Chemical Engineering Journal 495 (2024) 153524

Contents lists available at ScienceDirect

Chemical Engineering Journal

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

Machine learning-supported solvent design for lignin-first biorefineries and lignin upgrading

Laura König-Mattern [a], Edgar I. Sanchez Medina [b], Anastasia O. Komarova [c], Steffen Linke [b], Liisa Rihko-Struckmann [a], Jeremy S. Luterbacher [c,*], Kai Sundmacher [a,b,*]

[a] Max Planck Institute for Dynamics of Complex Technical Systems, Process Systems Engineering, Sandtorstraße 1, Magdeburg 39106, Germany
[b] Otto von Guericke University Magdeburg, Chair for Process Systems Engineering, Universitätsplatz 2, Magdeburg 39106, Germany
[c] École Polytechnique Fédérale de Lausanne, Laboratory of Sustainable and Catalytic Processing, Station 6, Lausanne 1015, Switzerland

ARTICLE INFO

Keywords:
Computer-aided Molecular Design
Graph Neural Networks
Solvents
Lignin
Biorefineries

ABSTRACT

Solvent selection is a difficult task for lignin-first biorefineries and lignin upgrading as the solvent must satisfy multiple complex technical requirements, while remaining extremely stable to allow recycling. High lignin solubility is a common selection criterion, but the ideal solvent for lignin-first biorefineries also requires non-reactivity towards acids and stabilising reagents encountered in the reaction liquor. To facilitate the search for promising solvents, we developed a computational solvent design framework. The framework consists of a graph-based genetic algorithm for molecular design wherein a graph neural network is used for lignin solubility predictions. Based on these predictions, the genetic algorithm iteratively optimises the molecular structures, inspired by evolutionary strategies, such as selection, cross-over, and mutation. The developed framework designed numerous solvents with high potential for application in lignin-first biorefineries and lignin upgrading. For these solvents, experiments confirmed solubilities between 20 and 60 wt.% across different types of lignin. Notably, several solvents were stable under typical biorefinery process conditions. Furthermore, the explainability of graph neural networks enabled us to link the lignin solubility predictions with structural features of the solvents, providing a clear rationale for solvent selection.

1. Introduction

Lignocellulosic biomass, an abundant source of renewable carbon, is a promising feedstock for the production of bio-based commodities [1]. During organosolv processing, lignocellulosic biomass is treated with organic solvents, water, and acid under elevated temperatures to separate the three major biomass fractions: lignin, cellulose and hemicellulose sugars. Harsh process conditions are required to liberate lignin from the lignin-carbohydrate complex of the recalcitrant biomass. However, the given process conditions contribute to the cleavage of ether and ester motifs of the native lignin and promote undesired condensation reactions [2]. The structurally altered lignin contains stable interunit C-C bonds which impede lignin upgrading to aromatic monomers. This problem gave rise to lignin-first biorefineries that apply approaches for active lignin stabilisation [2–7]. One such strategy is aldehyde-assisted fractionation (AAF), which exploits aldehyde protection chemistry to stabilise lignin and prevent its condensation. The aldehyde forms a stable acetal with the α- and γ-hydroxyl groups of the β-O-4 linkage in lignin preventing protonation of α-carbon, the most vulnerable position for condensation. As a consequence, the formation of interunit C-C bonds between the side-chain α-carbons and neighbouring aromatic rings is hindered (see SI for the detailed reaction mechanism) [2]. Both, the condensed and aldehyde-protected lignin, can be upgraded to a variety of applications such as lignin-based coatings, films, resins, nanoparticles, or thermoplastics [8–11]. However, the acetal-stabilised lignin can be depolymerised to aromatic monomers with high near-theoretical yields [7]. Another lignin-first approach is reductive catalytic fractionation (RCF). In RCF, the biomass is treated with an organic solvent combined with a transition metal catalyst. In the presence of a hydrogen donor, which could be an external H_2 source or the solvent itself, the catalyst directly converts the dissolved lignin into aromatic monomers by cleaving the β-O-4 aryl ether motifs [3].

Solvent selection is crucial for both lignin isolation from biomass and further lignin upgrading. For lignin-first approaches, important solvent properties are high lignin solubility, and stability towards the reaction liquor. Moreover, environmental, health and safety (EHS) criteria as

[*] Corresponding authors.
E-mail addresses: jeremy.luterbacher@epfl.ch (J.S. Luterbacher), sundmacher@mpi-magdeburg.mpg.de (K. Sundmacher).

https://doi.org/10.1016/j.cej.2024.153524
Received 10 April 2024; Received in revised form 31 May 2024; Accepted 25 June 2024
Available online 26 June 2024
1385-8947/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

well as solvent recyclability and price influence the feasibility of the process, rendering solvent selection a trade-off between different target properties.

Alcohols (ethanol, 2-propanol, 1-butanol) and acetone are frequently applied in organosolv processing due to their low price and benign EHS properties. However, these solvents suffer from low lignin solubility and are not applicable to AAF, as they react with aldehydes to form hemiacetals or hemiketals. Halogenated solvents are not recommended for application in industrial scale due to their health hazards and low lignin solubility. 2-methyl tetrahydrofuran (2-MeTHF) is a bio-based ether suitable for AAF with mediocre lignin solubility, which is prone to peroxide formation, posing a risk for explosions [12,13]. 1,4-dioxane offers high lignin solubility and is also stable under AAF process conditions [14]. However, peroxide formation, carcinogenicity and negative environmental effects of 1,4-dioxane [13] prompt the search for alternative solvents, especially for large-scale processes. In addition, solvents with high lignin solubility could be of immense interest for the processing of cellulose pulp and hemicellulose sugars to remove lignin impurities.

For RCF, solvents such as ethers and alcohols are a common choice [15]. Here, solvents with high lignin and hydrogen solubilities are preferred. When no external hydrogen is added to the process, the solvent acts as a hydrogen source itself and should therefore have sufficient hydrogen donating capacity.

Lignin upgrading commonly requires solvents with high lignin solubility, such as DMSO [11]. For lignin nanoparticle formation, antisolvent precipitation is a commonly applied method that relies on relative differences in lignin solubility between the applied solvents [9,10].

Solvent selection in lab settings is laborious and resource-consuming. Therefore, computer-guided methods were developed to facilitate the search [16–18]. We recently published a COSMO-RS-based [19–22] solvent screening analysing a database, containing more than 8000 potential solvents, for their applicability in lignocellulose processing [23]. Sulfoxides, azines, oxazolines, and phosphonates were computationally identified, and lignin solubilities up to 33 wt.% were experimentally confirmed. In the solvent screening, the search for ideal solvents was focused on a limited database. However, expanding the search space appears crucial to harness the full potential of computer-guided solvent selection.

Solvent design algorithms, such as variational autoencoders, generative adversarial networks, and evolutionary algorithms, enable *de novo* generation of solvents with tailored properties [24,25]. While the data-driven variational autoencoders and generative adversarial networks are black-box models that hardly allow to rationalise the designed structures, evolutionary algorithms allow for more insights into the structural changes performed on the molecule and to follow distinct structural patterns. State-of-the-art solvent selection for lignin dissolution and biomass fraction requires extensive expert knowledge, experimentally determined solvent parameters [26,27], or time-consuming quantum mechanical (QM) calculations [23].

Here, we present a computational solvent design framework for lignin-first biorefineries and lignin upgrading that operates independently from experimental parameters. We coupled a graph neural network (GNN) for lignin solubility predictions with the newly developed graph-based genetic algorithm *PSEvolve* for solvent design. The GNN was trained on COSMO-RS solubility data and eliminated the need for time-consuming QM calculations. In this way, the GNN acts as a surrogate model of COSMO-RS for lignin solubility predictions. In addition to significantly speeding up the lignin solubility predictions, GNNs can be coupled with attribution techniques to gain insights into the explainability of their predictions [28]. This feature was used to identify the most influential molecular substructures for the solubility predictions. Consequently, functional groups associated with high lignin solubility can be identified from expert knowledge using the GNN explainability as an extra toolkit to guide rational solvent selection. The molecular graph representation of the solvents was not only applied in the GNN solubility predictions but also exploited in the developed genetic algorithm *PSEvolve*. Genetic algorithms perform random mutations on molecules, and require a robust molecular representation, such as graphs, to ensure structural validity [29–31]. To allow for efficient exploration of the chemical space, *PSEvolve* combines graph and valence theory with a well-studied measure for synthetic accessibility [32]. In this manner, *PSEvolve* generates only structurally feasible molecules that are easily synthesisable or even commercially available.

Unlike many other molecular design algorithms, the developed genetic algorithm *PSEvolve* generates structurally feasible molecules that are easily synthesisable or even commercially available by applying graph and valence theory in combination with a measure for synthetic accessibility [32]. Therefore, the developed algorithm allows for fast and efficient exploration of the chemical space.

We demonstrate the solvent design framework for two test cases: In the first case, lignin dissolution is the main objective, which is an important mechanism in lignin isolation from biomass (e.g. in organosolv processing, AAF, and RCF) and lignin upgrading (e.g. production of lignin films or nanoparticles). The exploration of the chemical space is purely guided by maximising the lignin solubility. The second case focuses on solvent design for AAF. There, the objective is to maximise the lignin solubility under constraints of acid- and aldehyde-stability of the solvents. Finally, the most promising solvent candidates are selected for lignin solubility measurements and aldehyde-assisted biomass pretreatment.

2. Results and discussion

2.1. Workflow

In the proposed framework (see Fig. 1), we coupled a GNN with the newly developed graph-based genetic algorithm *PSEvolve* for the design of tailor-made solvents for lignocellulose processing and lignin upgrading. The GNN was trained and tested on COSMO-RS-generated solubilities of a representative lignin fragment in more than 3300 solvents.

Genetic algorithms are inspired by biological evolution processes and were frequently applied in computer-aided design of drugs, solvents, and catalysts [33–35]. The developed graph-based genetic algorithm *PSEvolve* is the core of the solvent design framework. In this study, *PSEvolve* iteratively modifies chemical structures to maximise their lignin solubility based on GNN solubility predictions. We initialised *PSEvolve* with a start population containing 1000 n-hexane molecules, a solvent with low solubility across different types of lignin [12,36] and optimised their structures over 1000 generations. In each generation, the GNN predicted lignin solubilities served as a measure of fitness for each chemical structure. In analogy to Darwin's "survival of the fittest", fitness-appropriate selection, cross-over, and random mutations aim to drive the population of chemical structures towards maximal fitness [37].

