ACS Sustainable Chemistry & Engineering

pubs.acs.org/journal/ascecg

Research Article

Machine Learning Optimization of Lignin Properties in Green Biorefineries

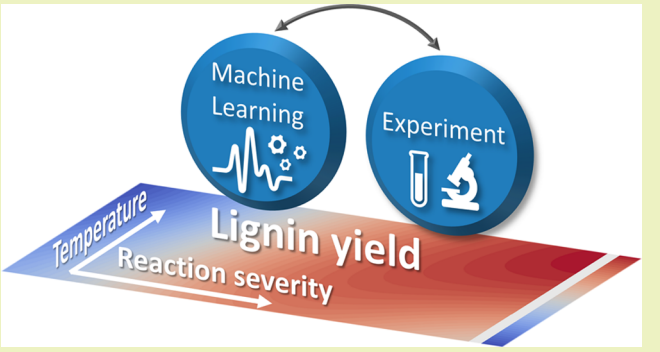
Joakim Löfgren, Dmitry Tarasov, Taru Koitto, Patrick Rinke, Mikhail Balakshin,* and Milica Todorović*

Cite This: *ACS Sustainable Chem. Eng.* 2022, 10, 9469−9479

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Novel biorefineries could transform lignin, an abundant biopolymer, from side-stream waste to high-value-added byproducts at their site of production and with minimal experiments. Here, we report the optimization of the AquaSolv omni biorefinery for lignin using Bayesian optimization, a machine learning framework for sample-efficient and guided data collection. This tool allows us to relate the biorefinery conditions like hydrothermal pretreatment reaction severity and temperature with multiple experimental outputs, such as lignin structural features characterized using 2D nuclear magnetic resonance spectroscopy. By applying a Pareto front analysis to our models, we can find the processing conditions that simultaneously optimize the lignin yield and the amount of β-O-4 linkages for the depolymerization of lignin into platform chemicals. Our study demonstrates the potential of machine learning to accelerate the development of sustainable chemical processing techniques for targeted applications and products.



KEYWORDS: *Biomass, Valorization, Lignin, Biorefinery, Machine learning, Bayesian optimization*

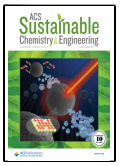
INTRODUCTION

The transition to a green and sustainable economy requires efficient utilization of our natural resources. Lignin, as a part of lignocellulosic biomatter, is an example of a naturally abundant but under-utilized resource.[1,2] Lignin is currently produced in large quantities as a side stream of pulping processes. Valorization of lignin into high-value-added industrially relevant byproducts, materials,[3,4] or chemicals[5−8] can therefore substantially increase both the sustainability and revenue of biorefineries.[4,9] Recently, some of us have developed AquaSolv Omni (AqSO), a green biorefinery for the integrated utilization of all biomass components, with special focus on the lignin-containing streams.[10,11] In the AqSO process, hydrothermal treatment (HTT) is first applied to the biomass feedstock, and the resulting solids are subsequently washed with a solvent to extract lignin. The aqueous environment of HTT eliminates the need for extra chemicals in the reaction medium, which makes AqSO a green process given a suitable choice of solvent.

Key strengths of the AqSO biorefinery are the highly tunable processing conditions, which enable high output versatility in terms of lignin composition, structure, and physicochemical properties. To make the biorefinery efficient and more profitable, the processing conditions must be optimized to ensure that the properties of the extracted lignin are ideal for the intended end product. In general, optimizing a biomass valorization process is a difficult task since it requires knowledge of how the processing conditions relate to the product properties. The same issue is encountered during the development of new valorization approaches: learning the connections between processing conditions and product properties requires significant time and resources. In particular for the valorization of lignin, the situation is further exacerbated by its complex chemical structure.[12]

To solve experimental optimization tasks, design of experiment (DOE) methods[13,14] are frequently employed. DOE methods provide general strategies for planning data collection and for modeling experimental output. In this context, we refer to the experimental output being modeled as the design target, abbreviated as target, and the range of processing conditions being considered as the design space. Since laboratory-based experiments are often time consuming and costly, DOE also strives for efficient sampling of the design space. Conventional approaches to DOE include space-filling designs,[15,16] factorial designs,[17] and response surface methods,[18,19] where the latter category is frequently regarded as an industry standard. A shortcoming of many conventional DOE

Received: March 31, 2022
Revised: June 23, 2022
Published: July 12, 2022



approaches is the predetermined fashion in which they sample design space, which does not allow for information from new measurements to be utilized to improve the sampling strategy. Predetermined sampling strategies also tend to sample excessively in regions where the design target is relatively uniform and insufficiently in regions with significant fluctuations.

We can avoid the shortcomings of classical DOE methods by turning to machine learning methods, which have solved a wide range of challenging modeling and optimization problems in recent decades.[20−22] Of particular relevance in the DOE context is Bayesian optimization (BO), an algorithm for autonomous complex optimization tasks. In addition to providing efficient optimization of arbitrary functions, BO constructs so-called surrogate models of design targets in the space of the design variables.[23,24] The surrogate models are constructed sequentially through a data collection policy known as the acquisition function, which ensures that the new samples are selected to be as informative as possible given the current state of the model. Consequently, BO is more sample efficient compared to traditional DOE methods and supports the incorporation of experimental noise, prior knowledge, and parallel experiments through batch acquisition functions.[25] Recently, BO has become a popular tool in computational materials science, where it has been used both as a means of efficient global optimization[26−28] and to guide materials discovery.[29] BO is also increasingly being used for optimization in DOE applications.[30−34] Here, we highlight the power of BO's surrogate model aspect, which provides us with insight into lignin chemistry and facilitates easy multiobjective optimization for green chemistry applications.

The overarching goal of our work is to optimize the AqSO biorefinery by making it more resource efficient and capable of producing lignin with structural properties tailored to specific applications in lignin valorization. Our first step is to use BO to obtain surrogate models that relate the key AqSO processing conditions, namely, the hydrothermal pretreatment temperature and reaction severity, to multiple experimental targets. These consist of the yield of the extracted lignin and the content of various structural properties obtained using 2D nuclear magnetic resonance (NMR) characterization. For the latter, we focus on the β-O-4 content, the ratio of syringyl and guaiacyl (S/G) units, and the content of carbohydrates present as lignin−carbohydrate complexes (LCCs). To accomplish our stated goal, we then apply a Pareto front analysis[35] to the surrogate models, which allows us to simultaneously maximize the lignin yield and optimize structural properties for selected applications. By ensuring that the lignin yield is maximized, we achieve better resource efficiency in the AqSO process.

The interplay between laboratory experiments and machine learning in our approach is conceptually illustrated in Figure 1. Processing conditions for new experiments are first selected by the BO algorithm. After the recommended experiments have been performed and experimental targets recorded, the results are used to update the BO model so that the cycle can start over.

Another important objective in this study is to understand how to efficiently apply a BO-guided workflow to optimize multiple experimental targets. Specifically, we consider solutions where the data collection is performed in batches to be more compatible with experimental workflows. With traditional BO, we would collect data sequentially and build surrogate models for each experimental target independently.

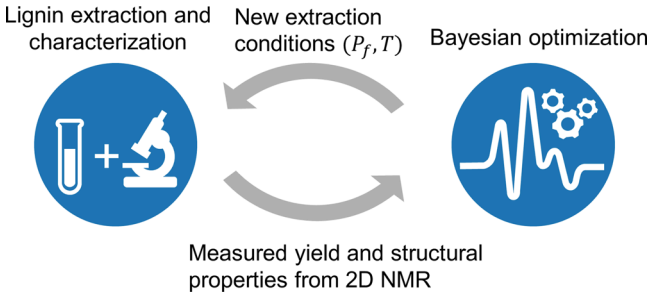


Figure 1. Machine learning-driven workflow for designing the AqSO biorefinery. Values for the HTT reactor's temperature and P-factor are selected using BO. After performing an extraction at these conditions, the lignin yield and structural properties are measured and fed back to the BO model. The new data (extraction conditions and measured properties) are used to update the program, and then, the cycle starts over.

This is not convenient for experimental data gathering, especially since each sample characterization produces information on multiple targets. We also consider a strategy for shared data collection where the experiments recommended to advance different targets are combined into one batch and performed at the same time. This conveniently drives the data collection toward different areas of the design space, and we deploy two acquisition strategies in parallel to further boost exploration. Batch characterization outcomes are then used to advance the models over all experimental targets.

