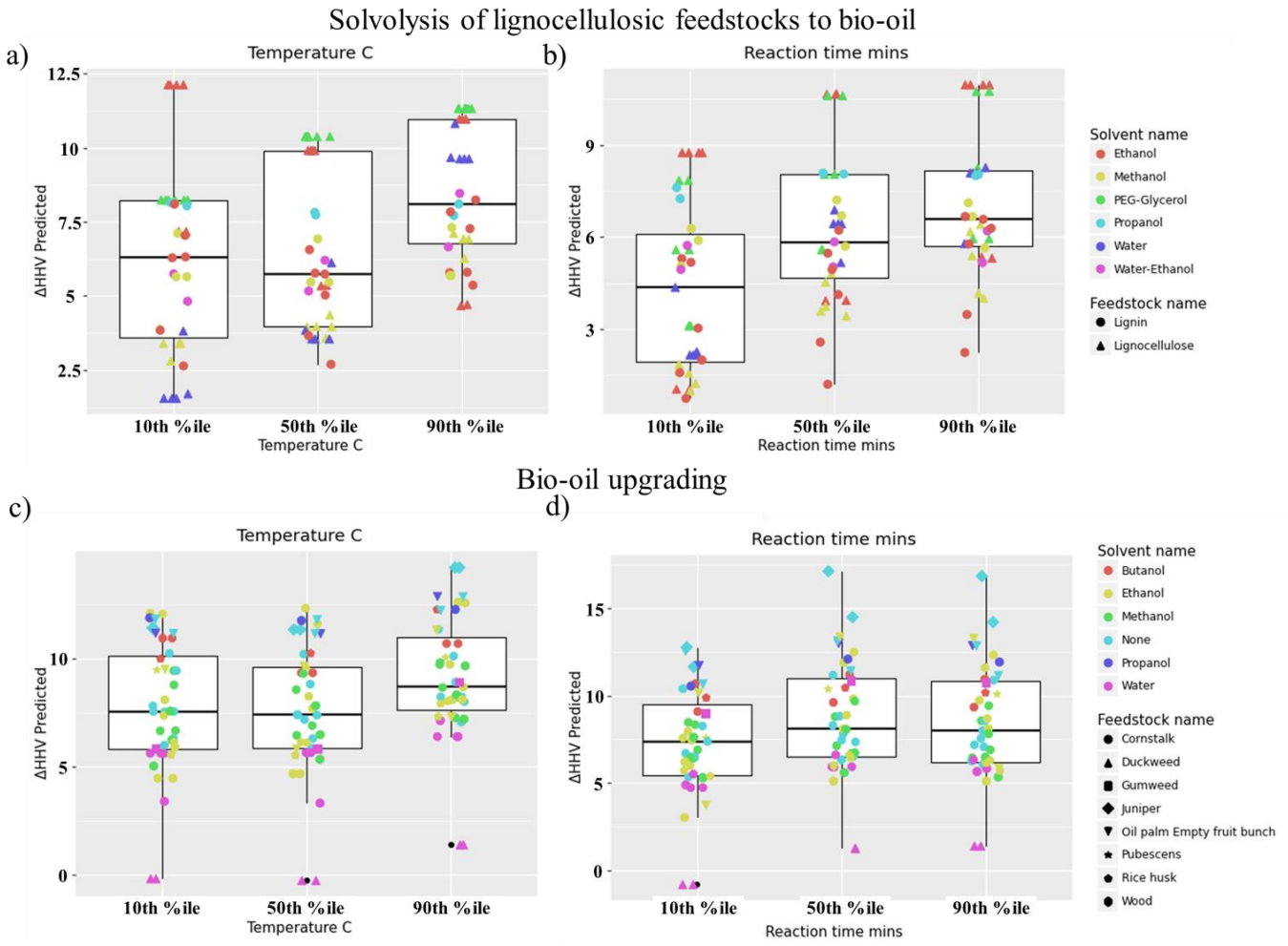


Figure 4-6. Partial dependency plot for ΔHHV changes at different temperatures and reaction time values for solvolysis of biomass to bio-oil in a) and b) , with solvents: ethanol, methanol, polyethylene (PEG)-Glycerol, propanol, water and water-ethanol mixture and feedstocks divided into either lignin or lignocellulose. Bio-oil upgrading in c) and d), with solvents: butanol, ethanol, methanol, propanol, water or no solvent, and feedstocks: cornstalk, duckweed, gumweed, juniper, oil palm empty fruit bunch, pubescens, rice husk and wood.

In figure 4-6 a), a significant increase in resulting ΔHHV of bio-oil produced can be observed when using higher temperatures. Similar to figure 4-6 b), increasing the reaction time results in higher expected ΔHHV. However, beyond a certain point, the resulting ΔHHV in the produced bio-oil does not appear to increase further. In the context of the gathered data and the reactions involved, it stands to reason that if hydrodeoxygenation reactions are ultimately responsible for the ΔHHV value obtained, all of the removable oxygen-containing species will have reacted at a certain point in the process, thus no further increase in ΔHHV should be possible.

It is also interesting to note that this pattern is largely echoed in figure 4-5 c) and d). In the case of figure 4-6 c), the increase of ΔHHV appears to be only marginally higher at temperatures in the

90th percentile. figure 4-6 d) also shows a similar pattern as in figure 4-6 c) which is presumably related to the same hydrodeoxygenation reactions previously mentioned.

Because of the heterogeneity of the dataset in this research in terms of solvents and catalysts, it must be noted that some of these simulated results may stray significantly from what a real experiment would result in. From the perspective of the solvent used, this is related to the fact that organic solvents may interact differently with the catalyst at a higher temperature, in ways that do not necessarily contribute to the expected change of ΔHHV. Notably, the short-chain alcohols used in the experiments in the dataset can undergo aldol condensation at different temperatures [27] which can compete with the arguably more favored alkylation reactions [28] or hydrogen transfer [29], which would contribute to higher ΔHHV.

The reactions involved in bio-oil upgrading fall under the umbrella term of hydrodeoxygenation (HDO) reactions, these include cracking, hydrocracking, decarboxylation, decarbonylation and hydrogenation, which are shown in Figure 4-7. These are among the reactions that represent the underlying chemistry behind the models seen in this chapter. However, it is important to emphasize that the models presented in this chapter do not directly reflect any particular HDO reaction, instead, predicting HHV values serving as a proxy of the "magnitude of oxygen removed" by these HDO reactions. Studying the HDO mechanism of would require conducting experiments and gathering data in relation to different oxygen-containing molecules, ideally working on those that are known to be difficult to deoxygenate and studying a particular aspect of the experiment, suc as the catalyst in detail, without too much variation of solvents or process parameters.

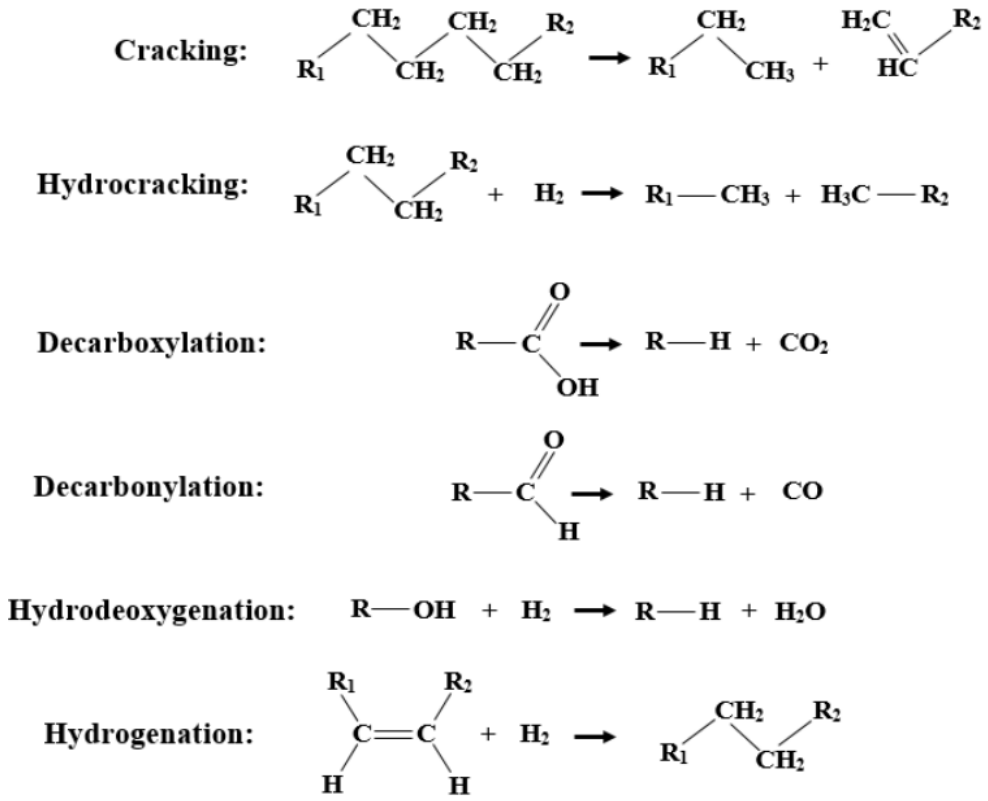


Figure 4-7. Reactions associated with catalytic HDO processes [11].

4.3.3 Significance of results and comparison to other methodologies for calculating HHV

The models developed in this research displayed a significant prediction performance in spite of the sparsity of the dataset. Because of this, using a dataset consisting of experiments only involving the use of a particular solvent or gathering more detailed information regarding the properties of the catalyst used should result in a model with much higher prediction performance. Using SHAP values offers the possibility of interpreting variable importance in a way that could allow us to connect the results with specific reaction mechanisms or phenomena that are responsible for the changes in bio-oil upgrading or solvothermal biomass conversion. This is especially valuable because biomass sources and the bio-oil resulting from their conversion can have a very extensive range of properties. Coupling experimental work with machine learning modeling and its associated explainable variable importance will undoubtedly become a powerful tool in the study of biomass conversion in general.

The models shown in this study can estimate final HHV or expected increase in HHV as a function of the process parameters, for the first time, as far as the published literature is concerned, in contrast to the various existing formulas derived from Dulong's formula for HHV [30, 31] that make use exclusively of the elemental composition of the resulting fuel. Other recent works related to biomass, HHV and machine learning share parallels with Dulong's formula in that they use the

properties of the biomass or biomass conversion products for calculating HHV [32, 33], but no process parameters. Those works that do involve process parameters tend to center on arguably simpler processes such as gasification [34], pyrolysis [35] and hydrothermal treatment [13], where the number of potential chemical interactions is lower on account of the absence of catalysts or reactive reaction media, such as various heterogeneous catalysts and organic solvents seen in this study. The models for bio-oil upgrading seen in this study can be used to research the cost-benefit-energy relation between proposed upgrading processes that differ in terms of catalysts, solvents and starting bio-oil. However, it is ultimately important to note that our model is not based on the entirety of the possible experimental space (all the possible combinations of variables), but rather limited only to the scope of currently available published literature, which may limit its applicability to never-before-tested combinations of reactants or catalyst species.