To ensure diversity among the population and to reduce the risk of convergence to local minima, *PSEvolve* features a broad range of mutation operations, such as the addition, deletion and substitution of bonds and atom, the relocation of molecular fragments, or the addition of functional groups. In contrast to several other molecular design algorithms, *PSEvolve* ensures generating structurally feasible molecules by combining implicit valence count and graph theory (see methods for details). Furthermore, *PSEvolve* allows for imposing constraints on the molecular structure. In this study, an important constraint was the synthetic accessibility score (SAS) [32]. The SAS is an estimate for the ease of synthesis of a molecule and prevents the algorithm from exploring structurally feasible molecules that are, however, hardly commercially available or involve difficult synthesis. Acid- and aldehyde-instable functional groups were excluded during solvent design tailored towards AAF. *PSEvolve* can be also adapted to other

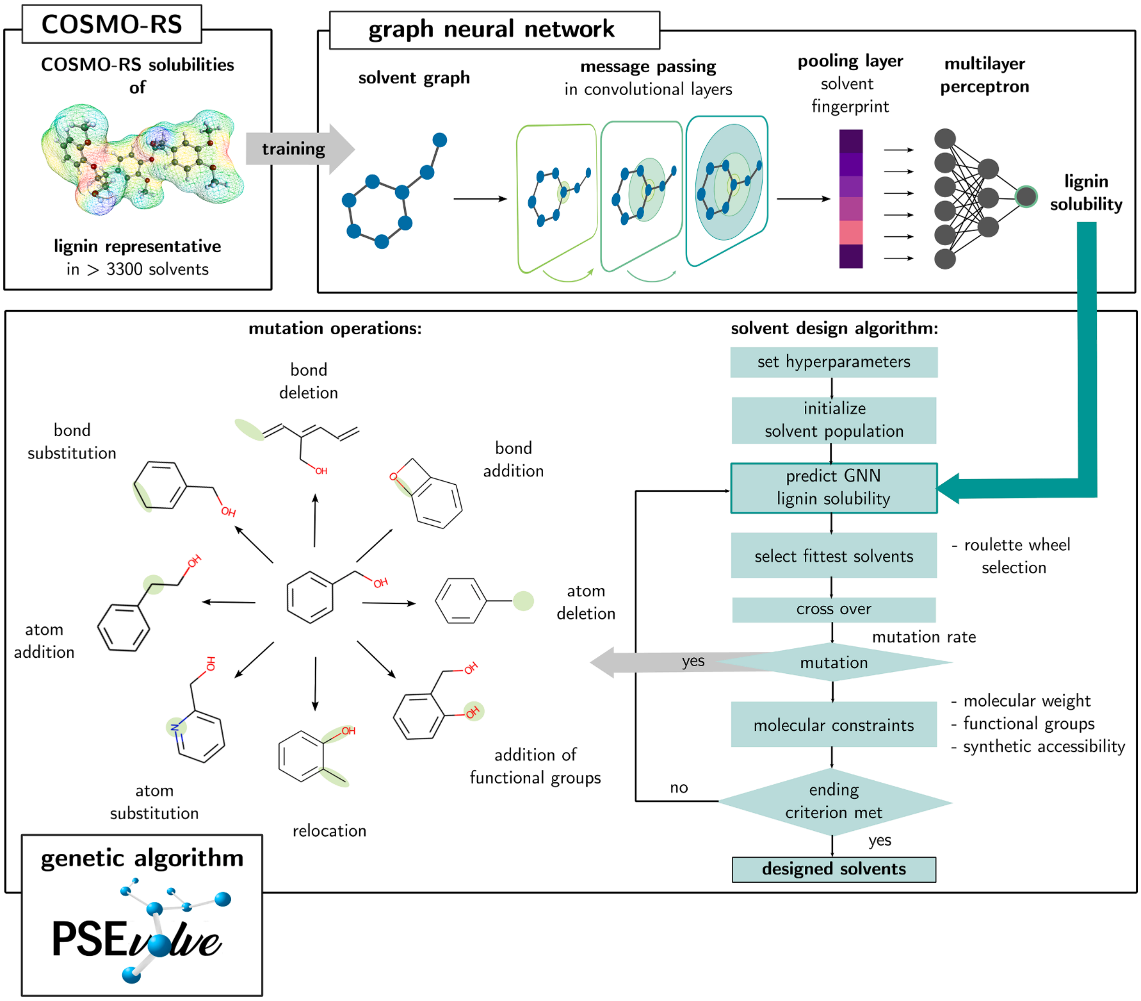

Fig. 1. Workflow for the proposed solvent design framework. The genetic algorithm *PSEvolve* optimises the structure of a molecule population by iteratively performing evolutionary operations, such as selection of the fittest individuals for cross-over, and mutation. The genetic algorithm performs the operations on the graph of the molecule to guarantee structural feasibility. The fitness is given by the lignin solubility as predicted by a GNN. The GNN was trained on COSMO-RS generated lignin solubility prior to coupling with the genetic algorithm. Constraints on the molecular structures, such as molecular weight, restriction of functional groups, and synthetic accessibility can be introduced for the design of tailor-made solvents.

molecular design problems with broad application in chemical engineering.

The GNN enabled fast and accurate predictions of the lignin solubility which was crucial for the fitness evaluation within the genetic algorithm. First, the chemical structure of the solvent molecules was transformed into a molecular graph. This graph was defined by a set of nodes and edges, representing atoms and chemical bonds, respectively. We assigned several atom and bond features to the graph (see methods for details). In the message-passing step, the node features of the graph were updated by using the information about the neighbouring nodes and the connecting edges. Then, by performing this message passing operation multiple times, the node embeddings were effectively enriched with information of their neighbourhood. The resulting updated graph was subsequently passed through a pooling operation to yield the “molecular fingerprint” of the solvent. This solvent fingerprint acts as a tailor-made vectorial representation of the solvent optimised for predicting lignin solubility This tailor-made fingerprint served as an input to a multilayer perceptron which finally predicted the lignin solubility.

As experimental data on lignin solubility is scarce, the GNN was trained and tested with COSMO-RS solubility predictions of a representative lignin structure in more than 3300 solvents. A trimer of guaiacyl (G)-units connected *via* β-O-4 bonds was used as a representative lignin structure for solubility predictions. The representative structure was modelled on a quantum chemical level (molecular weight: 530.57 g/mol, see SI for molecular structure). The β-O-4 motif is the predominant bond pattern in lignin, constituting approximately 50 % of the linkages in softwood, 60 % in hardwood, and 80 % in grasses [38]. The abundancy and ease of cleavage renders the β-O-4 motif, a key target for lignin depolymerisation [5]. In contrast to *p*-hydroxyphenyl (H)-and syringyl (S)-units, only G-units are produced across all hardwood, softwood, and herbaceous biomass sources [39]. In grasses and hardwoods, G-units are less abundant compared to S- and H-units. However, G-units only differ by a methoxy group from H- and S-units and can be therefore seen as intermediate structures. This structural similarity renders G-units excellent representative units. A recent computational analysis revealed that lignin solubility is rather depending on solvent parameters than on structural features of lignin [26], rendering the choice of the lignin representative less influential.

Overall, the GNN and COSMO-RS predictions are in good agreement ($R^2 = 0.896$, MAE = 0.322 for the test set), with slight deviations in the upper solubility ranges with $\log(x_{\mathrm{sol,\ lignin}}) > -1$ (Fig. 2 a). To obtain accurate lignin solubility predictions for a structurally different chemicals as generated by the solvent design algorithm, we applied a training set that reflects a high structural diversity (Fig. 2 b). Our previous study [23] showed, that COSMO-RS predictions are useful for qualitative

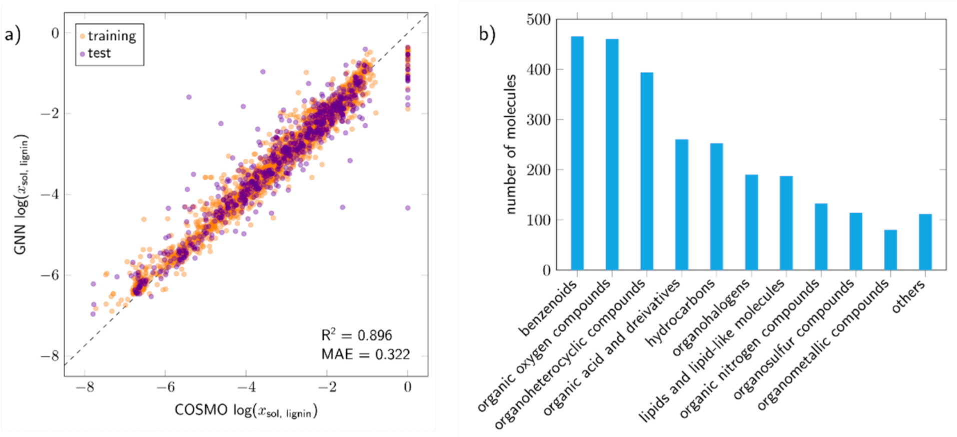

Fig.2. GNN-training with COSMO-RS solubility data for a representative lignin fragment. a) Parity plot for GNN vs. COSMO-RS predictions of the training and the test set. The coefficient of determination (R^2) and the mean absolute error (MAE) are given for the test set. b) Chemical classes of the training set as computed by the Classyfire toolbox.

solubility comparison of different solvents, rather than accurate absolute solubility predictions. Being trained on COSMO-RS data, the GNN consequently allows to qualitatively compare various solvents in their ability to dissolve lignin. Further details regarding applicability ranges of the GNN are provided in the SI. Furthermore, the accuracy of the COSMO-RS solubility predictions is reduced for solubilities of log($x_{\mathrm{sol,\ lignin}}$) > -1 [40]. The varying COSMO-RS accuracy might explain the disparity between the COSMO-RS and the GNN predictions in this region (see Fig. 2 a). Unfortunately, the region of high lignin solubility is of particular importance when designing solvents for the same purpose. However, the main objective of this study is to identify a broad range of so-far unexplored solvent classes, rather than identifying a single optimal one. Therefore, the deviations are less impactful for the scope of this work and do not outweigh the advantages of using the GNN as a surrogate model of COSMO-RS. Indeed, solvent design was only made possible by the low computational time and suitable accuracy of the GNN.

2.2. *Tailored solvents for lignin dissolution*

Many strategies for lignin upgrading, such as the fabrication of lignin films, coatings, or nanoparticles, require efficient dissolution of lignin. Therefore, in this first test case, the main objective was maximising the lignin solubility. We initialised the algorithm with a start population of hexane molecules, which are known for a low lignin solubility, without fixing any constraints on functional groups.

For analysing the most promising designed molecules, we selected those with the highest solubilities (log($x_{\mathrm{sol,\ lignin}}$) > -1.5; around 21,000 molecules). We studied the relation between the molecular structure and the GNN predicted lignin solubilities by applying t-distributed stochastic neighbour embedding (t-SNE) to the GNN-generated solvent fingerprints. t-SNE reduces dimensionality of the GNN fingerprint, a vectorial representation of the molecular graph, to a 2-dimensional space in which molecules with similar GNN fingerprints are located within proximity of each other. Structurally similar molecules were predicted to have similar lignin solubilities (Fig. 3 a). Therefore, the GNN can be considered as a quantitative structure–property relationship (QSPR) model that was trained “end-to-end” from the molecular graph to the lignin solubility. Additionally, the GNN was able to generate tailor-made optimised molecular fingerprints. We found regions with especially high lignin solubility predictions (log($x_{\mathrm{sol,\ lignin}}$) > -0.60) corresponding to sulfoxides, compounds with P = O motif, sulfones, triazines, diazines, and azoles (Fig. 3 a). Other solvent classes promising for lignin upgrading are morpholines, cyclic ethers, and cyclic ketones. The overall fittest solvent was dimethyl sulfoxide (DMSO) with log($x_{\mathrm{sol,\ lignin}}$) = -0.35, followed by 1,3,5-triazine (log($x_{\mathrm{sol,\ lignin}}$) = -0.36), N,N-dimethylpyrimidin-5-amine (log($x_{\mathrm{sol,\ lignin}}$) = -0.37), and dimethyl methyl phosphonate (DMMP) (log($x_{\mathrm{sol,\ lignin}}$) = -0.38).

