Lignin Biorefinery Optimization Through Machine Learning

Joakim Löfgren,[1] Dmitry Tarasov,[2] Taru Koitto,[2] Patrick Rinke,[1] Mikhail Y. Balakshin,[2] and Milica Todorović[1,3,*]

[1] *Department of Applied Physics, Aalto University, Espoo, Finland*
[2] *Department of Bioproducts and Biosystems, Aalto University, Espoo, Finland*
[3] *Department of Mechanical and Materials Engineering, University of Turku, Turku, Finland*

Lignin is an abundant biomaterial that currently emerges as a low value byproduct in the pulp and paper industry but could be repurposed for high-value products as part of the ongoing global transition to a sustainable society. To increase lignins value, rational and efficient approaches to optimizing lignin biorefineries to produce high value bioproducts are required. Here, we report the optimization of the AquaSolv Omni (AqSO) Biorefinery, a newly introduced biorefinery concept based on hydrothermal pretreatment and solvent extraction. We employ a machine-learning framework based on Bayesian optimization, to provide sample-efficient and guided data collection as well as surrogate model building. The surrogate models allow us to map multiple experimental outputs, including the extracted lignin yield and main structural properties obtained by 2D NMR, as functions of the hydrothermal pretreatment reaction severity and temperature. Our results show that with Bayesian optimization, predictive models can be converged with only 21 data points to within a margin of error comparable to the underlying experimental error. By applying a Pareto front analysis, we demonstrate how the predictive models can be used in tandem to identify optimal extraction conditions for concrete applications in lignin valorization.

I. INTRODUCTION

To achieve a sustainable economy, efficient and green utilization of natural resources is of paramount importance. Lignin, as a part of lignocellulosic biomatter, is an example of a naturally abundant, but under-utilized, resource. Lignin is currently produced in large quantities as a residual in pulp and papermaking as well as in biorefinery processes. Valorization of lignin into, e.g., materials [1–3] or chemicals, [4, 5] can therefore substantially increase both the sustainability and revenue of biorefineries [6]. The development or refinement of lignin valorization approaches, however, requires effective techniques for engineering lignin with targeted structures and properties.

Recently, we have suggested a green, lignin-first biorefinery process for the integrated utilization of all biomass components, with special focus on the lignin-containing streams (Tarasov et al., in preparation). [7] Our process consists of hydrothermal treatment (HTT) of biomass followed by a solvent wash of the resulting solids and is accordingly named AquaSolv Omni (AqSO). HTT is a facile and flexible biorefinery process, [8, 9] that has traditionally been used for cost-efficient extraction and valorization of hemicelluloses from biomass [10, 11] leaving the lignin component as a by-product. One of the foremost advantages of HTT is the use of water, which removes the need for extra chemicals in the reaction medium and renders the process environmentally friendly. The AqSO biorefinery couples highly tunable processing conditions with output versatility in terms of lignin composition, structure, and physicochemical properties. To take full advantage of the flexibility, the AqSO process needs to be tailored and optimized, in terms of its processing conditions, for specific end-products such as high value biomaterials or chemicals.

In industry, experimental optimization tasks are often approached with design of experiment (DOE) methods, [12, 13] which provide general strategies for planning data collection and modeling the experimental output. Since experiments are often time consuming and costly, DOE methods benefit from efficient sampling of the design space. Conventional approaches to DOE include space-filling designs, [14, 15] factorial designs, [16] and response surface methods, [17, 18] where the latter category is often regarded as an industry standard. A shortcoming of these approaches is that they largely employ a sampling of design space that is pre-determined. Since experiments must typically be performed sequentially or in batches, however, information from new measurements cannot be utilized to optimize the sampling strategy. Another issue concerns complex processes where many design variables lead to a high-dimensional design space that cannot be efficiently explored using traditional DOE methods.

The last decade has seen machine learning methods enter a variety of natural science domains to solve challenging optimization and modeling problems [19–21]. For applications in DOE, Bayesian optimization (BO) is a promising machine learning method that couples model building, data collection and global optimization of black-box functions. [22, 23] A key feature of BO is an integrated 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 sample efficient and provides better scaling with the dimension of the design space than traditional DOE methods. The BO process can also incorporate experimental errors, prior knowledge as well as batch experiments, [24] further increasing its potential for DOE tasks. BO can be extended to deal with a large category of design problems including

* milica.todorovic@utu.fi

both constraints and multiple objectives. [25]

Recently, BO has become a popular tool in computational materials science, where it has been used both as a means of efficient optimization [26–28] and to guide materials discovery. [29] BO is also increasingly being used for DOE, [30–34] although the overall number of studies remains modest. This holds true in particular for applications in the biomaterials community, where the adoption of machine learning methods is still in its infancy. Existing applications of BO to experiments tend furthermore to focus on optimization problems rather than mapping experimental outputs to design variables. As a result, limited attention is paid to BOs potential for predictive modeling and the insight into experiments and materials this can yield.

In the present work we evaluate the use of BO as a tool for guiding and modeling in the chemical engineering field by applying it to lignin extraction through the AqSO biorefinery process. Specifically, we aim to use BO to establish predictive models that map lignin properties as a function of the extraction conditions, namely the hydrothermal pretreatment temperature and reaction severity. As a part of the design process, we explore how to perform informative batch acquisitions for multiple output properties. We furthermore demonstrate, using a Pareto front analysis, how the resulting models can be used to simultaneously optimize sets of lignin properties for applications in lignin valorization. Our work describes a holistic approach to BO-based experimental design that emphasizes model building and interpretation and is suitable for a wide variety of design problems.

II. EXPERIMENTAL MATERIALS AND METHODS

A. Materials

We debarked, chipped, and ground a Birch wood (Betula sp.) stem into sawdust (0.55-0.125 mm particle were size selected). Prior to the HTT, we subjected the sawdust to acetone extraction to remove the lipophilic extractives.

B. Chemicals

We purchased acetone (C3H6O, 95 vol %) from Sigma-Aldrich and used it without purification. We also purchased sulfuric acid (H2SO4, 98 wt%), xylose, arabinose, rhamnose, glucose, galactose, mannose for ion chromatography analysis and chromium (III) acetylacetonate, (Cr(acac)3), endo-n-hydroxy-5-norbornene-2,3-dicarboximide, 1,3,5-Trioxane and 2-chloro-4,4,5,5-tetramethyl-1,3,2-dioxaphospholane (TMDP) (all analytical grades) for nuclear magnetic resonance (NMR) spectroscopy from Sigma-Aldrich.

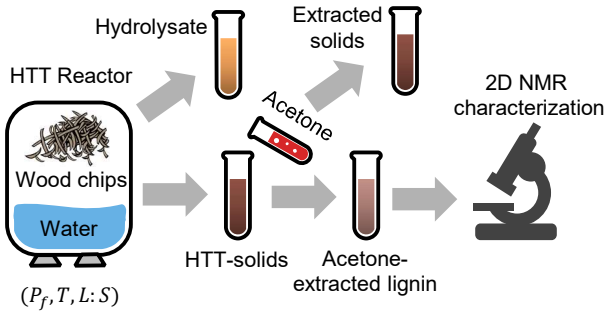

FIG. 1. A schematic illustration of the experimental setup employed for the AqSO biorefinery. A mixture of birch chips and water is subjected to hydrothermal treatment (HTT) in a reactor whose state can be described by the reactor temperature (T), liquid-to-solid ratio ($L:S=1$) and P-factor (P_f). The reaction produces a hydrolysate and HTT-solids. Addition of acetone to the HTT-solids results in acetone-extracted lignin and extracted solids. The structural properties of the acetone extracted lignin are finally characterized by 2D NMR. The hydrolysate and extracted solids are not used in this study.