4.3.4 Future research direction and recommendations

Machine learning tools can accelerate the development of biomass conversion processes that normally require extensive work, with findings that cannot always be extrapolated to other kinds of biomass. This is both due to the large number of process variables in biomass conversion processes but also to the wide variability in properties of different biomass feedstocks. There are a number of bottlenecks that have to be addressed in order for this to be realized. Firstly, clear guidelines for reporting experimental procedures and minimal suggested characterization of feedstocks and catalysts used should be developed. This is a sentiment shared by other authors with regard to lignin-first biorefining [36], though it does not address the usage of machine learning. Secondly, the experimental work related to thermochemical biomass conversion is, in general, very cumbersome and tends to require high-temperature and pressure-resistant equipment that can be costly. This is a matter that other experimental disciplines of science, such as biology [37] and organic synthesis [38] do not struggle with (as much), where high-throughput experimentation via robotic tools is already available [39] or upcoming [40]. A possible solution to this matter could be the development of new experimental methods that require less workup and reactors that can be deployed in large numbers simultaneously. Simple reactors made of high-pressure tubing and caps [41] are examples of alternatives that could be used in high-throughput experimentation of biomass conversion. However, these reactors may have mass transfer limitations that complicate the extrapolation of the obtained results. Model compounds or mixtures of model compounds could be used to deploy extensive arrays of experiments that can be then modelled and analyzed to obtain insight into what can be expected to happen in a given biomass conversion process.

4.4 Conclusions

The interpretable XGBoost models for the prediction of HHV of bio-oil from the solvothermal conversion of lignocellulosic biomass and bio-oil upgrading were performed. The R^2 scores ranging from 0.77 to 0.86 could be achieved despite the large diversity of reaction conditions,

solvents, and types of catalysts found in the dataset. SHAP values also provided the interpretable variable importance that coincides with findings found in the literature and highlighted the useful correlations that may allow for useful prediction of expected bio-oil quality given a set of process variables, minimizing the experimental work needed to obtain meaningful results. This work demonstrates that few variables dictate the possible increase in HHV in a given bio-oil to be upgraded or the conversion of lignocellulosic biomass to bio-oil in terms of its characteristics such as elemental composition. Statistically speaking, variables such as choice of solvent, initial moisture concentration in bio-oil and catalyst active phase were shown to be of little importance compared to reaction time and temperature, within the context of this dataset.

References:

[1] Alper, K., Tekin, K., Karagöz, S., & Ragauskas, A. J. (2020). Sustainable energy and fuels from biomass: a review focusing on hydrothermal biomass processing. Sustainable Energy & Fuels, 4(9), 4390-4414. https://doi.org/10.1039/D0SE00784F

[2] Abu-Omar, M. M., Barta, K., Beckham, G. T., Luterbacher, J. S., Ralph, J., Rinaldi, R., ... & Wang, F. (2021). Guidelines for performing lignin-first biorefining. *Energy & Environmental Science*, *14*(1), 262-292. https://doi.org/10.1039/D0EE02870C

[3] Adamovic, T., Zhu, X., Perez, E., Balakshin, M., & Cocero, M. J. (2022). Understanding sulfonated kraft lignin re-polymerization by ultrafast reactions in supercritical water. *The Journal of Supercritical Fluids*, 191, 105768. https://doi.org/10.1016/j.supflu.2022.105768

[4] Bepari, S., & Kuila, D. (2020). Steam reforming of methanol, ethanol and glycerol over nickel-based catalysts-A review. *International Journal of Hydrogen Energy*, *45*(36), 18090-18113. https://doi.org/10.1016/j.ijhydene.2019.08.003

[5] Biswas, B., Kumar, A., Krishna, B. B., & Bhaskar, T. (2021). Effects of solid base catalysts on depolymerization of alkali lignin for the production of phenolic monomer compounds. *Renewable Energy*, 175, 270-280. https://doi.org/10.1016/j.renene.2021.04.039

[6] Chuntanapum, A., & Matsumura, Y. (2010). Char formation mechanism in supercritical water gasification process: a study of model compounds. *Industrial & Engineering Chemistry Research*, *49*(9), 4055-4062. https://doi.org/10.1021/ie901346h

[7] Chen, P., Xie, Q., Du, Z., Borges, F. C., Peng, P., Cheng, Y., ... & Ruan, R. (2015). Microwave-assisted thermochemical conversion of biomass for biofuel production. Production of biofuels and chemicals with microwave, 83-98. DOI: 10.1007/978-94-017-9612-5_5

[8] Garcia, A. C., Shuo, C., & Cross, J. S. (2022). Machine learning based analysis of reaction phenomena in catalytic lignin depolymerization. *Bioresource Technology*, 345, 126503. https://doi.org/10.1016/j.biortech.2021.126503

[9] Hill, J., Nelson, E., Tilman, D., Polasky, S., & Tiffany, D. (2006). Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. *Proceedings of the National Academy of Sciences,* 103(30), 11206–11210. https://doi.org/10.1073/pnas.0604600103

[10] Heide, D., von Bremen, L., Greiner, M., Hoffmann, C., Speckmann, M., & Bofinger, S. (2010). Seasonal optimal mix of wind and solar power in a future, highly renewable Europe. *Renewable Energy*, *35*(11), 2483–2489. https://doi.org/10.1016/j.renene.2010.03.012

[11] Huang, X., Korányi, T. I., Boot, M. D., & Hensen, E. J. (2015). Ethanol as capping agent and formaldehyde scavenger for efficient depolymerization of lignin to aromatics. *Green Chemistry*, *17*(11), 4941-4950. https://doi.org/10.1039/C5GC01120E

[12] Han, X., Guo, Y., Liu, X., Xia, Q., & Wang, Y. (2019). Catalytic conversion of lignocellulosic biomass into hydrocarbons: A mini review. *Catalysis Today*, 319, 2-13. https://doi.org/10.1016/j.cattod.2018.05.013

[13] Katongtung, T., Onsree, T., & Tippayawong, N. (2022). Machine learning prediction of biocrude yields and higher heating values from hydrothermal liquefaction of wet biomass and wastes. *Bioresource Technology*, 344, 126278. https://doi.org/10.1016/j.biortech.2021.126278

[14] Kozlowski, J. T., & Davis, R. J. (2013). Heterogeneous catalysts for the Guerbet coupling of alcohols. *ACS Catalysis*, *3*(7), 1588-1600. https://doi.org/10.1021/cs400292f

[15] Lee, H. L., Boccazzi, P., Ram, R. J., & Sinskey, A. J. (2006). Microbioreactor arrays with integrated mixers and fluid injectors for high-throughput experimentation with pH and dissolved oxygen control. *Lab on a Chip*, *6*(9), 1229-1235. https://doi.org/10.1039/B608014F

[16] Lian, X., Xue, Y., Zhao, Z., Xu, G., Han, S., & Yu, H. (2017). Progress on upgrading methods of bio-oil: a review. *International Journal of Energy Research*, 41(13), 1798-1816. https://doi.org/10.1002/er.3726

[17] Michailof, C. M., Kalogiannis, K. G., Sfetsas, T., Patiaka, D. T., & Lappas, A. A. (2016). Advanced analytical techniques for bio-oil characterization. *Wiley Interdisciplinary Reviews: Energy and Environment*, 5(6), 614-639. https://doi.org/10.1002/wene.208

[18] McClelland, D. J., Galebach, P. H., Motagamwala, A. H., Wittrig, A. M., Karlen, S. D., Buchanan, J. S., ... & Huber, G. W. (2019). Supercritical methanol depolymerization and hydrodeoxygenation of lignin and biomass over reduced copper porous metal oxides. *Green Chemistry*, *21*(11), 2988-3005. https://doi.org/10.1039/C9GC00589G

[19] Mendes, F. L., da Silva, V. T., Pacheco, M. E., de Rezende Pinho, A., & Henriques, C. A. (2020). Hydrotreating of fast pyrolysis oil: A comparison of carbons and carbon-covered alumina as supports for Ni2P. *Fuel*, 264, 116764. https://doi.org/10.1016/j.fuel.2019.116764

[20] Molnar, C. (2020). Interpretable machine learning.

[21] Muangsuwan, C., Kriprasertkul, W., Ratchahat, S., Liu, C.-G., Posoknistakul, P., Laosiripojana, N., & Sakdaronnarong, C. (2021). Upgrading of light bio-oil from solvothermolysis liquefaction of an oil palm empty fruit bunch in glycerol by catalytic hydrodeoxygenation using NiMo/Al2O3 or CoMo/Al2O3 Catalysts. ACS Omega, 6(4), 2999–3016. https://doi.org/10.1021/acsomega.0c05387

[22] *Maps and data - global ethanol production by country or region*. Alternative Fuels Data 370 Center: Maps and Data - Global Ethanol Production by Country or Region. (2021, 371 June). Retrieved January 8, 2022, from https://afdc.energy.gov/data/10331

[23] Naimoli, S.; Ladislaw, S. Decarbonizing Heavy Industry. Csis. October 2020. Available online: https://www.csis.org/analysis/climate-solutions-series-decarbonizing-heavy-industry (accessed on 27 November 2022).

[24] Nieto, P. J. G., García–Gonzalo, E., Paredes–Sánchez, B. M., & Paredes–Sánchez, J. P. (2022). Forecast of the higher heating value based on proximate analysis by using support vector machines and multilayer perceptron in bioenergy resources. *Fuel*, 317, 122824. https://doi.org/10.1016/j.fuel.2021.122824

[25] Ong, H. C., Chen, W. H., Singh, Y., Gan, Y. Y., Chen, C. Y., & Show, P. L. (2020). A state-of-the-art review on thermochemical conversion of biomass for biofuel production: A TG-FTIR approach. *Energy Conversion and Management*, 209, 112634. https://doi.org/10.1016/j.enconman.2020.112634

[26] Onsree, T., Tippayawong, N., Phithakkitnukoon, S., & Lauterbach, J. (2022). Interpretable machine-learning model with a collaborative game approach to predict yields and higher heating value of torrefied biomass. *Energy*, 249, 123676. https://doi.org/10.1016/j.energy.2022.123676

[27] Prajitno, H., Insyani, R., Park, J., Ryu, C., & Kim, J. (2016). Non-catalytic upgrading of fast pyrolysis bio-oil in supercritical ethanol and combustion behavior of the upgraded oil. *Applied Energy*, 172, 12-22. https://doi.org/10.1016/j.apenergy.2016.03.093

[28] Rachel-Tang, D. Y., Islam, A., & Taufiq-Yap, Y. H. (2017). Bio-oil production via catalytic solvolysis of biomass. *RSC advances*, 7(13), 7820-7830. https://doi.org/10.1039/C6RA27824H

[29] Shevlin, M. (2017). Practical high-throughput experimentation for chemists. *ACS medicinal chemistry letters*, *8*(6), 601-607. https://doi.org/10.1021/acsmedchemlett.7b00165

[30] Hosokai, S., Matsuoka, K., Kuramoto, K., & Suzuki, Y. (2016). Modification of Dulong's formula to estimate heating value of gas, liquid and solid fuels. Fuel Processing Technology, 152, 399–405. https://doi.org/10.1016/j.fuproc.2016.06.040

[31] Sheng, C., & Azevedo, J. L. T. (2005). Estimating the higher heating value of biomass fuels from basic analysis data. Biomass and Bioenergy, 28(5), 499–507. https://doi.org/10.1016/j.biombioe.2004.11.008

