Science of the Total Environment 935 (2024) 173234

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Modeling lignin extraction with ionic liquids using machine learning approach

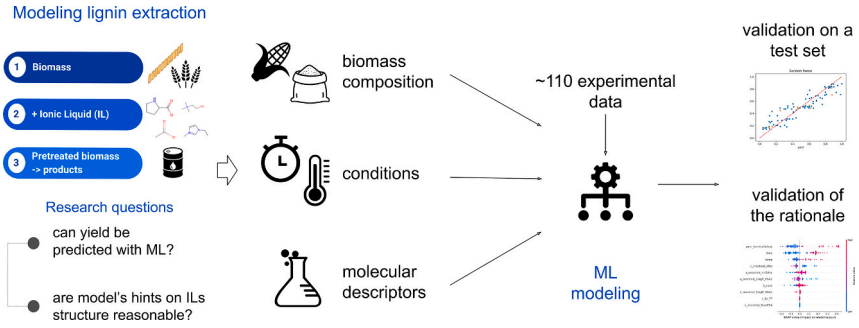
Karol Baran [a,*], Beata Barczak [b], Adam Kloskowski [a]

[a] Department of Physical Chemistry, Faculty of Chemistry, Gdansk University of Technology, Narutowicza 11/12, 80-233 Gdansk, Poland
[b] Department of Energy Conversion and Storage, Faculty of Chemistry, Gdansk University of Technology, Narutowicza 11/12, 80-233 Gdansk, Poland

HIGHLIGHTS

- Ionic Liquids usage for lignin extraction studied from a data analysis perspective
- Machine Learning approach for modeling extraction yield
- Molecular descriptors, biomass composition, and conditions serve as model input.
- Models are validated using test sets and interpretation of models' predictions.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Huu Hao Ngo

Keywords:
Quantitative Structure-Property Relationship (QSPR)
Lignin extraction
Designer solvents

ABSTRACT

Lignin, next to cellulose, is the second most common natural biopolymer on Earth, containing a third of the organic carbon in the biosphere. For many years, lignin was perceived as waste when obtaining cellulose and hemicellulose and used as a biofuel for the production of bioenergy. However, recently, lignin has been considered a renewable raw material for the production of chemicals and materials to replace petrochemical resources. In this context, an increasing demand for high-quality lignin is to be expected. It is, therefore, essential to optimize the technological processes of obtaining it from natural sources, such as biomass. In this work, an investigation of the use of machine learning-based quantitative structure-property relationship (QSPR) modeling for the preliminary processing of lignin recovery from herbaceous biomass using ionic liquids (ILs) is described. Training of the models using experimental data collected from original publications on the topic is assumed, and molecular descriptors of the ionic liquids are used to represent structural information. The study explores the impact of both ILs' chemical structure and process parameters on the efficiency of lignin recovery from different bio sources. The findings give an insight into the extraction process and could serve as a foundation for further design of efficient and selective processes for lignin recovery using ionic liquids, which can have significant implications for producing biofuels, chemicals, and materials.

* Corresponding author.
E-mail address: karol.baran@pg.edu.pl (K. Baran).

https://doi.org/10.1016/j.scitotenv.2024.173234
Received 11 February 2024; Received in revised form 25 April 2024; Accepted 12 May 2024
Available online 18 May 2024

1. Introduction

The utilization of fossil fuels has been closely linked to an escalation in environmental deterioration and pollution (Barczak et al., 2023). This association has prompted the exploration of alternative and sustainable energy sources, such as hydroelectric, wind, and solar power, and has fostered a growing interest in renewable resources, particularly lignocellulosic biomass (Paraschiv and Paraschiv, 2023). Lignocellulosic materials, encompassing hardwood, softwood, and agricultural residues, have emerged as pivotal renewable assets, constituting a substantial reservoir of carbon. Consequently, these resources have found successful applications in energy and chemical production processes (Kamm et al., 2008).

Plant biomass, a composite material, consists of three fundamental biopolymers: lignin, hemicellulose, and cellulose, in different proportions (Agbor et al., 2011; Kazimierski et al., 2022). The distribution and spatial arrangement of these constituents within cell walls are not uniform and are contingent upon factors such as plant species, tissue type, and the stage of cell wall maturation (Barakat et al., 2013; Isikgor and Becer, 2015). Each biopolymer exhibits distinctive properties that contribute to the overall recalcitrance of the raw material. The carbohydrate fraction, termed holocellulose, which encompasses cellulose and hemicellulose, imparts structural fortitude and rigidity to cell walls (Agbor et al., 2011). In contrast, lignin functions as an adhesive within cell walls, conferring resistance to compression, protection against pests and pathogens, and structural integrity to plant tissues (Rubin, 2008).

Broadly, plant biomass can be categorized into three principal groups predicated on the relative composition of holocellulose and lignin: softwood (coniferous; holocellulose 64.5 ± 4.6 %, lignin 28.8 ± 2.6), hardwood (deciduous; holocellulose 71.7 ± 5.7, lignin 23.0 ± 3.0), and herbaceous biomass, wherein the polymer composition varies substantially depending on the specific plant part (e.g., corn stover: holocellulose 59.3 %, lignin 18.1 %; corn cobs: holocellulose 63.9 %, lignin 21.2 %) (Liu et al., 2018a; Stolarski et al., 2018; Weerachanchai and Lee, 2013). The efficient utilization of residual herbaceous biomass, particularly in the context of expanding urbanization and heightened food demands, presents a significant technological and environmental challenge. Prioritizing the utilization of renewable resources, such as lignocellulosic biomass, stems from a heightened emphasis on environmental conservation. Nonetheless, the fractionation of biomass remains a persistent challenge, particularly concerning the isolation of lignin. Given its prevalence in forestry materials, energy crops, and agricultural residues, along with its substantial untapped potential for cost-effective utilization, refining the lignin extraction process is imperative. Failure to employ environmentally sound practices in biomass valorization could impede the transition from conventional petrochemical methods. Therefore, integrating green chemistry techniques is crucial to advancing the objectives of sustainable development (Hasanov et al., 2020).

The initial treatment of lignocellulosic biomass constitutes a pivotal preliminary phase, with the principal objectives of exposing the carbohydrate fraction and facilitating subsequent processing (Kumar and Sharma, 2017). The separation of individual biopolymers within the biomass matrix presents a formidable and resource-intensive challenge, primarily owing to the remarkable recalcitrance of cellulose to hydrolysis, the influence of oxidizing agents and harsh alkaline conditions, and the cohesive attribute of lignin. These properties of lignin entail the formation of hydrogen bonds among polymer molecules (Deng et al., 2022). However, other types of bonds between the biopolymer moieties should also be considered, including α ether linkage in carbohydrates-lignin complexes and covalent bonds. Besides hydrogen bonding, other interactions like van der Waals, ion-dipole attractions, and hydrophobic interactions are significant when examining interactions with lignin (Singh, 2022). There is a growing interest in green solvents like ionic liquids (IL) and deep eutectic solvents (DES) because of their minimal environmental impact, characterized by features such as low emissions due to extremely low vapor pressure, customizable chemical structures to reduce solvent usage, and the ease of their recycling and reusing (Zhu et al., 2018). The separation process is inaugurated with a pretreatment stage, implementable through both physical and thermochemical methodologies (Kumar and Sharma, 2017). This preliminary phase is specifically engineered to enhance the overall digestibility of biomass by means of lignin removal, thereby streamlining the subsequent processing of holocellulose. The acquired lignin has various applications, serving as a valuable fuel source, an effective emulsifying agent, or a binder. Meanwhile, following hydrolysis and fermentation, the carbohydrate fraction undergoes a conversion into high-value commodities, including biohydrogen, methane, bioalcohols, and carboxylic acids (Deng et al., 2022; Li and Takkellapati, 2018).

Conventional separation technologies are often beset by multifarious constraints, principally related to their selectivity and, consequently, their overall process efficiency. Moreover, in this epoch marked by an inexorable surge in environmental pollution, there is an earnest pursuit of technologies that leverage not only waste materials but also environmentally benign solvents. This quest is exemplified by the application of two, partially overlapping, group of compounds, i.e. DES and ILs. This quest is exemplified by the application of ILs (Pin et al., 2021). Current work is focused, solely, on the former one. Although some compounds can be included in both groups, there are fundamental differences in the chemical composition of ILs and DES. This might be the case since ILs are only composed of ions in equimolar ratios, and descriptors can be calculated separately for cation and anion. In the case of DESs, the fact that the eutectic mixture might be obtained by mixing two or more components in different molar ratios must be considered. In consequence, the modeling of DES systems requires a different approach. ILs have garnered significant attention due to their exceptional attributes, notably high thermal stability and low vapor pressure, which serve to mitigate exposure risks in separation processes, affording a substantial advantage over conventional volatile solvents (Flieger and Flieger, 2020). Recent years have witnessed a proliferation of research endeavors dedicated to the extraction of natural polymers in ionic liquids, thereby accentuating their noteworthy potential in pioneering alternative methods for the extraction and processing of biomass constituents (Rieland and Love, 2020). The most proficient lignin yield has been observed within the category referred to as herbaceous biomass. For instance, bagasse subjected to [Emim][ABS] at 190 °C for 1.5 h yielded an impressive lignin yield of 97 % (Tan et al., 2009). In contrast, when dealing with softwood and hardwood biomass, the efficiency of lignin separation has exhibited considerable variability, ranging from as low as 7.5 % for eucalyptus exposed to [Bmim][Cl] at 120 °C for 0.5 h to as high as 81.7 % for pine sapwood treated with [Bmim][HSO4] (20 % water) at 120 °C for 22 h (Brandt et al., 2011; Li et al., 2016). These results underscore the considerable potential of employing ionic liquids for the isolation of lignin from biomass. Nevertheless, the judicious selection of an appropriate ionic liquid, along with the optimization of temperature and contact time, remains a multifaceted issue that necessitates individualized analyses tailored to each biomass type or group. The industrial use of ionic liquids must, also, meet economic and environmental requirements. Due to their high price and not neutral impact on the environment, this is the main factor limiting their spread in practice (Sharma et al., 2023). In this context, conducting research based on building the theoretical predictive models is a desirable way to search for their optimal structure.