MATERIALS AND METHODS

Materials and Chemicals. We debarked, chipped, and ground a birch wood (Betula sp.) stem into sawdust (0.5−1.5 mm particle were size selected). Prior to the HTT, we subjected the sawdust to acetone (100%) extraction to remove the lipophilic extractives and eliminate their influence on the determination of lignin yield and structure. We purchased acetone (C_3H_6O, 95 vol %) and used it without purification, as well as deuterated dimethyl sulfoxide (DMSO-d6) for NMR spectroscopy. Both chemicals were of analytical grade and purchased from Sigma-Aldrich.

Hydrothermal Treatment. We applied HTT to the extractive-free sawdust (4 g) in a swing reactor equipped with temperature controls both in the heating block and inside the reactor. A schematic overview of the AqSO process is shown in Figure 2a. Several adjustable parameters influence the outcome of the lignin extraction and its structure, most importantly the reactor temperature (T), residence time (t_f), and liquid to solid ratio L:S. Since the heating period significantly affects the extraction process, we work with the reaction severity instead of the residence time. The HTT reaction severity is quantified in terms of the P-factor, which is calculated according to[36]

$$P_f[T(t), t_f] = \int_0^{t_f} \frac{k(T(t))}{k(100°)}\mathrm{dt} = \int_0^{t_f} e^{40.48-15106/T(t)}\mathrm{dt} \quad (1)$$

Here, the residence time is measured in hours, and the temperature is given in kelvin. The rate constant k of the reaction is assumed to obey the Arrenhius equation with an activation energy of 125.6 kJ mol^{-1}.[36] We employed a fixed liquid-to-solid ratio L:S = 1, motivated by our earlier investigation of the AqSO biorefinery that indicated this ratio to be favorable for the yield and structural properties of the lignin-containing products.[11]

Once the desired severity was reached, we immediately transferred the reactor into cold water. We subsequently separated the HTT solids and hydrolysate by filtration using a glass crucible (pore size 3 μm) and exhaustively washed the solids with deionized water. We then extracted lignin from the washed HTT solids with 90% (v/v)

https://doi.org/10.1021/acssuschemeng.2c01895
ACS Sustainable Chem. Eng. 2022, 10, 9469−9479

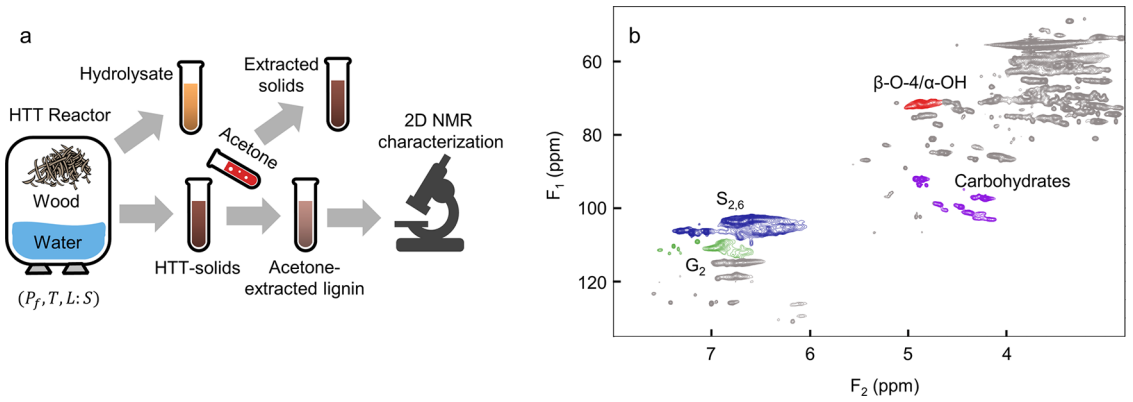


Figure 2. (a) Schematic of experimental setup employed for the optimization of AqSO biorefinery. A mixture of birch sawdust and water is subjected to hydrothermal treatment (HTT) in a reactor whose state can be described by the reactor temperature (T), liquid-to-solid ratio ($L{:}S$ = 1), and P-factor (P_f). The reaction produces a hydrolysate and HTT solids. Extraction of the HTT solids with aqueous acetone results in acetone-extracted lignin and extracted solids. The structural properties of the acetone-extracted lignin are characterized with 2D NMR spectroscopy. The hydrolysate and extracted solids are not addressed in this study. (b) Expanded region of interest in a heteronuclear single quantum coherence (HSQC) spectrum of acetone-extracted lignin produced at P_f = 500 and T = 180 °C. The amounts of β-O-4 linkages, S/G ratio, and LCC content were quantified by volume integration of the signal regions highlighted in the spectrum.

aqueous acetone. We rotary-evaporated the solution (at 40 °C) to produce acetone-extracted lignin, which was finally vacuum dried at 40 °C to constant weight to determine the yield.

2D HSQC NMR Analysis. We recorded HSQC NMR spectra using a Bruker AVANCE 600 NMR spectrometer equipped with a CryoProbe. We dissolved approximately 80 mg of each sample in 0.6 mL of dimethyl sulfoxide-d_6 (DMSO). We set an acquisition time of 77.8 ms for the 1H-dimension and collected 36 scans per block using 1024 complex data points. For the 13C-dimension, we set the acquisition time to 3.94 ms and recorded 256 time increments. We then processed the 2D HSQC NMR data (1024 × 1024 data points) by applying a QSINE window function to both the 1H and 13C dimensions. For calibration, we used the DMSO peak at δC = 39.5 ppm and δH = 2.49 ppm. To quantify the specific lignin moieties, we carried out volume integration of the HSQC spectra of the acetone-extracted lignin (Figure 2b). This procedure is described in greater detail elsewhere.[37−40] The intensities of the signals are expressed in mol %, i.e., per 100 aromatic units (Ar) assuming the sum of the signals of guaiacyl (G) and syringyl (S) units as 100% from their characteristic CH signals at G_2 and $S_{2,6}$ positions, correspondingly: $G_2 + S_{2,6}/2$ = 100%.

The characterized moieties analyzed in this paper consist of the number of β-O-4 linkages, the ratios of syringyl to guaiacyl (S/G) units, and the carbohydrate content (present as LCCs). β-O-4 linkages are the main reactive centers in native lignin, and the extent to which they are present in lignin/LCCs after the HTT extraction characterizes the degree of lignin transformation. In addition, good correlations between the amounts of β-O-4 linkages and other lignin characteristics have previously been reported for the AqSO process.[11] We note that 2D HSQC NMR analysis also lets us determine the content of several other moieties, the analysis of which has been omitted for brevity in the present work.

Bayesian Optimization. Here, we provide a short overview of Bayesian optimization and refer readers to the SI and extensive literature for more detailed accounts.[24,41] BO involves two main components, namely, a surrogate model that approximates the target function and an acquisition function that provides a data collection policy. During a BO iteration, the surrogate model is fit to the current data set using Gaussian process regression. The posterior mean of the Gaussian process represents the most probable approximation of the target, and the posterior variance provides a measure of the model uncertainty. By minimizing the acquisition function, a new sampling location is determined and used to augment the existing data set.

Acquisition functions come in many flavors that provide different trade-offs between exploitation and exploration. Exploitation refers to sampling regions of design space where the target is likely to achieve a minimum or maximum, while exploration visits regions of high model uncertainty where data have not been acquired before. In this work, we generate acquisitions with the exploration-modified lower confidence bound (eLCB) function[41,42] as well as through the standard deviation of the model (which we term the pure exploration function).

We carried out BO using the recently released BOSS code,[28,43] which provides a Python-based implementation of BO for applications in materials science. BOSS has previously been applied to a range of different problems in materials modeling.[44−47] The Gaussian processes used by BOSS were assigned uninformative zero priors for the mean functions, and radial basis set kernels, which reflect the smoothness of the materials properties, were used for the covariance functions. We initialized the kernel hyperparameters using inverse gamma priors and subsequently updated them during the BO process by maximizing the marginal likelihood. Initial data for the surrogate models were obtained from a batch of five Sobol points.[48] To account for the measurement uncertainty in the targets and achieve a better statistical fit to the observed data, we incorporated Gaussian noise terms with zero mean in the surrogate models. The standard deviations of the noise terms were chosen to reflect the estimated measurement errors of 5% for the lignin yield and 5%−10% for the structural properties.