[32] Ighalo, J. O., Adeniyi, A. G., & Marques, G. (2020). Application of linear regression algorithm and stochastic gradient descent in a machine-learning environment for predicting biomass higher heating value. Biofuels, Bioproducts and Biorefining, 14(6), 1286–1295. https://doi.org/10.1002/bbb.2140

[33] Yaka, H., Insel, M. A., Yucel, O., & Sadikoglu, H. (2022). A comparison of machine learning algorithms for estimation of higher heating values of biomass and fossil fuels from Ultimate Analysis. Fuel, 320, 123971. https://doi.org/10.1016/j.fuel.2022.123971

[34] Mutlu, A. Y., & Yucel, O. (2018). An artificial intelligence-based approach to predicting syngas composition for downdraft biomass gasification. Energy, 165, 895–901. https://doi.org/10.1016/j.energy.2018.09.131

[35] Leng, E., He, B., Chen, J., Liao, G., Ma, Y., Zhang, F., Liu, S., & E, J. (2021). Prediction of three-phase product distribution and bio-oil heating value of biomass fast pyrolysis based on machine learning. Energy, 236, 121401. https://doi.org/10.1016/j.energy.2021.121401

[36] Tabanelli, T. (2021). Unrevealing the hidden link between sustainable alkylation and hydrogen transfer processes with alcohols. *Current Opinion in Green and Sustainable Chemistry*, *29*, 100449. https://doi.org/10.1016/j.cogsc.2021.100449

[37] van Holst Pellekaan, N., Walker, M. E., Watson, T. L., & Jiranek, V. (2021). 'TeeBot': A High Throughput Robotic Fermentation and Sampling System. *Fermentation*, *7*(4), 205. https://doi.org/10.3390/fermentation7040205

[38] Wang, H., Feng, M., & Yang, B. (2017). Catalytic hydrodeoxygenation of anisole: an insight into the role of metals in transalkylation reactions in bio-oil upgrading. *Green Chemistry*, 19(7), 1668-1673. https://doi.org/10.1039/C6GC03198F

[39] Wang, Z., Zhao, W., Hao, G. F., & Song, B. A. (2020). Automated synthesis: current platforms and further needs. *Drug Discovery Today*, *25*(11), 2006-2011. https://doi.org/10.1016/j.drudis.2020.09.009

[40] Yang, Q., Wang, Q., & Yu, Z. (2015). Substitution of alcohols by N-nucleophiles via transition metal-catalyzed dehydrogenation. *Chemical Society Reviews*, *44*(8), 2305-2329. https://doi.org/10.1039/C4CS00496E

[41] Zhang, X., Wang, K., Chen, J., Zhu, L., & Wang, S. (2020). Mild hydrogenation of bio-oil and its derived phenolic monomers over Pt–Ni bimetal-based catalysts. *Applied Energy*, 275, 115154. https://doi.org/10.1016/j.apenergy.2020.115154

Chapter 5:

Dissertation summary and future research work

5.1 Dissertation summary and future research work

As stated in Chapter 1, the main theme through this dissertation was the usage of ML to clarify and understand the phenomena involved in different kinds of solvothermal lignin depolymerization approaches, the focus being on this particular subset of depolymerization processes due to the authors' perception that it is the most economically feasible and scalable lignin depolymerization process.

Chapter 2 focused on the development of a ML model that aimed to test whether the surface properties of heterogeneous catalysts (surface area, pore diameter and pore volume) played a significant role in the prediction of experimental outcomes from data from heterogeneously catalyzed lignin solvolysis experiments. The results indicate that these properties do not seem to play a major role in the prediction of bio-oil yield, but surface area does predict to some extent the yield of solid residues. For bio-oil yield, process parameters such as temperature, reaction time and choice of solvent seem to play a much larger role than others. This indicates that it is clear that heterogeneous catalysts clearly do play a role in the process, perhaps the majority of the depolymerization that happens is due to the effects of the solvent interacting with lignin at a given temperature for a particular amount of time. It must be noted that the contribution of these heterogeneous catalysts is not only increasing the bio-oil yield, but also changing the product distribution of the process, which is beyond the scope of this dissertation. As stated at the end of Chapter 2, it must be emphasized that the data used to train this model is arguably only a small fraction of the total possible experimental space, and attention should be paid to the data that was used to train the models.

Chapter 3 shifted the focus from heterogeneous to homogeneous catalysis, strictly focusing on studies that use water as the solvent of choice (hydrothermal), in attempt to further zoom into how the comparatively few process parameters seen in this kind of process interact. Once again ML models were trained using data from literature obtaining prediction performance similar to that seen in the models in Chapter 2. In this case, unsurprisingly temperature and reaction time were the two process parameters that contributed the most towards the prediction of experimental outcomes, with the catalyst playing a relatively minor role in comparison. This chapter also attempted to experimentally test the predictions made by the models, in order to further validate them. However, the yield of bio-oil displayed at times an extremely large deviation from the prediction, while the yield of solid residues obtained fell within the expected error range from the

model. In the chapter, possible explanations as to why the experimental results for bio-oil deviated so largely from the model's prediction were offered. The main possible explanation were offered, first, that the chemistry work-up steps followed after hydrothermal lignin depolymerization may differ across different studies, such as using different solvents to extract a solvent-soluble fraction of lignin from the solid residues, and whether this should or should not be reported as part of the total bio-oil yield (dubbed "Heavy oil" in some papers). Second is that the majority of the studies did not attempt to characterize the lignin that they used in their studies, which likely contributes to the error of the trained models. Elemental composition, ash content and solubility in different solvents is known to differ across different lignins, yet this is rarely acknowledged in the studies. Due to the abovementioned reasons, the possibility of using alternative metrics for reaction evaluation are suggested, such as the usage of measurable, useful properties from the bio-oil, such as higher heating value (HHV)

Having analyzed the two major groups of studies about lignin solvolysis, Chapter 4 is focused instead in predicting the final HHV of the bio-oil obtained from heterogeneously catalyzed solvolysis of lignocellulosic feedstocks, and also the changes in HHV when said bio-oil is upgraded hydrogenation or thermal treatment. Here once again, processes parameters such as temperature and reaction time proved to be good predictors of the HHV values. Elemental composition, oxygen content in the original feedstock or bio-oil to be upgraded in particular also largely contributed to the prediction. This falls within line of what is currently understood in literature for prediction of HHV from fossil and biofuels, where as per Dulong's formula, oxygen content in the fuel negatively impacts the resulting HHV, therefore the lower the oxygen content is, the larger the proportion of carbon and hydrogen in the bio-oil, resulting in larger HHV values obtained. Future work on this matter could focus on the technoenergetic evaluation of different solvothermal biomass upgrading processes, focusing both on the inputs (time, heat, solvents and catalysts) and outputs (yield of bio-oil and HHV).

Over the course of this dissertation, it is clear that procuring meaningful amounts of data for experimental work can be a difficult task, and that using data from literature only has clear limitations, such as scarcity and tendency from researchers to report only "good" results, that do not provide useful information to train a ML model. In addition to this, noise in the form of mismeasurement and non-explicit differences in methodology may reduce the quality of the data. To overcome this weakness, the possibility of miniaturizing and simplifying biomass conversion experiments is put forward, by planning experiments in such way that the results obtained may be representative of the phenomena involved in larger experiments, yet, faster to execute in large quantities. On the other hand, advances in the field of data science are allowing researchers to slowly move away from "big data" approaches that are simply incompatible with practical difficulties seen in experimental work. Although the idea of renewable aromatic chemicals and liquid biofuels sounds attractive and the chemistry involved is interesting, it is the authors belief that due to currently existing competing technologies we are unlikely to see lignin or lignocellulose liquefaction replacing the processes seen in petrochemistry, for the foreseeable future.

I believe that advances in ML will contribute towards clarifying the phenomena involved in all biomass conversion processes (particularly thermochemical ones), which will in turn lead to

obtaining more meaningful results that can be scaled up and applied, thus positively contributing towards the goals of sustainability and the development of a circular bio-economy.

Appendix 1

On the limitations of heterogeneous catalysis in solvothermal lignin depolymerization

1.0 Introduction

Understanding of lignin depolymerization has progressed largely since the early studies in the late 20th century that mostly used analogies from existing petrochemical industry. During the early stages of this field the structure of lignin and it's properties were poorly understood. These early studies employed drastic reaction conditions (temperature, hydrogen pressure) to depolymerize the lignin, and were ultimately successful in producing various aromatic monomers, despite the clear economic unfeasibility of using these reaction conditions to produce chemicals that were arguably cheaply available as side products of gasoline production through fluid catalytic cracking [7].

Because the seeming lack of urgency, the field progressed slowly and only started gaining traction in the early 21st century (Figure A-1), where various new approaches were put forward by research groups, mostly in the form of solvolysis by using different combinations of organic solvents, water, and catalysts. each highlighting the merits of their proposed method and using different methodologies to quantify their success, but still remaining unpractical at commercial scale. A representative scheme of this reaction's reactants, solvent, catalyst and expected products is shown in Figure A-2.

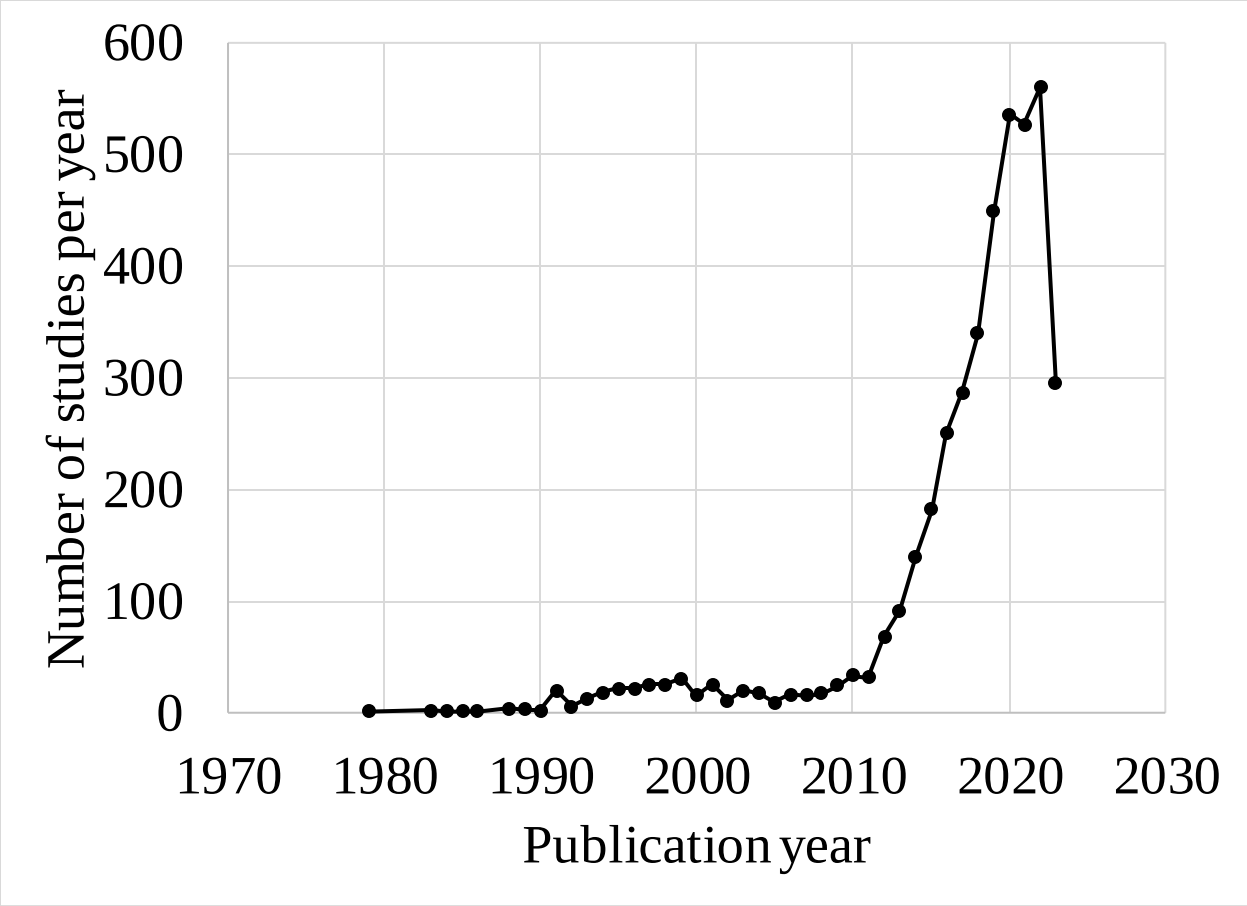


Figure A-1. Interest in lignin depolymerization over time. Results from Web of Science using the search string: lignin depolymerization. (As of August of 2023)