Due to the complexity of the problem stated it might be valuable to use the approach that might predict lignin yield value while also providing some insights into the importance of all the factors involved in the process of lignin extraction. This kind of predictive modeling is well addressed with Machine Learning (ML) methods that are data-driven approaches. Models are trained using experimental data from previous studies on the topic. Machine Learning in the context of chemical engineering entails the application of computational algorithms and statistical models to analyze and interpret complex datasets inherent to

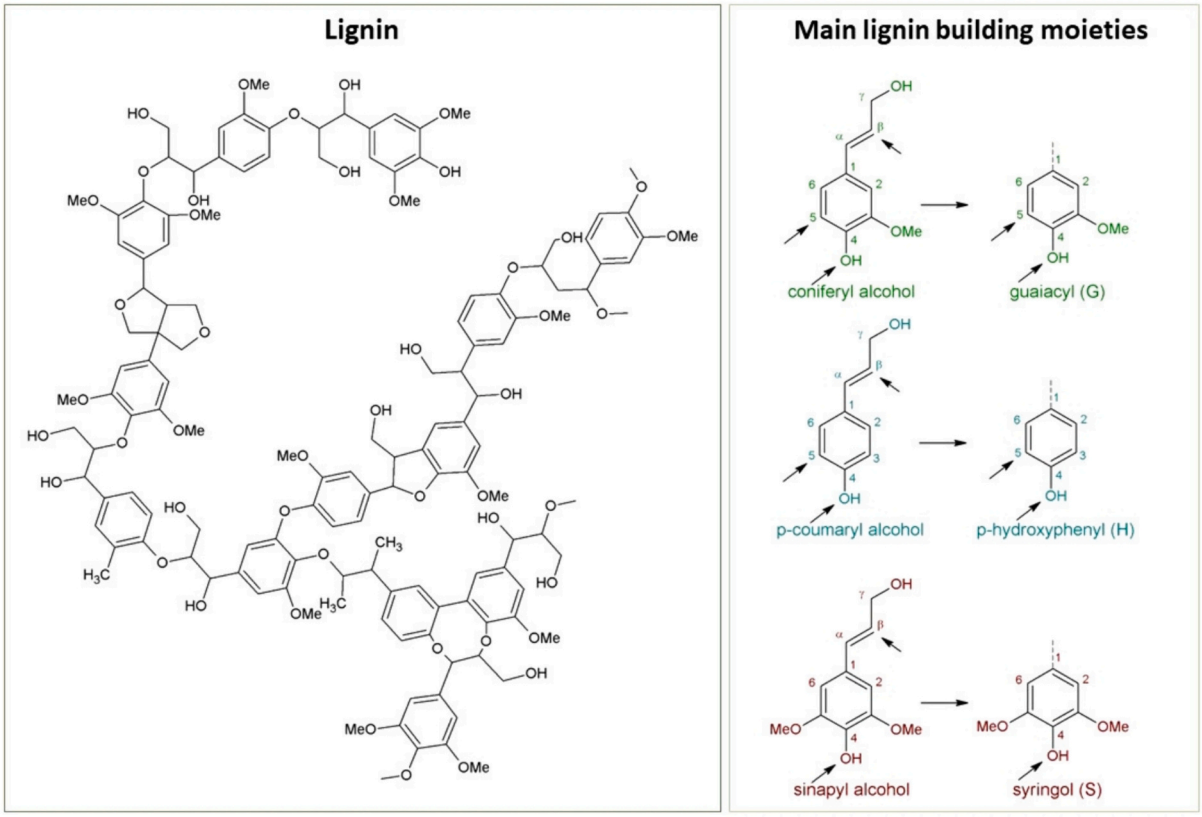


Fig. 1. Chemical structure of lignin and main lignin building moieties a) monolignols p-coumaryl alcohol and unit H, b) sinapyl alcohol, and unit S, and c) coniferyl alcohol and corresponding unit G.

chemical processes. By leveraging advanced mathematical techniques, practitioners in this field can derive insights, optimize processes, and make data-driven decisions. The integration of machine learning within chemical engineering facilitates the development of predictive models, enhancing efficiency and fostering a deeper understanding of intricate relationships within chemical systems. In this work, modeling incorporating structure-property correlations is used. The main goal of the work is to propose ML models that can predict lignin extraction yield based on parameters describing the composition of biomass, conditions at which the process occurs, and structural descriptors of ionic liquids. This endeavor represents a novel contribution to the field of chemical engineering as it addresses the pressing need for accurate and efficient prediction tools in biomass processing. The integration of machine learning techniques specifically tailored to the intricate interplay of biomass composition and ionic liquid characteristics is a unique aspect of this research, aiming to fill the gap in the field of predicting experimental results on biomass samples using an easy-to-use model. The significance of this work lies in its potential to enhance the overall efficiency of biomass processing, optimize resource utilization, and pave the way for sustainable and economically viable practices in the chemical industry. Furthermore, the exploration of this novel approach opens avenues for future research, fostering a deeper understanding of the underlying mechanisms governing biomass conversion processes and providing a basis for the development of advanced, data-driven strategies.

2. Background and problem foundation

2.1. Chemistry of lignin

Lignin, abundant and ubiquitous, constitutes a significant renewable energy source, comprising approximately one-third of the Earth's organic carbon, totaling around 1 billion metric tons (Martínez et al., 2009). It ranks as the second most prevalent biopolymer, following holocellulose (Achyuthan et al., 2010). As a binding agent, lignin reinforces the structure of plant cell walls by infiltrating the gaps between cellulose and hemicellulose, imparting strength and rigidity. Structurally, lignin is a three-dimensional polymer formed from the oxidative coupling of p-coumaryl alcohol, sinapyl alcohol, and coniferyl alcohol (see Fig. 1) (Isikgor and Becer, 2015). Its composition and properties can vary depending on factors such as plant species, botanical variety, and even within individual trees.

Lignin consists of monolignols, characterized by phenolic groups and propyl side chains, with variations in the number of methoxy groups attached to the aromatic residue. The prevalence of each monolignol type in lignin is determined by the plant species' taxonomic classification (Gillet et al., 2017), resulting in lignin types such as softwood lignin (G type), hardwood lignin (GS type), pressed wood lignin (HG type), and herbaceous lignin (HGS type) (G. Calvo-Flores et al., 2015).

Lignin forms complex covalently linked structures with hemicellulose within biomass, contributing to its structural integrity. In herbaceous biomasses, ferulic acid initially bonded to arabinoxylan via ester linkages facilitates this interaction, integrating into lignin as it matures through radical polymerization reactions. Softwood and hardwood exhibit direct interactions between lignin and carbohydrates during lignification, driven by interactions between carbohydrate hydroxyl groups and electrophilic intermediates in developing lignin chains (Brandt et al., 2013). The composition of individual polymers significantly influences material properties and interactions among lignocellulosic biomass constituents. Herbaceous biomasses, with lower lignin content, are more susceptible to pretreatment, while softwoods, rich in lignin, are less so (Weerachanchai and Lee, 2013), while softwoods, rich in lignin, evince lower susceptibility (Brandt et al., 2011). Higher hemicellulose or cellulose content may enhance delignification efficacy

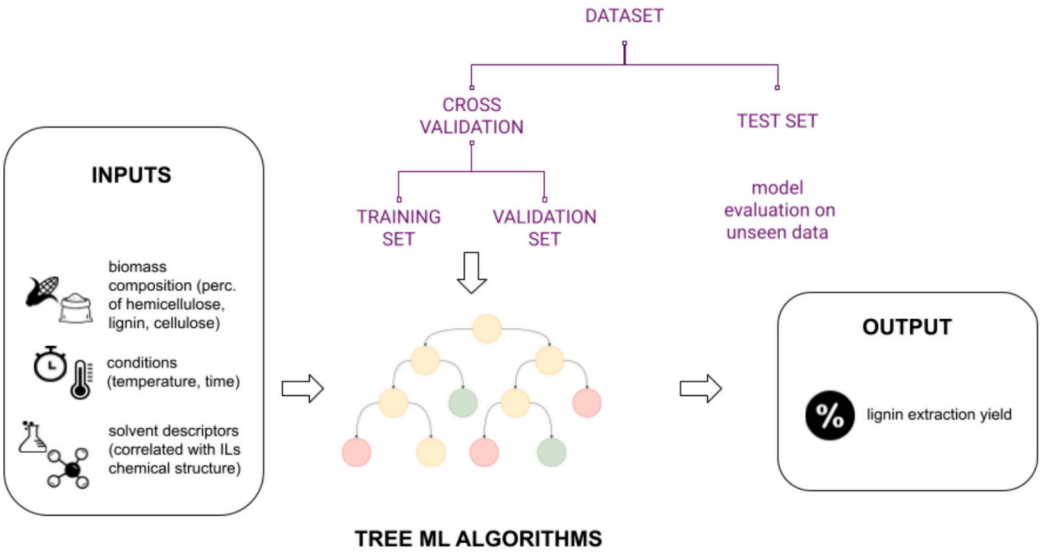


Fig. 2. Modeling schema incorporated in the study.

due to reduced lignin percentage and restricted lignin access attributed to hemicellulose structure, although comprehensive studies on other biopolymers' impact on lignin extraction efficiency are lacking.

2.2. Biomass pretreatment for lignin separation

The extraction of lignin from biomass can be achieved through physical, chemical, and physicochemical pretreatment or the isolation of native lignin (Yau et al., 2011). The selection of an appropriate biomass pretreatment method is of utmost importance as it dictates the structure of the extracted biopolymers (Hasanov et al., 2020). To efficiently isolate lignin from lignocellulosic biomass, pretreatment procedures should prioritize processes such as cellulose decrystallization, partial holocellulose depolymerization, and enhancement of the enzymatic digestibility of the initial biomass (Hasanov et al., 2020; Maurya et al., 2015). Lignin separation represents a costly and intricate procedure, often resulting in alterations to the native lignin structure, yielding what is known as technical lignins (Rinaldi et al., 2016).

Chemical extraction stands out as the most prevalent category of pretreatment methods, encompassing techniques like organosolv, sulfite, and soda processes (Galbe and Wallberg, 2019; Kumar and Sharma, 2017). Recently, ILs and DES are extensively studied as novel promising solvents for biomass pretreatment (Lopes, 2021) due to limitations of traditional processes. The mechanisms underlying these processes commonly involve the disruption of the bonds between lignin and hemicellulose within the lignin-hemicellulose complex, subsequently leading to lignin isolation through further treatment (Pinto et al., 2022). It is noteworthy that while most chemical methods primarily entail lignin modification via ester bond hydrolysis, a limited subset demonstrates the capability to effectively extract lignin from the cellulose matrix. Mentioned above methods come with a set of drawbacks, including the application of high pressures, susceptibility to equipment corrosion, elevated energy consumption, and difficulties in solvent recovery (Jönsson and Martín, 2016). Consequently, these processes are considered environmentally unfriendly, driving the quest for solvents with higher lignin solubility, such as ILs (Achinivu et al., 2014) and DESs (Fernandes et al., 2021; Magalhães et al., 2022).

In recent years, there has been a growing focus on ionic liquids as an alternative to conventional solvents due to their distinctive characteristics, which include a low melting point, non-flammability, low vapor pressure, and non-volatility (Claus et al., 2018). Since ILs are considered to be designer solvents, their chemical structure can be optimized so that biocompatible, biodegradable, and low toxic solvents can be obtained (Cvjetko Bubalo et al., 2013; Magina et al., 2021). It should be emphasized that supporting the optimization process through the use of modeling techniques using artificial intelligence tools should additionally reduce the energy and material costs of the process while minimizing the negative impact on the environment. These remarkable properties facilitate the recovery of ionic liquids in excess of 99 % in numerous processes, resulting in significant cost reductions (Yau et al., 2011).

The pretreatment of lignocellulosic biomass using ILs initiates with the dissolution of biomass at atmospheric pressure, typically at temperatures ranging from 90 to 160 °C, over durations spanning from a few minutes to 24 h (Samayam and Schall, 2010). Subsequently, the biomass is reprecipitated by introducing water and subjected to multiple washes before undergoing enzymatic hydrolysis and subsequent procedures. The mechanism underlying lignin extraction and regeneration in ionic liquids remains incompletely understood (Lindman et al., 2010). It is postulated that ionic liquids, characterized by diverse interaction types such as ionic, hydrogen bonding, dipolar, or π-π interactions, may compete with the hydrogen bonds present in lignocellulosic biomass (Brandt et al., 2013). This competitive interaction effectively disrupts the three-dimensional network of the lignin-hemicellulose complex, thereby enhancing the digestibility of the initial material (Besombes et al., 2004; Moulthrop et al., 2005). Moreover, research has shown that the IL cation has a minimal impact on reducing the molecular weight of lignin, while the IL anion plays a pivotal role (George et al., 2011). To date, research on the extraction of lignin from herbaceous lignocellulosic biomass has predominantly centered on the application of [Emim][OAc] and [Bmim][OAc] as ionic liquids (Halder et al., 2019). The effectiveness of lignin recovery has demonstrated variability contingent upon the specific biomass and process conditions. For instance, utilizing [Emim][OAc] during a 4-h treatment at 110 °C, the lignin recovery efficiency for rice husks amounted to 47 %, with an extension of the process to 8 h, resulting in nearly 100 % recovery (Lynam et al., 2012). In contrast, the lignin recovery efficiency for wheat straw, following a 2-h process at 120 °C, reached 87 % (da Costa Lopes et al., 2013), and switchgrass, subjected to a 3-h process at 160 °C, achieved a 69 % lignin recovery (Li et al., 2009). These findings emphasize the necessity to explore the combined impact of three main group of variables i.e. ionic liquid constituents, the biopolymer composition of biomass and the parameters of pretreatment processes, on the efficiency of lignin extraction. For this purpose, it was decided to develop and compare a number of ML tools as shown in Fig. 2.