C. Hydrothermal Treatment

A qualitative overview of the entire HTT process is shown in Fig. 1. We carried out the HTT of the extractive-free sawdust (4g) in a swing reactor at a liquid to solid (L/S) ratio of 1. The swing reactor was equipped with temperature control both in the heating block and inside the reactor. As the heating period has a significant effect on the extraction process, we chose to work with reaction severity instead of the residence time. HTT severity can be expressed in terms of the P-factor, which is calculated according to [35]

$$P_f = \int_0^t \frac{k(T(t'))}{k(100°)}dt' = \int_0^t \exp\left(40.48 - \frac{15106}{T(t')}\right)dt'. \quad (1)$$

Here, k is the rate constant, t the residence time (in hours) and T the reaction temperature (K). 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) aqueous acetone. We rotary-evaporated the solution (at 40 °C) to produce acetone-extractable lignin, which we finally vacuum-dried at 40 °C to constant weight to determine the yield.

D. 2D HSQC NMR analysis

We recorded the 2D Heteronuclear Single Quantum Coherence (HSQC) NMR spectra using a Bruker AVANCE 600 NMR spectrometer equipped with a Cry-

oProbe. We dissolved about 80 mg of each sample in 0.6 mL 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 2D HSQC NMR data (1024×1024 data points) by applying a QSINE window function to both the 1H and 13C dimensions. We used the DMSO peak at δC/δH 39.5/2.49 ppm/ppm for calibration. To quantify the specific lignin moieties, we carried out volume integration of the HSQC spectra of the acetone-extractable lignin (Fig. S2). We describe this procedure in greater detail elsewhere (Tarasov et al. in preparation). The intensity of the signals is expressed in mol%, i.e., per 100 aromatic units (Ar) assuming the sum of the signals of G- and S-units as 100% from their characteristic CH signals at G_2 and $S_{2,6}$ positions, correspondingly: $G_2 + S_{2,6}/_2 = 100\%$.

E. Bayesian Optimization

For brevity, we provide only a short overview of Bayesian optimization and refer readers to Sect. I of the SI and the literature [23, 36] for more detailed accounts. BO involves two main components, namely a surrogate model that approximates the objective 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 objective, and the posterior variance provides a measure of the model uncertainty. By minimizing the acquisition function, a new sampling location is subsequently determined and used to augment the existing data set. Acquisition functions come in many flavors and provide trade-offs between exploitation and exploration. Here, the former refers to the sampling regions of design space where the objective is likely to achieve a minimum or maximum, while the latter refers to the sampling of regions where model uncertainty is high and where data has not been sampled before. Common choices of acquisitions function include the lower confidence bound (LCB) function [37] and expected improvement (EI) function. In this work we also consider acquisitions made from the model standard deviation, which we shall refer to as the pure exploration function.

We carried out the BO using the recently released BOSS code [28], which provides a Python-based implementation of BO. Among a large selection of features, BOSS implements the exploration-modified lower confidence bound (eLCB) function used in this study. [36, 38] BOSS has previously been applied successfully to a range of different problems in materials modeling [39–42]. We defined the Gaussian processes by uninformative zero priors for the mean functions, and radial basis set (RBF) kernels to reflect the smoothness of the objectives.

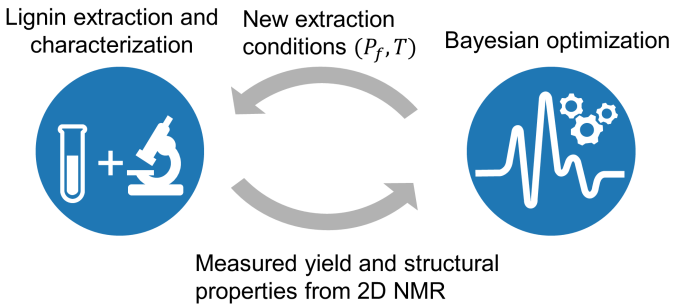

FIG. 2. Workflow for optimizing and planning AqSO biorefinery experiments using machine learning. The Bayesian Optimization program (right) suggests new lignin extraction conditions in terms of new values for the HTT reactors temperature and P-factor. After performing an extraction at these conditions, the lignin yield and structural properties are measured and fed back to the program. The new data (extraction conditions and measured properties) is used to update the program and the cycle starts over.

We initialized the kernel hyperparameters using inverse gamma priors and subsequently updated them during the BO process by maximizing the marginal likelihood. To initialize the surrogate models, we employed a batch of 5 Sobol points. To account for measurement errors in our objectives, we incorporated Gaussian noise terms with zero mean into the surrogate models.

F. Applying Bayesian Optimization to Experiments

The interplay between experiment and BO during the data collection process is illustrated in Fig. 2. In an experimental context, the objective can be any measurable output from the experiment such as the yield or a 2D NMR peak. The objective depends on several design variables that represent adjustable factors of the experiment, including precursor properties, processing conditions and apparatus settings. In the present work, we considered two design variables relevant for the HTT-treatment of the birch sawdust: the final reactor temperature (T), and the P-factor (P_f) . The liquid to-solid ratio of the birch biomass and water (L:S) in the HTT-reactor also plays an important role in determining the experimental output. When investigating a new experimental setup described by a high-dimensional design space, however, it is often convenient to limit the initial search to a smaller subspace. In this spirit, we employed a fixed liquid-to-solid ratio $L : S = 1$ and investigated the two-dimensional (P_f, T) design space, for which predictive models can be easily visualized. The decision to fix L:S was further motivated by a preliminary investigation of the AqSO biorefinery, which indicated that $L : S = 1$ optimal for the yield and properties of the lignin-containing stream (Tarasov et al., in preparation). The design space was then determined

based on feasible operating ranges for the two variables: $500 \leq P_f \leq 2500,\ 180 \leq T \leq 210\,(°C)$.

For the objectives, we considered the yield of acetone extractable lignin and several structural properties obtained through 2D NMR characterization. From the large number of different lignin structural characteristics revealed by 2D NMR, we chose to model the most important ones that provide information transferable to other lignin characteristics. The modelled structural properties include the content of β-O-4 linkages, the ratio syringyl and guaiacyl (S/G) units and the carbohydrate content as a part of lignin-carbohydrate complex. β-O-4 linkages are the main reactive centers of the native lignin; their amount after the process characterize the degree of lignin transformation. In addition, good correlation between the amounts of β-O-4 linkages and other lignin characteristics have previously been reported for the AqSO process (Tarasov et al. in preparation). To account for the experimental error in measuring these objectives, the standard deviations of the noise terms incorporated into the surrogate models was chosen to reflect estimated measurement errors of 5% for lignin yield and 5-10% for the structural properties.

A key feature of our approach is that the experimental data collection was performed by applying BO to two objectives, namely the extracted lignin yield and β-O-4 content. After the data collection is concluded, it is then possible to train surrogate models for any other objectives that were measured, even if those objectives were not actively employed to generate new acquisitions.