Various catalysts, solvents and combinations of reaction conditions proved to be "successful" but stating which one is superior is difficult due to the large availability of studies, incomplete or partial reporting of results as well as poor reasoning as to why the experiments are conducted in a particular way beyond "this combination of variables has not been tried yet". All of this contributes to the problem of comparing lignin depolymerization studies. The focus of these studies can be roughly split into catalysts and reaction media, usually in search of the optimal reaction conditions for one or the other.

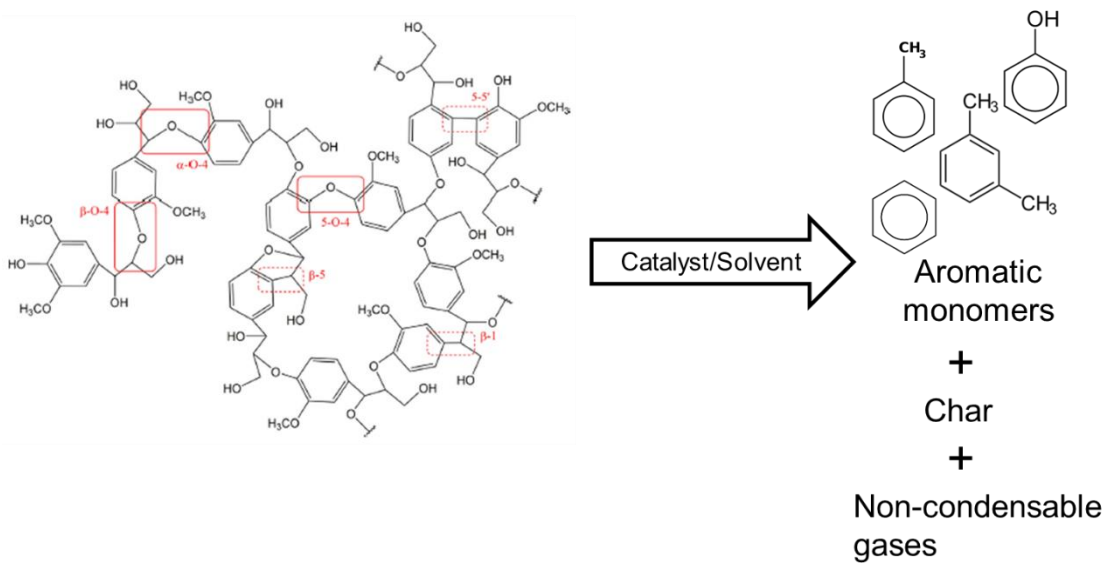


Figure A-2. Description of lignin depolymerization reactants, catalyst, solvent and products conducted at elevated temperature.

1.1 Catalysts in lignin depolymerization

In terms of catalysts, many successful examples can be seen in the existing literature. It is clear that the use of transition metal catalysts, both noble and non-noble exert a catalytic effect in the reaction: increasing the yield of bio-oil, reducing the formation of char and biasing the product distribution towards less oxygenated products. Because of the many combinations of catalyst, solvent and lignin properties, performance of catalysts can only be compared with other studies that share similar conditions. The performance of the catalyst is measured by comparing the fraction of lignin that was transformed into liquid products (bio-oil yield), gas or char. In some cases, performance is further defined as the quantity of aromatic monomers obtained from a particular lignin.

Of the noble metal catalysts, Ru [8,9] and Rh [10] have displayed outstanding performance compared to the other noble metal catalysts. Non-noble metals have been extensively used in many chemical processes, notably those related to petrochemistry, such as removal of oxygen, nitrogen or sulfur from hydrocarbons by hydrotreating [11]. Catalysts containing Ni, Mo, Co and W have been used in lignin depolymerization either by themselves or in bi-metallic catalysts [12, 13].

On the other hand, the role of support materials in lignin depolymerization catalysts has also been explored, where studies can be largely categorized by their use of acidic, neutral or basic supports (though the latter is rarely seen). Zeolites are the most prominent example of acidic supports, for example, [14] used Ni-Cu supported in various zeolites (H-beta, ZSM-5, MAS-7,

MCM-41, SAPO-11) for Kraft lignin depolymerization in presence of isopropanol as the reaction medium at 330 °C, obtaining bio-oil yields above 85% for all used zeolites. In the same vein, ref. [13] used conventional hydrotreatment metals (Ni, Mo, W, Co) supported on ZSM-5 in the presence of methanol as the reaction medium under hydrogen pressure. The resulting product selectivity is biased towards the formation of alkylphenols but also the formation of 30 wt.% char, due to the high acidity in the ZSM-5 support. It is clear that moderate acidity in the support can contribute positively to the result of the reaction and bias the products towards alkylphenols, which are good compounds for fuel purposes.

Yet even studies that do not involve acidic catalysts can perform very well, neutral support materials lack the acid sites that acidic support materials possess, their purpose being mostly to provide a medium for the active phase to disperse. Activated carbon is the principal example of this, with many studies involving transition metal catalysts in a variety of reaction media, most of them obtaining over 80% bio-oil yield, for example, [10] used Pt, Pd, Ru and Rh supported in activated carbon in the presence of isopropanol as the reaction medium for Kraft lignin (KL) depolymerization, obtaining over 100% bio-oil yield in some their experiments. In [15]'s work, Pt, Ni, Ru, Pd supported in activated carbon was used in the presence of ethanol–water mixture (1:1 wt.%) for kraft lignin depolymerization without additional hydrogen. The obtained product distribution contains mostly guaiacolic compounds, which is in line with the metals used and the absence of additional hydrogen.

1.2 Role of reaction media in lignin depolymerization

While the role of the catalyst used in lignin depolymerization had a more straightforward, well-understood purpose, the solvent used as reaction media seemed like an open-ended question, because of this, many studies have disregarded the role of the solvent used as reaction media or focused on a specific one that showed synergy with the catalysts they were studying. The solvents used can be largely divided in: water, alcohols and other organic solvents. Solvents used usually display optimal reaction performance close to their critical point, especially for the case of alcohols and water, usually around 250 to 350 °C.

Depolymerization of KL in sub- and supercritical water is usually carried out in conjunction with basic homogeneous catalysts of varying strengths such as NaOH, KOH, K_2CO_3 and Na_2CO_3, the addition of a basic catalyst seems to improve the reaction performance through at least two mechanisms: Firstly, by improving the solubility of KL in the aqueous solution; and secondly, by preventing to a certain degree the repolymerization that happens during the reaction. While the use of basic homogeneous catalysts appears to be the norm in hydrothermal depolymerization, there have been a few instances of hydrothermal studies using heterogeneous catalysts, notably zeolites ZSM-5 [16] and SBA-15, in conjunction with Na_2CO_3 [17].

The prevalence of studies employing short-chain alcohols (C1-C4) as reaction media for lignin depolymerization is high, the motivations for these are: lignin displays significantly higher solubility in short-chain alcohols than in pure water, facilitating the reaction, additionally the occurrence of low-temperature dry alcohol reforming [18] and hydrogen donating capacity of alcohols [19], reduces the oxygen content of the resulting bio-oil and prevents the formation of char. The occurrence of these two hydrogen forming reactions is mediated by the temperature, reaction time and catalyst used, with various reduced transition metals in diverse supports displaying notable performance.

As for other organic solvents, Dioxane has been used as a reaction medium in several KL depolymerization studies in conjunction with other solvents such as water [19], ethanol [20] and methanol [21]. Acetone [8] and dodecane [9] have been used as reaction media in KL depolymerization in a few studies. Acetone by itself does not dissolve KL extensively, but when combined with water its solubilizing capabilities increase drastically [22]. A bio-oil yield of 93% is reported in [8]; it is important to note that this study employs direct hydrogen pressure, as acetone does not possess hydrogen donating capacity. Dodecane as reaction medium does not possess hydrogen donating properties; thus, its only role in the reaction is to facilitate the interaction of lignin with the catalyst.

2.0. Methodology

Based on the studies published in the past decade, it is the authors belief that doing a purely experimental approach that focuses exclusively on a particular type of catalyst or reaction media without any particular reasoning or mindset would not further our understanding of lignin depolymerization. One of the biggest concerns found in these studies' reaction time varies dramatically among studies, ranging from a few minutes to 24 hours in some cases, all in which relatively high bio-oil yield was obtained. After further searching the literature for an explanation, no clear indication of why this variation in reaction time seems to exist. Considering that the work in catalysts until now has focused almost exclusively on the active species deposited in the catalyst, and the large gap in reaction time between heterogeneous and homogeneous catalysts, it would appear that no comprehensive work has been done in regard to the mass transfer phenomena involved in the reaction with regards to lignin's interaction with heterogeneous catalysts.

Lignin is a heterogeneous polymer that can display a large and varied molecular weight distribution and polydispersity index, depending on the lignocellulosic feedstock and method used to isolate it. This can be exemplified by Table A-1, where the average molecular weight (Mn), weight averaged molecular weight (Mw) and polydispersity index (D) for various biorefinery lignins is shown.