Applying Bayesian Optimization to Experiments. When applying BO in an experimental context, the target function can be any measurable output from the experiment. We use the BO process to determine how the target function depends on the design variables, such as adjustable experimental parameters, while performing as few experiments as possible. During a BO iteration, new experiments are performed using the values for the design variables suggested by the acquisition function. The measured experimental target values are then used together with the processing conditions to update the surrogate model (Figure 1). The next iteration then proceeds using the updated surrogate model, and this cycle is repeated until convergence is observed. Next, we consider the practical details of implementing this workflow.

We focus on several target lignin properties that include the extracted lignin yield, the number of β-O-4 linkages, the S/G ratio, and the carbohydrate content. The design variables correspond to the HTT processing conditions, namely, the P-factor and temperature (Figure 2). The P-factor and temperature ranges we investigated were determined based on a combination of the feasible operating range of the experimental apparatus and processing conditions that were expected to give appreciable lignin yield based on previous work:[10] $500 \le P_f \le 2500$ and $180 \le T \le 210$ (°C). The two variables and

https://doi.org/10.1021/acssuschemeng.2c01895
ACS Sustainable Chem. Eng. 2022, 10, 9469−9479

their corresponding limits make up the design space of our optimization problem.

We consider two different ways of adjusting the sampling of the design space to a setting with multiple experimental targets: pure and combined acquisition strategies. In the pure acquistions (PA) strategy, we carry out separate BO processes for each target function. This means that independent acquisitions are made for each target function and used to update the corresponding surrogate model at every iteration (Figure 3a) much like in conventional BO. With the combined acquistions (CA) strategy, our aim is to investigate if the optimization can be accelerated by allowing data exchange between two simultaneous BO processes. Here, data acquired for different target functions are combined and used to update all existing surrogate models at every BO iteration (Figure 3b).

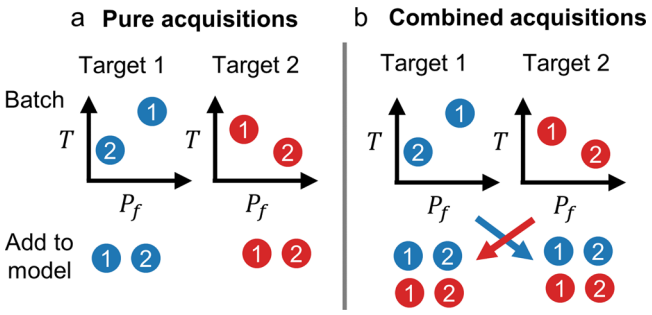


Figure 3. Two strategies for updating surrogate models for different target functions. (a) In the pure acquisitions strategy, only acquisitions (red and blue circles) made for a certain target are used to update the surrogate model for that target. (b) In the combined acquisitions strategy, acquisitions for different targets are pooled together so that each surrogate model is updated using the same set of acquisitions.

Another important consideration is which, and how many, target functions to include in the BO acquisition process. For simplicity, we elected to focus on two important targets: lignin yield and β-O-4 content. Given the important role that β-O-4 moieties play in chemical reactions, this will help us produce high yields of chemicals to target different lignin applications. We emphasize that, in addition to the lignin yield and β-O-4 content, each experiment performed also provided data for the S/G ratio and the carbohydrate content. This enabled us to train surrogate models for all these targets as well, even if they were not actively employed to generate new acquisitions.

We must also select appropriate acquisition functions for sampling. Our aim is to fit surrogate models that are predictive over the entire design space, which necessitates a certain degree of exploration. However, we are also interested in quickly identifying extremal regions (minima or maxima) since many applications require, for example, a high yield. To accomplish both these goals, we chose to perform two data acquisitions at each iteration: one data point suggested by the pure exploration function and one by the eLCB function. Using two different acquisition functions with two different experimental targets means that one BO iteration entails a batch of four experiments in total, which can be carried out in parallel in the laboratory to save time.

Pareto Front Analysis. Once we obtain converged surrogate models for the experimental targets, we can analyze them to establish the different processing conditions at which each individual lignin property is optimal. In this work, we go one step further and consider which experimental conditions optimize multiple lignin properties at once. In general, a single solution does not exist for such a multitarget optimization problem. Instead, we must look for optimal trade-offs between the targets involved.

Mathematically, the notion of an optimal trade-off is formalized by the concept of Pareto optimality (see SI for an extended description).[35] The Pareto theory tells us that a combination of target values constitutes an optimal trade-off, if an improvement in one target is always detrimental to at least one other target. Consider, for instance, the lignin yield and the number of β-O-4 linkages. If we

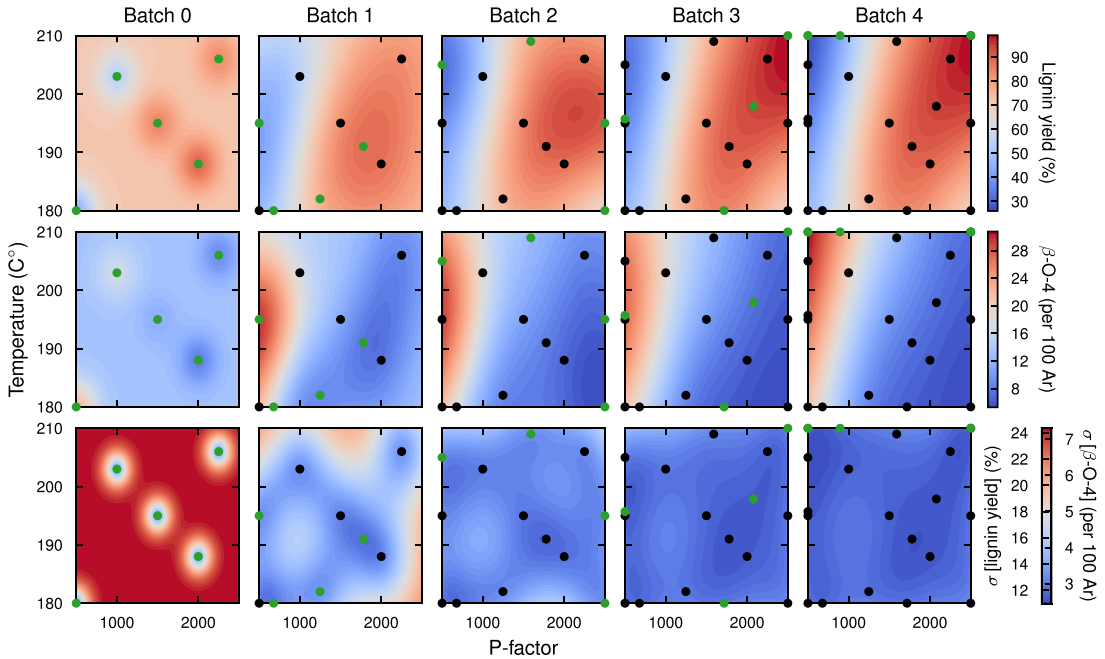


Figure 4. Successive improvement of the surrogate models for the lignin yield (top) and β-O-4 content (middle) as new acquisitions (green circles) are added to the existing data set (black circles). The panels display the surrogate models after adding batches 1, 2, 3, and 4 (fitted with 9, 13, 17, and 21 data points, respectively). As the landscapes evolve, the models' capability of predicting the yield in unknown regions of design space increases. The prediction uncertainties are quantified by the model standard deviation (bottom) which decreases as more data are collected. The acquisition strategy balances exploitation of regions where the yield is large and exploration of regions with high uncertainty.

https://doi.org/10.1021/acssuschemeng.2c01895
ACS Sustainable Chem. Eng. 2022, 10, 9469−9479