Since the experimental sample preparation and characterization typically takes several days, acquisitions were made in batches to make the process more efficient. Rather than obtaining multiple acquisitions from the same acquisition function, we acquired once from the exploration-modified lower confidence bound (eLCB) function and once from the pure exploration function. Here, the term pure exploration to refers to acquisitions made using the model standard deviation as acquisition function. The four-fold split of acquisitions between eLCB and pure exploration acquisitons functions as well as two different objectives emphasizes exploration of the design space and ensures that the collected data can be used to train predictive models for the remaining objectives. Note, however, that the presence of eLCB acquisitions in our approach means that our approach is not purely exploratory. This is motivated by our aim to quickly identify extremal regions, e.g., with high lignin yield, to efficiently design the AqSO biorefinery. With two different objectives and two different acquisition functions, each batch of acquisitions comprised four suggested experiments in total. The straightforward strategy for proceeding with the BO would then be to simply conduct two separate, isolated, BO processes, one for the lignin yield and one for the β-O-4 content. We shall refer to this as the pure acquisitions (PA) strategy (Fig. 3a). With this strategy, however, we do not make use of the fact that each sample yields a measured value

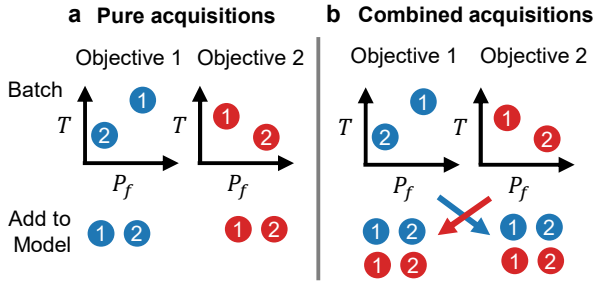

FIG. 3. Two strategies for updating surrogate models for different objectives. (a) In the pure acquisitions strategy, only acquisitions (red and blue circles) made for a certain objective are used to update the surrogate model for that objective. (b) In the combined acquisitions strategy, acquisitions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

for both objectives, and we might as well perform the two separate BO processes consecutively rather than simultaneously. Therefore, it is worthwhile investigating if a more efficient acquisition strategy can be obtained by sharing acquisitions between two simultaneous BO processes. To this end, we consider a strategy where data acquired from different surrogate models are combined and used to update all existing surrogate models at every BO iteration. This means that the final surrogate models for the two objectives are constructed using the same set of acquisitions and consequently we refer to this as the combined acquisitions (CA) strategy (Fig. 3b).

G. Pareto Front Analysis

In this work we also consider the simultaneous optimization of several experimental outputs that have been mapped as functions of the design variables via BO. In general, a single solution does not exist for such a multi-objective optimization problem, rather we must look for optimal trade-offs between the objectives involved. Mathematically, the notion of an optimal trade-off is formalized by the concept of Pareto optimality (see SI for an extended description). The theory tells us that a combination of objective values that can be feasibly obtained constitutes an optimal trade-off if an improvement in one objective is always detrimental to at least one other objective. We refer to an optimal trade-off as a Pareto point, and the set of all Pareto points is known as the Pareto front. The extraction conditions corresponding to a Pareto point is known as a Pareto optimal solution. Once the Pareto front is known, a particular Pareto point can be chosen as the preferred solution to the optimization problem, e.g., based on the relative importance of the design criteria.

III. RESULTS AND DISCUSSION

The presentation of our results is organized as follows: We start from a machine learning perspective and give a high-level illustration of the BO-driven data collection and its most salient features, such as the iterative process acquisitions and surrogate model updates. We then consider in greater detail the acquisition strategy and how it can be tuned to select informative data points when we are interested in more than one objective. To conclude the machine learning-focused parts of the results, we study the convergence of the surrogate models and assess their predictive power. We then turn to surrogate model predictions for the extracted lignin yield and key structural properties and the insight that these can provide into the extraction process and lignin chemistry. Finally, we show how to simultaneously optimize combinations of the yield and structural properties to meet a given set of design criteria for high-value applications in lignin valorization.

A. Data Collection Driven By 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, 4, where 0 corresponds to the initial batch. To visualize a surrogate model, we use 2D contour plots of the design space variables (P_f, T) versus the predicted objective values and refer to such plots as landscapes. The evolution of the extracted lignin yield and β-O-4 content landscapes as new batches of data are acquired using the CA strategy is visualized in Fig. 4. The figure also includes the predicted standard deviation, which quantifies the uncertainty in the objective value predictions. We note that for the CA strategy, the standard deviations of all objectives 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 both the evolution of the lignin yield and β-O-4 content standard deviation (Fig. 4, bottom).

The initial batch of Sobol points yields landscapes that, due to the overall data scarcity, is dominated by the mean of the observed values, and the surrogate models have essentially no predictive power. The lack of data is, at this stage, also reflected in the predicted standard deviation, which is uniformly high away from the Sobol points. As the dataset is further extended in batches two to four, the landscapes become smoother and the surrogate models predictive power, i.e., ability to accurately interpolate between acquisitions increases. Indeed, already in batch one, where four new acquisitions have been added to the initial data, both the lignin yield and β-O-4 content landscapes have resolved into two regions of predicted high and low values. This qualitative feature remains in the next three batches, but the size and shape of the regions, including the extrema, shift in the process. The refinement of the landscapes is accompanied by a corresponding decrease in the predicted standard deviation. In batch four a uniformly low standard deviation is obtained where the lower limit set by the experimental error built into the surrogate model. At this stage, the landscapes are converging as suggested qualitatively by the relatively small feature changes between the third and fourth batch. We furthermore observe that locations of the acquisitions in batches one to four are quite different from, e.g., what a space-filling design would yield in the sense that the sampling density of the design space is much less uniform. This is due to the exploitation-exploration trade-off in the acquisition strategy, which means that regions of potentially high objective values or high variance are preferentially sampled to maximize the relevant information contributed towards the design process. The result is a rapid improvement of the surrogate models with added data, which is one of the key features that makes BO an effective approach to DOE problems.

B. Choosing Informative Acquisitions

In the previous section we focused on the overarching picture of using BO to guide and model experiments. We demonstrated what the BO process looks like for the lignin yield and provided qualitative arguments for why it works. The fact that Fig. 4 was based on CA acquisition strategy was only briefly mentioned, however, and no detailed analysis of the acquisitions was given. In this section, we therefore compare the CA and PA strategies in terms of the informativeness of their respective generated acquisitions. In this context, an informative acquisition is one that provides the surrogate model which useful information, i.e., information on objective values in regions that have not been probed (exploration) or where a maximum is likely to be found (exploitation). 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 from the evolution of the lignin yield landscape is shown in Fig. 5. Note that for β-O-4 acquisitions we only show the CA strategy, since in the PA scheme they are not considered for the purpose of constructing the lignin yield surrogate model. In Fig. 5, acquisitions are labeled according to whether they are (1) acquired for the lignin yield or β-O-4 content and (2) generated by the eLCB or pure exploration functions. Corresponding snapshots for the evolution of the β-O-4 content are shown in Fig. S3. For both strategies, 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 acquisi-