Table A-1. The Mn and Mw of various lignins isolated from pre-treated biomasses. [from 23]

Biomass	Pre-treatment	Mn (g mol^{-1})	Mw (g mol^{-1})
Cotton Stalk	MWL	700	1520
	Ammonia Hydrothermal	560–890	1250–1740
Bamboo (Bambusa rigida sp)	MWL	1680	3260
	AL	1860	2840
Birch (Betula alnoides)	MWL	5860	10,860
	Microwave	3830	7290
	Heat	5000	11,450
Beech (Fagus sylvatica)	MWL	3690	5510
	Heat	2790	4020
Loblolly Pine (Pinus taeda)	MWL	989	7790
	Ceriporiopsis subvermispora	743–770	5147–6330
	MWL	7590	13,500
	OS-MWL	6530	16,800
	EOL	3070	5410
Poplar (Populus trichocarpa)	MWL	–	8550
	DAP	–	7500–8280
Switchgrass (Panicum virgatum var. Kanlow)	MWL	2070	5100
	EOL	980	4200
	Ethanol organosolv + ball mill	1580	5750
Poplar (Populus albaglandulosa)	MWL	4176	13,250
	Supercritical H2O	1042–1357	1655–4429
	Supercritical H2O + catalyst	949–1097	1526–2753
Tamarix ramosissima	MWL	2155	3750
	LHW	1380–2250	2690–3950
Lodgepole Pine Wood Chips	SPORL (LS-SP165)	810	1440
Commercial Softwood	SPORL (LSD-748)	4800	14,000

* AL: alkali lignin; DAP: dilute acid pre-treatment; EOL: ethanol organosolv lignin; LHW: liquid hot water; OS-MWL: organosolv milled wood lignin; SPORL: sulfite pre-treatment to overcome recalcitrance of lignocellulose.

As can be observed, the mass average molecular weight (Mw) can range from a few 1000 g mol^{-1} to as much as over 10000 g mol^{-1}, and values as high as 30000 g mol^{-1} have been reported for Kraft lignin.

Because catalysts used in lignin depolymerization had been largely synthesized based-off analogies from other previously existing, successful catalysts from hydrodeoxygenation, hydrogenation and hydrogenolysis processes no attention has been paid to the fact that lignin as a

reactant is very different than those found in the processes these catalysts were used in. In particular, the size of the pores found in heterogeneous catalysts tends to be either mesoporous (50 to 2 nm) or microporous (less than 2 nm), by comparing this pore size with, for example, the kinetic diameter of benzene that is 0.585 nm, it stands to reason that a larger, aliphatic branch containing molecules composed of multiple aromatic units may result in a much larger kinetic diameter that may result in very slow or impossible diffusion through the pores of the catalyst.

With this in mind the hypothesis that *"molecular weight distribution, and by extension kinetic diameter of lignin is correlated with mass transfer limitations in heterogeneously catalyzed depolymerization reactions"* is put forward. In this given context, our null hypothesis is that mass transfer phenomena do not limit the reaction rate, we aim to prove the alternative hypothesis which if true, would reveal the limitations associated with large reaction times due to slow mass transfer limitations in a commercial scale setting, i.e. very large reactor sizes, compared to when using homogeneous catalysts. Until now, studies in heterogeneously catalyzed lignin depolymerization have not explored the impact of lignin molecular weight in the reaction to detail.

2.1 Experimental testing of mass transfer limitations in lignin depolymerization

The goal of these experiments is to test the magnitude of mass transfer limitations that take place in heterogeneously catalyzed lignin depolymerization reactions. For simple chemical species, it is possible to calculate the kinetic diameter through the following formula (1) [26].

$$\Phi_k = 0.841 V_c^{1/3} \qquad (1)$$

Where V_c corresponds to the critical volume of the chemical species. However, due to the heterogeneous nature of lignin, it is not possible to measure its critical volume. To overcome this we propose a holistic approach to estimate how the molecular weight of a lignin fragment correlates to the mass transfer limitations.

When using a heterogeneous catalyst, we assume that lignin has to interact with the active sites found on the surface of the catalyst, usually in the form metal particles. These active sites can be located on the outer surface or within the pores of the catalyst. Outer surface-active sites make up for a comparatively smaller fraction of the total available sites, especially in high surface area materials such as zeolites and activated carbons.

Depending on the size of the lignin fragment, it is possible that diffusion through the pores of the catalyst may not be possible at the beginning, only becoming possible after the lignin fragment has decreased in size by reacting on the outer active sites. This could, hypothetically, create a bottleneck effect in the available active sites of the catalyst, resulting in an overall slower reaction. This is illustrated in Figure A-3.

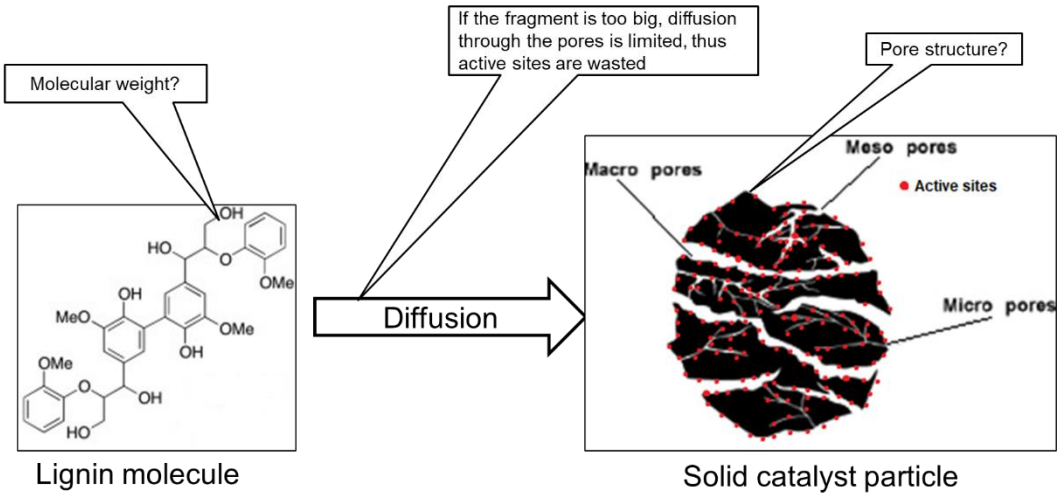


Figure A-3. Relation between lignin fragment size and diffusion through catalyst pores.

To test to what degree this bottleneck effect happens (if and when), we propose to use selectively deactivated zeolites whose outer layers have been coated with silica to inactivate the outer active sites. This kind of zeolite material has been used in the past to intensify a shape-selectivity effect in various reactions [27], with various methods existing for their synthesis [28].

2.2 Materials

Dealkaline lignin (DL) was bought from Asahi Kasei chemicals, Japan.. Other chemicals include methanol (MeOH) (99%), guaiacol (GUA) (99%), durene (99%), acetone (98%) and tetraethyl orthosilicate (99.8%) (TEOS) all bought from Wako chemicals, Japan. 3 different aluminosilicates were chosen on the basis of their surface and pore properties, these are: MFI, ZSM-5 and FAU, with theorical pore diameters of 4.7 Å, 5.95 Å and 7.35 Å, respectively.

2.3 Catalyst synthesis and characterization

Premade zeolites used in the experiments will be bought and their outer layers will be passivated according to the method outlined in [29]. Zeolite modification will be carried out by chemical liquid deposition of Tetraethyl orthosilicate (TEOS) by following the next procedure:

1. 1.0 g of pre-dried aluminosilicate is measured and suspended in 25.0 mL of hexane.
2. The mixture of zeolite and hexane is heated until reflux under vigorous stirring (500 rpm) for 1 hour.
3. 0.15 mL of TEOS is added to the aluminosilicate and hexane mixture and is stirred for another hour.
4. Hexane is removed by evaporation.
5. The dry zeolite is then calcined at 823 K for 4 hours.

This process may be repeated multiple times to guarantee complete coverage of the outer surface of the zeolite, according to [30], repeating the process 3 times is necessary for a mesoporous ZSM-5 zeolite.

Once the catalysts have been synthetized characterization by SEM microscope to confirm the presence of the silica layer in the zeolite, Brunauer–Emmett–Teller (BET) nitrogen adsorption-desorption to measure surface and pore properties before and after the deposition of silica and temperature programmed desorption of ammonia to measure the acidity of the parent zeolites used in the experiments.

2.4 Testing of catalysts

The testing of the catalysts was carried out in 50 mL Taiatsu (Japan) TPR-5 reactors equipped with a pressure gauge, thermocouple and gas line.

In experiments involving guaiacol as the reactant the GUA was introduced into the reactor a as a solution containing 250 mg of GUA/20 gr of MeOH, afterwards 50 mg of the selected catalyst was fed into the reactor, then sealed. In experiments with DL as the reactant, 250 mg of DL were fed into the reactor, followed by 20 gr of MeOH and 50 mg of the selected catalyst.

Reactors were first purged with nitrogen 3 times at 5 bars to remove any oxygen for all experiments. All experiments were carried out at 250 °C, with reaction time of 1 to 5 hours. Reactors were then heated via heating jackets at approximately 10 °C/min.

After the selected reaction time had elapsed, the reactors were cooled down via convection in a fume hood. After reaching ~50 °C the reactors were then degassed, and the liquid fraction was dumped into a glass beaker. The reactors were washed with 10 mL of MeOH to ensure no solid residues remained in the walls or bottom of the reactor. The liquid and solid products (including the catalyst) were filtered, and the liquid fraction was then used to prepare samples for GC-MS analysis, with durene added as internal standard.

To evaluate catalyst performance, in reactions with guaiacol, conversion % of GUA into other products was used as the performance metric. For reactions with lignin as the reactant, yield of MeOH soluble products was used as one of reaction performance metrics, yield of aromatic GC-MS-detectable species was used as the other reaction performance metric. These are shown in equations (2) and (3).

$$GUA\ conversion(\%) = \frac{starting\ GUA\ -GUA\ in\ liquid\ products}{starting\ GUA} * 100 \qquad (2)$$

$$Yield\ of\ aromatics(\%) = \frac{total\ aromatics\ in\ liquid\ products(gr)}{starting\ lignin\ (gr)} \qquad (3)$$

3.0 Expected results

Initial experiments aim to clarify first whether the proposed reaction conditions would not obfuscate the target of this study; the interaction between surface and pore properties of the catalysts chosen and the kinetic diameter (in the case of GUA) or molecular weight of lignin. Because MeOH is known to form aromatic compounds at high temperatures in presence of acidic catalyst, a blank reaction was ran, heating 20 gr of MeOH at 250 °C for 5 hours. This resulted in the formation of small quantities of 2-pentanone-5-methoxy (figure A-4), a compound that is known to form as a side reaction resulting from self-coupling of MeOH.

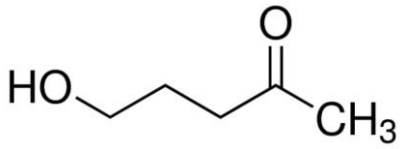


Figure A-4. 2-pentanone-5-methoxy resulting from self-coupling of MeOH at 250 °C for 5 hours.

Subsequently, catalysts ZSM-5 and FAU were then tested as per the reaction conditions mentioned in the methodology section. ZSM-5 and FAU are both aluminosilicates that differ only in surface area, pore properties and acidity. Below in Table A-2 we can observe their differences in these properties:

Table A-2. Average range of properties of FAU and ZSM-5 aluminosilicates.