3. Methods

3.1. Data collection and curation

The process of assembling the dataset was meticulously designed to capture pivotal information essential for this study. It encompassed defining the most extensive set among the distinguished lignocellulosic biomass groups, namely softwood, hardwood, and herbaceous biomass. Following an initial analysis of literature data, the decision was made to focus on the most abundant group - herbaceous biomass, which is expected to be the most amenable to pretreatment using ILs. A comprehensive database was prepared for individual lignin extraction experiments, concentrating on the characterization of the raw material composition—its polymer composition—and process parameters such as process efficiency, temperature, duration, ionic liquid concentration, as well as the type of cation and anion. Various studies reported differing values of lignin recovery efficiency for duplicated process parameters. In instances of incongruent information, source articles were scrutinized to select more credible data based on the evaluation of the experimental protocol used. The resulting dataset comprised 45 values for rice husks, 18 values for sugarcane bagasse, 15 values for wheat straw, 13 values for corn stover, 8 values for miscanthus, 6 values for switchgrass, and 4 values for corn cobs, totaling a diverse compilation of 109 data points. For comparison, an analogous dataset was compiled for woody biomass, consisting of 10 values for softwood biomass and 21 values for hardwood biomass. Detailed information regarding the dataset is available in Table S1-S3 in the supplementary electronic information (ESI).

Firstly, it was checked whether, in the dataset, there are clusters impacting its uniformity and, therefore capability for modeling. It was assessed visually using the FreeViz algorithm (Demsar et al., 2008). The FreeViz allows visualizations to be obtained with clear class separation, which is especially important in exploratory data analysis. This method allows for finding optimal two-dimensional projection to differentiate between samples in a dataset. The rationale behind this algorithm might be explained by a physical metaphor: there is a mathematical construct similar to the physical potential applied that allows for attraction and repulsion between data points. Similar datapoints attract while instances varying substantially repulse other points, analogous to physical forces between particles. As a result, optimal projection is found, and effective class separation is revealed in a scatterplot. The optimization objective, achieving effective class separation, is akin to identifying the configuration (projection) with minimal potential energy in this metaphorical framework.

The present study addresses a challenge marked by the constraint of a relatively small dataset comprising approximately 110 samples. Given this limitation, machine learning Gradient Boosting (GB) and Random Forest (RF) algorithms were prioritized over deep learning approaches. For this kind of modeling, it is adequate to represent chemical structures in the form of molecular descriptors (MDs) or molecular fingerprints (MFs). These are computational tools used in the field of cheminformatics and computational chemistry to characterize and represent chemical compounds in a format suitable for quantitative analysis and modeling. Both MDs and MFs play crucial roles in the development of predictive models, virtual screening, and quantitative structure-property relationship (QSPR) studies (Xue and Bajorath, 2000). Chemical representation, both MDs and MFs, is calculated for both cation and anion. MDs and MFs related to cations are presented with the prefix 'c', while those related to anions with the prefix 'a' followed by a name depicting the family of molecular representation used - "mordred" for MDs calculated in Mordred (Moriwaki et al., 2018) software and "fp" for Morgan MFs calculated using RDKIT (Landrum, 2013). Both MDs and MFs were calculated separately for cation and anion-forming ionic liquid. The entirety of the coding process was implemented in Python with help of commonly used machine learning packages namely scikit-learn (Pedregosa et al., 2012), SHAP (Lundberg et al., 2020a), LIME (Ribeiro, 2016), ELI5 (Korobov and Lopuhin, 2017), Orange Data Mining (Demšar et al., 2013).

MDs are numerical representations that encode structural, physicochemical, or topological information about a molecule. These descriptors provide a quantitative way to describe the structural features of a molecule in a numerical manner. MDs are diverse and can include properties like molecular weight, polarizability, bond angles, or constitutional indices. Some of the MDs are not easily interpretable, limiting the possibility of obtaining physical or chemical insights based on performed modeling. In that case, additional effort is paid during the modeling phase so that model interpretation can be feasible (Eichenlaub et al., 2023). MDs are particularly useful in QSPR studies, where the objective is to correlate the molecular structure of compounds with their physical or chemical properties. On the contrary, MFs are binary bit-string representations of molecular structures. These fingerprints encode the presence or absence of specific structural features or substructures within a molecule. Each bit in the fingerprint corresponds to the presence (1) or absence (0) of a particular moiety.

Since the number of structural features is exceptionally high, some feature space reduction is needed. Feature space reduction, in the context of ML methods, refers to the process of reducing the number of input variables or features in a dataset. The goal is to reduce the model's complexity, improve computational efficiency, and potentially enhance its ability to generalize. This reduction can be achieved through various techniques, including feature selection and dimensionality reduction. In this study, several steps were taken to remove redundant features. Firstly, constant features (that have the same value for every sample in the dataset) were excluded since they cannot differentiate between the points. Then, features that are correlated with other features with a Pearson correlation higher than 0.9 were excluded. Finally, the multicollinearity test was performed using the VIF criterium (Kim, 2019) to ensure that no information was provided to the model twice. Values of features were then normalized to the range 0–1. Finally, number of features for modeling was reduced using feature selection method, namely mutual information criterium.

Outliers in machine learning refer to data points that deviate significantly from the majority of the dataset and may introduce noise or distort the learning process of a model. The removal of outliers from the modeling training set is often desirable to enhance the robustness and generalization ability of the model, as outliers can disproportionately influence parameter estimation (Baran and Kloskowski, 2023). In order to remove data points that might be outlying from others, the Isolation Forest technique was incorporated, and hyperparameters for the method were selected using cross-validation.

3.2. Modeling

Two prominent algorithms, Gradient Boosting (GB) and Random Forest (RF), were chosen for the development of predictive models. These algorithms share a common foundation in that they build a collection, or ensemble, of decision trees to make predictions. Decision trees are individual models that recursively split the dataset into subsets based on features, forming a tree-like structure that facilitates decision-making. Their tree-based nature makes them particularly effective in scenarios with complex and non-linear relationships.

A dataset-splitting strategy was employed to ensure robust evaluation and generalizability of these models. This involves dividing the dataset into distinct subsets, typically comprising training, validation, and test sets. Each subset serves a unique purpose in the model development and evaluation process.

The training set is utilized to adjust the model's weights, exposing it to the patterns and relationships present in the data. The validation set, employing a 5-fold cross-validation strategy in this instance, serves as an intermediary check during the model development process. Cross-validation is a technique that partitions the dataset into multiple folds, training the model on different combinations of these folds and assessing its performance iteratively. This approach helps mitigate the variability

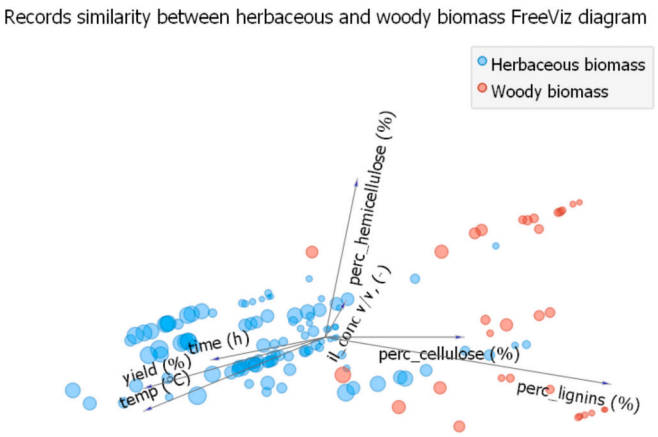


Fig. 3. FreeViz diagram showing records similarity in multidimensional space of process parameters.

in model evaluation that may arise from a single train-test split. Furthermore, a distinct test set, comprising 10 % of the original dataset, is reserved for final model evaluation. This test set, independent of both the training and validation sets, provides an unbiased assessment of the model's ability to generalize to new, unseen data. Intermediate steps in model building are assessed using the R^2 metric, while for the final model, additionally, the mean squared error (MSE) and mean average error (MAE) are provided. As a result, for train subset 72 % of the original dataset was allocated, while for validation and test accordingly 18 % and 10 % of the original dataset.

By reporting the mean metrics values computed over four distinct splits (i.e., repeated 4 times using different random seeds to randomly select test sets from the database), this study ensures a comprehensive and reliable evaluation of model performance.

3.3. Model interpretation

In order to enhance the interpretability of model predictions, various methods were employed. Specifically, SHAP (SHapley Additive exPlanations) (Lundberg et al., 2020b; Lundberg and Lee, 2017), LIME (Local Interpretable Model-agnostic Explanations) (Visani et al., 2021), and ELI5 (Saha, 2018) techniques were applied to elucidate the underlying rationale of the models' predictions.

The SHAP method was utilized to quantify the contribution of each feature to individual predictions, offering a comprehensive understanding of the impact of different variables on the model's outcomes. SHAP method utilizes game theory concepts to reveal how feature value contributes to both increasing or decreasing target value. The LIME (Local Interpretable Model-agnostic Explanations) technique was implemented to offer localized, easily understandable explanations for specific instances. LIME generates perturbations around a given prediction and observes the resulting changes in the model's output, providing insights into the local behavior of the model. In the ELI5 method, feature weights are computed by tracing decision paths within the ensemble of trees. In this process, each node within a tree yields an output score, and the attribution of a feature on the decision path is determined by the extent of score alteration from the parent node to its respective child node.

These interpretative tools contribute to a nuanced understanding of the relationships captured by the developed models, fostering transparency and insight into the predictive mechanisms employed.

4. Results and discussion

4.1. Preliminary analysis on selected biomass type

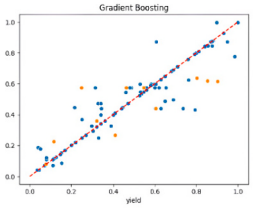
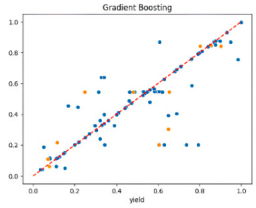
As previously described, there are two main groups of biomasses, namely woods (including softwoods and hardwoods) and herbaceous biomass. For the latter, more experimental results were reported in the literature on the topic. To obtain a dataset that is valuable for modeling, it must be ensured that all the records (data points) follow a uniform distribution in the multidimensional space created by features describing the records. To check that, parameters regarding the percentage of hemicellulose (perc_hemicellulose), cellulose (perc_cellulose), lignin (perc_lignin), as well as time, temperature, concentration of ionic liquid (il_conc), and yield are used. Relations in multidimensional space are then mapped to the image using the FreeViz algorithm. The result of such projection is shown in Fig. 3.