A common method to demonstrate the applicability of molecular design frameworks is the rediscovery of molecules with the desired target property [41]. DMSO and pyridine were recently reported as the most effective solvents for lignin dissolution [42,43], all of which were rediscovered by the presented solvent design framework. DMSO is frequently applied in the formation of lignin nanoparticles and film formation, which underlines the applicability of the solvent design framework for lignin upgrading.

In addition to the already established lignin solvents such as DMSO and pyridine, we discovered commercially available azoles, such as thiazole or isoxazole. Thiazole is only slightly toxic (LD_{50} oral rat: 938 mg/kg) [13] while toxicity data for isoxazole is currently lacking. Further aromatic N-heterocycles were designed, including triazines, diazines, pyridines, and bicyclic compounds. Common side chain motifs were methoxy-, alkyl-, and NH_2-groups. Pyridines and many diazines have benign EHS properties [13] and are readily commercially available. Most triazines, but also sulfones and phosphonates are solid at room temperature, limiting their applicability for lignin upgrading. Cyclic ethers and ketones were associated with lower GNN-predicted lignin solubilities compared to the aforementioned solvents, however, they were predicted to have a higher lignin solubility compared to the usually applied 1,4-dioxane.

During the solvent design, functional groups associated with low lignin solubilities (e.g. alkanes) were gradually replaced by functional groups associated with higher lignin solubilities (e.g. aromatic N-atoms, Fig. 3 b) leading to a gradually increasing mean lignin solubility of the population (Fig. 3 c). All designed solvents with a lignin solubility of log ($x_{\mathrm{sol,\ lignin}}$) > -1.5 are summarised in the Supplementory Information (SI).

2.3. *Connecting structural solvent features with lignin solubility*

GNNs can be coupled with attribution techniques to explain the predictions. Integrated gradients [28,44] (IG) is one of such attribution methods specifically developed to comply with the sensitivity and implementation invariance axioms. In the IG method, the integral of the gradients of the model’s output with respect to its input is computed, while gradually changing the input values from a baseline to the actual input of interest. This process effectively assigns importance scores to each input feature by attributing their contribution to the final

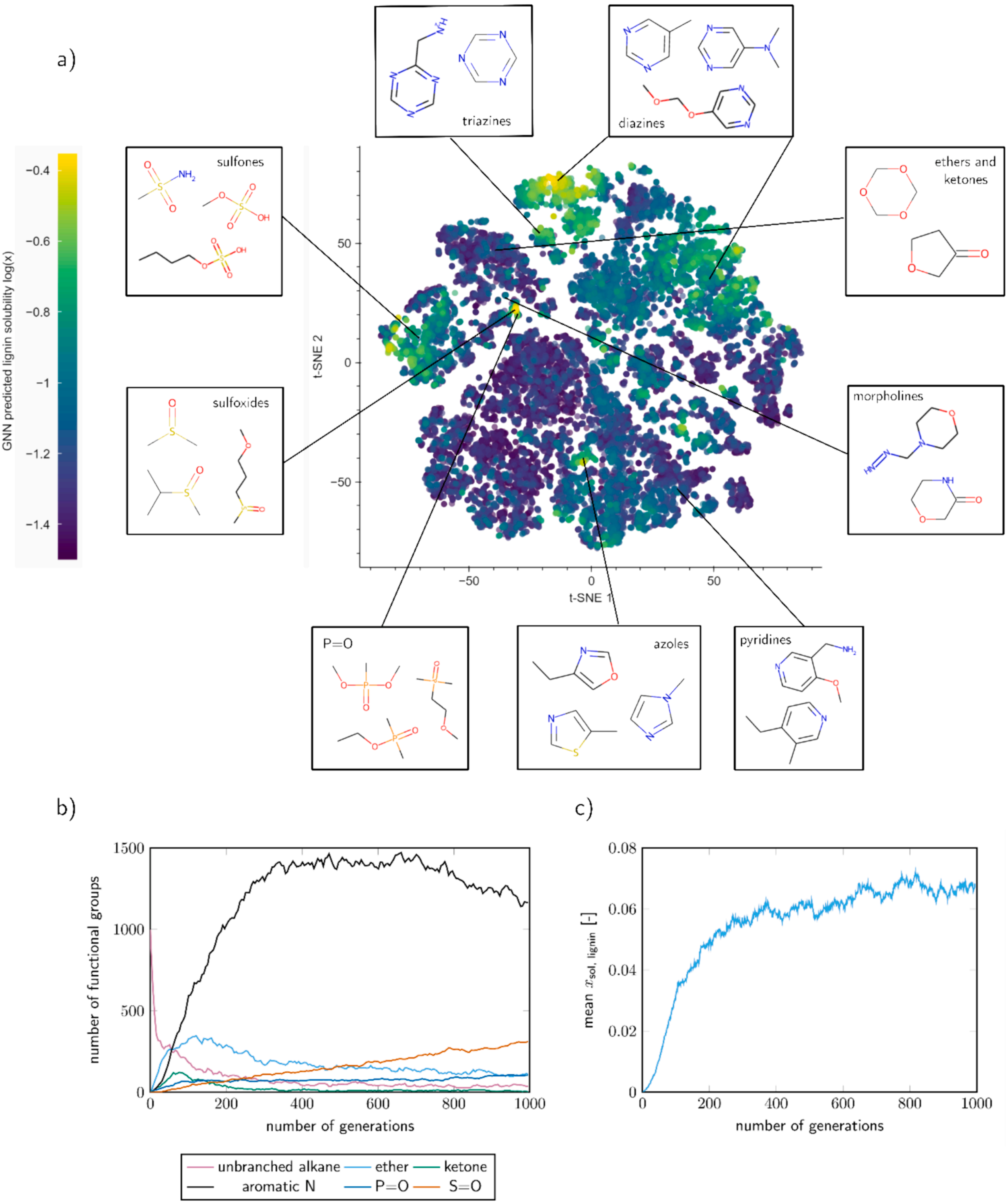

Fig.3. Application of the solvent design framework for lignin upgrading. a) t-SNE plot of the designed molecules with highlighted lignin solubility. b) Exploration of chemical space during molecule optimisation. c) Evolution of the molar lignin solubility during molecular optimisation.

prediction. The IG method identifies nodes and edges with the highest impact on a given prediction. Therefore, the contribution of each atom and bond within the solvent to the predicted lignin solubility can be visualised and potentially used as a guide to explain the results. To enable reliable interpretations of the results beyond theoretical predictions, we first experimentally measured the solubility of different types of lignin in the designed solvents. Subsequently, we employed IG, to attribute the predicted lignin solubility to structural features of the solvent.

For experimental validation, we selected 27 commercially available potential solvents with identical or similar structures to the solvent candidates designed by the genetic algorithm. Subsequently, these solvents were used to measure the solubilities for three different types of lignin ($T = 85$ °C): Kraft lignin isolated from softwood species, FABIOLA^{TM} organosolv lignin [45] isolated from hardwood, and herbaceous lignin isolated from corn cob by mild acidolysis (MAL) (see SI for 2D HSQC NMR of the wood and additional data points with lignin isolated from birchwood). N-heptane and dibutyl ether were chosen as a control solvents with low lignin solubility, and 2-MeTHF as a control for mediocre solubility, based on the results of our previous work [12]. The lignin solubilities ranged between 20-60 wt.% in most of the selected solvents (Fig. 4 a). Highest solubilities were measured for DMSO (≥ 60 wt.%), and isoxazole (≥ 50 wt.%), 2-picoline-n-oxide (≥ 49 wt.%), 2,5-dimethyl-pyrazine (≥ 49 wt.%), and thiazole (≥ 49 wt.%). In general, the experiments confirmed high solubilities for the solvents designed by the genetic algorithm. The differences in solubility between the different types of lignin were rather low, implying that the solvents tested are universally effective. Note that for most solvents, lignin saturation was not completely reached as the solutions became increasingly viscous with higher amounts of dissolved lignin and imposed challenges for

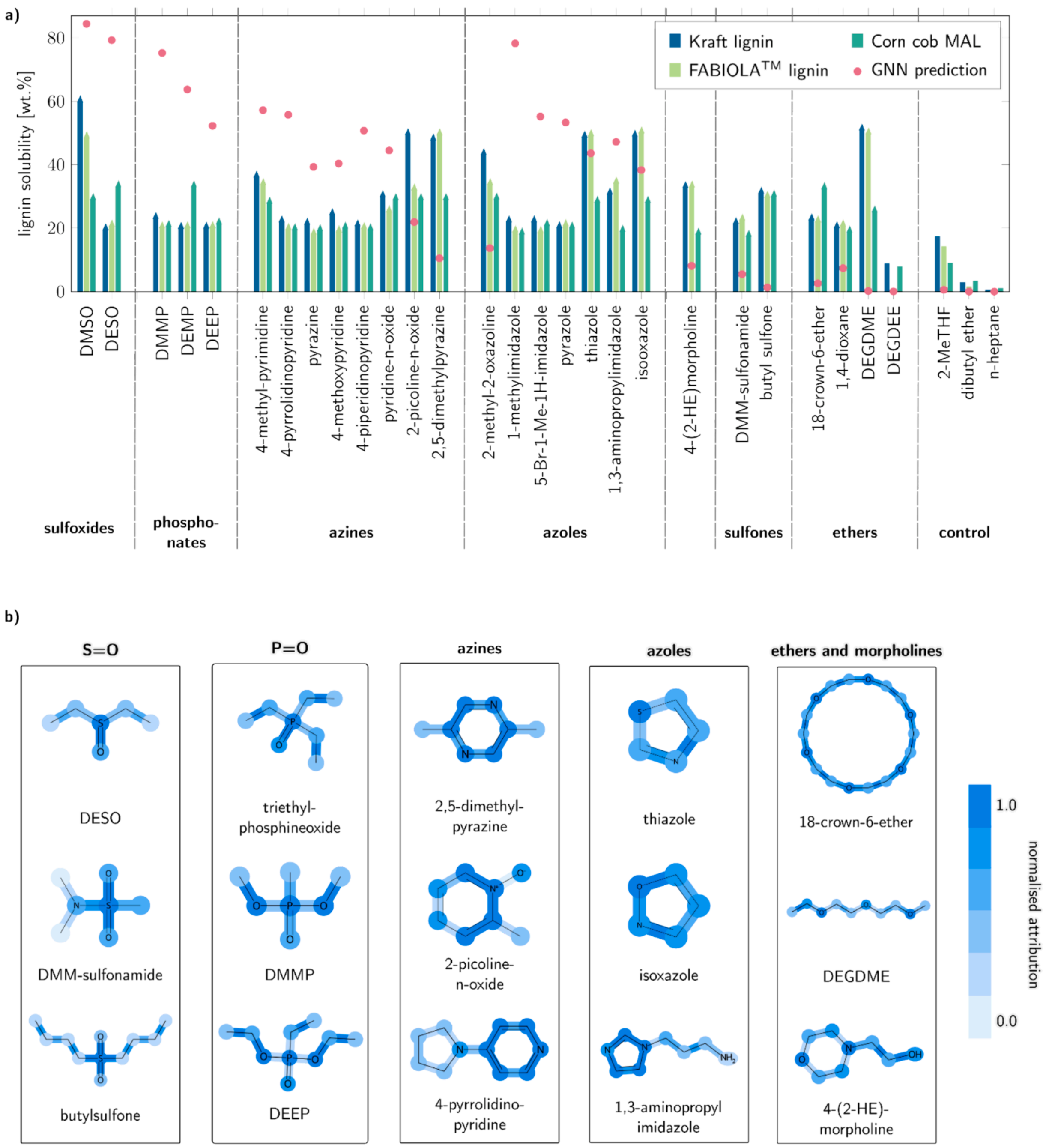

Fig.4. Experimental validation of the GNN predictions and attribution of structural features to the predicted lignin solubilities. a) Experimental lignin solubilities in the designed solvents for Kraft lignin, FABIOLA^{TM} lignin, and mild acidolysis lignin (MAL) isolated from corn cob. Most solvents were commercially available as originally designed by the genetic algorithm. Otherwise, structurally similar molecules were purchased. Arrows indicate that lignin saturation was not yet reached, however, the high viscosity of the solution hindered measurements with higher lignin loadings. The numerical data is provided in SI. b) Normalised attributions for each discovered solvent class. A higher attribution score of the highlighted structural feature indicates higher importance for the lignin solubility prediction. Abbreviations: DMSO – dimethyl sulfoxide, DESO – diethyl sulfoxide, DMM-sulfonamide – dimethyl methane sulfonamide, DMMP – dimethyl methyl phosphonate, DEMP – diethyl methyl phosphonate, DEEP – diethyl ethyl ethyl phosphonate, 5-Br-1-Me-1H-imidazole – 5-bromo-1-methyl-1H-imidazole, 4-(2-HE)morpholine – 4-(2-hydroxyethyl)morpholine, DEGDME – diethylene glycol dimethyl ether, DEGDEE – diethylene glycol diethyl ether, 2-MeTHF – 2-methyl tetrahydrofuran.

filtering even with specialised filters designed for viscous samples.