Property	FAU	ZSM-5
Total acidity (mmol/g)	0.2~1.5	0.1~1
Cage size (Å)	7~13	3.2~8.2
Pore mouth diameter (Å)	4~8	5.5~5.6
Surface area (m^2/gr)	400~700	200~600

While their specific properties can vary depending on the synthesis method used to obtain, the magnitude of pore mouth diameter is the property of interest for the initial experiments. GUA with a kinetic diameter 6.68 Å should be unable to fully interact with the acid active sites found in ZSM-5, only having access to the ones on the outer surface of the catalyst particle. In the other hand FAU allows for at least the partial diffusion of GUA through the pores, thus increasing the conversion of GUA for a given reaction time. In Figure 5-5, the distribution of chromatographic area is shown for the conversion of GUA at 250 °C for 5 hours, it can be appreciated that almost 50% of the chromatogram area are non-GUA products, this is in contrast to the reaction with ZSM-5, that displayed only 10% of non-GUA products using the same reaction conditions.

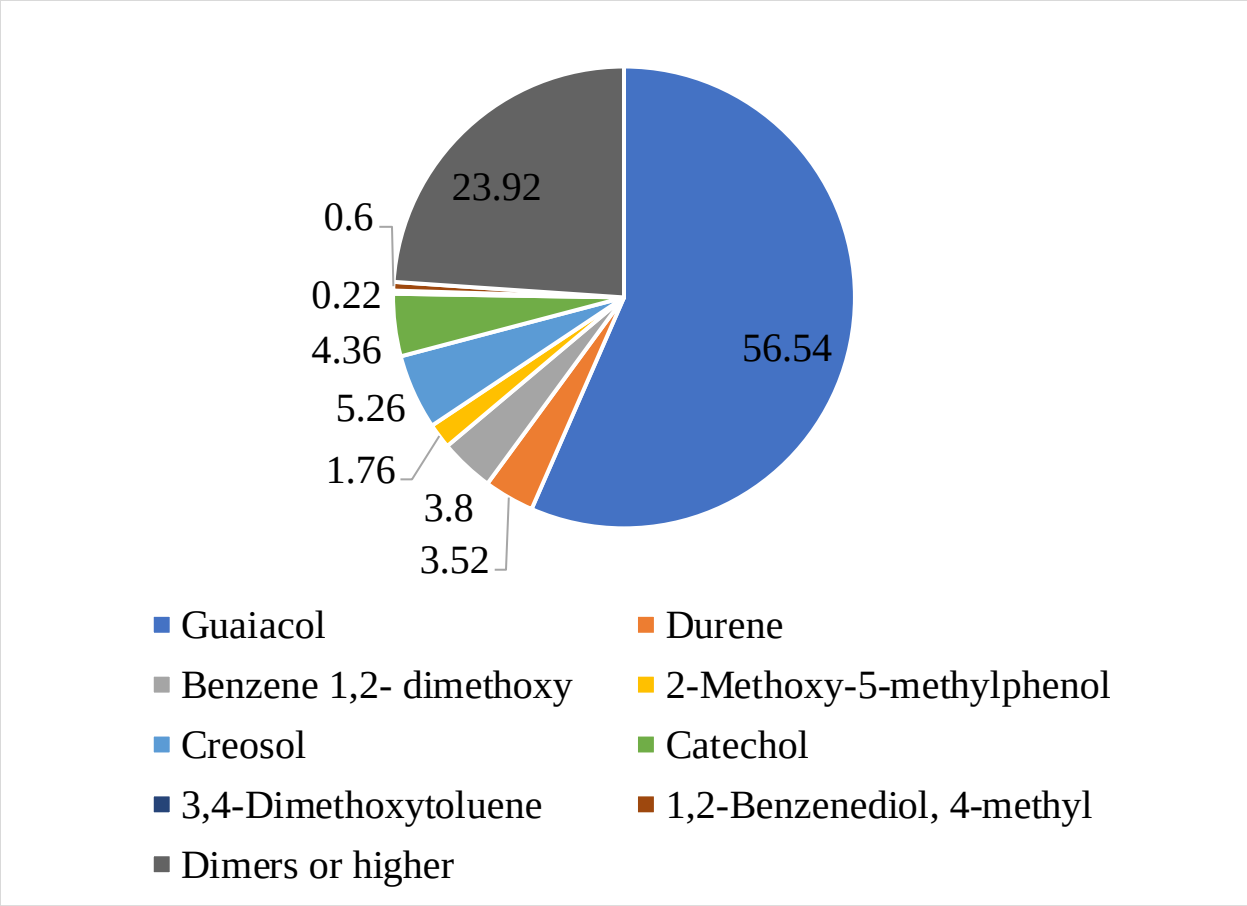


Figure A-5. Chromatogram area distribution for reaction of GUA with FAU catalyst at 250 °C for 5 hours.

These initial results fall within the expectations based on what is currently understood from literature, that unsurprisingly, for a reactant molecule to transform into a product, it must be able to reliably reach the active site of the chosen heterogeneous catalyst. Because GUA has a well-defined kinetic diameter, it is expected that by coating ZSM-5 and FAU with silica the conversion of GUA should drop due to the loss of outer active sites in the catalyst particles. However, this drop will be comparatively higher for ZSM-5 than for FAU, due to the fact that ZSM-5's GUA conversion relies almost exclusively on the outer active sites of the catalyst particle.

By first clarifying this interaction with GUA, further experiments with lignin will serve to highlight that indeed this interaction pattern between catalyst and reactant is not exclusive to well-defined molecules, but also heterogeneous polymer reactants such as lignin. Thus, bringing light to the issue of whether the overall reaction speed in heterogeneously catalyzed lignin depolymerization is governed by speed of the reaction mechanism, or the ability (or inability) for potentially large lignin fragments to reach the inner active sites of the catalyst particle.

References:

[1] Heide, D., von Bremen, L., Greiner, M., Hoffmann, C., Speckmann, M., & Bofinger, S. (2010). Seasonal optimal mix of wind and solar power in a future, highly renewable Europe. *Renewable Energy*, *35*(11), 2483–2489. https://doi.org/10.1016/j.renene.2010.03.012

[2] *Climate Solutions Series: Decarbonizing Heavy Industry*. Climate Solutions Series: Decarbonizing Heavy Industry | Center for Strategic and International Studies. (2021, July 12). https://www.csis.org/analysis/climate-solutions-series-decarbonizing-heavy-industry.

[3] Alonso, D. M., Hakim, S. H., Zhou, S., Won, W., Hosseinaei, O., Tao, J., Garcia-Negron, V., Motagamwala, A. H., Mellmer, M. A., Huang, K., Houtman, C. J., Labbé, N., Harper, D. P., Maravelias, C., Runge, T., & Dumesic, J. A. (2017). Increasing the revenue from lignocellulosic biomass: Maximizing feedstock utilization. *Science Advances*, *3*(5). https://doi.org/10.1126/sciadv.1603301

[4] Lynd, L. R., Liang, X., Biddy, M. J., Allee, A., Cai, H., Foust, T., Himmel, M. E., Laser, M. S., Wang, M., & Wyman, C. E. (2017). Cellulosic ethanol: status and innovation. *Current Opinion in Biotechnology*, *45*, 202–211. https://doi.org/10.1016/j.copbio.2017.03.008

[5] Sun, Z., Fridrich, B., de Santi, A., Elangovan, S., & Barta, K. (2018). Bright Side of Lignin Depolymerization: Toward New Platform Chemicals. *Chemical Reviews*, *118*(2), 614–678. https://doi.org/10.1021/acs.chemrev.7b00588

[6] Garcia, A. C., Cheng, S., & Cross, J. S. (2020). Solvolysis of Kraft Lignin to Bio-Oil: A Critical Review. *Clean Technologies*, *2*(4), 513–528. https://doi.org/10.3390/cleantechnol2040032

[7] Oh, Y., Shin, J., Noh, H., Kim, C., Kim, Y.-S., Lee, Y.-K., & Lee, J. K. (2019). Selective hydrotreating and hydrocracking of FCC light cycle oil into high-value light aromatic hydrocarbons. *Applied Catalysis A: General*, *577*, 86–98. https://doi.org/10.1016/j.apcata.2019.03.004

[8] Yuan, Z., Tymchyshyn, M., & Xu, C. C. (2016). Reductive Depolymerization of Kraft and Organosolv Lignin in Supercritical Acetone for Chemicals and Materials. *ChemCatChem*, *8*(11), 1968–1976. https://doi.org/10.1002/cctc.201600187

[9] Dong, L., Lin, L., Han, X., Si, X., Liu, X., Guo, Y., Lu, F., Rudić, S., Parker, S. F., Yang, S., & Wang, Y. (2019). Breaking the Limit of Lignin Monomer Production via Cleavage of Interunit Carbon–Carbon Linkages. *Chem*, *5*(6), 1521–1536. https://doi.org/10.1016/j.chempr.2019.03.007

[10] Yang, J., Zhao, L., Liu, C., Wang, Y., & Dai, L. (2016). Catalytic ethanolysis and gasification of kraft lignin into aromatic alcohols and H2-rich gas over Rh supported on

La2O3/CeO2–ZrO2. *Bioresource Technology*, *218*, 926–933. https://doi.org/10.1016/j.biortech.2016.07.052

[11] Mortensen, P. M., Grunwaldt, J.-D., Jensen, P. A., Knudsen, K. G., & Jensen, A. D. (2011). A review of catalytic upgrading of bio-oil to engine fuels. Applied Catalysis A: General, 407(1–2), 1–19. https://doi.org/10.1016/j.apcata.2011.08.046

[12] Qiu, S., Li, M., Huang, Y., & Fang, Y. (2018). Catalytic Hydrotreatment of Kraft Lignin over NiW/SiC: Effective Depolymerization and Catalyst Regeneration. *Industrial & Engineering Chemistry Research*, *57*(6), 2023–2030. https://doi.org/10.1021/acs.iecr.7b04803

[13] Narani, A., Chowdari, R. K., Cannilla, C., Bonura, G., Frusteri, F., Heeres, H. J., & Barta, K. (2015). Efficient catalytic hydrotreatment of Kraft lignin to alkylphenolics using supported NiW and NiMo catalysts in supercritical methanol. *Green Chemistry*, *17*(11), 5046–5057. https://doi.org/10.1039/c5gc01643f

[14] Kong, L., Liu, C., Gao, J., Wang, Y., & Dai, L. (2019). Efficient and controllable alcoholysis of Kraft lignin catalyzed by porous zeolite-supported nickel-copper catalyst. *Bioresource Technology*, *276*, 310–317. https://doi.org/10.1016/j.biortech.2019.01.015

[15] Zakzeski, J., Jongerius, A. L., Bruijnincx, P. C., & Weckhuysen, B. M. (2012). Catalytic Lignin Valorization Process for the Production of Aromatic Chemicals and Hydrogen. *ChemSusChem*, *5*(8), 1602–1609. https://doi.org/10.1002/cssc.201100699

[16] Qi, S.-C., Hayashi, J.-ichiro, Kudo, S., & Zhang, L. (2017). Catalytic hydrogenolysis of kraft lignin to monomers at high yield in alkaline water. *Green Chemistry*, *19*(11), 2636–2645. https://doi.org/10.1039/c7gc01121k

[17] Rana, M., Islam, M. N., Agarwal, A., Taki, G., Park, S.-J., Dong, S., Jo, Y.-T., & Park, J.-H. (2018). Production of Phenol-Rich Monomers from Kraft Lignin Hydrothermolysates in Basic-Subcritical Water over MoO3/SBA-15 Catalyst. *Energy & Fuels*, *32*(11), 11564–11575. https://doi.org/10.1021/acs.energyfuels.8b02616