Fig. 3 clearly shows that there is a clear separation between herbaceous biomass (represented as blue points) and woods (red points). Even without taking into account molecular representation, it can be observed that experiments differ significantly on the type of biomass used. Therefore, it could not be stated that the two groups could be potentially incorporated in a single model. The distinct physicochemical properties of wood and herbaceous biomass emphasize the necessity of considering them as separate product categories in the investigation of lignin extraction in ionic liquids (see Fig. 3). Wood is characterized by higher lignin content and intricate structural arrangements, demanding tailored pretreatment strategies due to its elevated resistance to

Table 1
Reduction of the number of features related to biomass composition.

No.	Biomass feature(s) included in modeling	Overall number of parameters	Ranking of features according to GB feature importance	GB-MDs model R^2 / MAE metric on dataset subset		
				Train	Valid.	Test
M1	cellulose	8	perc_cellulose, a_mordred_MDEO-11, temp, c_mordred_MATS3pe, il_conc, time, c_mordred_AATSC3i, c_mordred_AXp-5d	0.98 / 0.01	0.63 / 0.12	0.75 / 0.09
M2	hemicellulose	8	perc_hemicellulose, a_mordred_MDEO-11, time, temp, il_conc, c_mordred_AATSC3i, c_mordred_MATS3pe, c_mordred_AXp-5d	0.98 / 0.01	0.69 / 0.11	0.77 / 0.09
M3	lignin	9	perc_lignins, c_mordred_AATSC3i, a_mordred_MDEO-11, il_conc, temp, time, c_mordred_nHBAcc, c_mordred_MATS3pe, c_mordred_AXp-5d	0.98 / 0.01	0.67 / 0.11	0.77 / 0.09
M4	cellulose + hemicellulose	9	perc_hemicellulose, perc_cellulose, c_mordred_AATSC3i, temp, c_mordred_MATS3pe, time, a_mordred_MDEO-11, il_conc, c_mordred_AXp-5d	0.98 / 0.01	0.70 / 0.11	0.81 / 0.08
M5	cellulose + lignin	9	perc_lignins, c_mordred_AATSC3i, a_mordred_MDEO-11, temp, il_conc, perc_cellulose, time, c_mordred_MATS3pe, c_mordred_AXp-5d	0.98 / 0.01	0.68 / 0.11	0.80 / 0.08
M6	hemicellulose + lignin	9	perc_lignins, c_mordred_AATSC3i, a_mordred_MDEO-11, temp, il_conc, perc_hemicellulose, time, c_mordred_MATS3pe, c_mordred_AXp-5d	0.98 / 0.01	0.69 / 0.11	0.83 / 0.09
M7	cellulose + hemicellulose + lignin	15	perc_lignins, c_mordred_AATSC3i, a_mordred_MDEO-11, perc_cellulose, perc_hemicellulose, temp, il_conc, a_mordred_TopoShapeIndex, a_mordred_EState_VSA9, c_mordred_MATS3pe, time, a_mordred_mZagreb1, a_mordred_SLogP, a_mordred_MDEC-33, c_mordred_AXp-5d	1.00 / 0.02	0.63 / 0.12	0.71 / 0.10

Table 2
Results of comparison between scenarios M2 and M6.

Model	M2	M6
Loss	0.032	0.027
Bias	0.022	0.016
Variance	0.010	0.011
Variance as the fraction of loss	31 %	41 %
predicted target vs. target (yield) plot		

extraction. Conversely, herbaceous biomasses exhibit lower lignin content and notably simpler structures, rendering them more amenable to pretreatment processes. Treating wood and herbaceous biomasses as distinct units allows for a targeted approach, thereby facilitating the development of refined strategies for lignin valorization in biomass conversion processes. Following that conclusion, molecular representation was calculated only for records obtained from experiments on herbaceous biomass samples.

4.2. Preliminary analysis on feature space reduction

In the dataset, as provided in Table S2, several features describing biomass composition were reported, namely percentage of cellulose, percentage of hemicellulose, and percentage of lignin. While the focus of research has predominantly been on three biopolymers, it's important to note that biomass also comprises ash (from <1 % in debarked wood to up to 25 % in herbaceous materials) (Yan et al., 2020), moisture (Haldar and Purkait, 2020), and other constituents, such as proteins (e.g. corn stover and poplar contain 2.5 % and 1.3 % protein respectively) (MacLellan et al., 2017). However, literature discussing their presence in lignin isolation tests is limited, primarily due to their significance in biomass utilization as a fuel in thermochemical conversion processes such as combustion or pyrolysis. Additionally, the chemical composition of plants varies depending on the species, maturity, and environmental conditions in which they were cultivated (Langsdorf et al., 2021). The statistical rationale for their exclusion, based on their high correlation with the utilized parameters, is provided in Table S5, ESI. However, often in modeling, it is beneficial to include as few features as possible. Reducing the number of features, resulting in lowering the dimensionality of a problem, often results in obtaining models that are less probable to overfit data (Teixeira et al., 2013). Therefore, all possible combinations of biomass composition–related features were tested for the model's performance to obtain the balance between the best metric value and the probability of the model being overfitted to data. Table 1 shows the results of a comparison between all these possible scenarios. Overall, the number of parameters was established by iteratively increasing the number of descriptors used for modeling until the performance on the validation set stopped rising with the increasing number of descriptors. Graphs justifying the selection are shown in Fig. S6.

Table 3
Blackbox models metrics for different molecular representations.

Model	R^2 Metrics of Blackbox Models Performance (on train / validation / test sets) in accordance with molecular representation		
	Molecular Descriptors (MDs)	Molecular Fingerprints (MFs)	Both Representations (MFs + MDs)
Gradient Boosting (GB)	0.98 / 0.69 / 0.77	0.98 / 0.58 / 0.77	1.00 / 0.66 / 0.76
Random Forest (RF)	0.88 / 0.63 / 0.71	0.87 / 0.58 / 0.75	0.88 / 0.62 / 0.73

As can be seen in Table 1, models built upon different scenarios of used biomass-related features differ in the case of both the overall number of parameters needed for modeling and metrics. The number of parameters in each scenario was established using cross-validation search. Probing was performed in range between 2 and 15 parameters. Therefore, metrics are presented for best performing model with optimal number of parameters under each scenario and differ between models. Among the 7 models, variants M2 and M6 seem to be the most promising. They both have comparatively good performance on the test set and comparable metric values on the validation set. However, it is hard to differentiate between the two. Scenario M6 resulted in a higher R^2 value on the test set. On the contrary, M2 might be more resistant to possible overfitting since it uses fewer parameters. Furthermore, most of the models including M2 and M6 rely on the same set of MDs, namely c_mordred_AATSC3i, c_mordred_MATS3pe, c_mordred_AXp-5d, a_mordred_MDEO-11. From the chemical point of view one might be concerned if lignin percentage being additional (when compared to M2) parameter describing biomass composition is in fact providing any useful information to the model. To further disclose which scenario should be used in further analysis, the loss decomposition method, as implemented in mlxtend (Domingos, 2002; Raschka, 2018) is used. Results are shown in Table 2.

Loss decomposition allows for evaluating part of MSE loss that corresponds to model's bias (error related to systematic incorrectness in predicted values) and variance (changes in predicted output with accordance to small change in input parameters). Even though M6 is characterized by a lower value of the expected loss, the variance was responsible for a higher percentage of the loss. Furthermore, variance increased while moving from scenario M2 to M6. This fact is related to M6 utilizing more features than M2. This fact might further support the hypothesis on relatively high model's variance. Since higher variance is related to the higher probability of model overfitting to data, M2 seemed to be a more reliable choice. Moreover, in M2 model by incorporating less biomass-related features, impact of structural features of ILs are expected to be of higher importance. Since it was one of the objectives of this study to provide validation of the models by comparison with literature findings, M2 was found to be better corresponding with this objective and selected for further studies.

Surprisingly, models incorporating a percentage of lignin as an input parameter did not necessarily perform better than models that are not given that feature in the input matrix. Even though this study concentrates on lignin isolation, the description of biomass does not need to explicitly include information on the percentage of lignin. This can be justified by poorer models' metrics but also by the fact that the percentage of hemicellulose has a relatively higher standard deviation in

Table 4
Descriptors used in the models.

Descriptor	Description / interpretation	R^2 for fit with a set of interpretable descriptors
Blackbox model		
c_mordred_AATSC3i	Autocorrelation of lag 3 weighted by ionization potential	1.00
c_mordred_MATS3pe	Moran coefficient of lag 3 weighted by pauling electronegativity	1.00
c_mordred_AXp-5d	5-Ordered averaged Chi path weighted by sigma electrons	1.00
a_mordred_MDEO-11	Molecular distance edge between primary O and primary O	0.95
Set of interpretable descriptors		
c_mordred_nAromAtom / a_mordred_nAromAtom	Aromatic atoms count	
c_mordred_nHBDon	Number of hydrogen bond donor	
a_mordred_nHBAcc	Number of hydrogen bond acceptors	
c_mordred_SlogP_VSA1 / a_mordred_SlogP_VSA1	Estimated partition coefficient in with accordance to atoms' contribution to the molecular surface area	
c_mordred_nRot / a_mordred_nRot	Rotatable bonds count	
c_mordred_TopoPSA / a_mordred_TopoPSA	Topological polar surface area	
c_mordred_MW / a_mordred_MW	Exact molecular weight	

the dataset, leading to more information being processed by the model.

4.3. Blackbox models

For modeling, it was decided to use one parameter describing biomass composition (percentage of hemicellulose i.e. scenario M2 as described earlier), three describing conditions under which the process occurs (temperature, time, and concentration of ionic liquid), as well as molecular descriptors (MDs) /fingerprints (MFs). Firstly, the impact of molecular representation was tested. In order to check the influence of that factor, three possible scenarios were tested, namely the usage of solely MDs, solely MFs, and a combination of both (MDs + MFs). The results of the comparison are shown in Table 3.

The superiority of MDs-based models over other scenarios could be explained by taking into account the statistical characteristics of the variables used in each of the scenarios. MDs provide detailed quantitative information about a molecule's structure, and they are often numerical values with high variance. MFs, on the contrary, represent a binary encoding of structural patterns, being classified as categorical features (only a set of discrete values are allowed). While MFs are efficient for capturing molecular similarity and facilitating quick comparisons, they may lack the nuanced information embedded in MDs. Combining MDs and MFs seeks to leverage the strengths of both approaches, but the efficacy of this hybrid strategy depends on the compatibility between the two types of features. In the studied example, it seems that all substantial information was already captured by MDs.

In order to check whether the model could provide some insights into the process being modeled, careful analysis of used descriptors is needed. Even though the MFs-based model seems to be interpretable, its predictive power is relatively poor (validation $R^2 < 0.6$), and the ability to obtain insights from this model is limited. Therefore, interpretability options are evaluated for the MDs-based model. A list of descriptors used for the models' building is shown in Table 4. Descriptions are based on official Mordred documentation (Moriwaki et al., 2018).

As can be seen, descriptors incorporated in the blackbox model were not easily interpretable for humans. Even though features like weighted autocorrelation or path in molecular graphs might describe a molecule well, it is challenging to obtain useful chemical intuition based on their value. However, it was tested whether they can be correlated with descriptors that have simple chemical interpretation. In order to establish that, several interpretable descriptors related to molecular features that are expected to be of importance in the process were selected. Multiple linear regression was incorporated, and the R^2 metric was used to assess linear correlation. For all descriptors, correlation with a set of interpretable alternatives was found to be exceptionally high. Therefore, it might be stated that almost all the information they were carrying would still be capture into the model built using only interpretable descriptors. Statistical analysis confirmed that the models constructed in this way were of similar quality in terms of metrics (Table S8), and at the same time provided better insight into the desired structural features of ionic liquids.