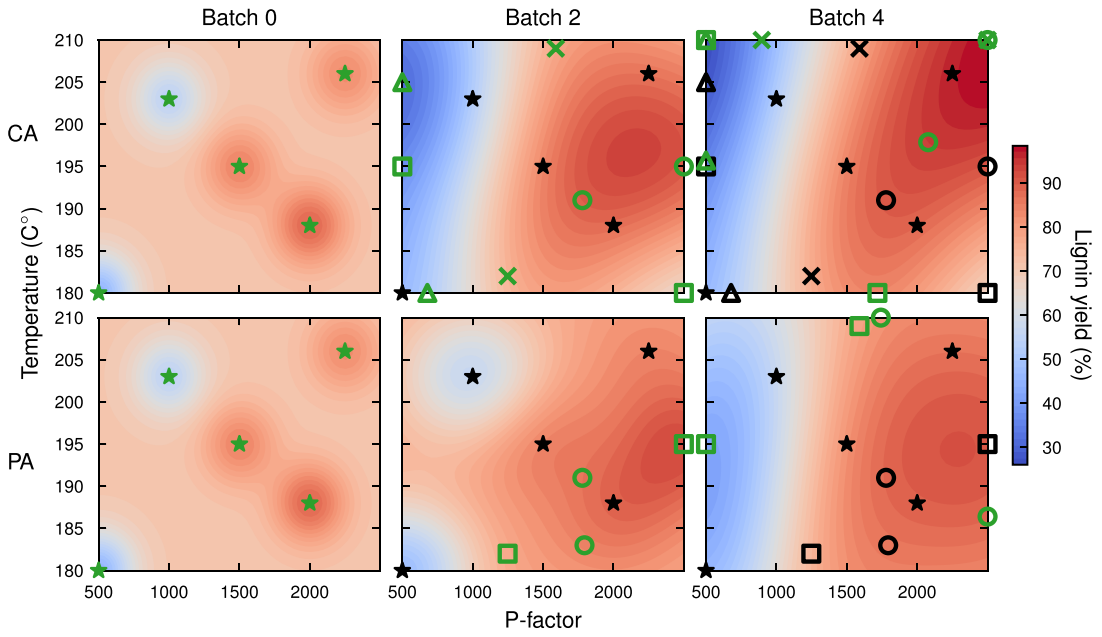


Figure 5. Comparison of the combined acquisitions (CA, top) and pure acquisitions (PA, bottom) strategies. Snapshots of the lignin yield surrogate models are shown for batches 0, 2, and 4 with new (green) and existing (black) acquisitions labeled according to the experimental target and acquisition function from which they were obtained: circles/squares, lignin yield eLCB/pure exploration; triangles/crosses, β-O-4 content eLCB/pure exploration. In the CA scheme, combining acquisitions from different targets before the model updates leads to more informed acquisitions in subsequent batches and more rapid exploration of relevant regions of the design space. Notably, PA fails to discover the region of high yield around P_f = 2500 and T = 210 K.

are interested in achieving both high yield and a high β-O-4 content, we have an optimal trade-off for fixed processing conditions (P_f, T) if an increase in the yield always leads to a decrease in the β-O-4 content and vice versa. We refer to an optimal trade-off as a Pareto point and to the set of all Pareto points as the Pareto front. Similarly, the lignin extraction conditions corresponding to the Pareto points are known as the Pareto optimal solutions. Once the Pareto front is known, a particular Pareto point can be chosen as the preferred solution to the optimization problem based on some additional design criterion, for example, a minimum required yield.

RESULTS AND DISCUSSION

The presentation of our results is organized as follows. First, we describe the experimental BO data collection and the qualitative convergence of the surrogate models. We furthermore consider how useful our two different data collection strategies (Figure 3) were in this task. Next, we analyze the surrogate models to gain insight into the extraction process and its underlying chemistry. Last, we show how to simultaneously optimize lignin yield and structural properties to obtain extraction conditions tailored for specific applications in lignin valorization.

Experimental Data Collection with Bayesian Optimization. The experimental data collection was carried out iteratively in five batches of acquisitions, until convergence, to train two surrogate models representing the extracted lignin yield and the β-O-4 content, respectively. For convenience, we label the batches 0, 1, 2, 3, and 4, where 0 corresponds to the initial batch of Sobol points. To visualize a surrogate model, we use 2D contour plots of the predicted target in the (P_f, T) space and refer to such plots as landscapes. The evolutions of the extracted lignin yield and β-O-4 content landscapes as new batches of data are acquired using the CA strategy are visualized in Figure 4. The figure also includes the predicted standard deviation, which quantifies the prediction uncertainty. We note that for the CA strategy the standard deviations of both targets are identical, up to differences in scale, since the surrogate models are all based on the same acquisition set. Hence, one set of contours is sufficient to represent the standard deviation of both the lignin yield and β-O-4 content (Figure 4, bottom).

The initial batch of Sobol points yields landscapes that, due to the overall data scarcity, are dominated by the mean of the observed values, and the surrogate models have essentially no predictive power. The lack of data at this stage is also reflected in the predicted standard deviation, which is uniformly high at a distance from the Sobol points. As the data set is extended in batches 2−4, the landscapes become smoother, and the surrogate model's predictive power, i.e., ability to accurately interpolate between acquisitions, increases. In the converged model, the lignin yield and β-O-4 content landscapes have resolved into regions of high and low values. The refinement of the landscapes is accompanied by a corresponding decrease in the predicted standard deviation. In batch 4, we obtain a uniformly low standard deviation and approach the lower limit set by the experimental noise built into the surrogate models. At this stage, the landscapes are converging, as qualitatively suggested by the relatively small feature changes observed between the batches 3 and 4. We furthermore observe that the design space is sampled nonuniformly, as exemplified by the locations of the acquisition points in batches 1−4. This nonuniformity results from the exploitation−exploration trade-off in the acquisition functions and leads to a rapid improvement of the surrogate models with added data, which is one of the key features that makes BO an effective approach to experimental design problems.

https://doi.org/10.1021/acssuschemeng.2c01895
ACS Sustainable Chem. Eng. 2022, 10, 9469−9479

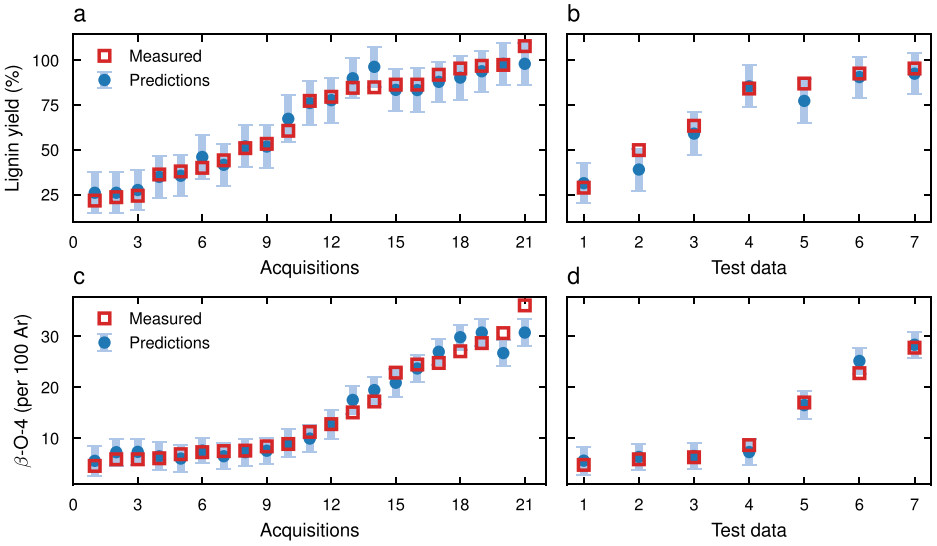


Figure 6. Validation of the surrogate models trained for the lignin yield (a, b) and β-O-4 content (c, d). To assess the accuracy of the models' predictions, the measured target values for a set of independently collected test data are used (right). The predictions for the test set are contrasted to the predictions for the set of acquisitions used to train the models (left). The error bars indicate the predicted standard deviations. Both models provide accurate predictions that are comparable across the acquisitions and test data, indicating that the models strike a good balance between generalizing to new data and reproducing the training data.

Performance of Different Acquisition Strategies. In this section, we compare the CA and PA strategies in terms of the information contents of their respective acquisitions. While the relative information content of an acquisition is difficult to quantify directly, we associate higher information content with faster convergence of the surrogate model. We carry out the comparison between the strategies by studying the batch-by-batch development of the landscapes, as well as the locations of acquisitions made with the eLCB and pure exploration acquisition functions.

A comparison of the two strategies for selected snapshots of the lignin yield landscape is shown in Figure 5. Here, acquisitions are labeled according to whether they are acquired for the lignin yield or β-O-4 content and whether they are generated by the eLCB or pure exploration function. Corresponding snapshots for the β-O-4 content are shown in Figure S2. For both CA and PA, the lignin yield eLCB acquisitions are primarily exploitative in nature and probe the region where $P_f \geq 1500$. Similarly, the β-O-4 eLCB acquisitions all occur in the vicinity of the $P_f \approx 500$ region with a high β-O-4 content. As expected, the batch 2 and 4 landscapes reveal that the CA-generated surrogate model is more developed compared to its PA equivalent, since it utilizes more of the available data points.