[18] Kim, K. H., Brown, R. C., Kieffer, M., & Bai, X. (2014). Hydrogen-Donor-Assisted Solvent Liquefaction of Lignin to Short-Chain Alkylphenols Using a Micro Reactor/Gas Chromatography System. *Energy & Fuels*, *28*(10), 6429–6437. https://doi.org/10.1021/ef501678w

[19] Jin, L., Li, W., Liu, Q., Wang, J., Zhu, Y., Xu, Z., Wei, X., & Zhang, Q. (2018). Liquefaction of kraft lignin over the composite catalyst HTaMoO6 and Rh/C in dioxane-water system. *Fuel Processing Technology*, *178*, 62–70. https://doi.org/10.1016/j.fuproc.2018.05.014

[20] Wu, Z., Zhao, X., Zhang, J., Li, X., Zhang, Y., & Wang, F. (2019). Ethanol/1,4-dioxane/formic acid as synergistic solvents for the conversion of lignin into high-value added phenolic monomers. *Bioresource Technology*, *278*, 187–194. https://doi.org/10.1016/j.biortech.2019.01.082

[21] Zhang, B., Li, W., Dou, X., Wang, J., Jin, L., Ogunbiyi, A. T., & Li, X. (2020). Catalytic depolymerization of Kraft lignin to produce liquid fuels via Ni–Sn metal oxide catalysts. *Sustainable Energy & Fuels*, *4*(3), 1332–1339. https://doi.org/10.1039/c9se01089k

[22] Domínguez-Robles, J., Tamminen, T., Liitiä, T., Peresin, M. S., Rodríguez, A., & Jääskeläinen, A.-S. (2018). Aqueous acetone fractionation of kraft, organosolv and soda lignins. *International Journal of Biological Macromolecules*, *106*, 979–987. https://doi.org/10.1016/j.ijbiomac.2017.08.102

[23] Tolbert, A., Akinosho, H., Khunsupat, R., Naskar, A. K., & Ragauskas, A. J. (2014). Characterization and analysis of the molecular weight of lignin for biorefining studies. *Biofuels, Bioproducts and Biorefining*, *8*(6), 836–856. https://doi.org/10.1002/bbb.1500

[24] Zhu, X., Wang, X., & Ok, Y. S. (2019). The application of machine learning methods for prediction of metal sorption onto biochars. *Journal of Hazardous Materials*, *378*, 120727. https://doi.org/10.1016/j.jhazmat.2019.06.004

[25] Chuquin-Vasco, D., Chuquin-Vasco, N., Chuquin-Vasco, J., & Lo-Iacono-Ferreira, V. (2021). Prediction of Methanol Production in a Carbon Dioxide Hydrogenation Plant Using Neural Networks. *Energies*, *14*(13), 3965. https://doi.org/10.3390/en14133965

[26] Bird, R. B., Muttzall, K. M. K., & Heuven, J. W. van. (2000). *Transport phenomena*. Wiley.

[27] Kobayashi, T., Furuya, T., Fujitsuka, H., & Tago, T. (2019). Synthesis of Birdcage-type zeolite encapsulating ultrafine Pt nanoparticles and its application in dry reforming of methane. *Chemical Engineering Journal*, *377*, 120203. https://doi.org/10.1016/j.cej.2018.10.140

[28] Pirngruber, G. D., Laroche, C., Maricar-Pichon, M., Rouleau, L., Bouizi, Y., & Valtchev, V. (2013). Core–shell zeolite composite with enhanced selectivity for the separation of branched paraffin isomers. *Microporous and Mesoporous Materials*, *169*, 212–217. https://doi.org/10.1016/j.micromeso.2012.11.016

[29] Yue, Y.-H., Tang, Y., Liu, Y., & Gao, Z. (1996). Chemical Liquid Deposition Zeolites with Controlled Pore-Opening Size and Shape-Selective Separation of Isomers. *Industrial & Engineering Chemistry Research*, *35*(2), 430–433. https://doi.org/10.1021/ie9502648

[30] Losch, P., Boltz, M., Bernardon, C., Louis, B., Palčić, A., & Valtchev, V. (2016). Impact of external surface passivation of nano-ZSM-5 zeolites in the methanol-to-olefins reaction. *Applied Catalysis A: General*, *509*, 30–37. https://doi.org/10.1016/j.apcata.2015.09.037

Appendix 2

Explaining permutation importance (PI) feature importance

Permutation importance method

The permutation importance method is used to evaluate the importance of features in ML models for regression problems [1]. The method is based on the idea of how the models' performance is affected by changing the values of a given feature. The models performance is evaluated by using a statistical metric such as mean square error (MSE) or coefficient of determination (R^2). This method assumes that the features with the highest importance influence the performance of the model more significantly. By this reasoning, if a given important feature is taken and its values are randomized, the model's performance should dramatically decrease, due to the model's inability to said feature to make predictions.

Permutation importance for a feature can be calculated as follows:

$$PI = \frac{1}{N}\sum_{i=1}^{N}\left(Score_{original} - Score_{randomized}\right)$$

Where:

- PI is the permutation importance of the feature.
- $N$is the number of permutations.
- $Score_{original}$ is the model's performance score on the original dataset.
- $Score_{randomized}$ is the model's performance score on the dataset where the values of the feature have been randomly shuffled.

A higher permutation importance value indicates that the feature is more important for making accurate predictions. If randomizing the feature's values leads to a significant decrease in model performance, it can be said that the feature is important in the regression model. Comparing feature's permutation importance allows for grasping their relative importance in the context of the dataset used for the regression model.

In the context of this dissertation, as it has been stated in previous chapters, permutation importance of a given model may not be representative of the entirety of the possible experimental space, thus, whatever conclusions are extrapolated from this importance must acknowledge this.

Reference:

[1] Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. Bioinformatics, 26(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Appendix 3

Explaining SHapley Additive exPlanations (SHAP)

Explaining SHapley Additive exPlanations (SHAP)

How SHAP works

SHAP (Shapley Additive explanations) stands as the current state-of-the-art method for machine learning model expandability. The method was first introduced in 2017 in a paper published by Lundberg and Lee [1] and essentially allows for reverse-engineer any output from a machine learning model. SHAP values are "model agnostic", which means they can be used with any complex model regardless of its type (for example, decision trees, neural networks) providing a way for the user to understand how the model is taking decisions or calculating the outputs. Although the actual implementation of SHAP values is not difficult (being directly available as a library for Python), understanding how it works is important to better grasp whether the results obtained by using it can be trusted or when it may be used for better effect.

SHAP values are based on Shapley values, a concept borrowed from game theory which involves two parts: a game and players. In the context of SHAP values a "game" is an individual prediction from a model and the "players" are the features or variables used in said model. Shapley values attempt to quantify the contribution made by each "player" in each game, on the other hand SHAP values quantify the contribution made by each feature in a model. It is important to note that in the context of ML each "game" refers to only one prediction attempt.

As an example, a hypothetical ML that predicts income as a function of: age, gender and job of a given person is explained next []. Shapley values are based on the idea that the result of every possible combination of players must be considered to evaluate the importance of a given player. In this example, this corresponds to the all the possible combinations of f characteristics (where f can go from 0 to F, F being the total number of existing features, 3 in this case). This is termed "power set" in math and can be represented as a tree, as shown in Figure A-6 below.

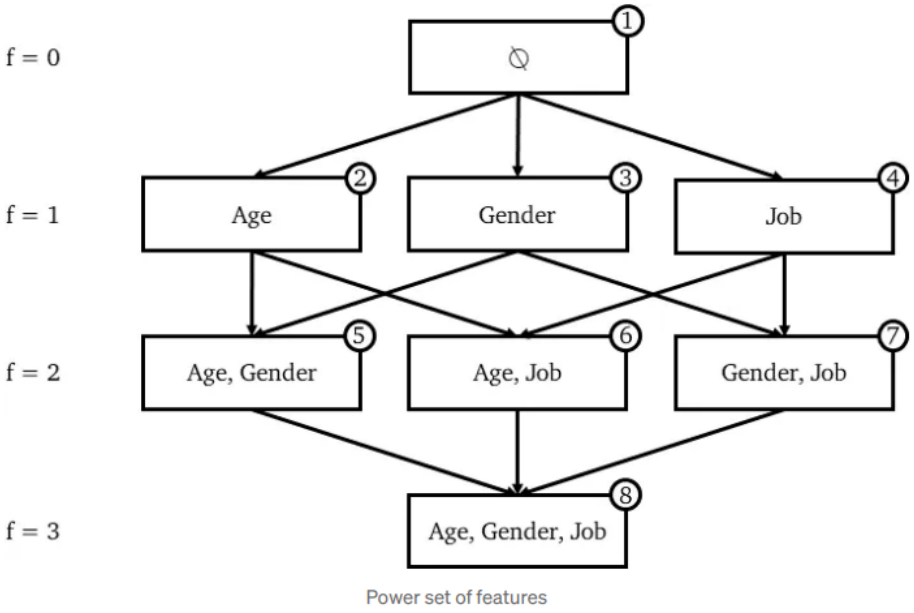


Figure A-6. Representation of a power set.

Each node represents a combination of features, and each border represents the inclusion of a feature that was not present in the previous combination. In this case, there are 8 possible combinations. SHAP trains a predictive model for each possible combination, resulting in 8 models trained in the case of data with 3 features. These models are, of course, the same in terms of hyperparameters and the data they are trained with. The only thing that changes across models is which features are included in them. For a new prediction x_0 the 8 models represented as a tree would look like Figure A-7 below.

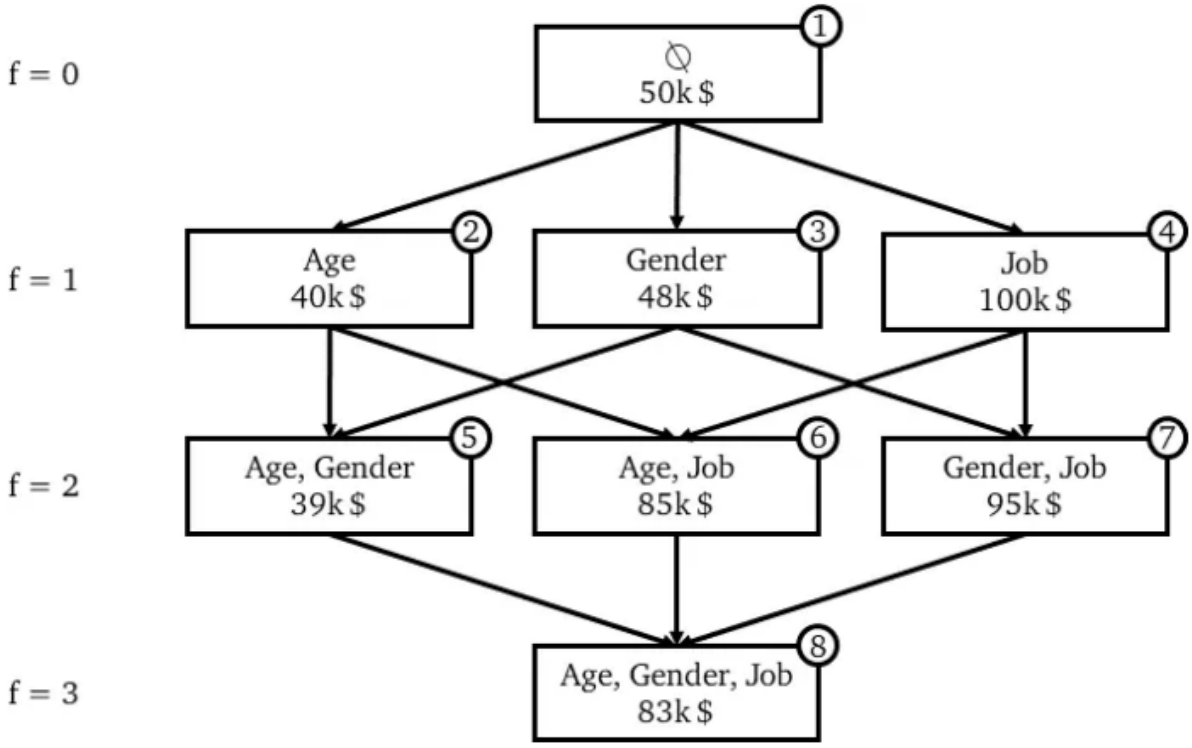


Figure A-7. Predictions made by different models for x_0. In each node, the first row reports the coalition of features included in the model, the second row reports the income predicted for x_0 by that model.