We noticed deviations between predicted and experimental solubility data for some of the selected ethers. We measured lignin solubilities of up to 51 wt.% for diethylene glycol dimethyl ether (DEGDME), whereas the structurally similar diethylene glycol diethyl ether (DEGDEE) dissolved maximally 8.9 wt.% of lignin. The GNN was not able to discriminate the differences between the ether structures and predicted for both solvents nearly no lignin dissolution. Additional experiments showed that this difference in solubility seems to be influenced by additional CH_3-groups that reduce the polarity (see SI for more

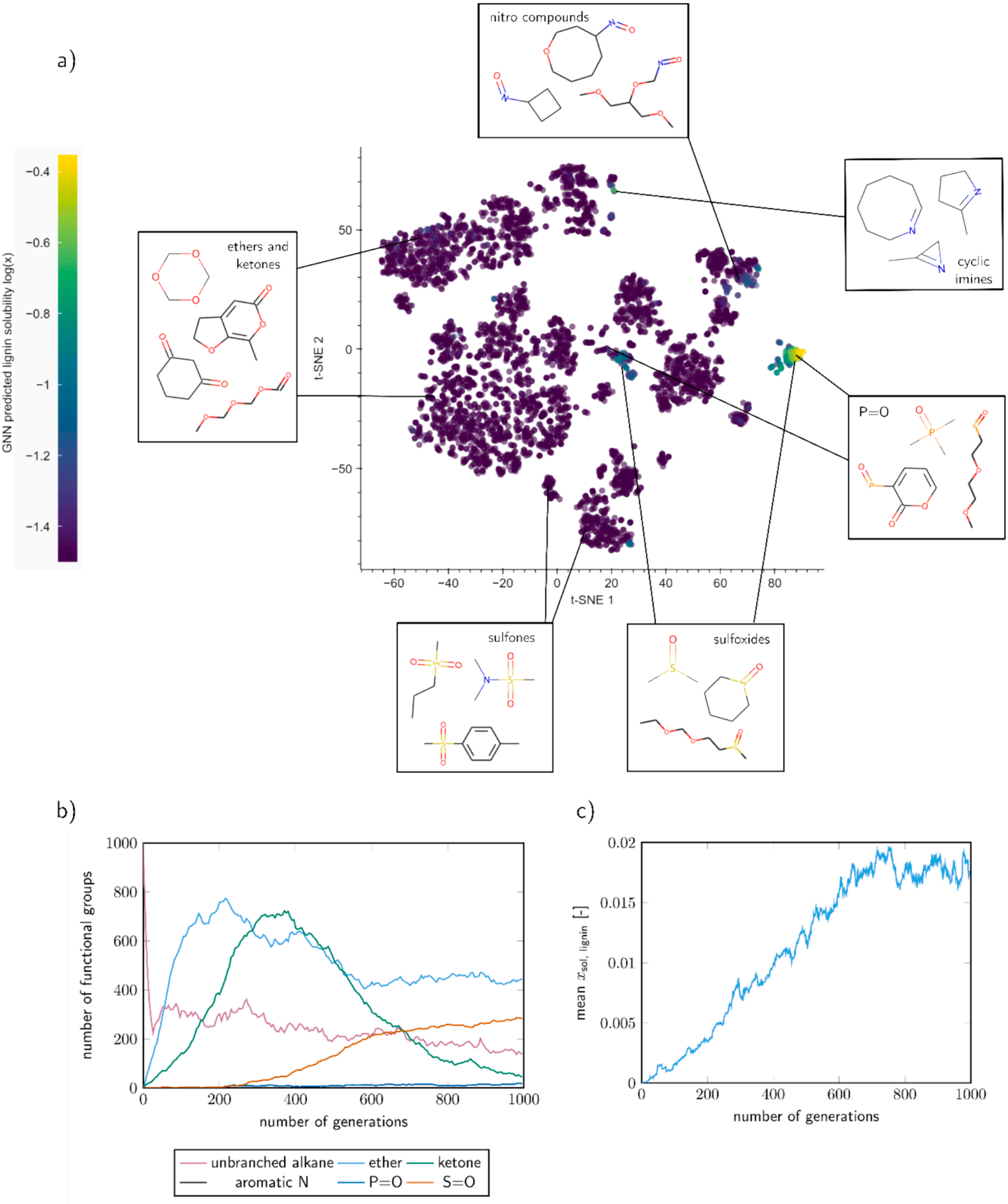

Fig.5. Application of the solvent design framework for AAF. a) t-SNE plot of the designed molecules with highlighted lignin solubility. b) Exploration of chemical space during molecule optimisation. c) Evolution of the molar lignin solubility during molecular optimisation.

details).

Subsequently, we applied the IG method to identify the structural features with the highest influence on the GNN-predicted lignin solubility. In a comprehensive study of different attribution methods [28], IG was suggested as the most suitable alternative of graph attribution when the last layer of the GNN framework is not a pooling mechanism, similar to the setup in the present study. We computed normalised attributions that indicated the importance of nodes and bonds for the solubility predictions (Fig. 4 b). A higher attribution score indicates a higher importance on the predictions. Notice that since the attribution scores were normalised, they only allowed for a relative comparison within the same molecule. The high lignin solubility in sulfoxides and sulfones was mainly attributed to the presence of S-atoms and the adjacent double bonds in solvent structures. Similarly, the P=O motif in phosphine oxides was the most influential in promoting lignin solubility. Additionally, in phosphonates, phosphorus (P) and oxygen (O) atoms received the highest attribution scores. The high lignin solubility in aromatic heterocycles was in general attributed to the N-atoms and the neighbouring aromatic bonds. However, for specific classes like oxazoles and thiazoles, the sulfur (S-) and oxygen (O-) atoms within the aromatic rings had a higher attribution score than the aromatic N. In contrast, nitrogen atoms within side chains as well as alkyl chains and saturated rings found in compounds such as butyl sulfone, 4-pyrrolidinopyridine, and 1,3-aminopropyl imidazole, had less impact on solubility predictions. In the case of ethers, the oxygen atoms (O-atoms) had a more significant impact on the predicted solubility of lignin than the carbon–carbon (C-C) bonds. In general, solvents containing S, N, P, O and/or aromatic bonds were associated with high lignin solubility predictions, indicating consistency with general chemical intuition. Indeed, lignin with numerous benzene rings and oxygen atoms could engage in π-interactions with solvent molecules and form hydrogen bonds with these heteroatoms. However, as predictions and experiments diverge for ethers, we expect additional important features affecting solubility that could not be captured by the GNN. In line with the presented results, the computational Kamlet-Taft-parameter analysis of Sumer and van Lehn revealed that a solvent requires good hydrogen bond-accepting ability and intermediate to high polarity for efficient lignin dissolution [26].

2.4. Tailoring molecular solvent structures for aldehyde-assisted fractionation

To identify solvents compatible with the AAF process that also includes an acid and an aldehyde, we searched for solvents that provided high lignin solubility while being stable towards the reaction liquor. For this purpose, we excluded several functional groups due to their potential reactivity: primary and secondary amines, aldehydes, aromatic N-heterocycles, isocyanates, amides, esters, and hydrazides. Although ketones can undergo aldol condensation with aldehydes under acidic conditions, we did not exclude keto-groups, as they are easily mutatable by the algorithm to other functional groups, such as ether or C=C groups.

During the design, sulfoxides, sulfones, ethers, ketones, P=O compounds, and cyclic ethers and ketones were explored (Fig. 5 a). In addition, the functional group restrictions spurred the exploration of non-excluded nitrogen-containing patterns, such as nitro-groups or cyclic imines, with high GNN-predicted lignin solubilities.

Similar to the solvent design for lignin dissolution, the number of alkane groups decreased rapidly within the first generations which were replaced by functional groups associated with higher lignin solubilities (Fig. 5 b).

However, unlike for lignin dissolution only, the search became more targeted, concentrating on solvent classes that are effective at dissolving lignin while taking into account the functional group restrictions. Among the 100 fittest solvent candidates, nearly 90 % were sulfoxides, with DMSO being the overall fittest designed solvent ($\log(x_{\text{sol, lignin}}) =$ -0.35). Finally, due to the functional group constraints, the algorithm designed fewer solvents with high lignin solubilities compared to the run for lignin upgrading. As a consequence, the average lignin solubility of the solvent population was lower (Fig. 5 c).

2.5. Experimental aldehyde-assisted pretreatment of birch wood

AAF leads to the separation of three main components of the biomass: cellulose-rich pulp, uncondensed acetal-stabilised lignin, and aldehyde-protected xylose (a product of hemicellulose depolymerisation). We experimentally evaluated the designed solvents for their applicability in AAF, using milled birch wood, propionaldehyde, and $HCl_{37\%}$ added to selected solvent candidates, such as sulfoxides, sulfones, phosphonates, and ethers. The suitability of the tested solvents for AAF was assessed by measuring the weight of isolated cellulose pulp, the yield of lignin monomers after hydrogenolysis of the pretreatment liquor, and the yield of xylose protected by propionaldehyde (dipropylxylose, DPX). These metrics can serve as indicators of the effectiveness of biomass depolymerisation into these three fractions.

Notably, the tested sulfoxides, sulfones (except the DMM-sulfonamide), phosphonates, and ethers were resistant to acidic conditions (0.4 M HCl) and elevated temperature ($T = 85$ °C), showing no visible signs of degradation. Upon completion of the pretreatment reaction, we observed that the biomass was uniformly disrupted to smaller fragments in samples containing butyl sulfone, DEGDEE, and DMM-sulfonamide. The filtrated pulp in these samples appeared as a fine powder of light colour (Fig. 6a), constituting around 40 wt.% of the biomass. In contrast, in the samples, containing DEGDME, 18-crown-6 ether, DMSO, and phosphonates, the biomass retained its original form of wood chips. The filtrated pulp in these cases constituted up to 90 wt.% of the biomass, indicating a less effective extraction of biomass components in the liquor, with a significant portion remaining retained within the cellulose fibers. We speculate that these solvents were too polar to attack the nonpolar faces of cellulose that are typically disrupted by hydrophobic stacking interactions within cellulose in solvent–H_2O mixtures [46].