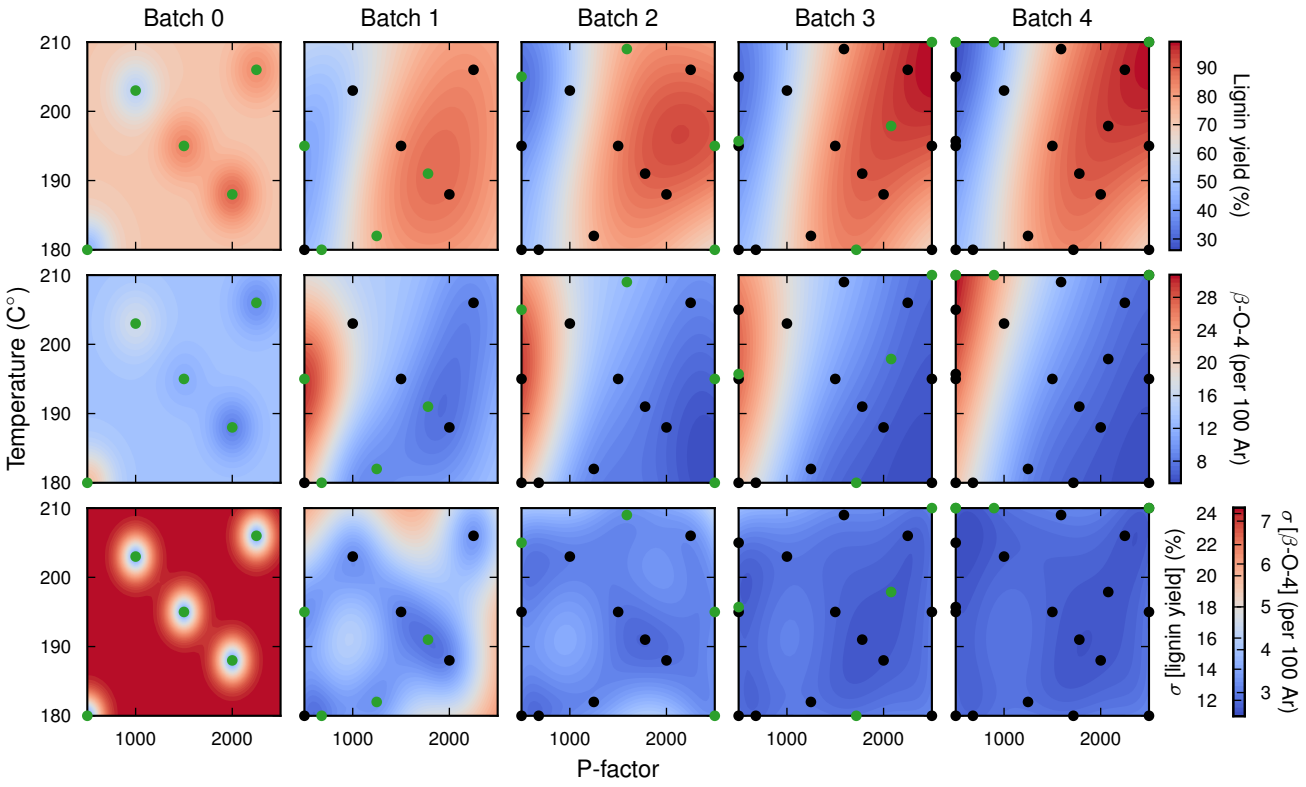

FIG. 4. Successive improvement of the surrogate models for the lignin yield (top) and β-O-4 content (middle) as new acquisitions (green circles) are added to the existing dataset (black circles). As the landscapes evolves, the model's capability of predicting the yield in unknown regions of design space increases. The prediction uncertainties are quantified by the model standard deviation (bottom) which simultaneously decreases as more data is collected. The acquisition strategy balances exploitation of regions where the yield is large and exploration of regions with high uncertainty.

tions all occur in the vicinity of the $P_f \approx 500$ region with 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 Fig. 5 clearly illustrates how acquisitions from the eLCB and pure exploration functions relate to features in the landscape, it does not allow us to directly compare the performance of the CA and PA strategies since the PA surrogate models use fewer data point than their CA counterparts. To accomplish this, we instead need to compare surrogate models constructed from the entire PA dataset, including β-O-4 acquisitions (Figs. S4 and S5). When comparing these landscapes, the differences between the strategies then lie solely in the location of the acquisitions, rather than their number. 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, allowing for more informative points to be picked during both exploitation and exploration. We can thus conclude that the CA is a more effective acquisition strategy for BO-driven DOE targeting multiple objectives than PA.

In generating eLCB acquisitions for more than one objective, our results further highlight the importance of choosing objectives 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 tend to have little overlap since high objective values are attained for high and low P_f, respectively (Fig. 5). In contrast, a high degree of similarity between the landscapes would lead to a high degree of overlap between the eLCB acquisitions generated for the different objectives and consequently less informative batches. Using either strategy, a significant fraction, 62% for CA, of the acquisitions are made on the design space boundary. This is a result of the increased exploration and non-periodicity of the problem, since points close to the design space boundary have, on average, fewer neighboring acquisitions that can be used by the surrogate model to effectively interpolate. In applications of BO to experiments, design space bounds thus need to be chosen carefully such that experiments are still feasible for any, potentially more extreme, processing conditions implied by a boundary acquisition.

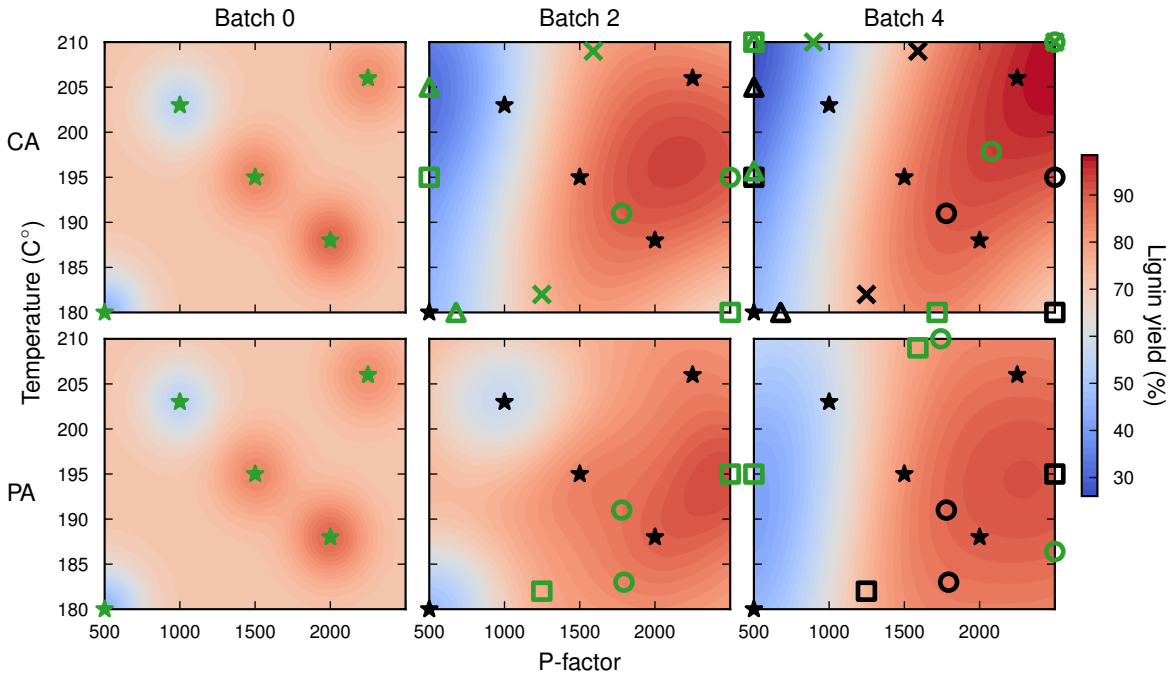

FIG. 5. Comparison of the combined acquisitions (CA, top) and pure acquisitions (PA, bottom) strategies. Snapshots of the lignin yield surrogate models are shown for batches 0,2 and 4 with new (green) and existing (black) acquisitions labeled according to the objective and acquisitions function from 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 objectives 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 $T = 210\,\mathrm{K}$.