4.4. Interpretable models and interpretation of features importance

Proper understanding of model's rationale is crucial for assessing its predictive usefulness and correctness. Analyzing Shapley diagrams, one can notice certain regularities common to all built models (see Fig. 4): the efficiency of the lignin extraction process is determined to the greatest extent by the composition of the biomass expressed as the percentage of hemicellulose (perc_hemicellulose), and two parameters of the process, i.e., time and temperature. The fourth variable not related to the structure of ionic liquids is their concentration (il_conc), and this factor is still one of the most statistically significant: the fourth in both models obtained using RF, the sixth in the GB models, and the seventh in the model-based simultaneously on molecular descriptors and molecular fingerprints (GB).

However, one might want to assess whether the impact of molecular descriptors is significant for the model's prediction. To further address this issue, an additional sensitivity analysis was performed incorporating two feature importance scores: GB-based and permutation analysis (see Table 5). The GB method is related to how often feature is used to make decision to create another split in decision trees (Hastie, 2009), while permutation method involves tracking the reduction in the model's score when a single feature's values are changed at random (Breiman, 2001). It can be observed that the impact of features related to biomass composition and process parameters is of high importance, as it was observed based on Shapley values. On the contrary, it can be clearly seen that molecular descriptors' impact does not seem insignificant or minor.

For instance, descriptor a_mordred_SlogP_VSA1 is comparable in importance with time and temperature, according to GB score. Similarly, c_mordred_nHBAcc and concentration of IL have similar scores in both GB and permutation analysis. The most important features related to the solvent structure have 26 % and 21 % scores for the most important feature for GB and permutation methods, respectively.

Even though their scores are lower, they are vital for predicting the variance of a target variable. This fact can be further examined by comparing Pearson correlations with lignin yield values. None of the features are strongly correlated with the target property, implying that a combination of them is needed to explain the variance in the dataset.

The prevailing impact of variables not directly related to the chemical structure of ionic liquids will, to some extent, suppress the possibility of optimization of the chemical structure of ionic liquids. On the other hand, attempts made to omit some of the above-mentioned variables in constructing models led to a significant deterioration of the model's predictive abilities and statistical metrics. In the case of hemicellulose content, duration of the extraction process, and temperature, the obtained models clearly indicate that increasing the values of these parameters is beneficial for the efficiency of the lignin extraction process, which is consistent with literature reports (Brandt et al., 2013). In the case of the last two, the interpretation is rather simple as both factors are related to the kinetics of the extraction process, time directly, while

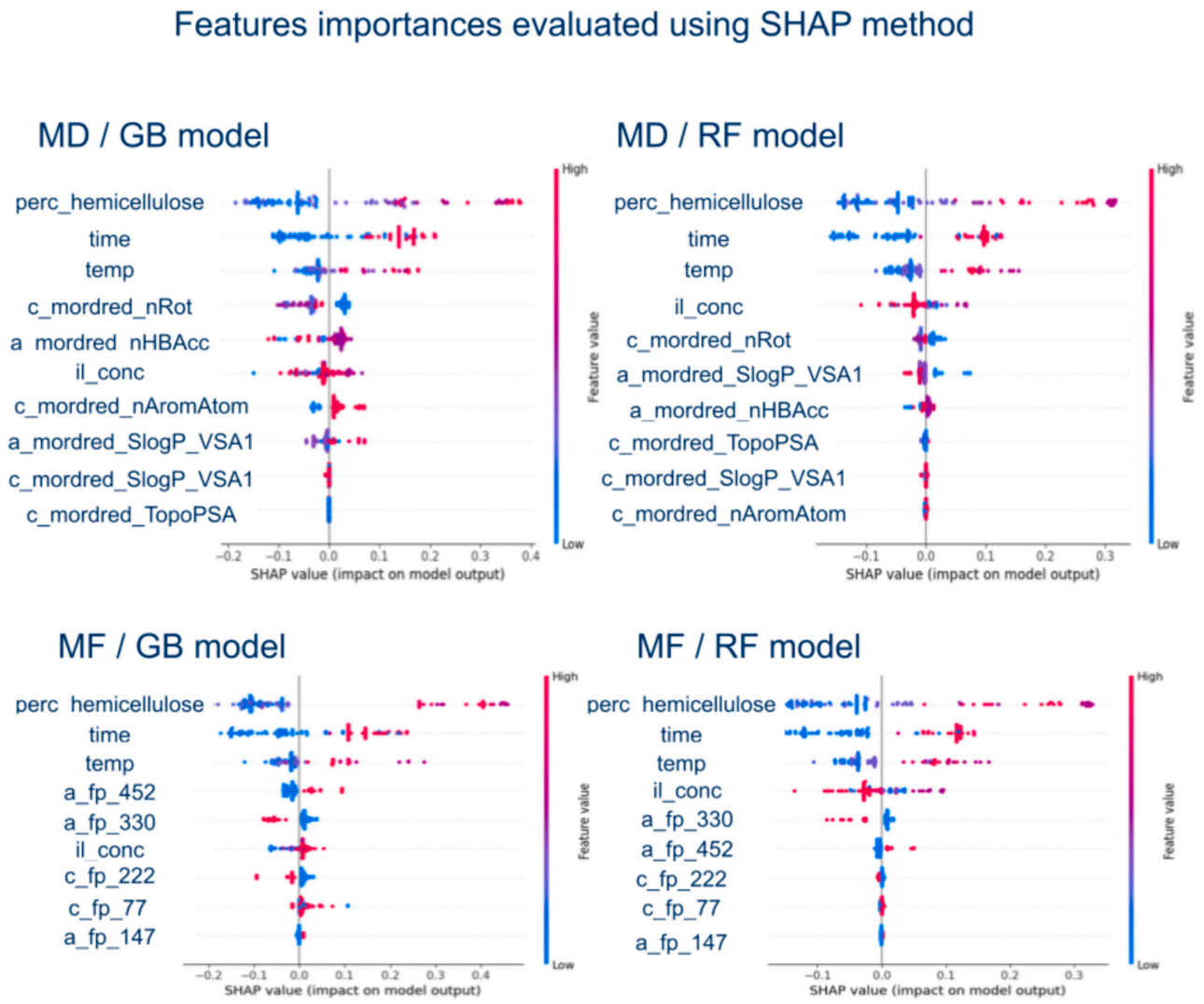


Fig. 4. SHAP diagrams for RF and GB models.

Table 5
Feature importances according to GB scores and permutation method for interpretable models.

Feature	GB feature importance score	Permutation feature importance score	Pearson correlation with lignin yield
perc_hemicellulose	0.425	0.71 ± 0.20	0.49
time	0.145	0.50 ± 0.13	0.11
a_mordred_SlogP_VSA1	0.112	0.089 ± 0.040	- 0.33
temp	0.101	0.37 ± 0.10	0.42
il_conc	0.086	0.27 ± 0.10	- 0.15
c_mordred_nRot	0.069	0.148 ± 0.050	0.11
a_mordred_nHBAcc	0.036	0.069 ± 0.031	0.12
c_mordred_SlogP_VSA1	0.018	0.0017 ± 0.0030	0.00
c_mordred_nAromAtom	0.009	0.0007 ± 0.0023	0.14

the temperature influences both by increasing the reaction rate constant and by lowering the viscosity of the reaction mixture (Amini et al., 2021). In the case of hemicellulose content, it is worth noting that its biomass content is not statistically significantly correlated with the lignin content (R^2 = 0.10). Consequently, the increased efficiency of lignin extraction is not likely to be the result of its lower content in the biomass and is caused by other factors. It can be assumed that the hemicellulose content affects the structure of the biomass as a whole, which translates into a more or less limited availability of lignin exposed to the solvent (ionic liquid) (Geng et al., 2019).

However, a comparison of the model's rationale regarding ILs structural features and previously reported relationships might serve as additional validation of the model. Despite the relatively low statistical significance of both molecular descriptors and molecular fingerprints in the obtained models, their interpretation is still possible. In models based on molecular descriptors, the set of descriptors was limited to six due to the small impact on the model's performance of descriptors with more distant positions. In both cases (GB and RF), the same set of descriptors was obtained, which can be ordered as follows: c_mordred_nRot > a_mordred_nHBAcc > a_mordred_SlogP_VSA1 > c_mordred_nAromAtom > c_mordred_SlogP_VSA1 = c_mordred_TopoPSA. When interpreting the meaning of descriptors on the potentially optimal structure of ionic liquids, it should be borne in mind that the models are built on a relatively small number of different cations (7) and anions (26). As a consequence, the indicated importance of individual descriptors for the obtained models may be significantly disturbed by both the number of representatives and the range of variability of the descriptor itself. Nevertheless, the observed relationships are consistent with current knowledge. The c_mordred_nRot descriptor, the value of which varies from 0 to 5 for the set of ionic liquids used, actually indicates the presence of long alkyl chains in the cation. Consequently, its contrary effect on the extraction efficiency (the higher

Table 6
Detailed metrics of the obtained models.

Model	R^2 metrics			MAE metrics			MSE metrics		
	Train	Valid.	Test	Train	Valid.	Test	Train	Valid.	Test
GB	0.98	0.71	0.73	0.01	0.11	0.10	0.01	0.02	0.02
RF	0.87	0.63	0.63	0.08	0.13	0.12	0.01	0.03	0.03

it is, the lower the efficiency) can be associated, in principle, with the higher viscosity of ionic liquids composed of cations with long alkyl substituent (Amini et al., 2021).

A clearly proportional influence of the descriptor value on the extraction efficiency can be seen for the second, most important descriptor a_mordred_nHBAcc, describing atoms that are hydrogen bond acceptors in the anion. The descriptor value varies in the range from 0 to 4 and depends on the number of nitrogen and oxygen atoms. The ability to participate in the formation of hydrogen bonds between hydroxyl groups present in the lignin structure, and the solvent is considered to be a key mechanism supporting the lignin extraction process, which is reflected in the obtained models (Ocreto et al., 2022). A significant impact on the lignin extraction efficiency can also be attributed to the number of aromatic atoms in the cation structure, which, according to literature reports, is the result of the π-π interactions between the cation and aromatic rings present in the lignin structure. The importance of this descriptor is higher in the case of the GB model, while in the RF model, it is the last of the six used. The probable reason is the small variability of this descriptor, which for the set of cations has a value of 5 for cations with an imidazolium ring and 0 for all others. Another descriptor that has a large impact on the efficiency of lignin extraction is a_mordred_SlogP_VSA1, the value of which depends on the van der Waals surface area of atoms influencing the value of the octanol/water partition coefficient, which, in effect, is a measure of the polarity of the molecule and the ability to accept a hydrogen bond (Labute, 2000). In both models, the descriptor has a large impact on the outcome variable, but in the RF model, a lower descriptor value enhances the extraction process, while in the GB model, this effect is less clear. The remaining descriptors, however, improve the model statistics, but their quantitative interpretation is difficult to perform. Models obtained using MFs as an independent variable have properties that indicate the effect of the presence of specific groups of atoms in cation and anion molecules. The problem with their use is the number of atoms constituting the fingerprint: too small a radius (1–2 atomic radii) does not allow for structural interpretation, while too large radii significantly reduce the diversification of the dataset. In the latter case, the resulting model is very dependent on the range of structural variability of the input data, which is visible in the examined case. The obtained statistical parameters of models based on fingerprints are significantly lower (approximately 10 %) than models using molecular descriptors. However, despite these limitations, the calculated models are consistent with the results obtained from GB models and with literature reports. Moreover, in some cases, they are complementary to GB models. For example, it can be clearly seen that ionic liquids with the Cl^- anion (a_fp_330) are unfavorable for the extraction efficiency of lignin. The detrimental impact of chloride anions present in ionic liquids on lignin extraction primarily arises from elevated extraction temperatures. These elevated temperatures can induce structural alterations in polysaccharides, consequently fostering the generation of undesired by-products. Additionally, the heightened viscosity of Cl- anion-containing IL's poses barriers to efficient mass and heat transport, thereby constraining the effectiveness of lignin extraction (Naz et al., 2021). Such a conclusion would be difficult to draw out based solely on molecular descriptor values. The remaining anion fingerprints present in the models indicate a positive impact of the presence of primary amine groups (a_fp_147), bisulfates, and anions with a carboxyl group (a_fp_295) as well as anions with oxygen bound to sulfur (overlapping fingerprints: a_fp_285, a_fp_452, a_fp_285), which

Table 7
Analysis of feature contributions to models' predictions for a few examples based on ELI5 methods.