After pulp separation, the filtrate contained extracted lignin with propionaldehyde-protected β-O-4 linkages as confirmed by the HSQC NMR (Fig. 8, SI). Hydrogenolysis of such uncondensed lignin over Ru/C at 250 °C produces valuable aromatic monomers and their quantification provides insights into the effectiveness of lignin extraction and quality. The benchmark solvent 1,4-dioxane allowed to produce near-theoretical 7.8 wt.% of monomers on a raw biomass basis, followed by DEGDEE and butyl sulfone, each providing around 5 wt.% of monomers, and DEGDME and 18-crown-6 ether, yielding less than 4 wt.% (Table 5, SI).

Aldehyde-protected sugars including DPX demonstrated potential as sustainable solvents and versatile platform chemicals [47–50]. The yield of DPX exceeded 20 wt.% (based on the raw biomass) in 1,4-dioxane, and the designed solvents DEGDEE and butyl sulfone. The DPX yield for the crown ether, DEGDME and DMM-sulfonamide was lower than 15 wt.%. DMM-sulfonamide demonstrated signs of degradation during the pretreatment, while two other solvents did not provide sufficient biomass disruption as mentioned above. We detected low amounts of DPX ($<$ 1 wt.%) produced after pretreatment in phosphonates and sulfoxides. Interestingly, in control experiments with dibutyl ether and heptane, we obtained DPX yields comparable to 1,4-dioxane, suggesting that the interactions of non-polar solvents and the biomass enable sufficient contact between the reaction components (e.g. aldehyde) and xylan that closely interacts with cellulose in a plant cell. However, these solvents could not provide effective delignification and a high-quality lignin as the lignin forms a globule with reduced surface area in such highly nonpolar solvents [51], preventing its protection by the aldehyde and leading to condensation.

The pretreatment experiments showed that the designed glycol ethers, and sulfones are promising solvent candidates for the AAF procedure. Notably, DEGDEE and butyl sulfone provided effective

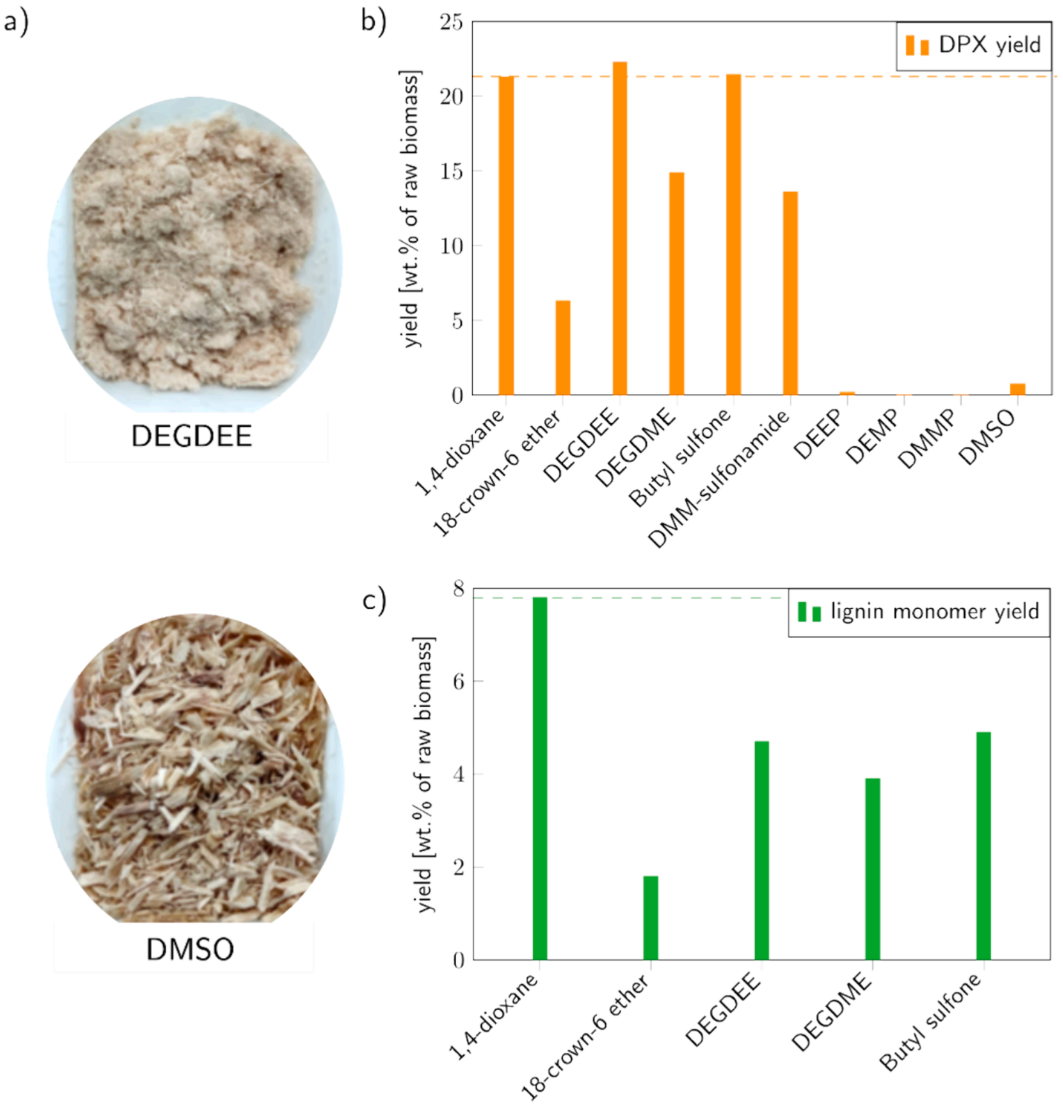

Fig.6. a) Cellulose-rich pulp after pretreatment with degdee (above) and dmso (below). b) DPX yield on raw biomass basis c) lignin monomer yield on raw biomass basis in the liquor after propionaldehyde-assisted pretreatment of birch wood at 85 °C for 3 h using the selected solvents.

fractionation of cellulose pulp, and successful extraction of PA-protected lignin and xylose in the pretreatment liquor with yields comparable to the carcinogenic benchmark solvent 1,4-dioxane. Using high boiling solvents that are solid at room temperature, such as butyl sulfone, in the AFF requires adaptions of current procedures for the isolation of fractions from the liquor and for solvent recovery. Additionally, this solvent is currently produced in limited quantities and its toxicological profile is not well studied, opening opportunities for future research.

The solvents tested in this work were specifically designed to target the lignin fraction of the biomass. Therefore, most candidates are highly polar since this characteristic feature is necessary to ensure lignin solubilisation in the liquor, thereby increasing its surface area exposed for a reaction with aldehyde in the mixture (see SI for details) [51]. However, we found that the overall quality and quantity of the extracted lignin are also significantly influenced by other solvent properties. In particular, in addition to polar groups, the solvent should have sufficient non-polar domains that can disrupt the cellulose microfibrils and facilitate the extraction of lignin and xylan. This is likely the case of butyl sulfone that has both a highly polar SO_2-group and two distinct aliphatic chains in its structure, showing superior results, as well as DEGDEE which significantly outperformed DEGDME despite the only difference being the presence of two additional methyl groups in the DEGDEE structure. Lastly, the stability of the solvent under acidic conditions at high temperatures is crucial to maintain a consistent and effective chemical environment throughout the process.

3. Conclusions

The presented computational solvent design framework generated so-far unexplored solvent classes for lignocellulose biomass fractionation and lignin upgrading. Besides DMSO, which is frequently applied in lignin upgrading, we designed azoles, and six-membered aromatic N-heterocycles. Due to their high, experimentally validated lignin solubilities (20-60 wt.%) and their low toxicity, these solvents are highly interesting for application in the fabrication of lignin-based films, nanoparticles, or resins.

Solvent selection for AAF is a complex task that requires a solvent that not only effectively dissolves lignin but also disrupts cellulose fibers and remains unreactive and stable under acidic conditions. In addition, EHS criteria and commercial availability immensely narrow the space of potential solvent candidates. Despite these challenges, our computational framework successfully designed promising solvents for AAF such as glycol and cyclic ethers as well as sulfones. Solvents from these groups showed acid stability under AAF conditions and provided performance nearly on par with 1,4-dioxane, while being potentially less toxic. We found that stable aprotic solvents possessing heteroatoms (e.g. oxygen, nitrogen, or sulfur), giving them sufficient polarity, and hydrocarbon motifs, representing the nonpolar domain of the solvent, could be suitable for application in AAF and in lignocellulose processing in general. The challenge for finding the right balance of these characteristics remains opened for future research.

The use of GNNs was computationally fast, serving as a surrogate model for COSMO-RS lignin solubility predictions. Due to the applicability of explainability methods to the GNN, such as IG, we were able to analyse the structural solvent features impacting the predicted lignin solubility. In line with experimental results, solvents containing sulfur, aromatic nitrogen, phosphorous or oxygen were connected to high lignin solubility predictions. Therefore, the presented solvent design framework not only facilitated the exploration of promising solvents but also provided valuable insights into the specific molecular characteristics essential for achieving high lignin solubility. Furthermore, the presented solvent design approach could serve as a blueprint for other types of molecules, e.g. for cellulose for which only few effective solvents were reported.

4. Methods

4.1. COSMO-RS lignin solubility predictions

The molecular weight of lignin ranges between 2,500 and 15,000 g/mol. COSMO-RS lignin solubility predictions require QM calculations of lignin but are infeasible for the whole molecule due to its size. Similar to our previous study [12], we used a representative lignin structure that captures the most prominent structural features of the original polymer. We used a trimer of G-units connected *via* the most common bond pattern, the β-O-4 motif (molecular weight: 530.57 g/mol, see SI for molecular structure). The lignin solubility was predicted by COSMO-RS predictions with the G-trimer at a temperature of 70 °C, matching with mild processing conditions. The QM calculations and the COSMO-RS solubility predictions were performed as described earlier [12]. For this study, we used an iterative algorithm to improve the accuracy (see SI for details).

4.2. Data set splitting

The final data set contained COSMO-RS lignin solubility predictions in 3314 different solvents. The data set was split into a model developing set and a test set with a proportion of 80/20. The Butina clustering algorithm, as implemented in *rdkit* [52], was first used to generate clusters of similar molecules. Each cluster was then randomly split with a proportion of 80/20 to generate the model's developing set and the test set. In this way, we ensured that the model was trained and tested on similar chemical structures. In this way, the estimation of the model's performance reflects its accuracy within the domain defined by the chemical classes used during the model's development. Furthermore, this type of splitting ensures an even distribution of chemical classes within the model's developing and test sets.

The binary Morgan fingerprint with a radius of two and a bitsize of 2048 was used to calculate the Tanimoto similarity of each pair of molecules. With this information a matrix of molecular distances was constructed by subtracting the corresponding similarity value from one. This matrix of distances was utilized to perform the Butina clustering. A threshold of five molecules was used to differentiate between a large and a small cluster. All large clusters were randomly split as mentioned above while all small clusters were directly put into the model's developing set. The test set was reserved for model assessment while the model's developing set was further split with a proportion of 85/15 to constitute the train and validation sets. All results shown in the paper correspond to the test set unless mentioned otherwise.