C. 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 note that the size of the test set is restricted by the significant cost of performing additional experiments. The predictive power of the surrogate models for lignin yield and β-O-4 content was subsequently evaluated for both the test set as well as the acquisitions used to train the model. A comparison of the predictions with the experimentally measured values is displayed in Fig. 6. It can be seen that, qualitatively, both surrogate models perform well on the test set and that the measured values all fall within one standard deviation from the predicted values (Fig. 6b,d). We show the corresponding results for the PA strategy in Fig. S6.

Based on the validation data, several quantitative performance metrics were also calculated (I). One important metric is the mean absolute percentage error (MAPE), which measures the mean prediction error on a data set. For the acquisitions set, the MAPE is 6.8% and 11.0% 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. This noise ensures that the surrogate models are not overfitted to the acquisition data since this typically is detrimental to the predictive power of the models. The corresponding MAPEs for the test data are 7.8% for the lignin yield and 8.6 for the β-O-4 content. The fact that these numbers are comparable to the MAPEs for the acquisitions set, and thus by extension the estimated experimental error, implies that the surrogate models are well converged and can provide accurate predictions.To assess the convergence of the predicted maxima of the lignin yield and β-O-4 content were also monitored (Fig. S7).

We emphasize the role of exploratory acquisitions in obtaining this level of surrogate model accuracy over all of design space despite using small dataset consisting of only 21 points. Incorporating additional exploration is crucial since traditional improvement-based acquisition functions are constructed to preferentially converge the surrogate model in regions around extrema. An additional benefit of promoting exploration is that the final dataset is better suited for fitting surrogate models for other objectives that were measured during the BO-

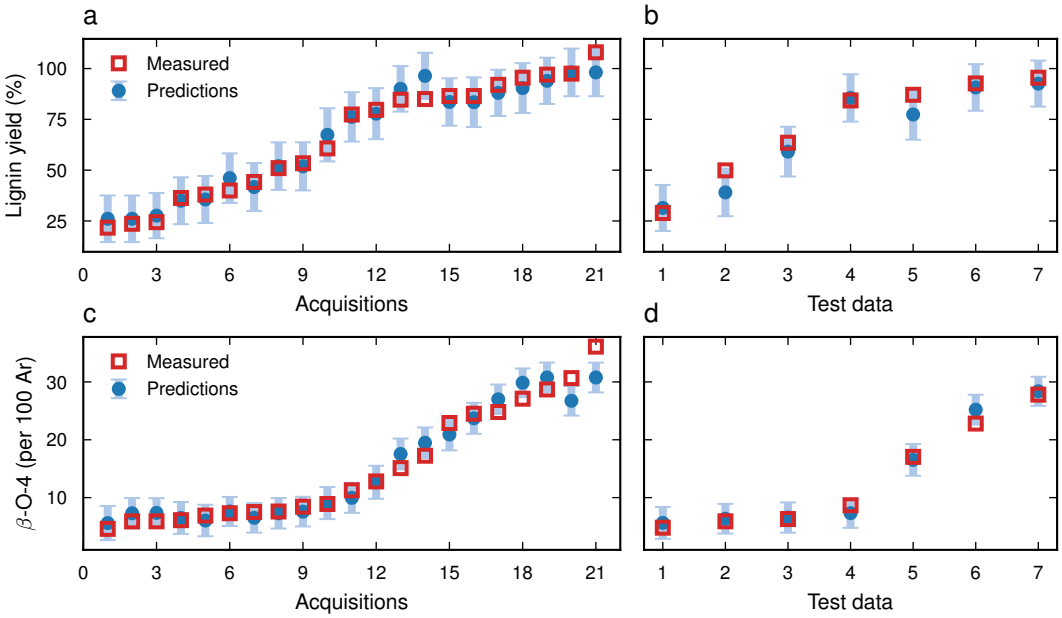

FIG. 6. Validation of the surrogate models trained for the lignin yield (a-b) and β-O-4 content (c-d). To assess the accuracy of the models' predictions, the measured objective values for a set of independently collected test data is used (right). The predictions for the test set are contrasted to the predictions for the set of acquisitions used to train the models (left). The errorbars 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.

TABLE I. Summary of model validation metrics for the lignin yield and β-O-4 content. The estimated experimental error is followed by the root mean square error (RMSE), mean absolute error (MAE), mean average percentage error (MAPE) and mean predicted standard deviation (MPSD). The MAPE obtained for 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.

Objective	Exp. error	Data set	RMSE	MAE	MAPE	MPSD
Lignin yield	5%	Acquisitions	4.7	3.8	6.8	11.8
(%)		Test	6.0	4.8	7.8	11.7
β-O-4 content	5–10%	Acquisitions	2.0	1.6	11.0	2.6
(per 100 Ar)		Test	1.1	0.9	8.6	2.6

guided data collection, since exploration, unlike exploitation, is not inherently objective-specific.

D. Model Predictions for Key Lignin Properties

Up until this point, we have studied the surrogate models and data collection process from a machine learning perspective. Now, we consider how the surrogate model can be used to learn about the experimental process. To this end, we consider the surrogate model predictions for a selected set of the measured lignin properties that are of key importance for lignin chemistry and valorization. To complement the already established surrogate models for lignin yield andβ-O-4 content we fit additional surrogate models for the S/G ratio and the total carbohydrate content.

For the lignin yield, the general trend is that a higher P-factor and a higher temperature leads to increased yield, with a predicted maximum yield of $98 \pm 13\%$ at $(P_f, T) = (2500, 207\,^{\circ}\mathrm{C})$. The increase in lignin yield with P-factor is related to formation of more acetone-soluble lignin fragments, as well as cleavage of lignin-carbohydrate linkages under more severe reaction conditions. The average measured lignin yield of 108% at $(P_f, T) = (2500, 210\,^{\circ}\mathrm{C})$ also indicates the formation of polyfurans from xylan degradation products (furfural) that is quantified as lignin [7] (so-called pseudo-lignin [43, 44]) . We note that the measured yield at these conditions exceeds the predicted maximum yield. The deviation is due to the inclusion of experimental error in the surrogate model, and the measured value is still within one standard deviation of the predicted maximum. The β-O-4 content exhibits a behavior antagonistic to the lignin yield in the sense that high β-O-4 content can only be achieved at low P-factor. Indeed, at lower reac-