GB model		RF model	
Feature	Impact	Feature	Impact
Example with low yield (sample 65)			
IL concentration	+0.005	[bias]	+0.5
c_mordred_nRot	−0.084	IL concentration	+0.001
time	−0.091	perc_hemicellulose	−0.127
perc_hemicellulose	−0.165	Time	−0.131
Example with medium yield (sample 1)			
time	+0.153	[bias]	+0.5
a_mordred_SlogP_VSA1	+0.093	time	+0.122
a_mordred_nHBAcc	+0.021	a_mordred_nHBAcc	−0.001
perc_hemicellulose	−0.115	perc_hemicellulose	−0.086
Example with high yield (sample 46)			
perc_hemicellulose	+0.321	[bias]	+0.5
IL concentration	+0.057	perc_hemicellulose	+0.239
a_mordred_nHBAcc	+0.031	IL concentration	+0.068
temperature	−0.019	temperature	−0.006

can generally be interpreted as the presence of atoms that accept hydrogen bonds. One can also notice a change in the proportion of variables describing the structure of the cation and anion, which, in the case of MF, shifted towards the anion fingerprints. In the case of cation fingerprints, there is a significant reduction in the diversification of the dataset. In the case of fingerprints c_fp_5, c_fp_15, and c_fp_70, they describe the presence of only one type of cation, while for c_fp_77, there is only one cation in the data set for which the fingerprint value is 0. Consequently, the interpretability of the results is very limited. However, the negative impact on the extraction process of the presence of hydroxyl groups in the cation described by the high value of the c_fp_222 descriptor can be indicated. It is believed that the hydroxyl groups in the cation can form hydrogen bonds with the acceptor sites of the anion, thus limiting its potential to interact with the hydroxyl groups present in the biomass sample (Hasanov et al., 2020).

4.5. Models' comparison and modeling limitations

The final evaluation of the model's performance with additional metrics is provided in Table 6 which shows that the models have a good predictive power, even on a test set. MEA metric is about 0.10 for validation and test set, implying that, on average, model error is about 10 percentage points (pp). This value might be compared with uncertainties of obtaining yield as estimated from a few works in the field that reported multiple yield values for the same process condition. In the study (Liu et al., 2018b) value range, out of 4 repetitions, was 5.8 pp., and the standard deviation was about 2.4 pp., while in the other study (Wu et al., 2011) value range was 3.2 pp. It can be clearly seen that the average error of the model is just slightly higher than the uncertainty of the measurements as reported by experimental works in the field.

Further comparison between the RF and GB models is possible via the interpretation of models' predictions for certain data points. Interpretation of models' predictions for specific samples from the dataset might be an interesting way of performing meta-analysis, trying to capture the relative importance of parameters influencing process yield.

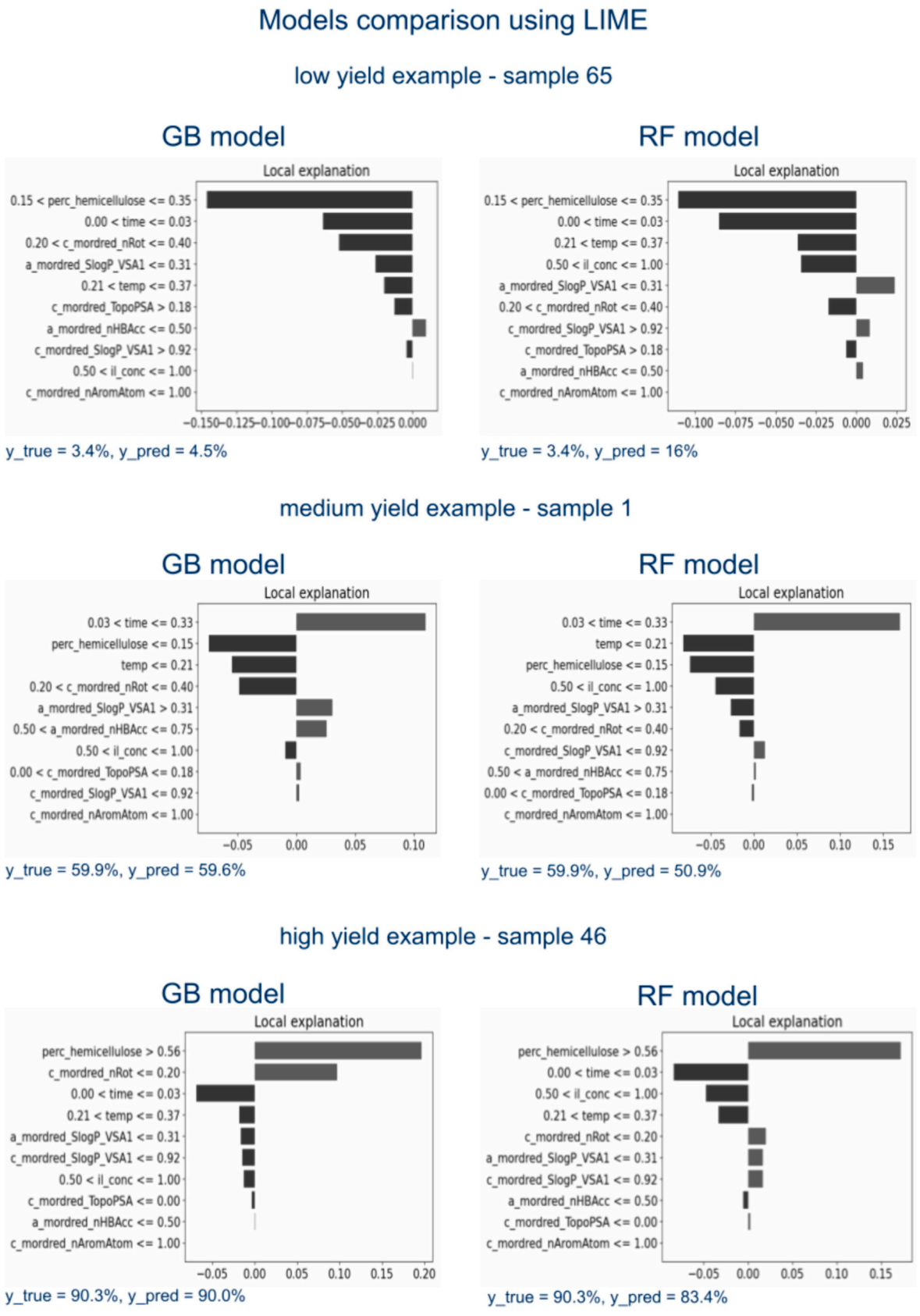


Fig. 5. LIME explanations for exemplary samples.

Detailed results of such analysis are shown in Table 7 and Fig. 5. Details on the input values for used datapoints are provided in Table 2, ESI.

Analysis of results shown in Table 7 regarding ELI5 analysis reveals significant differences between RF and GB models. For the GB model, MDs often have impacts similar to process parameters. This is the case for sample 65, where c_mordred_nRot and time have very similar importance and impact (both value and sign of calculated impact are comparable), or for sample 46, where the impact of a_mordred_nHBAcc leverages the impact of temperature. That is often not the case for RF models, which seem to be dominated solely by process parameters. Differences between the two models are also clear when comparing the sample with a medium yield value like sample 1. In that case, bias is the most important factor impacting predicted yield. It is a clear drawback from the perspective of model interpretation. Even though the predictive power of the RF model seems to be satisfactory, it has a higher bias. As a result, model interpretation is disturbed by the bias.

It should be mentioned that different model interpretation methods operate on significantly different approaches and, therefore, might lead to slightly different conclusions. In order to obtain as precise information as possible, one might want to use a few methods and compare the insights. Somehow, similar conclusions result from LIME analysis, as shown in Fig. 5. For all of the studied examples from the dataset, the impact of a few first features on predictions was significantly larger in the RF model than in the GB model. This is yet another consequence of the RF model having a higher bias than the GB model. However, details on the importance of specific descriptors differ from what was observed in the ELI5 method. For example, for sample 46, LIME assigns very high importance to descriptor c_mordred_nRot, which ELI5 did not find to be influential for that sample. This implies the impact of a low number of rotatable bonds, probably related to the rigid aromatic structure of the cation, and indirectly suggests the high relevance of π- π interactions with the aromatic imidazolium ring. The opposite impact occurs for samples 1 and 65, where lack of aromaticity in cation structures results in a lower target value, as shown by LIME graphs for the GB model for the two samples. The impact of that structural feature seems to be similar or even more relevant than the impact of temperature.

Finally, modeling limitations should be discussed to set assumptions that have to be met in order to use models properly. Modeling in the realm of machine learning is inherently constrained by several limitations that warrant careful consideration. One prominent concern is the challenge of overfitting, where a model becomes excessively tailored to the training data, capturing noise rather than genuine patterns. This phenomenon compromises the model's ability to generalize to new, unseen data, undermining its predictive reliability. Striking a balance between model complexity and generalizability is crucial to mitigate the risks associated with overfitting. In this study, this concern was addressed by using a test set and providing detailed explanations of the merits behind models' reasoning.

Another notable limitation stems from the nature of the training dataset. The quality and representativeness of the data used to train a model significantly influence its performance. Since the training dataset does not encompass the full spectrum of all possible scenarios, the model may exhibit biases and struggle when confronted with novel situations that are vastly different from what was reported up to date. Due to the relatively small size of the dataset, it cannot be assured that models' are not somehow biased by the design of experiments available in the literature. Additionally, the applicability domain of the models, or the range of conditions under which it reliably operates, is determined by the examples present in the training set. Large deviations from these conditions may lead to suboptimal or inaccurate predictions. Due to this limitation, wood biomasses were excluded from this study at the early stage of the models' development. Consequently, the model should be applied only to herbaceous biomass samples.

The inherently probabilistic nature of many real-world phenomena further complicates modeling efforts, as it might benefit from acknowledging uncertainties associated with predictions, which was not possible in this study due to a lack of enough experimental data on the topic.