4.3. GNN architecture

The graph of the solvent $G = (V, E)$ consists of a set of nodes V connected by a set of edges E, representing the corresponding atoms and bonds, respectively. Atom and bond features (see Table 1) were calculated by *rdkit* [52] (version 2021.03.1) and constitute the vectorial representation of the corresponding nodes and edges within the graph. A matrix of atom features $A \in \{0,1\}^{n_a \text{ x } n_{af}}$ and a matrix of bond features $B \in \{0,1\}^{n_b \text{ x } n_{bf}}$ was then defined for each molecule, where n_a and n_b denote the number of atoms and bonds in the solvent, and n_{af} and n_{bf} refer to the number of atom and bond features. The connectivity between atoms and bonds was given by the connectivity matrix $C \in \mathbb{N}^{2\text{x}2n_b}$ capturing the indices of the source and receiver nodes. The features, see Table 1, were selected to distinguish fundamental differences between atoms and bonds within a given molecule [53,54]. The cheminformatics package *rdkit* converted the solvents' SMILES representations into molecular objects, and calculated all atomic and bond features and the connectivity matrix. One-hot-encoding was used to encode the atomic and bond information into fixed-size vectors for all molecules according to the dimensions shown in Table 1.

PyTorch geometric (version 2.3.1) and PyTorch (version 1.10.2) were used for the GNN setup. The model consists of 3 message passing layers operating with a hidden-dimension of 50 and using the *NNConv* architecture. The message passing is based on the based on the continuous kernel-based convolutional operator from Gilmer et al. [55]

$$a_v^{(l+1)} = W^{(l)} a_v^{(l)} + \frac{1}{|N(v)|} \sum_{w \in N(v)} \left(\phi_e^{(l)}(b_{vw}) \bullet a_w^{(l)} + q^{(l)} \right),$$

where $a_v^{(l+1)}$ stands for the vector of updated node features for node v, $W^{(l)}$ corresponds to a matrix of learnable parameters at message passing layer (l), $N(v)$ stands for the cardinality of the set of neighbouring nodes of node v, $E(l)$ corresponds to the edge-transformation function (here implemented as a single-hidden layer neural network with the ReLU activation function), b_{vw} stands for the vector of edge features for the edge connecting node v and node w, $q(l)$ is a learnable bias vector at message passing layer (l). A single hidden-layer neural network with dimension 64 and the ReLU activation function was used as the edge-transforming function. The batch normalisation proposed by [56] was used after each message passing layer to enhance the training of the model. The Leaky ReLU activation was used after the first and second message passing layers to update the node embeddings. The molecular fingerprint was then obtained by using the max global pooling function on the final updated graph. A multi-layer perceptron (MLP) was later used to regress the final solubility prediction from the molecular fingerprint. This MLP contains 2-hidden layers with dimensions 50 and 25. A dropout ratio of 0.1 was used in the message-passing layers and the final MLP to prevent overfitting. The model was trained "end-to-end" from the molecular graph to the lignin solubility using the AdamW optimizer with a learning rate of 0.001 and batches of 32 graphs. The

Table 1
Atom and bond features incorporated in the GNN.

Atom or bond	Feature	Description	Dimension
Atom features	Atom type	(C, O, N, Cl, F, S, Si, Br, P, Se, I, B, As, Ge, Al)	15
	Ring	Is it in ring?	1
	Aromatic	Is it aromatic?	1
	Hybridisation	(sp, sp^2, sp^3, sp^{3}d)	4
	Bonds	Number of bonds attached (0,1,2,3,4)	5
	Charge	Formal charge (0,1,-1,3)	4
	H's attached	Number of bonded H's (0,1,2,3)	4
Bond features	Bond type	(single, double, triple, aromatic)	4
	Conjugated	Is it conjugated?	1
	Ring	Is it in ring?	1

mean squared error (MSE) was used as the loss function. The training was performed for 100 epochs. A learning rate scheduler was used to decrease the learning rate by a factor of 0.8 using a patience of three epochs. The training was performed independently on 5 different train/validation splits resulting in 5 independent models. The final predictions were made by the ensemble of these 5 models by averaging their individual predictions. All hyperparameters were determined based on ablation studies assessed on the validation set.

4.4. Explainability of GNN predictions

To gain further insights into structural groups that contribute to high compared to low solubilities, we applied the Integrated Gradients (IG) attribution method [28,44] which is only applicable to a binary classification GNN. Therefore, a second GNN (referred to as the "classification GNN") was trained to perform the classification of the molecules into "promising solvent" and "non-promising solvent". The binary classification threshold was set to $\log(x_{\mathrm{sol,\ lignin}})$ = -1.5 according to the solubility values predicted by COSMO-RS. The same train and test splits as for the regression task were used here. For developing the classification GNN the chosen message-passing scheme corresponds to the one proposed by [57] as implemented in *PyTorch geometric*. Two message-passing layers were used with a hidden dimension of 32. Then, a global sum pooling layer was used to obtain the molecular fingerprint. Finally, a MLP of 2-hidden layers with a hidden size of 32 and the ReLU activation function was used to map the fingerprints to the binary solubility classes. The final two neurons of the MLP used the log-softmax activation function. Drop-out with a probability of 0.5 was used after the first hidden layer of the MLP to prevent overfitting. The class with the predicted probability of belonging was selected as the predicted solvent class. The classification GNN was trained using the negative log-likelihood as the loss function and the Adam optimizer. The training ran for 100 epochs with a learning rate of 0.001 and a batch size of 128 graphs. Further information regarding the classification GNN can be found in the SI.

The classification GNN was coupled with the IG method to highlight the structural features of each input graph that were the most relevant for classifying the solvent as "promising" or "not-promising". For this, the IG implementation from *Captum* [58] (version 0.6.0) was used. The corresponding solvent graph with all node features equal to zero was used as a baseline for IG. The attribution scores were normalised for each graph so that they lie between 0 and 1. These scores reflect the least and most important substructures of the graph for predicting the corresponding class, respectively. The default Gauss-Legendre quadrature rule as implemented in *Captum* was used for computing the integral of the gradients. It is important to highlight that the intention of gathering explainability scores by using IG and the classification GNN is to support or guide the scientist in the overall explainability and interpretation tasks. The attribution techniques should not be used as the solely ground truth for scientific discovery. Therefore, the explainability scores described in this section should be taken more as an extra tool to support experimental discovery rather than as the scientific discovery *per se*.

4.5. Implementation of the genetic algorithm

The graph-based genetic algorithm *PSEvolve* iteratively performed evolutionary operations on a given population of start molecules to drive the population towards desired properties. In each iteration, also referred to as generation, the fitness of each molecule was evaluated and molecules with the lowest fitness values were deleted from the population to maintain a constant population size. In this study, the fitness was given by the molar lignin solubility $x_{\mathrm{sol,lignin}}$ as predicted by the GNN. Molecules for cross-over are selected based on their fitness value according to roulette-wheel selection. During cross-over, the selected molecules were fragmented, and the fragments were stored in a mating pool. Children were generated by combining randomly selected parent fragments from the mating pool. The probability of occurring mutations was determined by the mutation rate. In this study, we initialised the algorithm with a start population of 1000 hexane molecules and optimised the structures over the course of 1000 generations. In each generation, 50 parents were chosen for cross-over to produce 100 children. The mutation operation was randomly chosen from the possible operations summarised in Table 2.

The generation of structurally feasible molecules was guaranteed by adhering to the rules of graph and valence counts as implemented in *network* [59] (version 2.6.3) and *rdkit* [52] (version 2022.03.5).

Most of the molecular operations were based on a fragmentation- and a combination algorithm. The fragmentation algorithm split a given molecule into two or more fragments. The molecular graph was searched for so-called "bridges" whose removal would split the graph into unconnected fragments. From the set of identified bridges, one was randomly selected and the corresponding bond is deleted. However, molecules solely composed of rings, do not contain bridges. If no bridge could be identified, the molecule was checked for the occurrence of rings of which one was randomly selected. Bonds shared by multiple rings would not split the graph after their deletion. The shared ring bonds were identified by *rdkit* and excluded from the set of deletable ring bonds. Two deletable bonds were randomly chosen and deleted, leading to a fragmentation of the ring.

The combination algorithm randomly combined two or more fragments to one molecule. All fragments given as input to the combination algorithm were searched for atoms that are able to form additional bonds (implicit valence ≥ 1). In this way, for each fragment, a set of potentially bond-forming atoms is identified. From each set of potentially bond-forming atoms, one was randomly selected and a bond was created.

Atom deletion comes with the risk of deleting atoms that lead to fragmentation of the molecule. Therefore, safe atom deletion relies on the identification of articulation points. Similar to the concept of

Table 2
Overview of the graph-altering operations implemented in *PSEvolve.*

Graph-altering operation	Description
Atom addition	A given molecule is fragmented using the fragmentation algorithm and the atom to be added is treated as another fragment. All fragments are combined using the combination algorithm.
Bond addition	Use the adjacency matrix and implicit valences to identify atoms that are not yet connected and able to form further bonds. From the set of connectable atoms, randomly choose one pair. If the implicit valence is ≥ 2, the algorithm randomly decides whether a single or double bond is formed.
Atom substitution	A randomly chosen atom is substituted by another randomly chosen atom type.
Bond substitution	A randomly chosen bond is substituted by another bond type (single or double bond).
Atom deletion	Identify articulation points and remove the identified atoms from the set of deletable atoms. Randomly choose one deletable atom.
Bond deletion	Identify bridges and remove the identified bond from the set of deletable bonds. Randomly choose one deletable bond.
Relocation	The fragmentation algorithm is used to split a given molecule into several fragments. The combination algorithm randomly combines the fragments.
Addition of functional groups	A given molecule and a randomly chosen functional group (see SI for a list of functional groups) are combined using the combination algorithm.
Cross-over	Parent molecules are selected (roulette-wheel selection). The parents are subjected to the fragmentation and the resulting fragments are stored in the mating pool. Children are generated by randomly selecting two fragments from the mating pool which are subsequently subjected to the combination algorithm.

bridges, articulation points are atoms, whose deletion would split the molecule into fragments. Therefore, the identified articulation points were excluded from the set of deletable atoms. For safe atom deletion, we randomly chose one of the deletable atoms.

For all operations involving bond formation, the algorithm decided randomly between creating single or double bonds, if the implicit valence of the corresponding atoms was at least 2.

PSEvolve can incorporate constraints to further tailor the molecule design to the desired application. In this study, we limited the molecular weight to 200 g/mol. Furthermore, we used the SAS as an additional constraint. The SAS is an estimate for the ease of synthesis of a molecule and ranges from 1 (easily synthesizable) to 10 (difficult to synthesize). Within the genetic algorithm, the SAS score of each individual was constrained to ≤ 3.5. To prevent the design of halogenated solvents, or solvents containing metals, the search space was restricted to C-, O-, H-, N-, S-, and P-atoms only. The maximum molar weight was set to 200 g/mol. To tailor the generation of solvents towards AAF, acid- and aldehyde-instable groups were excluded. We ensured, that the molecular constraints align with the structural features within the GNN training data (e.g. the atom type restriction matches the atom types within the GNN training set) to stay within the applicability range of the GNN (see SI).

The validity of the molecules under the given constraints was verified each time the algorithm introduced new molecules to the population or altered a given structure. Unsuitable molecules were deleted and the operation was performed until a valid structure was generated. In this way, the population size remained constant. All hyperparameters and constraints are given in Table 3. Subsequently, the fitness is evaluated and the described steps are repeated until a defined end criterion is reached. For the setup, the calculation time was ca. 3 h on a standard notebook (Lenovo IdeaPad 5, AMD Ryzen 7 5700U, 16 GB RAM).