While Figure 5 clearly illustrates how acquisitions from the eLCB and pure exploration functions relate to features in the landscape, it does not allow us to compare the convergence rate of the CA and PA strategies since the PA surrogate models use fewer data points than their CA counterparts. This is a consequence of how we defined the PA acquisition strategy. That is, data are collected separately for the two targets and used to build independent surrogate models; hence, the β-O-4 acquisitions are not included in the bottom row of Figure 5. To obtain a better comparison for the convergence rates of CA and PA, we instead compare the CA surrogate models to corresponding PA models constructed from the entire PA data set, i.e., including all the β-O-4 acquisitions (Figures S3 and S4). When comparing these landscapes, the differences between the strategies thus lie solely in the locations of the acquisitions, rather than their numbers. The comparison reveals that CA does indeed yield a more accurate surrogate model, as evidenced by the fact that the PA model fails to predict the region surrounding maximum yield obtained for high P_f and T. This can be attributed to the fact that in the CA strategy, new acquisitions always take the full set of previous acquisitions into account, even if they were made for another target, allowing for more informative points to be picked during both exploitation and exploration. We can thus conclude that CA is a more effective acquisition strategy than PA for a BO-driven experiment design.

In generating eLCB acquisitions for more than one target, our results further highlight the importance of choosing targets with dissimilar landscapes to guide the data collection. In the case of lignin yield and β-O-4 content, we see that their respective eLCB acquisitions are complementary since high target values are attained for high and low P_f, respectively (Figure 5). In contrast, similar landscapes would lead to similar suggestions for new experiments and, consequently, slower model convergence.

Last, we see that, using either strategy, a significant fraction of the suggested experiments lies on the design space boundary. Accordingly, the design variable bounds need to be chosen carefully to avoid extreme processing conditions under which experiments are not feasible.

Surrogate Model Validation. Model validation refers to the process of assessing how well the model can make predictions for a set of test data that was not included in the model training and is therefore a crucial part of any machine learning application. To this end, we compiled a set of test data from seven experiments that were conducted independently from the CA acquisition strategy. We evaluated the predictive power of the converged surrogate models for lignin yield and β-O-4 content for both the test set as well as the acquisitions used to train the model. From the comparison of the predictions with the experimentally measured values (Figure 6b, d), we can qualitatively observe that both surrogate models

https://doi.org/10.1021/acssuschemeng.2c01895
ACS Sustainable Chem. Eng. 2022, 10, 9469−9479

perform well on the test set and that the measured values all fall within one standard deviation of the predictions.

On the basis of the validation data, we calculated averaged predictions errors on both the acquisitions and the test set Table 1. For the acquisitions set, the MAPE is 6.8% and 11.0%

Table 1. Summary of Model Validation Metrics for Lignin Yield and β-O-4 Content[a]

Target	Exp. error	Data set	RMSE	MAE	MAPE
Lignin yield (%)	5%	Acquisitions	4.7	3.8	6.8
		Test	6.0	4.8	7.8
β-O-4 content (per 100 Ar)	5%−10%	Acquisitions	2.0	1.6	11.0
		Test	1.1	0.9	8.6

[a]The estimated experimental error is followed by the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The mean absolute percentage error (MAPE) obtained for the test set is similar to the MAPE of the acquisitions and comparable to the estimated experimental error, indicating that the surrogate models are able to make accurate predictions.

for the lignin yield and β-O-4 content, respectively. These numbers reflect the estimated experimental error, which was encoded into the surrogate models as a fixed Gaussian noise. The corresponding MAPEs for the test data are 7.8% for the lignin yield and 8.6% for the β-O-4 content.

We notice that the prediction errors for the test set are comparable to the estimated experimental error of 5% for the lignin yield and 5%−10% for the β-O-4 content. Previously, we have also seen that the lignin yield and β-O-4 landscapes only undergo minor changes between batches 3 and 4 (Figure 4), which indicates that the surrogate models are converged. Coupled with this observation, the small prediction errors obtained in Table 1 imply that extending the data set beyond the existing 21 data points would not lead to a significant change in the landscapes nor an increase in the predictive power of the surrogate models.

We emphasize the role of exploratory acquisitions in obtaining this level of surrogate model accuracy over all of design space despite using a relatively small data set. An additional benefit of promoting exploration is that the final data set is better suited for fitting surrogate models for other targets that were measured during the BO-guided data collection, since exploration, unlike exploitation, is not specific to an experimental target. While the added emphasis on exploration impedes the convergence of the landscape maxima, we can observe from the evolution of the predicted maxima (Figure S3a, b) that changes between the third and fourth batches are small.

Additional insight into the surrogate models and their convergence can also be gleaned by studying the length scale λ and variance hyperparameters of the radial basis set kernels (eq S7) used for the GPs. In Figure S6, we observe that the surrogate model hyperparameters have converged with the number of data points in the model, although variance appears to be more sensitive to statistical fluctuations in model fitting. Of particular relevance to chemical interpretation are the length scales, which represent the characteristic distances over which the surrogate model varies. For the lignin yield, the length scales have converged to $(\lambda_T, \lambda_{P_f}) \approx (750, 18\ ^\circ\mathrm{C})$ (Figure S6a) and are thus comparable to ranges of our processing conditions. In other words, the surrogate models must vary slowly over design space, as confirmed by the fact that the minimum and maximum yields are separated by $P_f = 2000 \approx 2.67\lambda_{P_f}$ (Figure S6). An interesting consequence is that while we have no direct knowledge of the surrogate model outside the current design space limits, we would not expect to find significant differences in the predicted yields if we only look a distance $d \ll \lambda$ outside the limits. A similar analysis holds also for β-O-4 surrogate models (Figure S6b), see the Supporting Information (SI) for additional details.

Model Predictions for Key Lignin Properties. In this section, we use surrogate model predictions for the lignin yield, β-O-4 content, S/G ratio, and total carbohydrate content to learn more about the chemical reactions underlying the AqSO process.

The lignin yield increases with P-factor and temperature (Figure 7a), with a predicted maximum yield of 98 ± 13% at

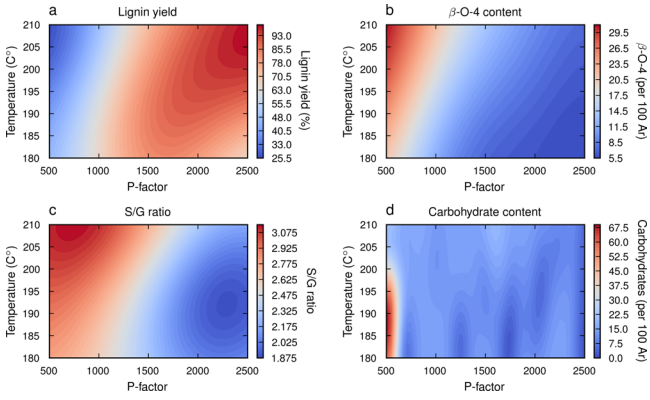


Figure 7. Predicted landscapes for key lignin properties. (a) Lignin yield increases with both temperature and P-factor. (b) β-O-4 content exhibits an antagonistic relationship with the lignin yield and is large when the P-factor is low. (c) Ratio of syringyl to guaiacyl units is also large when the P-factor is low but is less sensitive to changes in P-factor at higher temperatures. (d) Significant amounts of carbohydrates can only be obtained at low to intermediate temperatures and low P-factors. Large measurement uncertainties relative to the local landscape corrugation leads to fitting issues for surrogate model.

