Hydrothermal liquefaction data for use in machine learning models

Geert Haarlemmer a, Lucie Matricon a, Anne Roubaud a

a Univ. Grenoble Alpes, CEA, LITEN, DTCH, LRP, F-38000 Grenoble, France

[geert.haarlemmer@cea.fr](mailto:geert.haarlemmer@cea.fr)

Abstract

Hydrothermal liquefaction is a sustainable pathway to generate biogenic liquids from organic resources. The technology is compatible with a wide variety of resources such as ligno-cellulosic resources, organic waste, algae, and sewage sludge. The chemistry is complex and predictions of product yields are notoriously difficult. Understanding and modelling of hydrothermal liquefaction is currently mostly based on a simplified biochemical analysis, product yield data and for a small application field. This paper presents a large dataset of 2587 experiments in batch reactors that were extracted from 173 publications in the scientific literature. This paper presents how machine learning can contribute to the field of hydrothermal processes. The presented techniques allow a better understanding of the data and its accuracy.

**Keywords**: Hydrothermal liquefaction, Yield data, Data analysis, Machine learning.

* 1. Introduction

The hydrothermal liquefaction (HTL) of biomass is now an established technology, making it possible to produce high quality biofuels from low value biogenic waste. Hydrothermal processes operate with hot compressed water, typically above 200 °C and at high pressures converting bio-resources in a crude oil. The field of hydrothermal processes remains very empiric, mostly based on experiments. This is mainly due to the complexity of the resources as well as its chemistry.

Hydrothermal reactions are performed in reactors that are capable to resist to high pressures. There are two major categories of laboratory reactors. The vast majority of scientific papers present work using batch reactors, where the resource is placed in the reactor and the closed system is heated. A large number of experiments can be conducted under controlled conditions. The validity of batch experiments to represent a future industrial transformation is however limited as the heating and cooling phases play an important role on the experiments. Continuous reactors in the academic laboratories are less numerous and experiments are longer to conduct. Batch reactors do have an important role to play in contributing to the understanding of the chemistry.

Modelling of the hydrothermal conversion is an important subject in the literature ([Shahbaz et al. 2023](#_ENREF_6)). Modelling is done by (non)linear correlations ([Déniel et al. 2016](#_ENREF_1)) and kinetic models ([Hietala and Savage 2021](#_ENREF_2)). Most of these approaches have been limited in the past to specific resources. Exploitation at a larger scale can be done with machine learning algorithms as shown by ([Li et al. 2021](#_ENREF_4)) among others. The advantage of this approach is that the data can be less structured and more voluminous than is common practice in a design of experiment approach. The validity of this approach is however often limited due to the limited amount of data used, typically a few hundred or even less experiments.

There are many incomplete data sets in the literature, insufficient data to understand the resource, but also missing crucial details concerning the reactor and the experiments, making interpretation of the results difficult and reproduction impossible. Resource and product analysis also form an important obstacle in the interpretation of reported hydrothermal experiments.

This study presents results on the HTL of a wide variety of resources, compiled from 173 literature references. This paper aims to show how modern data analysis techniques can make sense of the experimental data in the literature. The data was used to create a model that allows an accurate prediction of the hydrothermal process. The objective is to evaluate the accuracy of the data, not for each experiment but the ability to predict an arbitrary experiment from the literature data.

* 1. Biomass analysis

The description of biomass solely by ultimate analysis (CHONS, Ash, Heating Value) is useful for high temperature gasification but not so useful for low temperature processes such as HTL as mentioned in the introduction. Conversion in hot compressed water starts with the hydrolysis of the biopolymers, followed by the conversion of the produced molecules. For this reason, it is essential to be able to distinguish different common biopolymers (lignin, cellulose, proteins, and lipids) from simpler molecules (sugars, fatty acids, amino acids and pigments). It has been well established that proteins tend to be depolymerised to amino acids and dissolved under hydrothermal conditions with the formation of bio-oil through Maillard reactions. Lipids (not limited to triglycerides) are hydrolysed into glycerol and fatty acids. The latter contributes mainly to bio-oil formation. Compositional data is however not always reported. Most papers present an ultimate analysis and around half of the papers present details compositional analysis. Literature on hydrothermal carbonisation rarely include compositional analysis data. A majority of publications present compositional data that does not add up to 100 %. While this in itself can be understood, it is a problem in the usage of the results.

The compositional analysis in the form of lignin, carbohydrates, proteins and lipids remains a gross simplification compared to the vast complexity of actual biomass. In addition, measurements are often uncertain and can easily produce inaccurate results. In practice however, the full analysis, including profiling amino acids, fatty acids, sugar, and pigments analysis, and others are complex and are not performed on a routine basis.