Table 3
Hyperparameters and constraints for the genetic algorithm.

Parameter	Description	Value	Unit
Population size	Constant value describing the size of the population	1000	Number of molecules
Molecule types in the start population	N-hexane	–	–
Fitness	Molar lignin solubility $x_{\text{sol, lignin}}$	predicted by GNN	–
Number of parent molecules	Number of parent molecules selected for cross-over	50	Number of molecules
Number of child molecules	Number of child molecules generated by cross-over	100	Number of molecules
Mutation rate	Constant value describing the probability of occurring mutations	0.1	–
Maximum molar weight	Upper bound for molar weight	200	g/mol
SAS	Synthetic accessibility score [32] ranging from 1 (easily synthesisable) to 10 (difficult to synthesize)	3.5	–
Atom type constraints	Only C-, H-, O-, N-, S-atoms used	–	–
Group constraints	For AAF only: acid- and aldehyde-instable groups were eliminated (primary and secondary amines, aldehydes, aromatic N-heterocycles, isocyanates, amides, esters, hydrazides)	–	–
End criterion	Stop after a certain number of generations	1000	Number of generations

4.6. Preparation of lignin samples

In this study, we tested three types of lignin. FABIOLA$^{\text{TM}}$ lignin was isolated from Rettenmaier beechwood by the acetone organosolv process and was generously provided by our collaborator TNO (the Netherlands). Kraft lignin originated from softwood species, was acquired from Berner Fachhochschule. Mild acidolysis lignin (MAL) was isolated from corn cob species using a technique described in detail here [12,60]. Corn cobs were obtained from IP-Suisse in Lausanne (Switzerland), milled using a 6 mm screen and sieved with a 0.45 mm mesh to isolate particles smaller than 0.45 mm, which were used for lignin extraction.

4.7. Experimental lignin solubility measurements

The solubility of lignin samples in the selected solvents was determined experimentally at 85 °C using gas chromatography as described in detail [12]. This method is applicable for solvents that are high-boiling, solid at room temperature and tend to solidify when using traditional gravimetric techniques for measuring solubility of the target solute. The details are provided in SI.

4.8. Aldehyde-assisted pretreatment experiments

Birch wood (*Betula pendula*) was procured from M. Studer of the Bern University of Applied Sciences. The wood chips were sorted to remove residual bark and leaves, then milled using a 6-mm screen and used directly for pretreatment experiments.

The propionaldehyde-assisted pretreatment of birch wood in the selected solvents was performed as described in detail in the past work [14]. Briefly, 4.5 g of milled birch wood, 4.8 ml of propionaldehyde, 0.85 ml of $HCl_{37\%}$, and 25 ml of solvent were added in a thick-walled glass reactor equipped with a stir bar. The reactor was placed in an oil bath heated to 85 °C and the reaction proceeded for 3 h while stirring at 600 rpm. The reaction was cooled to room temperature and 20 ml of 1,4-dioxane was added to the mixture. 0.5 ml aliquot of the reaction liquor was taken, diluted in DMSO and analysed by gas chromatography (GC, Agilent Technologies 7890B) equipped with flame ionisation detector (FID) and HP-5 Column (Agilent) to determine DPX yield using calibration curve method. The yield is provided on a raw biomass basis (non-dried, non-extracted birch wood) accounting for the weight of DPX derived from propionaldehyde [14]. The pretreatment reaction mixture was filtered using a Nylon filter of 0.8 μm to separate cellulose-rich pulp. The pulp was then dried in a vacuum desiccator for 48 h and then weighed. The filtrate was neutralised by gradually adding 0.86 g of sodium bicarbonate solid at room temperature until pH 6–7. The solution was diluted to 100 mL with 1,4-dioxane in a volumetric flask and centrifuged to remove precipitated salt. 20 mL of the 1,4-dioxane/lignin solution was taken for hydrogenolysis in a 50-ml Parr reactor (stainless steel) equipped with a magnetic stirrer and a K-type thermocouple. 100 mg of 5 % Ru/C was added to the solution and the reactor was pressurized to 40 bar of hydrogen gas. The reactor was heated to 250 °C for 3 h. Then, the reactor was cooled down to room temperature, depressurized, and the solution was filtered with 0.2 μm PTFE syringe filter to remove catalyst. 200 μl of an internal standard solution of decane (∼0.05 g/ml) was added to the filtered solution and 1 ml sample was analysed by GC-FID to determine monomer yield. The quantification of monomers was performed using the Effective Carbon Number (ECN) method and described in detail in the past work [14].

Author contributions

LKM performed COSMO-RS lignin solubility predictions and coded the genetic algorithm. EIS implemented the graph neural network and attributions. AOK and LKM contributed to the experiments and analysed the data. All authors contributed to writing the manuscript. JSL and KS supervised the project.

CRediT authorship contribution statement

Laura König-Mattern: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Edgar I. Sanchez Medina:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Anastasia O. Komarova:** Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Investigation, Formal analysis, Conceptualization. **Steffen Linke:** Writing – review & editing, Writing – original draft, Conceptualization. **Liisa Rihko-Struckmann:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Jeremy S. Luterbacher:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Kai Sundmacher:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

Jeremy Luterbacher reports a relationship with Bloom Biorenewables that includes: board membership. Jeremy Luterbacher has patent #WO2017178513A1 pending to EPFL. If there are other authors, they 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

PSEvolve as used in the manuscript and the GNN for lignin solubility predictions are available under: https://github.com/koenigmattern/PSEvolve_lignin_solvents. *PSEvolve* is available under: https://github.com/koenigmattern/PSEvolve.

Acknowledgements

We acknowledge funding by the Christiane Nüsslein-Volhard foundation. Open Access funding was provided by the Max Planck Society. This work was partly supported by the Research Initiative "SmartProSys: Intelligent Process Systems for the Sustainable Production of Chemicals" funded by the Ministry for Science, Energy, Climate Protection and the Environment of the State of Saxony-Anhalt. This publication was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation. The work was also supported by Grants 200021_182605 and CRSII5_180258 and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 945363. Edgar Ivan Sanchez Medina is affiliated with the International Max Planck Research School for Advanced Methods in Process and Systems Engineering–IMPRS ProEng at the Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg. We thank IP-Suisse for providing corn cobs, Dr. M. Studer for providing Birch wood, TNO and Berner Fachhochschule for providing lignin samples.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cej.2024.153524.

References

[1] R.C. Kuhad, A. Singh, Lignocellulose biotechnology: current and future prospects, Crit. Rev. Biotechnol. 13 (1993) 151–172, https://doi.org/10.3109/07388559309040630.

[2] L. Shuai, M.T. Amiri, Y.M. Questell-Santiago, F. Héroguel, Y. Li, H. Kim, R. Meilan, C. Chapple, J. Ralph, J.S. Luterbacher, Formaldehyde stabilization facilitates lignin monomer production during biomass depolymerization, Science 354 (2016) 329–333, https://doi.org/10.1126/science.aaf7810.

[3] S. Van den Bosch, W. Schutyser, R. Vanholme, T. Driessen, S.-F. Koelewijn, T. Renders, B. De Meester, W.J.J. Huijgen, W. Dehaen, C.M. Courtin, B. Lagrain, W. Boerjan, B.F. Sels, Reductive lignocellulose fractionation into soluble lignin-derived phenolic monomers and dimers and processable carbohydrate pulps, Energ. Environ. Sci. 8 (2015) 1748–1763, https://doi.org/10.1039/C5EE00204D.

[4] Y. Liu, N. Deak, Z. Wang, H. Yu, L. Hameleers, E. Jurak, P.J. Deuss, K. Barta, Tunable and functional deep eutectic solvents for lignocellulose valorization, Nat. Commun. 12 (2021) 5424, https://doi.org/10.1038/s41467-021-25117-1.

[5] W. Lan, M.T. Amiri, C.M. Hunston, J.S. Luterbacher, Protection group effects during α, γ-diol lignin stabilization promote high-selectivity monomer production, Angew. Chem. Int. Ed. 57 (2018) 1356–1360, https://doi.org/10.1002/anie.201710838.

[6] A. De Santi, M.V. Galkin, C.W. Lahive, P.J. Deuss, K. Barta, Lignin-first fractionation of softwood lignocellulose using a mild dimethyl carbonate and ethylene glycol organosolv process, ChemSusChem 13 (2020) 4468–4477, https://doi.org/10.1002/cssc.201903526.

[7] M.M. Abu-Omar, Guidelines for performing lignin-first biorefining, Environ. Sci. (2021) 31.

[8] C. Gioia, M. Colonna, A. Tagami, L. Medina, O. Sevastyanova, L.A. Berglund, M. Lawoko, Lignin-based epoxy resins: unravelling the relationship between structure and material properties, Biomacromolecules 21 (2020) 1920–1928, https://doi.org/10.1021/acs.biomac.0c00057.

[9] P. Figueiredo, M.H. Lahtinen, M.B. Agustin, D.M. De Carvalho, S. Hirvonen, P. A. Penttilä, K.S. Mikkonen, Green fabrication approaches of lignin nanoparticles from different technical lignins: a comparison study, ChemSusChem 14 (2021) 4718–4730, https://doi.org/10.1002/cssc.202101356.

[10] A. Manisekaran, P. Grysan, B. Duez, D.F. Schmidt, D. Lenoble, J.-S. Thomann, Solvents drive self-assembly mechanisms and inherent properties of kraft lignin nanoparticles (<50 nm), J. Colloid Interface Sci. 626 (2022) 178–192, https://doi.org/10.1016/j.jcis.2022.06.089.

[11] J. Ruwoldt, F.H. Blindheim, G. Chinga-Carrasco, Functional surfaces, films, and coatings with lignin – a critical review, RSC Adv. 13 (2023) 12529–12553, https://doi.org/10.1039/D2RA08179B.

[12] L. König-Mattern*, A.O. Komarova*, A. Ghosh, S. Linke, L.K. Rihko-Struckmann, J. Luterbacher, K. Sundmacher, High-throughput computational solvent screening for lignocellulosic biomass processing, Chemical Engineering Journal 452 (2023) 139476. https://doi.org/10.1016/j.cej.2022.139476.

[13] CompTox Chemicals Dashboard, (n.d.). https://comptox.epa.gov/dashboard/ (accessed September 27, 2023).

[14] M. Talebi Amiri, G.R. Dick, Y.M. Questell-Santiago, J.S. Luterbacher, Fractionation of lignocellulosic biomass to produce uncondensed aldehyde-stabilized lignin, Nat. Protoc. 14 (2019) 921–954, https://doi.org/10.1038/s41596-018-0121-7.

[15] W. Schutyser, S. Van Den Bosch, T. Renders, T. De Boe, S.-F. Koelewijn, A. Dewaele, T. Ennaert, O. Verkinderen, B. Goderis, C.M. Courtin, B.F. Sels, Influence of bio-based solvents on the catalytic reductive fractionation of birch wood, Green Chem. 17 (2015) 5035–5045, https://doi.org/10.1039/C5GC01442E.

[16] C. Balaji, T. Banerjee, V.V. Goud, COSMO-RS based predictions for the extraction of lignin from lignocellulosic biomass using ionic liquids: effect of cation and anion combination, J. Solut. Chem. 41 (2012) 1610–1630, https://doi.org/10.1007/s10953-012-9887-3.