$(P_f, T) = (2500, 207\ ^\circ\mathrm{C})$. We explain the relation between yield and P-factor by an increased cleavage of interunit lignin and lignin−carbohydrate linkages and the formation of more acetone-soluble lignin fragments under more severe reaction conditions. The experimentally measured lignin yield of 108% at $(P_f, T) = (2500, 210\ ^\circ\mathrm{C})$ is larger than the maximum yield predicted by the surrogate models but is still contained within one standard deviation of the prediction. The yield measured at these conditions furthermore exceeds 100 mass%, indicating the formation of polyfurans from xylan degradation products (furfural) that is quantified as lignin (so-called pseudolignin).[11]

The β-O-4 content is negatively correlated with the lignin yield in the sense that a high β-O-4 content can only be achieved at a low P-factor (Figure 7b). This behavior is due to a decrease in the break down of native lignin-rich moieties, such as β-O-4, at low reaction severity. The S/G ratio, which is important for specific high-value-added lignin applications, follows trends similar to those observed for the β-O-4 content (Figure 7c). We find the highest S/G ratio at low severity, which is likely due to the higher reactivity of S units in lignin

https://doi.org/10.1021/acssuschemeng.2c01895
ACS Sustainable Chem. Eng. 2022, 10, 9469−9479

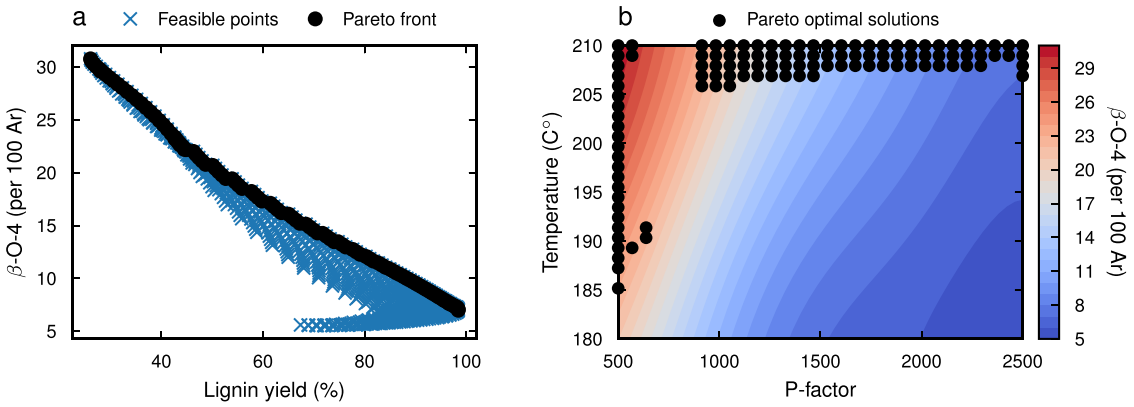


Figure 8. Tailoring the AqSO biorefinery for production of aromatic platform chemicals by simultaneously attempting to maximize lignin yield and β-O-4 content. The result is a set of optimal trade-offs for which an increase in one of the experimental targets always leads to a decrease in the other. The optimal trade-offs are known as the Pareto front (a) and can be projected into Pareto optimal solutions in (P_f, T) space (b). A single point on the Pareto front can be selected by assigning a desired value to one of the targets. The corresponding Pareto optimal solution then provides the optimal extraction conditions for the valorization.

fragmentation and therefore their predominant release during the initial stages of the AqSO process. As expected from our previous work,[11] we observe the highest carbohydrate content, corresponding to LCC moieties, at low severities where the formation of new LCC linkages is more favorable than the cleavage of native LCCs (Figure 7d). With increasing process severity, the equilibrium shifts to the degradation of LCCs, both due to cleavage of the linkages between lignin and carbohydrates as well as the degradation of carbohydrates (mainly xylan) themselves. Thus, while the current ranges of the processing conditions are not favorable for the generation of LCCs, our surrogate model captures important trends and suggests lower severities and temperatures as a starting point for future investigations. We furthermore see that the predicted carbohydrate content fluctuates between 0 to 30 units per 100 Ar several times as the P-factor increases under moderate temperatures. This behavior can be an indication of a larger measurement noise at lower P-factors than initially assumed for the carbohydrate content. Surrogate model landscapes are thus not only useful for predictions but can also highlight issues with the experimental process.

Our landscapes reveal that the lignin characteristics do not only depend on the P-factor, as observed for xylan extraction from wood,[36] but also on the reaction temperature at a fixed P-factor. The landscape shows that this dependence is surprisingly complex. For example, at a low P-factor (e.g., P = 500), the highest yield is observed for a low reaction temperature (180 °C), but at a high P-factor (e.g., P = 2000), it occurs for the highest reaction temperature (210 °C). This complex dependence can be explained by competing reactions[11,49] with different activation energies that consequently contribute differently to the overall process.

The (P_f, T) dependence of all the structural properties we have modeled can, to some extent, be qualitatively explained by well-known principles from lignin chemistry. However, the surrogate model landscapes provide the quantitative predictions necessary for large-scale applications in lignin valorization. As noted for the carbohydrate content, the surrogate models can also provide information on potential problems in the experimental characterization, such as larger than expected measurement noise.

Tailoring Extraction Conditions for Different Lignin Applications. Using our trained surrogate models, we can derive optimal extraction conditions for arbitrary design criteria associated with lignin-based products. In this context, the design criteria are a set of requirements placed on the extracted lignin by a potential application. For instance, the lignin yield should typically exceed some minimum threshold, and we might require high amounts of specific moieties. To solve this problem, we can employ a Pareto front analysis to our surrogate models to find optimal trade-offs between the design criteria. The practical implication of having multiple optimal trade-offs is that changing the extraction conditions to bring one target closer to its design criteria will result in at least one other target having a less optimal value.

As a concrete example of how to apply a Pareto front analysis, we consider optimizing the AqSO biorefinery for extracting lignin suitable as a feedstock in the production of aromatic platform chemicals. For this purpose, most processes require a maximal number of β-O-4 linkages to maximize the yield of the targeted monomers.[6−8] Hence, the maximum revenue per original biomass correlates with both high lignin yield and high β-O-4 content. Using our surrogate models, we can determine all feasible combinations of yield and β-O-4 content, i.e., those obtainable by varying the processing conditions in their allowed range. These feasible points forms a two-dimensional space, as indicated in blue in Figure 8a. By subsequently calculating the Pareto front, we see that it lies on the boundary of the space of feasible points; these are the optimal trade-offs. We observe from the shape of the Pareto front that high values of the lignin yield are correlated with low β-O-4 content, in agreement with Figure 7a and b. The corresponding Pareto optimal solutions, i.e., the processing conditions (P_f, T) that produce optimal trade-offs, are then found by projecting the Pareto front onto (P_f, T) space (Figure 8b).

Since every point on the Pareto front represents an optimal trade-off, no point is inherently preferred over another. This gives us the freedom to adjust the extraction conditions to particular applications and to take physicochemical constraints and cost−benefit considerations into account. For our aromatic platform chemicals example, we might choose to accept a slightly smaller yield of 60% in return for a higher β-O-4 content. We can then determine from Figure 8 that this trade-off results in 17.5 β-O-4 linkages per 100 Ar and can be achieved by extracting the lignin at (P_f, T) = (1227, 210 °C).

https://doi.org/10.1021/acssuschemeng.2c01895
ACS Sustainable Chem. Eng. 2022, 10, 9469−9479

As a second application case, we consider the use of lignin as an antioxidant, for which a high amount of phenolic OH groups would be beneficial.[50] While we have not measured the phenolic OH content directly in this work, we note that it is a typical product of β-O-4-unit cleavage, and thus, we expect the highest phenolic OH content at the lowest number of β-O-4 linkages.[11] We can then infer from Figure 7a and b that phenolic OH content correlates positively with the lignin yield and hence we should be able to obtain phenolic OH-rich lignin with high yield. To find suitable extraction conditions, we again apply a Pareto front analysis where we look for trade-offs involving low β-O-4 content and high lignin yield. We then find that the trade-off involving the lowest β-O-4 content equal to 5.5 units per 100 Ar can be obtained with 72% yield at (P_f, T) = (2397, 182 °C). We thus expect that these extraction conditions are favorable for extraction of lignin with antioxidant properties.

Further examples of applications could be given for the optimal S/G ratio and amounts of carbohydrates in lignin (present as LCC). For instance, higher proportions of S units should be beneficial for lignin depolymerization due to the higher reactivity of S units compared to that of G units.[7] In contrast, S units are not suitable for various cross-linking reactions as both ortho positions to the phenolic OH (typical reaction centers in cross-linking) in the aromatic ring of S units are occupied by OMe groups.[51] In surfactant applications, the presence of hydrophilic carbohydrates moieties (as LCC) can be advantageous, while, for example, the production of aromatic monomers requires high purity lignin.[7]

It is important to keep in mind that there is no all-purpose lignin; different applications call for optimization of different properties. In future work, we aim to establish surrogate models that correlate lignin structure with optimal product properties for specific applications. This would allow us to link the effects of the processing conditions on lignin structure and effects of those properties on the performance of lignin in specific applications.