In here, the gap in prediction across two connected nodes (models) can be ascribed to the effect of the extra feature present, this is called "marginal contribution". To obtain the overall effect of a single feature in the above example, it would be necessary to consider the marginal contribution of the feature in all models where said feature is present. These marginal contributions can be then aggregated a weighted average, which in the case of the above figure, for the "age" feature:

$$SHAP_{Age}(x_0) = w_1 \times MC_{Age,\{Age\}}(X_0) + w_2 \times MC_{Age,\{Age,Gender\}}(X_0) + w_3 \times MC_{Age,\{Age,Job\}}(X_0) + w_4 \times MC_{Age,\{Age,Gender,Job\}}(X_0)$$

Where $w_1 + w_2 + w_3 + w_4 = 1$, for this example.

The sum of the weights of all the marginal contributions of models of 1 feature should be equal to the sum of the weights of all marginal contributions of a 2-feature model, and so forth accordingly. All the weights of marginal contributions to f-feature-models should be equal to each other for each f.

$$SHAP_{Age}(x_0) = w_1 \times MC_{Age,\{Age\}}(X_0) + w_2 \times MC_{Age,\{Age,Gender\}}(X_0) + w_3 \times MC_{Age,\{Age,Job\}}(X_0) + w_4 \times MC_{Age,\{Age,Gender,Job\}}(X_0)$$

$$SHAP_{Age}(x_0) = \frac{1}{3} \times (-10k\$) + \frac{1}{6} \times (-9k\$) + \frac{1}{6} \times (-15k\$) + \frac{1}{3} \times (-12k\$) = -11.33k\$$$

The above equations can be generalized as:

$$SHAP_{feature}(x) = \sum_{set:feature \in set} [|set| \times (\frac{F}{|set|})]^{-1} [Predict_{set}(x) - Predict_{set \setminus feateure}(x)]$$

Applied to the above example it would yield:

SHAP_Age(x_0) = -11.33k$

SHAP_Gender(x_0) = -2.33k$

SHAP_Job(x_0) = 46.66k$

The sum of these three above would yield 33k$, which is the difference between the full model (83k$) and dummy model without features (50k$).

It is worth noting that as the number of features grows, these sorts of calculations can become too time and resource consuming, necessitating the creation of methods to approximate the results.

References:

[1] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30. https://doi.org/10.48550/arXiv.1705.07874

[2] Mazzanti, S. (2023, September 15). Shap explained the way I wish someone explained it to me. Medium. https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30 (Accessed on 2023/10/10)

Appendix 4:

On the addition of catalysts descriptors and testing of Gaussian process regression.

Introduction

Previous chapters did not attempt to describe the properties of the catalysts (heterogeneous or homogeneous) beyond inputting the name of the active phase as a categorical feature. The reasoning for not attempting to describe the catalysts properties is that, as per the literature available [1 and 2], most of the descriptors that can be used for a given catalyst are usually in the form of the properties of the element that is deposited in the support. The fact that a significant fraction of the data points in the data sets for Chapters 2 and 4 include non-single metal catalysts complicates things. For Chapter 2, only 47 of 102 datasets include a single metal- catalyst, for Chapter 4 for bio-oil upgrading dataset includes 109 datapoints out of 212 with single-metal catalysts and for solid feedstocks dataset upgrading is 80 out of 176. Thus, roughly 50% or more of the data is not compatible with the descriptors used for single metal catalysts. Among the catalysts/materials that fall out of the category of "single metal catalysts", are materials like zeolites, activated carbons, bi- or trimetallic catalysts.

In this Appendix, we will attempt to see the impact of adding the following features to describe the catalyst (where possible), by using both XGBoost and Gaussian process regression:

- Dopant atomic weight (umas)
- Dopant electron affinity (KJ/mol)
- Ionization energy of dopant (eV)
- Dopant melting point (K)
- Dopant boiling point (K)
- Dopant enthalpy of fusion (KJ/mol)
- Atomic radius of metal (nm)
- Dopant's number of electrons

Shown in the table below are the R^2 and RMSE scores for both training and testing for the models trained in Chapter 2 and Chapter 4, both with XGBoost and GPR, where possible. All model's hyperparameters were chosen by using 5-fold cross-validation.

Table A-3. Model's performance with and without catalysts descriptors.

Chapter	Label	Model	Catalyst descriptors (yes or no)	Training		Test	
				R2	RMSE	R2	RMSE
2	Bio-oil yield	XGBoost	Yes	0.99	0.36	0.81	8.72
2	Bio-oil yield	XGBoost	No	0.99	0.80	0.82	8.53
2	Solid residue yield	XGBoost	Yes	0.98	1.94	0.77	7.93
2	Solid residue yield	XGBoost	No	0.96	2.68	0.87	5.85
4	BO final HHV	XGBoost	Yes	0.95	1.04	0.87	1.97
4	BO final HHV	XGBoost	No	0.95	1.05	0.85	2.18
4	BO ΔHHV	XGBoost	Yes	0.94	1.10	0.78	2.11
4	BO ΔHHV	XGBoost	No	0.93	1.18	0.77	2.14
4	SF final HHV	XGBoost	Yes	0.94	1.12	0.82	1.98
4	SF final HHV	XGBoost	No	0.96	0.91	0.85	1.82
4	SF ΔHHV	XGBoost	Yes	0.95	1.07	0.71	2.08
4	SF ΔHHV	XGBoost	No	0.95	1.10	0.80	1.75
2	Bio-oil yield	GPR	No*	0.96	0.17*	0.66	79.73
2	Solid residue yield	GPR	No*	0.95	0.20*	0.21	78.50
4	SF final HHV	GPR	No*	0.94	0.24*	0.77	27.12
4	SF ΔHHV	GPR	No*	0.94	0.25*	0.72	7.75
4	BO final HHV	GPR	No*	0.93	0.24*	0.75	31.47
4	BO ΔHHV	GPR	No*	0.93	0.23*	0.80	10.33
2	Bio-oil yield	Linear regression	No[1]	0.86	6.55	0.59	12.93
2	Solid residue yield	Linear regression	No[1]	0.74	7.33	0.44	12.50
4	SF final HHV	Linear regression	No[1]	0.90	1.53	-0.95	8.29
4	SF ΔHHV	Linear regression	No[1]	0.91	8.29	-1.07	8.29
4	BO final HHV	Linear regression	No[1]	0.91	1.47	0.54	3.50
4	BO ΔHHV	Linear regression	No[1]	0.90	1.48	0.36	3.15

* Data was normalized.

[1] Missing values from data sets were removed.

It can be observed that for the XGBoost models based on Chapter 2's data, adding the catalyst descriptors results in a small drop in testing scores. For Chapter 4's XGBoost models a small improvement in testing scores can be seen.

It is difficult to assert why addition of these features to Chapters 2 and 4 results in these small changes. Although it is possible that the addition of these new features only further increases the

likelihood of overfitting, due to the small amount of data that is available. Understandably, because the descriptors only apply for a small fraction of the rows in the data, it may be difficult for the trained models to leverage this information in a way that results in better predictions, and unsurprisingly this also results in the features holding low importance, from what can be observed from permutation importance analysis. From the point of view of literature, we know that the features that we added hold little importance in other studies [1]. Other studies make use of density-functional theory (DFT) to calculate catalyst properties to some effect, though such calculations are beyond the scope of this dissertation [3].

Attempting to use the datasets with new catalyst descriptors with GPR resulted in many problems, mostly due to GPR's lack of a native way to handle missing values the way XGBoost can. In an attempt to fix this, K-nearest-neighbors method for imputation was tested, however, this resulted in terrible results (such as negative R^2 scores). The reasoning for this is that, the new catalyst descriptors do not apply to a large part of the datasets used, catalytic "species" inserted in the dataset in the form of a one-hot encoded words such as "zeolite", "activated carbon", bi- or tri-metallic catalysts cannot be described the same way that single metal-doped catalysts can, even if a number can be estimated for the missing feature value. Attempting to estimate the above-mentioned features by using K-nearest-neighbors probably results in non-sensical values, thus poor performance.

Additionally, linear regression algorithm was used as a benchmark to highlight the performance of the other models. Although the linear regression models achieve good training performance indicators, their testing scores are rather low in comparison to the ones seen using XGBoost.

Although the gap between training and testing for the XGBoost models is the lowest among the models trained, it is still far from ideal. Having more data points by executing experiments would preferable, although it is still possible that the resulting model would not improve purely by doing that. It is my belief that the complexities related to reaction kinetics are not properly explained with the current features available, and that the key for better models would be to include features that describe the interaction between the solid catalyst and the reactant (lignin, lignocellulose). This would be challenging, as neither lignin nor lignocellulose have a well-defined chemical structure or molecular size, complicating the creation of kinetic models. Still, estimating some of the not-explicitly-stated properties of the catalyst by using principles based on chemistry or material science could be an interesting future topic for research.

References:

[1] Wang, G., Fearn, T., Wang, T., & Choy, K.-L. (2021). Machine-learning approach for predicting the discharging capacities of doped lithium nickel–cobalt–manganese cathode materials in li-ion batteries. ACS Central Science, 7(9), 1551–1560. https://doi.org/10.1021/acscentsci.1c00611

[2] Toyao, T., Suzuki, K., Kikuchi, S., Takakusagi, S., Shimizu, K., & Takigawa, I. (2018). Toward Effective Utilization of Methane: Machine Learning Prediction of Adsorption Energies on

Metal Alloys. The Journal of Physical Chemistry C, 122(15), 8315–8326. https://doi.org/10.1021/acs.jpcc.7b12670

[3] Jinnouchi, R., & Asahi, R. (2017). Predicting catalytic activity of nanoparticles by a DFT-aided machine-learning algorithm. The Journal of Physical Chemistry Letters, 8(17), 4279–4283. https://doi.org/10.1021/acs.jpclett.7b02010