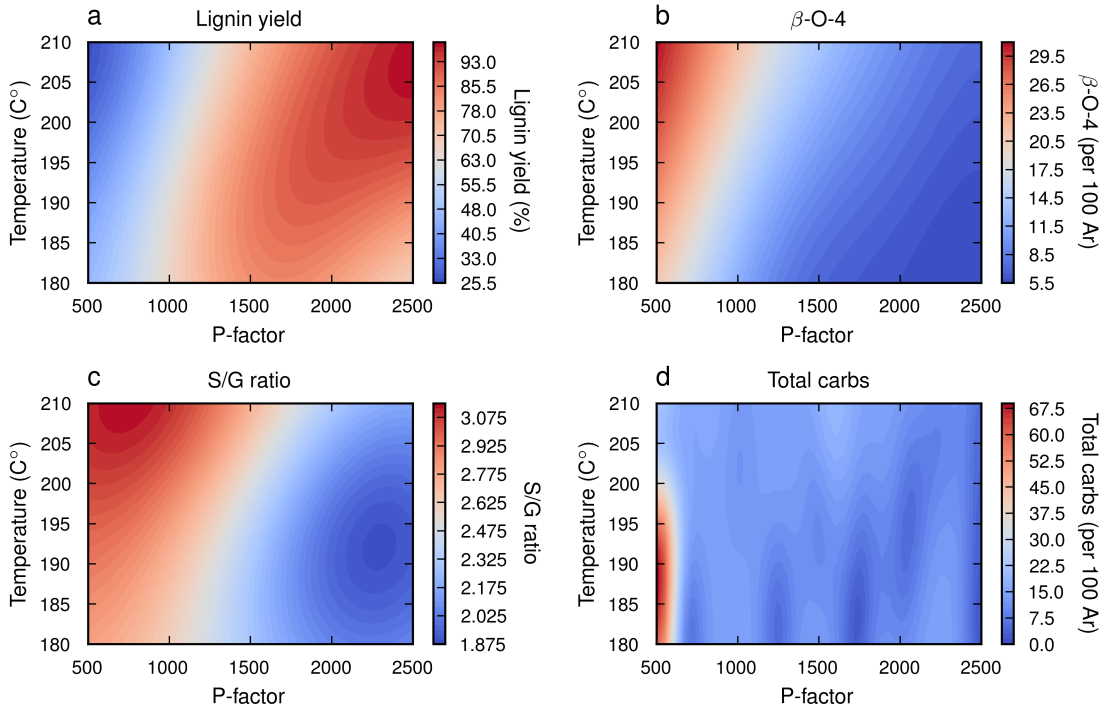

FIG. 7. Predicted landscapes for key lignin properties. a) The lignin yield increases with both temperature and P-factor. b) The β-O-4 content exhibits an antagonistic relationship with the lignin yield and is large when the P-factor is low. c) The ratio of syringyl to guaiacyl units is also large when 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.

tion severity, we can generally expect that the moieties which are rich in native lignin, such as β-O-4, are broken down to a lesser extent.

Another important property is S/G ratio, which is important in specific high-value lignin applications. The highest S/G ratio was found at low process severity, likely due to higher reactivity of S-units in lignin fragmentation and therefore their predominant release (extraction with the solvent) at the initial stage of the process. We see that while the (P_f, T)-dependence of these properties can to some extent be qualitatively explained by basic lignin chemistry considerations, the surrogate models provide the quantitative predictions which are necessary for large-scale applications in lignin valorization.

E. Tailoring Extraction Conditions for Different Lignin Applications

Having established surrogate models for a range of objectives of interest, we have all the information required to derive optimal extraction conditions for arbitrary design criteria associated with lignin-based products. Design criteria generally take the form of a set of constraints on the physicochemical properties of the lignin imposed by the application. In the present case, we only consider the structural properties for which we have trained surrogate models. In addition to constraints on the properties, the yield of the extracted lignin also needs to exceed some minimum threshold for an application to be financially viable. Thus, we are looking to find extraction conditions that give a high yield and match one or more other design criteria relating to the lignin properties. To solve this problem, a Pareto front analysis can be employed to find optimal trade-offs between the lignin yield and properties. The practical implication of having multiple optimal trade-offs is that changing the extraction conditions to bring one objective closer to its design criteria will result in at least one other objective 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. [4] Hence, the maximal revenue (per the original biomass, e.g., AqSO feed) correlates with both high AEL yield, and high β-O-4 content. Using our surrogate models we can determine the feasible combinations of yield and β-O-4 content, which form a two-dimensional area (Fig. 8a). By subsequently calculating the Pareto front, we see that it

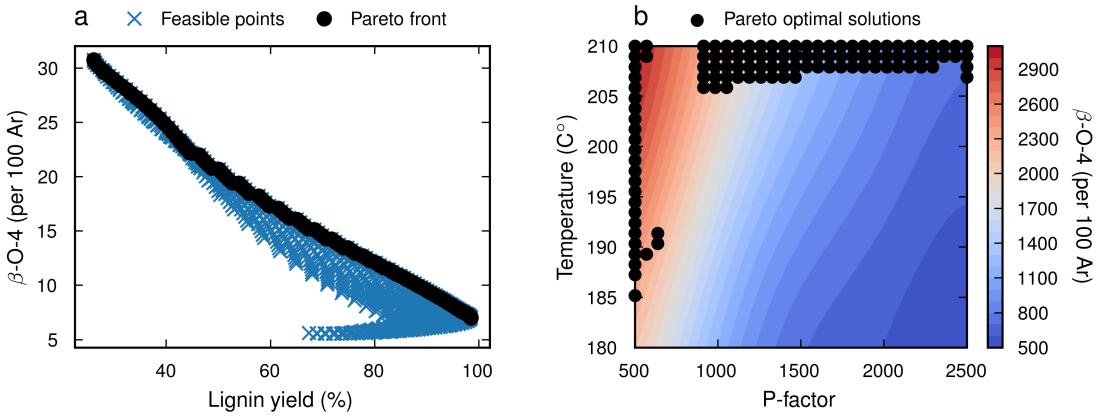

FIG. 8. Optimizing the AqSO biorefinery for production of aromatic platform chemicals by simultaneously attempting to maximize lignin yield and β-O-4 content. The Pareto front (a) and corresponding Pareto optimal solutions (b) represent optimal trade-offs for the objectives. A single point on the Pareto front can be selected by assigning a desired value to one of the objectives. The corresponding Pareto optimal solution then provides the optimal extraction conditions for the valorization.

constitutes part of the feasible points boundary (Fig. 8a). We observe from the shape of the Pareto front that high values of the AEL yield are correlated with low β-O-4 content, in agreement with Fig. 7a-b). The corresponding Pareto optimal solutions, i.e., (P_f, T)-pairs corresponding to optimal trade-offs are then found from the projection of the Pareto front in (P_f, T)-space (Fig. 8b). We notice that all trade-offs involving high yield are obtained at high temperature, whereas trade-offs involving high β-O-4content are obtained a low P-factor. To determine the processing conditions, we should perform the lignin extraction at, we need to narrow down the Pareto front to a single point by imposing an additional constraint. For instance, we might accept a slightly smaller yield of say 60%, in return for a higher β-O-4content. We can then determine from Fig. 8 that this trade-off results in 17.5 β-O-4 linkages per 100 Ar and can be achieved by extracting at $(P_f, T) = (1227, 210\,^{\circ}\mathrm{C})$. We note that this modeling-based design process is completely general and can be applied to any application with an arbitrary number of different properties, as long as the corresponding surrogate models have been obtained. We could envision an application where a high amount of phenolic OH groups is beneficial, e.g., for the use of lignin as an antioxidant. As phenolic OH is a typical product of β-O-4-unit cleavage, one should expect the highest phenolic OH content at the lowest number of β-O-4-linkages. By Fig. 8, a high phenolic OH content would then correlate positively with the lignin yield, and we can consequently expect to obtain trade-offs with high values for both quantities. Similar examples could be given for the optimal S/G ratio and amounts of carbohydrate in lignin (present as lignin carbohydrate complexes). For example, higher proportion of S-units should be beneficial for post-processing lignin depolymerization due to the higher reactivity of S-units compared to that of G-units. In contrast, S-units are not suitable for various crosslinking reactions as both ortho-positions to the phenolic OH (typical reaction centers in crosslinking) in the aromatic ring of S-units are occupied by OMe-groups. The presence of hydrophilic carbohydrates moieties (as lignin-carbohydrate complexes) can be advantageous in surfactant applications, but other lignin applications (e.g., production of aromatic monomers) require high purity lignin.