* 1. Hydrothermal experiments

HTL experiments in batch reactors follow a common pattern. The resource is placed in the autoclave reactor with the desired water content and heated. The products are recovered and quantified after opening the reactor. This makes it possible to compare the data and to exploit it. Reactors and experimental procedures are quite constant across the literature. Product recovery techniques do show some variations but can easily be classified in families.

A typical experiment consists of a biomass slurry with a determined dry matter to water ratio being placed in a batch reactor. The reactor is then heated to the reaction temperature with a certain heating rate of that often varies from 4 to 100 °C/min, depending on the size and the heating technology. Once the reactor reaches the desired reaction temperature, it is held during a specified holding time. The pressure inside the reactor is the result of the reaction temperature, due to the water vapour pressure, and to a minor extend the initial nitrogen pressure and the quantity of gas produced. After this holding time, the reactor is cooled to room temperature. This procedure is constant across most studies.

The oily or solid product is referred to as biocrude in this study; this is anything not in the aqueous or gaseous phases. Biocrude is extremely variable in aspect, ranging from a free flowing viscous liquid to a dry powder. Extraction with a solvent separates the biocrude in bio-oil and biochar. Reliable bio-oil and bio-char yields can be obtained relatively easily from the HTL experiments. The yield of the aqueous and the gaseous phases largely depend on the applied techniques. There different product recovery techniques and yield definitions. The solid yield definition is relatively standard as the product not dissolving in water and the organic solvent. The oil yield is the oil retrieved by extraction from the oily residue after liquefaction.

For each reported experiment in this study, the biochemical composition of the biomass is used as presented in the publication. Each experiment is further described by the dry matter content in the biomass slurry, additional chemicals, heating rate, reaction temperature, holding time, extraction order and solvent used. The reported results are bio-oil, char, gas and aqueous phase yields whenever they are available.

* 1. Data collection

All the data in this paper has been previously described in publications. The data was collected from the websites from scientific publishers (only peer reviewed articles were considered). In this study, we only consider fully presented and documented data sets, including an adequate analysed biomass, the experiments fully described and results are presented in an accessible way.

Table . Range of the data in the included dataset.

|  |  |
| --- | --- |
| **Independent variables** | **Range** |
| Temperature | 20 - 600 °C |
| Heating rate | 1 - 850 °C/min |
| Holding time | 0 - 1320 min |
| Dry matter concentration | 1 - 50 % |
| Composition | Ash, lignin, carbohydrates, proteins, lipids, guaiacol, amino acids, carboxylic acids, fatty acids, glycerol |
| Solvents | Ethyl acetate, ether, dichloromethane, chloroform, acetone, n-hexane, toluene, dimethyl sulfoxide, ethanol, isopropanol |
| Method | Order solvent extraction after or before water separation (values 1 and 2 respectively) |

* 1. Numerical methods used

The algorithms used in this study are well known algorithms from the SciKit-Learn library and are implemented in Python 3.9. The regressor used in this study is the random forest regressor. The random forest regressor is a robust ensemble method based on multiple decision trees. A decision tree is a graph that, based on tests, proposes an outcome starting from a set of input parameters. As a measurement of the quality of the fit, the coefficient of determination is used as calculated by the function r2\_score in the SciPy library. This formulation of R² ensures that the upper limit is 1 for a perfect fit.

Data from studies in any field are subject to uncertainties. In the machine learning and artificial intelligence field, process variables (independently modifiable parameters) are referred to as features. Uncertainties come from experimental errors variations in biomass analysis techniques and experimental practices. The accuracy that can be obtained from modelling with the data in this study is evaluated using MAPIE (Model Agnostic Prediction Interval Estimator). This Python library allows the identification of confidence intervals on data modelling with an arbitrary regressor. The theoretical basis of this library is described by ([Kim et al. 2020](#_ENREF_3)). Analysis of the data is also performed using the SHAP library that supplies algorithms for interpretable AI. The library uses a game theory approach initially proposed by Lloyd Shapely and developed by ([Lundberg et al. 2019](#_ENREF_5)) as a python library. The algorithms in this library allow the evaluation of individual variables and their interactions on the global results as well as individual experiments.

* 1. Data Analysis

The content of the dataset is difficult to visualise to check for coherency, as there are nine dimensions. Figure 1 shows how the composition of the resources in the database covers the full spectrum. Most real biomasses contain a well balances mixture of compounds required for life. In practice, the extremes and high lipid experiments are mainly covers by experiments on model mixtures and pure compounds.

|  |  |
| --- | --- |
|  |  |
| Figure 1 Ternary diagram with compositions, the size is proportional to the number of occurrences. The resource categories are colour coded, FW ▲, FPW ▼, Algae ◄, Sewage **x**, Mixture ■, Model ●, Herbs, and Wood **+**. | Figure 2 Ternary diagram with random forest predictions of bio-oil yields |

To improve the understanding of the data a random forest model was generated and asked to predict the oil yield as a function of the composition as shown in Figure 2. The model was build using 16 parameters, or features. This modelling only takes into account the compositional effect and ignores all other process parameters.