[17] Y. Chu, X. He, MoDoop: an automated computational approach for COSMO-RS prediction of biopolymer solubilities in ionic liquids, ACS Omega 4 (2019) 2337–2343, https://doi.org/10.1021/acsomega.8b03255.

[18] L. König-Mattern, S. Linke, L. Rihko-Struckmann, K. Sundmacher, Computer-aided solvent screening for the fractionation of wet microalgae biomass, Green Chem. (2021) 10.1039.D1GC03471E. https://doi.org/10.1039/D1GC03471E.

[19] A. Klamt, Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena, J. Phys. Chem. 99 (1995) 2224–2235, https://doi.org/10.1021/j100007a062.

[20] A. Klamt, V. Jonas, T. Bürger, J.C.W. Lohrenz, Refinement and parametrization of COSMO-RS, Chem. A Eur. J. 102 (1998) 5074–5085, https://doi.org/10.1021/jp980017s.

[21] A. Klamt, F. Eckert, COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids, Fluid Phase Equilib. 172 (2000) 43–72, https://doi.org/10.1016/S0378-3812(00)00357-5.

[22] COSMOtherm, Release 19; © 2019 COSMOlogic GmbH & Co. KG, a Dassault Systèmes company, (n.d.).

[23] L. König-Mattern, A.O. Komarova, A. Ghosh, S. Linke, L.K. Rihko-Struckmann, J. Luterbacher, K. Sundmacher, High-throughput computational solvent screening for lignocellulosic biomass processing, Chem. Eng. J. 452 (2023) 139476, https://doi.org/10.1016/j.cej.2022.139476.

[24] D.C. Elton, Z. Boukouvalas, M.D. Fuge, P.W. Chung, Deep learning for molecular design—a review of the state of the art, Mol. Syst. Des. Eng. 4 (2019) 828–849, https://doi.org/10.1039/C9ME00039A.

[25] V. Venkatasubramanian, K. Chan, J.M. Caruthers, Computer-aided molecular design using genetic algorithms, Comput. Chem. Eng. 18 (1994) 833–844, https://doi.org/10.1016/0098-1354(93)E0023-3.

[26] Z. Sumer, R.C. Van Lehn, Heuristic computational model for predicting lignin solubility in tailored organic solvents, ACS Sustain. Chem. Eng. 11 (2023) 187–198, https://doi.org/10.1021/acssuschemeng.2c05199.

[27] E.C. Achinivu, M. Mohan, H. Choudhary, L. Das, K. Huang, H.D. Magurudeniya, V. R. Pidatala, A. George, B.A. Simmons, J.M. Gladden, A predictive toolset for the identification of effective lignocellulosic pretreatment solvents: a case study of solvents tailored for lignin extraction, Green Chem. 23 (2021) 7269–7289, https://doi.org/10.1039/D1GC01186C.

[28] B. Sanchez-Lengeling, J. Wei, B. Lee, E. Reif, P. Wang, W. Qian, K. McCloskey, L. Colwell, A. Wiltschko, Evaluating attribution for graph neural networks, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020: pp. 5898–5910. https://proceedings.neurips.cc/paper_files/paper/2020/file/417fbbf2e9d5a28a855a11894b2e795a-Paper.pdf.

[29] M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N.C. Frey, P. Friederich, T. Gaudin, A.A. Gayle, K.M. Jablonka, R.F. Lameiro, D. Lemm, A. Lo, S.M. Moosavi, J. M. Nápoles-Duarte, A. Nigam, R. Pollice, K. Rajan, U. Schatzschneider, P. Schwaller, M. Skreta, B. Smit, F. Strieth-Kalthoff, C. Sun, G. Tom, G. Falk Von Rudorff, A. Wang, A.D. White, A. Young, R. Yu, A. Aspuru-Guzik, SELFIES and the future of molecular string representations, Patterns 3 (2022) 100588, https://doi.org/10.1016/j.patter.2022.100588.

[30] M. Krenn, F. Häse, A. Nigam, P. Friederich, A. Aspuru-Guzik, Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation, Mach. Learn.: Sci. Technol. 1 (2020) 045024, https://doi.org/10.1088/2632-2153/aba947.

[31] A.K. Nigam, R. Pollice, M. Krenn, G.D.P. Gomes, A. Aspuru-Guzik, Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES, Chem. Sci. 12 (2021) 7079–7090, https://doi.org/10.1039/D1SC00231G.

[32] P. Ertl, A. Schuffenhauer, Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions, J Cheminform 1 (2009) 8, https://doi.org/10.1186/1758-2946-1-8.

[33] J.O. Spiegel, J.D. Durrant, AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization, J Cheminform 12 (2020) 25, https://doi.org/10.1186/s13321-020-00429-4.

[34] R. Laplaza, S. Gallarati, C. Corminboeuf, Genetic optimization of homogeneous catalysts, Chemistry Methods 2 (2022) e202100107.

[35] T. Zhou, J. Wang, K. McBride, K. Sundmacher, Optimal design of solvents for extractive reaction processes, AIChE J 62 (2016) 3238–3249, https://doi.org/10.1002/aic.15360.

[36] K. Wang, F. Xu, R. Sun, Molecular characteristics of kraft-AQ pulping lignin fractionated by sequential organic solvent extraction, IJMS 11 (2010) 2988–3001, https://doi.org/10.3390/ijms11082988.

[37] Z. Michalewicz, M. Schoenauer, Evolutionary algorithms for constrained parameter optimization problems, Evol. Comput. 4 (1996) 1–32, https://doi.org/10.1162/evco.1996.4.1.1.

[38] R. Rinaldi, R. Jastrzebski, M.T. Clough, J. Ralph, M. Kennema, P.C.A. Bruijnincx, B. M. Weckhuysen, Paving the way for lignin valorisation: recent advances in bioengineering, biorefining and catalysis, Angew. Chem. Int. Ed. 55 (2016) 8164–8215, https://doi.org/10.1002/anie.201510351.

[39] J. Ralph, C. Lapierre, W. Boerjan, Lignin structure and its engineering, Curr. Opin. Biotechnol. 56 (2019) 240–249, https://doi.org/10.1016/j.copbio.2019.02.019.

[40] Dassault Systèmes company, COSMOtherm Reference Manual, (2019).

[41] J. Leguy, T. Cauchy, M. Glavatskikh, B. Duval, B. Da Mota, EvoMol: a flexible and interpretable evolutionary algorithm for unbiased de novo molecular generation, J Cheminform 12 (2020) 55, https://doi.org/10.1186/s13321-020-00458-z.

[42] A. Dastpak, T.V. Lourencon, M. Balakshin, S. Farhan Hashmi, M. Lundström, B. P. Wilson, Solubility study of lignin in industrial organic solvents and investigation of electrochemical properties of spray-coated solutions, Ind. Crop Prod. 148 (2020) 112310, https://doi.org/10.1016/j.indcrop.2020.112310.

[43] J. Sameni, S. Krigstin, M. Sain, Solubility of lignin and acetylated lignin in organic solvents, BioResources 12 (2017) 1548–1565, https://doi.org/10.15376/biores.12.1.1548-1565.

[44] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, (2017). https://doi.org/10.48550/ARXIV.1703.01365.

[45] A.T. Smit, M. Verges, P. Schulze, A. Van Zomeren, H. Lorenz, Laboratory- to pilot-scale fractionation of lignocellulosic biomass using an acetone organosolv process, ACS Sustain. Chem. Eng. 10 (2022) 10503–10513, https://doi.org/10.1021/acssuschemeng.2c01425.

[46] L. Petridis, J.C. Smith, Molecular-level driving forces in lignocellulosic biomass deconstruction for bioenergy, Nat. Rev. Chem. 2 (2018) 382–389, https://doi.org/10.1038/s41570-018-0050-6.

[47] A.O. Komarova, G.R. Dick, J.S. Luterbacher, Diformylxylose as a new polar aprotic solvent produced from renewable biomass, Green Chem. 23 (2021) 4790–4799, https://doi.org/10.1039/D1GC00641J.

[48] L.P. Manker, G.R. Dick, A. Demongeot, M.A. Hedou, C. Rayroud, T. Rambert, M. J. Jones, I. Sulaeva, M. Vieli, Y. Leterrier, A. Potthast, F. Maréchal, V. Michaud, H.-A. Klok, J.S. Luterbacher, Sustainable polyesters via direct functionalization of lignocellulosic sugars, Nat. Chem. 14 (2022) 976–984, https://doi.org/10.1038/s41557-022-00974-5.

[49] Y.M. Questell-Santiago, J.H. Yeap, M. Talebi Amiri, B.P. Le Monnier, J. S. Luterbacher, Catalyst evolution enhances production of xylitol from acetal-stabilized xylose, ACS Sustainable, Chem. Eng. 8 (2020) 1709–1714, https://doi.org/10.1021/acssuschemeng.9b06456.

[50] L. Huang, Z. Bian, D. Li, X. Cheng, X. Luo, L. Shuai, J. Liu, Catalytic conversion of diformylxylose to furfural in biphasic solvent systems, Front. Bioeng. Biotechnol. 11 (2023) 1146250, https://doi.org/10.3389/fbioe.2023.1146250.

[51] J.V. Vermaas, M.F. Crowley, G.T. Beckham, Molecular lignin solubility and structure in organic solvents, ACS Sustain. Chem. Eng. 8 (2020) 17839–17850, https://doi.org/10.1021/acssuschemeng.0c07156.

[52] RDKit: Open-source cheminformatics, (2022). https://www.rdkit.org (accessed June 13, 2022).

[53] E.I. Sanchez Medina, S. Linke, M. Stoll, K. Sundmacher, Graph neural networks for the prediction of infinite dilution activity coefficients, Digital, Discovery 1 (2022) 216–225, https://doi.org/10.1039/D1DD00037C.

[54] E.I. Sanchez Medina, S. Linke, M. Stoll, K. Sundmacher, Gibbs-Helmholtz graph neural network: capturing the temperature dependency of activity coefficients at infinite dilution, Digital, Discovery 2 (2023) 781–798, https://doi.org/10.1039/D2DD00142J.

[55] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, (2017). https://doi.org/10.48550/ARXIV.1704.01212.

[56] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, (2015). https://doi.org/10.48550/ARXIV.1502.03167.

[57] C. Morris, M. Ritzert, M. Fey, W.L. Hamilton, J.E. Lenssen, G. Rattan, M. Grohe, Weisfeiler and leman go neural: higher-order graph neural networks, (2018). https://doi.org/10.48550/ARXIV.1810.02244.

[58] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, O. Reblitz-Richardson, Captum: A unified and generic model interpretability library for PyTorch, (2020). https://doi.org/10.48550/ARXIV.2009.07896.

[59] A. Hagberg, P.J. Swart, D.A. Schult, Exploring network structure, dynamics, and function using NetworkX, in: United States, 2008. https://www.osti.gov/servlets/purl/960616.

[60] A. Das, A. Rahimi, A. Ulbrich, M. Alherech, A.H. Motagamwala, A. Bhalla, L. da Costa Sousa, V. Balan, J.A. Dumesic, E.L. Hegg, B.E. Dale, J. Ralph, J.J. Coon, S. S. Stahl, Lignin conversion to low-molecular-weight aromatics via an aerobic oxidation-hydrolysis sequence: comparison of different lignin sources, ACS Sustain. Chem. Eng. 6 (2018) 3367–3374, https://doi.org/10.1021/acssuschemeng.7b03541.