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, using the surrogate models presented here, to establish predictive models that link the entire value chain, starting from the effect of the processing conditions on lignin structure and properties, to the effect of these properties and their effect on lignin performance in specific applications.

IV. CONCLUSIONS

The AqSO biorefinery provides a flexible framework for obtaining lignin-based products in a facile and sustainable manner. To properly utilize AqSOs versatility, we have presented a machine learning approach to designing and optimizing the biorefinery based on BO. The BO approach to DOE can be distinguished from classical methods by the coupled model building and data collection policy, which ensures that planned experiments are always chosen to provide as much useful information as possible given the current state of the model. While previous studies on applying BO to experimental work focused largely on optimizing for a pre-determined design criterion, our approach also highlights capability of BO to provide predictive modeling.

We showed that predictive models with an accuracy

comparable to the experimental error could be obtained for the yield of the extracted lignin and structural properties such as the β-O-4 content and S/G ratio using only 21 data points. An important feature of our approach was the use of parallel exploratory and exploitative acquisitions performed simultaneously for multiple objectives to take advantage of batched experiments. The extra emphasis on exploration was crucial in obtaining models that are predictive over all of design space rather than just in the vicinity of a maxima as is commonly observed for conventional acquisition strategies. When splitting acquisitions between multiple objectives we furthermore demonstrated that models converge faster when the acquisitions are combined before the models are updated (CA strategy) compared to the baseline acquiring new data separately for the objectives (PA strategy).

Significant effort was also put toward model interpretation and analysis. Our models predict that lignin yield and several structural properties such as β-O-4 content, S/G ratio and carbohydrate content cannot simultaneously be obtained in high quantities. In addition to providing new insights, our findings bring a quantitative dimension to previously observed trends for the AqSO biorefinery. We furthermore describe how to use predictive models to find optimal extraction conditions corresponding to a set design criterion, expressed in terms of the objectives, using a Pareto front analysis. The process is illustrated for the case of breaking lignin down into value-added chemicals, which requires a both a high yield and S/G ratio. The result is a set of optimal tradeoffs between the yield and S/G ratio where a single solution can subsequently be chosen based on a quantitative requirement on either property.

In the larger context of experiment design, our work indicates that BO can play a bigger role than indicated by previous efforts focusing on global optimization. Indeed, new experimental methodology and material processing techniques are not always developed with a straightforward optimization problem in mind. In these cases, one should arguable focus more on obtaining globally predictive surrogate models that can later be used to a solve specific design problems via a Pareto analysis. While the underlying mathematics of BO remains relatively complex, newly developed codes such as BOSS are paving the way for more stream-lined and user-friendly applications of BO to experiments.

ASSOCIATED CONTENT

Supporting Information

The supporting information contains a more technical introduction to BO and the concept of Pareto optimality. A HSQC spectrum for the acetone-extractable lignin is also included, along with complementary results for the acquisition strategy comparisons and the model validation.

Data Availability

The data and code that was used in this study is available from the corresponding author, M.T., upon reasonable request.

AUTHOR INFORMATION

Corresponding Author

*E-mail: milica.todorovic@utu.fi

Orcid

Joakim Löfgren: 0000-0001-6968-5966
Milica Todorović: 0000-0003-0028-0105
Patrick Rinke: 0000-0003-1898-723X

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).

[1] S. Chatterjee and T. Saito, Lignin-Derived Advanced Carbon Materials, ChemSusChem **8**, 3941 (2015).
[2] W. Fang, S. Yang, X.-L. Wang, T.-Q. Yuan, and R.-C. Sun, Manufacture and application of lignin-based carbon fibers (LCFs) and lignin-based carbon nanofibers (LCNFs), Green Chemistry **19**, 1794 (2017).
[3] R. J. Li, J. Gutierrez, Y.-L. Chung, C. W. Frank, S. L. Billington, and E. S. Sattely, A lignin-epoxy resin derived from biomass as an alternative to formaldehyde-based wood adhesives, Green Chemistry **20**, 1459 (2018).
[4] Z. Sun, B. Fridrich, A. de Santi, S. Elangovan, and K. Barta, Bright Side of Lignin Depolymerization: Toward New Platform Chemicals, Chemical Reviews **118**, 614 (2018).
[5] H. Jørgensen, J. B. Kristensen, and C. Felby, Enzymatic conversion of lignocellulose into fermentable sugars: Challenges and opportunities, Biofuels, Bioproducts and Biorefining **1**, 119 (2007).
[6] A. J. Ragauskas, G. T. Beckham, M. J. Biddy, R. Chandra, F. Chen, M. F. Davis, B. H. Davison, R. A. Dixon, P. Gilna, M. Keller, P. Langan, A. K. Naskar, J. N. Saddler, T. J. Tschaplinski, G. A. Tuskan, and C. E. Wyman, Lignin Valorization: Improving Lignin Processing in the Biorefinery, Science **344**, 10.1126/science.1246843 (2014).
[7] T. V. Lourencon, L. G. Greca, D. Tarasov, M. Borrega, T. Tamminen, O. J. Rojas, and M. Y. Balakshin, Lignin-First Integrated Hydrothermal Treatment (HTT) and