It is obvious that the composition is not the only process parameter that is important. Figure 3 shows the feature importance on the oil yield prediction as predicted by the SHAP algorithm. The value on the horizontal axis give the influence of the parameter on the deviation from the global average.

The composition plays an important role. The most important process parameter is the temperature. The model compounds such as glycerol, lignin, carboxylic acids and amino acids do not play a significant role, mostly because they contribute little to the oil yield. Products from guaiacol experiments are classed as bio-oil. To check whether it is possible to reduce the dimensions of the modelling problem; a principal component analysis was conducted and presented in Figure 4.

|  |  |
| --- | --- |
|  |  |
| Figure 3 Feature importance | Figure 4 Principle component analysis |

The model accuracy (R²) of the bio-oil prediction is 97.7 % on the training data (70 % of the data) and 87.8 % on the test data (30 % of the data). This is a good score for this type of modelling ([Zhang et al. 2021](#_ENREF_7)). Typical values oscillate around 90 % for training data and can be as low as 75 % for test data.

|  |  |
| --- | --- |
| C:\Users\GH222598\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Publi_W2R_Temp_Oil.jpg | C:\Users\GH222598\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Publi_W2R_Temp_Oil.jpg |
| Figure 5 Model predictions with ● the training data and ▲ the test data | Figure 6 Model predictions coloured by temperature |

It is also possible to evaluate the quality of the random forest model by the evaluation of the confidence interval. This will put an accuracy on the model, rather than a goodness of fit. Figure 5 shows the predictions on the training set in black circles (●) and the test set in red triangles (▲). The confidence interval as calculated by the MAPIE algorithm is traced in the same graph. The 2σ confidence interval is around ±10 % absolute of the calculated yield.

Figure 6 shows the same data with the colour now indicating the temperature for which each experiment was performed. It shows that there is a light tendency for low temperatures below 250 °C to produce low oil yields. The 300 to 400 °C range favours medium oil yields in the 40 to 60 % yield range. High yields above 80 % are obtained by low and high temperatures alike; most likely, these are experiments with lipids or very high fat resources that produce oil no matter what other process parameters.

* 1. Conclusions

The large volume of data in the literature shows that there is a significant interest in hydrothermal experiments. This paper presents 2578 experiments from 173 scientific papers. Each individual experiments may be well done; their use for the scientific community is limited when the results cannot be exploited in larger meta studies. Meta studies are essential for the better understanding of the hydrothermal field as a whole.

Globally, taking all the data into account the expected absolute uncertainty of the data can be estimated at 10 % absolute. This value should be interpreted as the 2σ uncertainty of the prediction for a new resource, based on existing data.

References

Déniel, M., G. Haarlemmer, A. Roubaud, E. Weiss-Hortala and J. Fages (2016). "Modelling and Predictive Study of Hydrothermal Liquefaction: Application to Food Processing Residues." Waste and Biomass Valorization 8(6): 2087-2107.

Hietala, D. C. and P. E. Savage (2021). "A molecular, elemental, and multiphase kinetic model for the hydrothermal liquefaction of microalgae." Chemical Engineering Journal 407: 127007.

Kim, B., C. Xu and R. F. Barber (2020). Predictive Inference Is Free with the Jackknife+-after-Bootstrap. 34th Conference on Neural Information Processing Systems.

Li, J., W. Zhang, T. Liu, L. Yang, H. Li, H. Peng, S. Jiang, X. Wang and L. Leng (2021). "Machine learning aided bio-oil production with high energy recovery and low nitrogen content from hydrothermal liquefaction of biomass with experiment verification." Chemical Engineering Journal 425: 130649.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee (2019). "Explainable AI for Trees: From Local Explanations to Global Understanding." ArXiv.

Shahbaz, M., M. Alherbawi, P. Parthasarathy, G. McKay and T. Al-Ansari (2023). Process modelling of the hydrothermal liquefaction of oil-palm waste for biocrude and hydrochar production. Computer Aided Chemical Engineering. A. C. Kokossis, M. C. Georgiadis and E. Pistikopoulos, Elsevier. 52**:** 1475-1481.

Zhang, W., J. Li, T. Liu, S. Leng, L. Yang, H. Peng, S. Jiang, W. Zhou, L. Leng and H. Li (2021). "Machine learning prediction and optimization of bio-oil production from hydrothermal liquefaction of algae." Bioresource Technology 342: 126011.