CONCLUSIONS

We have improved the AqSO biorefinery by increasing its resource efficiency and by optimizing properties of the extracted lignin for valorization applications. We accomplished this by using BO to construct surrogate models capable of predicting experimental targets (lignin yield and moieties quantified using 2D NMR) in terms of the processing conditions (temperature, P-factor). Using the depolymerization of lignin into platform chemicals as an example application, we subsequently applied a Pareto front analysis to determine the processing conditions that provide optimal balance between high lignin yield and a high amount of β-O-4 linkages. The derived processing conditions thus allow for resource-efficient extraction of β-O-4-rich lignin suitable for depolymerization applications.

While we have presently focused on refining the AqSO biorefinery concept, our work also highlights how BO can function as the cornerstone of a holistic framework for developing sustainable materials technologies. The key advantage provided by BO in this context is the ability to establish surrogate models from a small set of data that is automatically curated by the algorithm. Here, we showed how the data can be efficiently collected in batches and shared between multiple surrogate models to accelerate convergence and provide predictive power over the entire design space. A Pareto front analysis can then be applied as a general tool to optimize products and increase resource efficiency, without the need to conduct further experiments. This will in turn help increase the long-term sustainability of the technology. Our work thus shows the potential of machine learning methods, such as BO, to both improve and expedite the development of sustainable materials processing and engineering.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssuschemeng.2c01895.

A more technical introduction to BO and the concept of Pareto optimality, complementary results for acquisition strategy comparisons, and model validation. (PDF)

AUTHOR INFORMATION

Corresponding Authors

Mikhail Balakshin − *Department of Bioproducts and Biosystems, Aalto University, 02150 Espoo, Finland*; Email: mikhail.balakshin@aalto.fi

Milica Todorović − *Department of Mechanical and Materials Engineering, University of Turku, 20014 Turku, Finland*; orcid.org/0000-0003-0028-0105; Email: milica.todorovic@utu.fi

Authors

Joakim Löfgren − *Department of Applied Physics, Aalto University, 02150 Espoo, Finland*; orcid.org/0000-0001-6968-5966

Dmitry Tarasov − *Department of Bioproducts and Biosystems, Aalto University, 02150 Espoo, Finland*

Taru Koitto − *Department of Bioproducts and Biosystems, Aalto University, 02150 Espoo, Finland*

Patrick Rinke − *Department of Applied Physics, Aalto University, 02150 Espoo, Finland*; orcid.org/0000-0003-1898-723X

Complete contact information is available at: https://pubs.acs.org/10.1021/acssuschemeng.2c01895

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from the Aalto University Internal Seed Fund, the Academy of Finland through project nos. 316601 and 341589, the FinnCERES BioEconomy flagship, and the Finnish Center for Artificial Intelligence (FCAI). The data and code[43] that were used in this study are freely available online.[52]

REFERENCES

(1) Ragauskas, A. J.; Beckham, G. T.; Biddy, M. J.; Chandra, R.; Chen, F.; Davis, M. F.; Davison, B. H.; Dixon, R. A.; Gilna, P.; Keller, M.; Langan, P.; Naskar, A. K.; Saddler, J. N.; Tschaplinski, T. J.; Tuskan, G. A.; Wyman, C. E. Lignin Valorization: Improving Lignin Processing in the Biorefinery. *Science* **2014**, *344*, 1246843.

(2) Holladay, J. E.; White, J. F.; Bozell, J. J.; Johnson, D. Top Value-Added Chemicals from Biomass - Volume II—Results of Screening for Potential Candidates from Biorefinery Lignin. *PNNL-16983* **2007**, DOI: 10.2172/921839.

(3) Berlin, A.; Balakshin, M. In *Bioenergy Research: Advances and Applications*; Gupta, V. K., Tuohy, M. G., Kubicek, C. P., Saddler, J.,