5. Summary

In this study, machine learning models for predicting the yield of lignin extraction in ionic liquids were presented. The basic assumption of work is based on distinguishing the modeling process according to its purpose. There is a slight distinction between modeling focused on interpreting phenomena, and modeling focused on predicting phenomena. In order to comprehend the relative relevance of features and the lignin extraction process, an additional model involving the user-aided selection of molecular descriptors was required. Interpretative models are indispensable for elucidating the fundamental mechanisms and interactions with intricate biomass matrices in lignin extraction. Combining these modeling techniques provides a synergistic potential to find new solvents and methods that maximize lignin quality and recovery. Molecular descriptors like the number of rotatable bonds, number of aromatic atoms, number of hydrogen bond donors and acceptors, and weighted estimated partition coefficient are identified as factors describing ILs molecule that contribute to the model predicting lignin extraction yield. Based on the anticipated significance of various variables, proposed hypotheses can assist in prioritizing experimental studies and direct future research efforts. It was shown that modeling lignin yield for wood and herbaceous biomass should be performed separately due to significant differences between the two categories. For the latter, it was demonstrated that ML models achieving R^2 metric for separate test set, disclosed from models' training, of over 0.75 can be built. Proposed models incorporated hemicellulose percentage, time, temperature, concentration of IL, and molecular descriptors as input features. Insights into models' rationale were, however, primary. Interpretable models were proposed to be additionally validated by comparison with literature findings in the field.

Code availability

Application serving as a tool for presented model and the underlying code for this study is available and can be accessed via the link to the GitHub repository at: https://github.com/kbarn411/Lignin-ILs-Modeling.

CRediT authorship contribution statement

Karol Baran: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft. **Beata Barczak:** Conceptualization, Data curation, Formal analysis, Investigation, Writing – original draft. **Adam Kloskowski:** Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my data and codes

Acknowledgments

The authors would like to acknowledge the Centre of Informatics Tricity Academic Supercomputer & Network for providing the computational resources used for carrying out calculations in this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2024.173234.

References

Achinivu, E., Howard, R., Li, G., Gracz, H., Henderson, W., 2014. Lignin extraction from biomass with protic ionic liquids. Green Chem. 16 https://doi.org/10.1039/C3GC42306A.

Achyuthan, K.E., Achyuthan, A.M., Adams, P.D., Dirk, S.M., Harper, J.C., Simmons, B.A., Singh, A.K., 2010. Supramolecular self-assembled Chaos: polyphenolic Lignin's barrier to cost-effective lignocellulosic biofuels. Molecules 15, 8641–8688. https://doi.org/10.3390/molecules15118641.

Agbor, V.B., Cicek, N., Sparling, R., Berlin, A., Levin, D.B., 2011. Biomass pretreatment: fundamentals toward application. Biotechnol. Adv. 29, 675–685. https://doi.org/10.1016/j.biotechadv.2011.05.005.

Amini, E., Valls, C., Roncero, M.B., 2021. Ionic liquid-assisted bioconversion of lignocellulosic biomass for the development of value-added products. J. Clean. Prod. 326, 129275 https://doi.org/10.1016/j.jclepro.2021.129275.

Barakat, A., de Vries, H., Rouau, X., 2013. Dry fractionation process as an important step in current and future lignocellulose biorefineries: a review. Bioresour. Technol. 134 https://doi.org/10.1016/j.biortech.2013.01.169.

Baran, K., Kloskowski, A., 2023. Graph neural networks and structural information on ionic liquids: cheminformatics study on molecular physicochemical property prediction. J. Phys. Chem. B. 127, 10542–10555. https://doi.org/10.1021/acs.jpcb.3c05521.

Barczak, B., Łuczak, J., Kazimierski, P., Klugmann-Radziemska, E., Lopez, G., Januszewicz, K., 2023. Exploring synergistic effects in physical-chemical activation of Acorus calamus for water treatment solutions. J. Environ. Manag. 347, 119000 https://doi.org/10.1016/j.jenvman.2023.119000.

Besombes, S., Utille, J.-P., Mazeau, K., Robert, D., Taravel, F., 2004. Conformational study of a guaiacylβ-O-4 lignin model compound by NMR. Examination of intramolecular hydrogen bonding interactions and conformational flexibility in solution. Magn. Reson. Chem. 42, 337–347. https://doi.org/10.1002/mrc.1317.

Brandt, A., Ray, M., Quỳnh, T., Leak, D., Murphy, R.J., Welton, T., 2011. Ionic liquid pretreatment of lignocellulosic biomass with ionic liquid–water mixtures. Green Chem. 13, 2489–2499. https://doi.org/10.1039/C1GC15374A.

Brandt, A., Gräsvik, J., Hallett, J., Welton, T., 2013. Deconstruction of lignocellulosic biomass with ionic liquids. Green Chem. 15, 550–583. https://doi.org/10.1039/C2GC36364J.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Claus, J., Sommer, F.O., Kragl, U., 2018. Ionic liquids in biotechnology and beyond. Solid State Ionics 314, 119–128. https://doi.org/10.1016/j.ssi.2017.11.012.

Cvjetko Bubalo, M., Radošević, K., Radojcic Redovnikovic, I., Halambek, J., Srček, V., 2013. A brief overview of the potential environmental hazards of ionic liquids. Ecotoxicol. Environ. Saf. 99 https://doi.org/10.1016/j.ecoenv.2013.10.019.

da Costa Lopes, A.M., João, K.G., Rubik, D.F., Bogel-Łukasik, E., Duarte, L.C., Andreaus, J., Bogel-Łukasik, R., 2013. Pre-treatment of lignocellulosic biomass using ionic liquids: wheat straw fractionation. Bioresour. Technol. 142, 198–208. https://doi.org/10.1016/j.biortech.2013.05.032.

Demsar, J., Leban, G., Zupan, B., 2008. FreeViz—an intelligent multivariate visualization approach to explorative analysis of biomedical data. J. Biomed. Inform. 40, 661–671. https://doi.org/10.1016/j.jbi.2007.03.010.

Demšar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Možina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Pretnar Žagar, A., Žbontar, J., Žitnik, M., Zupan, B., 2013. Orange: data mining toolbox in Python. J. Mach. Learn. Res. 14, 2349–2353.

Deng, W., Feng, Y., Fu, J., Guo, H., Guo, Y., Han, B., Jiang, Z., Kong, L., Li, C., Liu, H., Nguyen, P., Ren, P., Wang, F., Wang, S., Wang, Yanqin, Wang, Ye, Wong, S., Yan, K., Yan, N., Zhou, H., 2022. Catalytic conversion of lignocellulosic biomass into chemicals and fuels. Green Energy & Environ. 8 https://doi.org/10.1016/j.gee.2022.07.003.

Domingos, P., 2002. A Unified Bias-Variance Decomposition for Zero-One and Squared Loss.

Eichenlaub, J., Baran, K., Smiechowski, M., Kloskowski, A., 2023. Free volume in physical absorption of carbon dioxide in ionic liquids: molecular dynamics supported modeling. Sep. Purif. Technol. 313, 123464 https://doi.org/10.1016/j.seppur.2023.123464.

Fernandes, C., Melro, E., Magalhães, S., Alves, L., Craveiro, R., Filipe, A., Valente, A.J.M., Martins, G., Antunes, F.E., Romano, A., Medronho, B., 2021. New deep eutectic solvent assisted extraction of highly pure lignin from maritime pine sawdust (Pinus pinaster Ait.). Int. J. Biol. Macromol. 177, 294–305. https://doi.org/10.1016/j.ijbiomac.2021.02.088.

Flieger, J., Flieger, M., 2020. Ionic liquids toxicity—benefits and threats. Int. J. Mol. Sci. 21, 6267. https://doi.org/10.3390/ijms21176267.

G. Calvo-Flores, F., Dobado, J., Isac-García, J., Martin-Martinez, F., 2015. Lignin and Lignans as renewable raw materials: chemistry. Technology and Applications. https://doi.org/10.1002/9781118682784.

Galbe, M., Wallberg, O., 2019. Pretreatment for biorefineries: a review of common methods for efficient utilisation of lignocellulosic materials. Biotechnol. Biofuels 12, 294. https://doi.org/10.1186/s13068-019-1634-1.

Geng, W., Narron, R., Jiang, X., Pawlak, J., Chang, H.-M., Park, S., Jameel, H., Venditti, R., 2019. The influence of lignin content and structure on hemicellulose alkaline extraction for non-wood and hardwood lignocellulosic biomass. Cellulose 26. https://doi.org/10.1007/s10570-019-02261-y.

George, A., Tran, K., Morgan, T., Benke, P., Berrueco, C., Lorente, E., Wu, B., Keasling, J., Simmons, B., Holmes, B., 2011. The effect of ionic liquidcation and anion combinations on the macromole- cular structure of lignins. Green Chem. 13 https://doi.org/10.1039/C1GC15543A.

Gillet, S., Aguedo, M., Petitjean, L., Morais, A.R.C., da Costa Lopes, A.M., Łukasik, R.M., Anastas, P.T., 2017. Lignin transformations for high value applications: towards targeted modifications using green chemistry. Green Chem. 19, 4200–4233.

Haldar, D., Purkait, M., 2020. A review on the environment-friendly emerging techniques for pretreat- ment of lignocellulosic biomass: mechanistic insight and advancements. Chemosphere 264, 128523. https://doi.org/10.1016/j.chemosphere.2020.128523.

Halder, P., Kundu, S., Patel, S., Setiawan, A., Atkin, R., Parthasarthy, R., Paz-Ferreiro, J., Surapaneni, A., Shah, K., 2019. Progress on the pre-treatment of lignocellulosic biomass employing ionic liquids. Renew. Sust. Energ. Rev. 105, 268–292. https://doi.org/10.1016/j.rser.2019.01.052.

Hasanov, I., Raud, M., Kikas, T., 2020. The role of ionic liquids in the lignin separation from lignocellulosic biomass. Energies (Basel) 13. https://doi.org/10.3390/en13184864.

Hastie, T., 2009. The elements of statistical learning: data mining, inference, and prediction, the elements of statistical. Learning. https://doi.org/10.1007/978-0-387-84858-7.

Isikgor, F.H., Becer, C.R., 2015. Lignocellulosic biomass: a sustainable platform for the production of bio-based chemicals and polymers. Polym. Chem. 6, 4497–4559. https://doi.org/10.1039/C5PY00263J.

Jönsson, L.J., Martín, C., 2016. Pretreatment of lignocellulose: formation of inhibitory by-products and strategies for minimizing their effects. Bioresour. Technol. 199, 103–112. https://doi.org/10.1016/j.biortech.2015.10.009.

Kamm, B., Kamm, M., Gruber, P., Kromus, S., 2008. Biorefinery systems–an overview. Biorefineries-Industrial Processes and Products: Status Quo and Future Directions. 1–40. https://doi.org/10.1002/9783527619849.ch1.

Kazimierski, P., Januszewicz, K., Godlewski, W., Fijuk, A., Suchocki, T., Chaja, P., Barczak, B., Kardaś, D., 2022. The course and the effects of agricultural biomass pyrolysis in the production of high-calorific biochar. Materials 15. https://doi.org/10.3390/ma15031038.

Kim, J., 2019. Multicollinearity and misleading statistical results. Korean J. Anesthesiol. 72 https://doi.org/10.4097/kja.19087.

Korobov, M., Lopuhin, K., 2017. Overview ELI5 [WWW Document].

Kumar, A.K., Sharma, S., 2017. Recent updates on different methods of pretreatment of lignocellulosic feedstocks: a review. Bioresour. Bioprocess. 4, 7. https://doi.org/10.1186/s40643-017-0137-9.

Labute, P., 2000. A widely applicable set of descriptors. J. Mol. Graph. Model. 18, 464–477. https://doi.org/10.1016/S1093-3263(00)00068-1.

Landrum, G., 2013. RDKit documentation. Release 1, 1–79.