Synthesis of Low-Cost Biorefinery Particles, ACS Sustainable Chemistry & Engineering **8**, 1230 (2020).
[8] F. Carvalheiro, L. Duarte, and F. Gírio, Hemicellulose biorefineries: A review on biomass pretreatments, Journal of scientific and industrial research **67**, 849 (2008).
[9] S.-Y. Jeong and J.-W. Lee, Hydrothermal Treatment, in *Pretreatment of Biomass*, edited by A. Pandey, S. Negi, P. Binod, and C. Larroche (Elsevier, Amsterdam, 2015) pp. 61–74.
[10] A. Romaní, G. Garrote, and J. C. Parajó, Bioethanol production from autohydrolyzed Eucalyptus globulus by Simultaneous Saccharification and Fermentation operating at high solids loading, Fuel **94**, 305 (2012).
[11] A. Romaní, G. Garrote, J. L. Alonso, and J. C. Parajó, Bioethanol production from hydrothermally pretreated Eucalyptus globulus wood, Bioresource Technology **101**, 8706 (2010).
[12] R. A. Fisher, *The Design of Experiments*, The Design of Experiments (Oliver & Boyd, Oxford, England, 1935) pp. xi, 251.
[13] D. C. Montgomery, *Design and Analysis of Experiments* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2006).
[14] M. D. McKay, R. J. Beckman, and W. J. Conover, A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, Technometrics **21**, 239 (1979).
[15] J. C. Helton and F. J. Davis, Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, Reliability Engineering & System Safety **81**, 23 (2003).
[16] L. M. Collins, J. J. Dziak, K. C. Kugler, and J. B. Trail, Factorial Experiments: Efficient Tools for Evaluation of Intervention Components, American journal of preventive medicine **47**, 498 (2014).
[17] G. E. P. Box and K. B. Wilson, On the Experimental Attainment of Optimum Conditions, Journal of the Royal Statistical Society. Series B (Methodological) **13**, 1 (1951).
[18] R. H. Myers, D. C. Montgomery, G. G. Vining, C. M. Borror, and S. M. Kowalski, Response Surface Methodology: A Retrospective and Literature Survey, Journal of Quality Technology **36**, 53 (2004).
[19] W. Duch and G. H. F. Diercksen, Neural networks as tools to solve problems in physics and chemistry, Computer Physics Communications **82**, 91 (1994).
[20] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis, Improved protein structure prediction using potentials from deep learning, Nature **577**, 706 (2020).
[21] B. Tang, Z. Pan, K. Yin, and A. Khateeb, Recent Advances of Deep Learning in Bioinformatics and Computational Biology, Frontiers in Genetics **10**, 214 (2019).
[22] J. Mockus, On Bayesian Methods for Seeking the Extremum, in *Proceedings of the IFIP Technical Conference* (Springer-Verlag, Berlin, Heidelberg, 1974) pp. 400–404.
[23] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, Taking the Human Out of the Loop: A Review of Bayesian Optimization, Proceedings of the IEEE **104**, 148 (2016).
[24] D. Ginsbourger, R. Le Riche, and L. Carraro, Kriging Is Well-Suited to Parallelize Optimization, in *Computational Intelligence in Expensive Optimization Problems*, Adaptation Learning and Optimization, edited by Y. Tenne and C.-K. Goh (Springer, Berlin, Heidelberg, 2010) pp. 131–162.
[25] A. Shah and Z. Ghahramani, Pareto frontier learning with expensive correlated objectives, in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16 (JMLR.org, New York, NY, USA, 2016) pp. 1919–1927.
[26] E. Garijo del Río, J. J. Mortensen, and K. W. Jacobsen, Local Bayesian optimizer for atomic structures, Physical Review B **100**, 104103 (2019).
[27] M. K. Bisbo and B. Hammer, Efficient Global Structure Optimization with a Machine-Learned Surrogate Model, Physical Review Letters **124**, 086102 (2020).
[28] M. Todorović, M. U. Gutmann, J. Corander, and P. Rinke, Bayesian inference of atomistic structure in functional materials, npj Computational Materials **5**, 1 (2019).
[29] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, Accelerated search for materials with targeted properties by adaptive design, Nature Communications **7**, 11241 (2016).
[30] C. Li, D. Rubín de Celis Leal, S. Rana, S. Gupta, A. Sutti, S. Greenhill, T. Slezak, M. Height, and S. Venkatesh, Rapid Bayesian optimisation for synthesis of short polymer fiber materials, Scientific Reports **7**, 5683 (2017).
[31] P. B. Wigley, P. J. Everitt, A. van den Hengel, J. W. Bastian, M. A. Sooriyabandara, G. D. McDonald, K. S. Hardman, C. D. Quinlivan, P. Manju, C. C. N. Kuhn, I. R. Petersen, A. N. Luiten, J. J. Hope, N. P. Robins, and M. R. Hush, Fast machine-learning online optimization of ultra-cold-atom experiments, Scientific Reports **6**, 25890 (2016).
[32] A. Vahid, S. Rana, S. Gupta, P. Vellanki, S. Venkatesh, and T. Dorin, New Bayesian-Optimization-Based Design of High-Strength 7xxx-Series Alloys from Recycled Aluminum, JOM **70**, 2704 (2018).
[33] S. Sun, A. Tiihonen, F. Oviedo, Z. Liu, J. Thapa, Y. Zhao, N. T. P. Hartono, A. Goyal, T. Heumueller, C. Batali, A. Encinas, J. J. Yoo, R. Li, Z. Ren, I. M. Peters, C. J. Brabec, M. G. Bawendi, V. Stevanovic, J. Fisher, and T. Buonassisi, A data fusion approach to optimize compositional stability of halide perovskites, Matter **4**, 1305 (2021).
[34] Q. Liang, A. E. Gongora, Z. Ren, A. Tiihonen, Z. Liu, S. Sun, J. R. Deneault, D. Bash, F. Mekki-Berrada, S. A. Khan, K. Hippalgaonkar, B. Maruyama, K. A. Brown, J. Fisher III, and T. Buonassisi, Benchmarking the Performance of Bayesian Optimization across Multiple Experimental Materials Science Domains, arXiv:2106.01309 [cond-mat, physics:physics] (2021), arXiv:2106.01309 [cond-mat, physics:physics].
[35] H. Sixta, Kraft Pulping Kinetics, in *Handbook of Pulp* (John Wiley & Sons, Ltd, 2006) pp. I–XXXII.
[36] E. Brochu, V. M. Cora, and N. de Freitas, A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning, arXiv:1012.2599 [cs] (2010), arXiv:1012.2599 [cs].
[37] D. Cox and S. John, A statistical method for global optimization, in *[Proceedings] 1992 IEEE International Conference on Systems, Man, and Cybernetics* (1992) pp.

1241–1246 vol.2.

[38] N. Srinivas, A. Krause, S. Kakade, and M. Seeger, Gaussian process optimization in the bandit setting: No regret and experimental design, in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10 (Omnipress, Madison, WI, USA, 2010) pp. 1015–1022.

[39] J. Järvi, P. Rinke, and M. Todorović, Detecting stable adsorbates of (1S)-camphor on Cu(111) with Bayesian optimization, Beilstein Journal of Nanotechnology **11**, 1577 (2020).

[40] J. Järvi, B. Alldritt, O. Krejčí, M. Todorović, P. Liljeroth, and P. Rinke, Integrating Bayesian Inference with Scanning Probe Experiments for Robust Identification of Surface Adsorbate Configurations, Advanced Functional Materials **n/a**, 2010853 (2021).

[41] L. Fang, E. Makkonen, M. Todorović, P. Rinke, and X. Chen, Efficient Amino Acid Conformer Search with Bayesian Optimization, Journal of Chemical Theory and Computation **17**, 1955 (2021).

[42] S.-A. Jin, T. Kämäräinen, P. Rinke, O. J. Rojas, and M. Todorovic, Machine Learning as a Tool to Engineer Microstructures: Morphological Prediction of Tannin-Based Colloids Using Bayesian Surrogate Models, ChemRxiv 10.26434/chemrxiv.14477691.v1 (2021).

[43] J. Li, G. Henriksson, and G. Gellerstedt, Carbohydrate reactions during high-temperature steam treatment of aspen wood, Applied Biochemistry and Biotechnology **125**, 175 (2005).

[44] F. Hu, S. Jung, and A. Ragauskas, Pseudo-lignin formation and its impact on enzymatic hydrolysis, Bioresource Technology **117**, 7 (2012).