Xu, F., Eds.; Elsevier: Amsterdam, 2014; pp 315−336, DOI: 10.1016/B978-0-444-59561-4.00018-8.
(4) Balakshin, M. Y.; Capanema, E. A.; Sulaeva, I.; Schlee, P.; Huang, Z.; Feng, M.; Borghei, M.; Rojas, O. J.; Potthast, A.; Rosenau, T. New Opportunities in the Valorization of Technical Lignins. *ChemSusChem* **2021**, *14*, 1016−1036.
(5) Zakzeski, J.; Bruijnincx, P. C. A.; Jongerius, A. L.; Weckhuysen, B. M. The Catalytic Valorization of Lignin for the Production of Renewable Chemicals. *Chem. Rev.* **2010**, *110*, 3552−3599.
(6) Schutyser, W.; Renders, T.; Van den Bosch, S.; Koelewijn, S.-F.; Beckham, G. T.; Sels, B. F. Chemicals from Lignin: An Interplay of Lignocellulose Fractionation, Depolymerisation, and Upgrading. *Chem. Soc. Rev.* **2018**, *47*, 852−908.
(7) Abu-Omar, M. M.; Barta, K.; Beckham, G. T.; Luterbacher, J. S.; Ralph, J.; Rinaldi, R.; Roman-Leshkov, Y.; Samec, J. S. M.; Sels, B. F.; Wang, F. Guidelines for Performing Lignin-First Biorefining. *Energy Environ. Sci.* **2021**, *14*, 262−292.
(8) Questell-Santiago, Y. M.; Galkin, M. V.; Barta, K.; Luterbacher, J. S. Stabilization Strategies in Biomass Depolymerization Using Chemical Functionalization. *Nature Reviews Chemistry* **2020**, *4*, 311−330.
(9) Dessbesell, L.; Paleologou, M.; Leitch, M.; Pulkki, R.; Xu, C. C. Global Lignin Supply Overview and Kraft Lignin Potential as an Alternative for Petroleum-Based Polymers. *Renewable and Sustainable Energy Reviews* **2020**, *123*, 109768.
(10) Tarasov, D.; Schlee, P.; Pranovich, A.; Moreno, A.; Wang, L.; Sipponen, M. H.; Xu, C.; Balakshin, M. Towards a New Generation of Biorefinery: AqSO Process to Streamline the Development of High-Value Sustainable Products from All Lignocellulosic Biomass Components. Unpublished work, 2022.
(11) Tarasov, D.; Schlee, P.; Pranovich, A.; Moreno, A.; Wang, L.; Sipponen, M. H.; Xu, C.; Balakshin, M. Novel AqSO biorefinery for high-value sustainable products from all biomass components. In *Proceedings of the 16th European Workshop on Lignocellulosics and Pulp*, Gothenburg, Sweden, 2022; p. 43−46.
(12) Balakshin, M.; Capanema, E. A.; Zhu, X.; Sulaeva, I.; Potthast, A.; Rosenau, T.; Rojas, O. J. Spruce Milled Wood Lignin: Linear, Branched or Cross-Linked? *Green Chem.* **2020**, *22*, 3985−4001.
(13) Fisher, R. A. *The Design of Experiments*; Oliver & Boyd: Oxford, England, 1935.
(14) Montgomery, D. C. *Design and Analysis of Experiments*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2006.
(15) McKay, M. D.; Beckman, R. J.; Conover, W. J. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* **1979**, *21*, 239−245.
(16) Helton, J. C.; Davis, F. J. Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems. *Reliability Engineering & System Safety* **2003**, *81*, 23−69.
(17) Collins, L. M.; Dziak, J. J.; Kugler, K. C.; Trail, J. B. Factorial Experiments: Efficient Tools for Evaluation of Intervention Components. *American journal of preventive medicine* **2014**, *47*, 498−504.
(18) Box, G. E. P.; Wilson, K. B. On the Experimental Attainment of Optimum Conditions. *Journal of the Royal Statistical Society. Series B (Methodological)* **1951**, *13*, 1−45.
(19) Myers, R. H.; Montgomery, D. C.; Vining, G. G.; Borror, C. M.; Kowalski, S. M. Response Surface Methodology: A Retrospective and Literature Survey. *Journal of Quality Technology* **2004**, *36*, 53−77.
(20) Duch, W.; Diercksen, G. H. F. Neural Networks as Tools to Solve Problems in Physics and Chemistry. *Comput. Phys. Commun.* **1994**, *82*, 91−103.
(21) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577*, 706−710.
(22) Tang, B.; Pan, Z.; Yin, K.; Khateeb, A. Recent Advances of Deep Learning in Bioinformatics and Computational Biology. *Frontiers in Genetics* **2019**, *10*, 214.
(23) Mockus, J. On Bayesian Methods for Seeking the Extremum. In *Proceedings of the IFIP Technical Conference*, Berlin, Heidelberg, 1974; pp 400−404, .
(24) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE* **2016**, *104*, 148−175.
(25) Ginsbourger, D.; Le Riche, R.; Carraro, L. In *Computational Intelligence in Expensive Optimization Problems*; Tenne, Y., Goh, C.-K., Eds.; Adaptation Learning and Optimization; Springer: Berlin, Heidelberg, 2010; pp 131−162, DOI: 10.1007/978-3-642-10701-6_6.
(26) Garijo del Río, E.; Mortensen, J. J.; Jacobsen, K. W. Local Bayesian Optimizer for Atomic Structures. *Phys. Rev. B* **2019**, *100*, 104103.
(27) Bisbo, M. K.; Hammer, B. Efficient Global Structure Optimization with a Machine-Learned Surrogate Model. *Phys. Rev. Lett.* **2020**, *124*, 086102.
(28) Todorović, M.; Gutmann, M. U.; Corander, J.; Rinke, P. Bayesian Inference of Atomistic Structure in Functional Materials. *npj Computational Materials* **2019**, *5*, 1−7.
(29) Xue, D.; Balachandran, P. V.; Hogden, J.; Theiler, J.; Xue, D.; Lookman, T. Accelerated Search for Materials with Targeted Properties by Adaptive Design. *Nat. Commun.* **2016**, *7*, 11241.
(30) Li, C.; Rubín de Celis Leal, D.; Rana, S.; Gupta, S.; Sutti, A.; Greenhill, S.; Slezak, T.; Height, M.; Venkatesh, S. Rapid Bayesian Optimisation for Synthesis of Short Polymer Fiber Materials. *Sci. Rep.* **2017**, *7*, 5683.
(31) Wigley, P. B.; Everitt, P. J.; van den Hengel, A.; Bastian, J. W.; Sooriyabandara, M. A.; McDonald, G. D.; Hardman, K. S.; Quinlivan, C. D.; Manju, P.; Kuhn, C. C. N.; Petersen, I. R.; Luiten, A. N.; Hope, J. J.; Robins, N. P.; Hush, M. R. Fast Machine-Learning Online Optimization of Ultra-Cold-Atom Experiments. *Sci. Rep.* **2016**, *6*, 25890.
(32) Vahid, A.; Rana, S.; Gupta, S.; Vellanki, P.; Venkatesh, S.; Dorin, T. New Bayesian-Optimization-Based Design of High-Strength 7xxx-Series Alloys from Recycled Aluminum. *JOM* **2018**, *70*, 2704−2709.
(33) Sun, S.; Tiihonen, A.; Oviedo, F.; Liu, Z.; Thapa, J.; Zhao, Y.; Hartono, N. T. P.; Goyal, A.; Heumueller, T.; Batali, C.; Encinas, A.; Yoo, J. J.; Li, R.; Ren, Z.; Peters, I. M.; Brabec, C. J.; Bawendi, M. G.; Stevanovic, V.; Fisher, J.; Buonassisi, T. A Data Fusion Approach to Optimize Compositional Stability of Halide Perovskites. *Matter* **2021**, *4*, 1305−1322.
(34) Liang, Q.; Gongora, A. E.; Ren, Z.; Tiihonen, A.; Liu, Z.; Sun, S.; Deneault, J. R.; Bash, D.; Mekki-Berrada, F.; Khan, S. A.; Hippalgaonkar, K.; Maruyama, B.; Brown, K. A.; Fisher, J., III; Buonassisi, T. Benchmarking the Performance of Bayesian Optimization across Multiple Experimental Materials Science Domains. *npj Computational Materials* **2021**, *7*, 1−10.
(35) Ngatchou, P.; Zarei, A.; El-Sharkawi, A. Pareto Multi Objective Optimization. In *Proceedings of the 13th International Conference on Intelligent Systems Application to Power Systems*, 2005; pp 84−91, .
(36) Sixta, H. *Handbook of Pulp*; John Wiley & Sons, Ltd., 2006; pp I−XXXII, DOI: DOI: 10.1002/9783527619887.fmatter.
(37) Ralph, J.; Landucci, L. L. In *Lignin and Lignans*; Heitner, C., Dimmel, D., Schmidt, J., Eds.; CRC Press, 2010.
(38) Balakshin, M. Y.; Capanema, E. A.; Chen, C.-L.; Gracz, H. S. Elucidation of the Structures of Residual and Dissolved Pine Kraft Lignins Using an HMQC NMR Technique. *J. Agric. Food Chem.* **2003**, *51*, 6116−6127.
(39) Lancefield, C. S.; Wienk, H. L. J.; Boelens, R.; Weckhuysen, B. M.; Bruijnincx, P. C. A. Identification of a Diagnostic Structural Motif Reveals a New Reaction Intermediate and Condensation Pathway in Kraft Lignin Formation. *Chemical Science* **2018**, *9*, 6348−6360.

(40) Balakshin, M.; Capanema, E.; Gracz, H.; Chang, H.-m.; Jameel, H. Quantification of Lignin-Carbohydrate Linkages with High-Resolution NMR Spectroscopy. *Planta* **2011**, *233*, 1097−1110.
(41) Brochu, E.; Cora, V. M.; de Freitas, N. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *arXiv Preprint*, arXiv:1012.2599, 2010, DOI: 10.48550/ARXIV.1012.2599.
(42) Gutmann, M. U.; Corander, J. Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models. *Journal of Machine Learning Research* **2016**, *17*, 1−47.
(43) BOSS: Bayesian Optimization Structure Search. https://gitlab.com/cest-group/boss (accessed May 24, 2022_.
(44) Järvi, J.; Rinke, P.; Todorović, M. Detecting Stable Adsorbates of (1S)-Camphor on Cu(111) with Bayesian Optimization. *Beilstein Journal of Nanotechnology* **2020**, *11*, 1577−1589.
(45) Järvi, J.; Alldritt, B.; Krejčí, O.; Todorović, M.; Liljeroth, P.; Rinke, P. Integrating Bayesian Inference with Scanning Probe Experiments for Robust Identification of Surface Adsorbate Configurations. *Adv. Funct. Mater.* **2021**, *31*, 2010853.
(46) Fang, L.; Makkonen, E.; Todorović, M.; Rinke, P.; Chen, X. Efficient Amino Acid Conformer Search with Bayesian Optimization. *J. Chem. Theory Comput.* **2021**, *17*, 1955−1966.
(47) Jin, S.-A.; Kämäräinen, T.; Rinke, P.; Rojas, O. J.; Todorović, M. Machine Learning as a Tool to Engineer Microstructures: Morphological Prediction of Tannin-Based Colloids Using Bayesian Surrogate Models. *MRS Bull.* **2022**, *47*, 29−37.
(48) Sobol', I. M. On the Distribution of Points in a Cube and the Approximate Evaluation of Integrals. *USSR Computational Mathematics and Mathematical Physics* **1967**, *7*, 86−112.
(49) Li, S.; Lundquist, K.; Westermark, U. Cleavage of Arylglycerol SS-Aryl Ethers under Neutral and Acid Condition. *Nordic Pulp & Paper Research Journal* **2000**, *15*, 292−299.
(50) Dizhbite, T.; Telysheva, G.; Jurkjane, V.; Viesturs, U. Characterization of the Radical Scavenging Activity of Lignins—Natural Antioxidants. *Bioresour. Technol.* **2004**, *95*, 309−317.
(51) Sarkanen, K. V.; Ludwig, C. H. *Lignins: Occurrence, Formation, Structure and Reactions*; Wiley-Interscience: New York, 1971.
(52) Löfgren, J.; Tarasov, D.; Koitto, T.; Rinke, P.; Balakshin, M.; Todorovic, M. *Lignin Biorefinery Experimental Optimization Using BOSS*; Zendo, 2022. DOI: 10.5281/zenodo.6581802