Langsdorf, A., Volkmar, M., Holtmann, D., Ulber, R., 2021. Material utilization of green waste: a review on potential valorization methods. Bioresour. Bioprocess. 8, 19. https://doi.org/10.1186/s40643-021-00367-5.

Li, C., Knierim, B., Manisseri, C., Arora, R., Scheller, H., Auer, M., Vogel, K., Simmons, B., Singh, S., 2009. Comparison of dilute acid and ionic liquid pretreatment of switchgrass: biomass recal- citrance, delignification and enzymatic saccharification. Bioresour. Technol. 101, 4900–4906. https://doi.org/10.1016/j.biortech.2009.10.066.

Li, H.-Y., Chen, X., Wang, C.-Z., Sun, S.-N., Sun, R.-C., 2016. Evaluation of the two-step treatment with ionic liquids and alkali for enhancing enzymatic hydrolysis of Eucalyptus: chemical and anatomical changes. Biotechnol. Biofuels 9, 166. https://doi.org/10.1186/s13068-016-0578-y.

Li, T., Takkellapati, S., 2018. The current and emerging sources of technical lignins and their applications: sources of technical Lignins. Biofuels Bioprod. Biorefin. 0 https://doi.org/10.1002/bbb.1913.

Lindman, B., Karlström, G., Stigsson, L., 2010. On the mechanism of dissolution of cellulose. J. Mol. Liq. 156, 76–81. https://doi.org/10.1016/j.molliq.2010.04.016.

Liu, Z., Li, L., Liu, C., Xu, A., 2018a. Pretreatment of corn straw using the alkaline solution of ionic liquids. Bioresour. Technol. 260, 417–420. https://doi.org/10.1016/j.biortech.2018.03.117.

Liu, Z., Li, L., Liu, C., Xu, A., 2018b. Pretreatment of corn straw using the alkaline solution of ionic liquids. Bioresour. Technol. 260, 417–420. https://doi.org/10.1016/j.biortech.2018.03.117.

Lopes, A., 2021. Biomass delignification with green solvents towards lignin valorisation: ionic liquids vs deep eutectic solvents. Acta Innov. 64–78 https://doi.org/10.32933/ActaInnovations.40.5.

Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions.

Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020a. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2 https://doi.org/10.1038/s42256-019-0138-9.

Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020b. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2 https://doi.org/10.1038/s42256-019-0138-9.

Lynam, J., Reza, M.T., Vasquez, V.R., Coronella, C., 2012. Pretreatment of rice hulls by ionic liquid dissolution. Bioresour. Technol. 114, 629–636. https://doi.org/10.1016/j.biortech.2012.03.004.

MacLellan, J., Chen, R., Yue, Z., Kraemer, R., Liu, Y., Liao, W., 2017. Effects of protein and lignin on cellulose and xylan anaylses of lignocellulosic biomass. J. Integr. Agric. 16, 1268–1275. https://doi.org/10.1016/S2095-3119(15)61142-X.

Magalhães, S., Moreira, A., Almeida, R., Cruz, P.F., Alves, L., Costa, C., Mendes, C., Medronho, B., Romano, A., Carvalho, M. da G., Gamelas, J.A.F., Rasteiro, M. da G., 2022. Acacia wood fractionation using deep eutectic solvents: extraction, recovery, and characterization of the different fractions. ACS Omega 7, 26005–26014. https://doi.org/10.1021/acsomega.1c07380.

Magina, S., Barros-Timmons, A., Ventura, S., Evtuguin, D., 2021. Evaluating the hazardous impact of ionic liquids – challenges and opportunities. J. Hazard. Mater. 412, 125215 https://doi.org/10.1016/j.jhazmat.2021.125215.

Martínez, Á.T., Ruiz-Dueñas, F.J., Martínez, M.J., del Río, J.C., Gutiérrez, A., 2009. Enzymatic delignification of plant cell wall: from nature to mill. Curr. Opin. Biotechnol. 20, 348–357. https://doi.org/10.1016/j.copbio.2009.05.002.

Maurya, D.P., Singla, A., Negi, S., 2015. An overview of key pretreatment processes for biological conversion of lignocellulosic biomass to bioethanol. 3 Biotech 5, 597–609. https://doi.org/10.1007/s13205-015-0279-4.

Moriwaki, H., Tian, Y., Kawashita, N., Takagi, T., 2018. Mordred: A molecular descriptor calculator. J. Chem. 10 https://doi.org/10.1186/s13321-018-0258-y.

Moulthrop, J., Swatloski, R., Moyna, G., Rogers, R., 2005. High-resolution 13C NMR studies of cellulose and cellulose oligomers in ionic liquid solutions. Chem. Commun. (Camb.) 12, 1557–1559. https://doi.org/10.1039/b417745b.

Naz, S., Uroos, M., Muhammad, N., 2021. Effect of molecular structure of cation and anions of ionic liquids and co-solvents on selectivity of 5-hydroxymethylfurfural from sugars, cellulose and real biomass. J. Mol. Liq. 334, 116523 https://doi.org/10.1016/j.molliq.2021.116523.

Ocreto, J., Chen, W.-H., Rollon, A., Ong, H.C., Petrissans, A., Pétrissans, M., De Luna, M. D., 2022. Ionic liquid dissolution utilized for biomass conversion into biofuels, value-added chemicals and advanced materials: a comprehensive review. Chem. Eng. J. 445, 136733 https://doi.org/10.1016/j.cej.2022.136733.

Paraschiv, L.S., Paraschiv, S., 2023. Contribution of renewable energy (hydro, wind, solar and biomass) to decarbonization and transformation of the electricity generation sector for sustainable develop- ment. Energy Rep. 9, 535–544. https://doi.org/10.1016/j.egyr.2023.07.024.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., Louppe, G., 2012. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12.

Pin, T.C., Rabelo, S.C., Pu, Y., Ragauskas, A.J., Costa, A.C., 2021. Effect of Protic ionic liquids in sugar cane bagasse pretreatment for lignin valorization and ethanol production. ACS Sustain. Chem. Eng. 9, 16965–16976. https://doi.org/10.1021/acssuschemeng.1c05353.

Pinto, E., Aggrey, W.N., Boakye, P., Amenuvor, G., Sokama-Neuyam, Y.A., Fokuo, M.K., Karimaie, H., Sarkodie, K., Adenutsi, C.D., Erzuah, S., Rockson, M.A.D., 2022. Cellulose processing from biomass and its derivatization into carboxymethylcellulose: a review. Sci. Afr. 15, e01078 https://doi.org/10.1016/j.sciaf.2021.e01078.

Raschka, S., 2018. MLxtend: providing machine learning and data science utilities and extensions to Python's scientific computing stack. J. Open Source Softw. 3, 638. https://doi.org/10.21105/joss.00638.

Ribeiro, M.T., 2016. Local Interpretable Model-Agnostic Explanations (Lime) [WWW Document].

Rieland, J.M., Love, B.J., 2020. Ionic liquids: a milestone on the pathway to greener recycling of cellulose from biomass. Resour. Conserv. Recycl. 155, 104678 https://doi.org/10.1016/j.resconrec.2019.104678.

Rinaldi, R., Jastrzebski, R., Clough, M., Ralph, J., Kennema, M., Bruijnincx, P., Weckhuysen, B., 2016. Paving the way for lignin valorisation: recent advances in bioengineering, biorefining and catalysis. Angew. Chem. Int. Ed. Engl. 55 https://doi.org/10.1002/anie.201510351.

Rubin, E., 2008. Genomics of cellulosic biofuels. Nature 454, 841–845. https://doi.org/10.1038/nature07190.

Saha, S., 2018. A comprehensive guide to convolutional neural networks—the ELI5 way. Towards data science 15.

Samayam, I.P., Schall, C.A., 2010. Saccharification of ionic liquid pretreated biomass with commercial enzyme mixtures. Bioresour. Technol. 101, 3561–3566. https://doi.org/10.1016/j.biortech.2009.12.066.

Sharma, V., Tsai, M.-L., Nargotra, P., Chen, C.-W., Sun, P.-P., Singhania, R.R., Patel, A.K., Dong, C.-D., 2023. Journey of lignin from a roadblock to bridge for lignocellulose biorefineries: a comprehen- sive review. Sci. Total Environ. 861, 160560 https://doi.org/10.1016/j.scitotenv.2022.160560.

Singh, S.K., 2022. Ionic liquids and lignin interaction: an overview. Bioresour. Technol. Rep. 17, 100958 https://doi.org/10.1016/j.biteb.2022.100958.

Stolarski, M., Śnieg, M., Krzyżaniak, M., Tworkowski, J., Szczukowski, S., Graban, Ł., Lajszner, W., 2018. Short rotation coppices, grasses and other herbaceous crops: biomass properties versus 26 genotypes and harvest time. Ind. Crop. Prod. 119, 22–32. https://doi.org/10.1016/j.indcrop.2018.03.064.

Tan, S., MacFarlane, D., Upfal, J., Edye, L., Doherty, W., Patti, A., Pringle, J., Scott, J., 2009. Extraction of lignin from lignocellulose at atmospheric pressure using alkylbenzenesulfonate ionic liquid. Green Chem. 11 https://doi.org/10.1039/b815310h.

Teixeira, A., Leal, J., Falcão, A., 2013. Random forests for feature selection in QSPR models - an application for predicting standard enthalpy of formation of hydrocarbons. J. Chem. 5, 9. https://doi.org/10.1186/1758-2946-5-9.

Visani, G., Bagli, E., Chesani, F., Poluzzi, A., Capuzzo, D., 2021. Statistical stability indices for LIME: obtaining reliable explanations for machine learning models. J. Oper. Res. Soc. 73, 1–11. https://doi.org/10.1080/01605682.2020.1865846.

Weerachanchai, P., Lee, J.-M., 2013. Effect of organic solvent in ionic liquid on biomass pretreatment. ACS Sustain. Chem. Eng. 1, 894–902. https://doi.org/10.1021/sc300147f.

Wu, H., Pale, M., Miao, J., Doherty, T., Linhardt, R., Dordick, J., 2011. Facile pretreatment of Ligno- cellulosic biomass at high loadings in room temperature ionic liquids. Biotechnol. Bioeng. 108, 2865–2875. https://doi.org/10.1002/bit.23266.

Xue, L., Bajorath, J., 2000. Molecular descriptors in Chemoinformatics, computational combinatorial chemistry, and virtual screening. Comb. Chem. High Throughput Screen. 3, 363–372. https://doi.org/10.2174/1386207003331454.

Yan, J., Oyedeji, O., Leal, J.H., Donohoe, B.S., Semelsberger, T.A., Li, C., Hoover, A.N., Webb, E., Bose, E.A., Zeng, Y., Williams, C.L., Schaller, K.D., Sun, N., Ray, A.E., Tanjore, D., 2020. Characterizing variability in lignocellulosic biomass: a review. ACS Sustain. Chem. Eng. 8, 8059–8085. https://doi.org/10.1021/acssuschemeng.9b06263.

Yau, E., Badal, K., Collier, J., Ramachandran, K., Ramakrishnan, S., 2011. Chemical and Physico- chemical pretreatment of lignocellulosic biomass: a review. Enzyme Res. 2011, 787532 https://doi.org/10.4061/2011/787532.

Zhu, X., Peng, C., Chen, H., Chen, Q., Zhao, Z., Zheng, Q., Xie, H., 2018. Opportunities of ionic liquids for lignin utilization from biorefinery. ChemistrySelect 3, 7945–7962. https://doi.org/10.1002/slct.201801393.