Machine learning as a tool to engineer microstructures: Morphological prediction of tannin-based colloids using Bayesian surrogate models

Soo-Ah Jin,[†] Tero Kämäräinen,[‡] Patrick Rinke,[§] Orlando J. Rojas[‡,^,] and Milica Todorović[§,||,*]*

†Department of Chemical & Biomolecular Engineering, North Carolina State University, Raleigh, North Carolina 27695, United States.

‡Department of Bioproducts and Biosystems, Aalto University, Vuorimiehentie 1, Espoo, P.O. Box 16300, FI-00076 Aalto, Finland.

§ Department of Applied Physics, Aalto University, P.O. Box 11100, FI-00076 Aalto, Finland.

^ Bioproducts Institute, Departments of Chemical & Biological Engineering, Chemistry, and Wood Science, 2360 East Mall, The University of British Columbia, Vancouver, BC V6T 1Z3, Canada.

||Department of Mechanical and Materials Engineering, University of Turku, FI-20014 Turku, Finland.

*Corresponding author: milica.todorovic@aalto.fi, orlando.rojas@ubc.ca

KEYWORDS tannic acid, Gaussian process regression, morphology prediction

ABSTRACT

Oxidized tannic acid (OTA) is a useful biomolecule with a strong tendency to form complexes with metals and proteins. In this study we open the possibility to further the application of OTA when assembled as supramolecular systems, which typically exhibit functions that correlate with shape and associated morphological features. We use artificial intelligence (AI) to selectively engineer OTA into particles encompassing 1-dimensional (1D) to 3-dimensional (3D) constructs. We employed Bayesian regression to correlate colloidal suspension conditions (pH and pK_a) with the size and shape of the assembled colloidal particles. Fewer than 20 experiments were found to be sufficient to build surrogate model landscapes of OTA morphology in the experimental design space, which were chemically interpretable and endowed predictive power on data. We produced multiple property landscapes from the experimental data, helping us to infer solutions that would satisfy, simultaneously, multiple design objectives. The balance between data efficiency and the depth of information delivered by AI approaches testify to their potential to engineer particles, opening new prospects in the emerging field of particle morphogenesis, impacting bioactivity, adhesion, interfacial stabilization and other functions inherent to OTA.

INTRODUCTION

Tannic acid (TA) is an abundant and versatile bio-based material, which readily affords synthetic pathways for the isolation of its elementary building blocks. TA contains many hydroxyl groups, allowing it to form complexes with different macromolecules via hydrogen-bonding, hydrophobic and cation-π interactions [1,2]. Abundant hydroxyl groups make TA highly soluble and stable in aqueous solutions. In alkaline conditions, TA undergoes oxidation [3,4] and produces oxidized tannic acid (OTA) followed by oligomerization. Concomitant oligomerization of OTA leads to the formation of compounds with higher molecular weight and thereby decreases the solubility of the substance [4]. In this form, OTA can interact with different molecules and serve as coatings [3,5], surface modifiers [1,6] and emulsion stabilizers [1,3,6,7], or act as stabilizing and reducing agents to aid in inorganic nanoparticle growth [8–10], all the while imparting beneficial biological functionality [11,12]. For instance, tannic acid has recently been shown to suppresses SARS-CoV-2 as a dual inhibitor of the viral main protease [13]. All these favorable aspects of OTA and other phenolic particles have fueled research into a wide spectrum of applications [14].

OTA can also be crystallized into particles with structural properties that are highly sensitive to the experimental synthesis. Previously, Bhangu *et al.* [10] developed a sonochemical method to chemically transform amorphous tannic acid into nano/micro-sized crystalline particles without the use of reagents or organic solvents. They obtained OTA particles of different size and shape by simply varying ultrasonic parameters. Kämäräinen *et al.* [4] further presented a facile and scalable protocol to prepare OTA of varying morphologies by altering the TA oxidation conditions. The dimensions, shapes and the yield of these crystalline particles were highly sensitive to initial TA concentration, reaction time, initial pH and pK_a of the base.

While OTA particulate constructs can facilitate a range of new applications, particle morphology is a key consideration. In many high surface area systems that incorporate particulate matter, particle morphology and size are major contributors to their overall performance through, e.g., relationships between morphology and packing [15], percolation [16], rheology [17] and bioactivity [18]. Consequently, morphological concerns have been recognized to play an important role in many applications ranging from heterogeneous catalysts [19] and electrochemical cells [20,21] to drug delivery systems [22], among others.

In this work, we employ artificial intelligence (AI) to explore the morphology landscape of OTA particles in the chemical design space of processing conditions. As illustrated in **Figure 1**, we start with OTA synthesis experiments and digitalize them into data points for particle morphology. We apply Gaussian Process Regression (GPR) [23], an AI tool for supervised learning, to compute a surrogate model for OTA morphology. Based on the morphology model in the design space of chemical synthesis, we consider which particle shapes are achievable, and learn how to tune the processing conditions to achieve an optimal outcome for a targeted application.

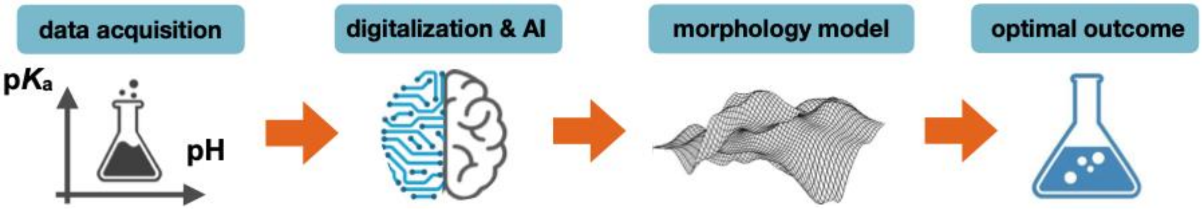

Figure 1. Workflow for AI-guided morphology control of synthesized OTA particles

GPR has been employed in materials science for experimental materials design [24–30], often in combination with Bayesian optimization [31–33]. Given data within the phase space of *N* design parameters, GPR produces the statistically most likely *N*-dimensional landscape, which serves as a surrogate model of a target property [23]. Gaussian processes (GPs) are capable of good data interpolation, allowing us to build good quality surrogate models with relatively few data points. They produce smooth and continuous landscapes, that reflect the continuous chemical process underpinning the data, and can account for experimental uncertainties as data noise. All these characteristics makes GPR well-suited to experimental applications.

The previous study of OTA particle synthesis employed principal component analysis [34,35] (PCA, an AI tool for unsupervised learning) on experimental data to ascertain that pH and pK_a used in the OTA solution correlate most strongly with particle shape. We proceed to consider OTA morphology in the 2-dimensional search space of pH and pK_a. Sample characterization was performed by scanning electron microscopy (SEM) imaging. To digitalize the particle shape information, we quantified the physical dimensions allowed by OTA simple crystalline habits and took note of experimental uncertainties.

While PCA is a versatile tool, it was unable to offer further insight into morphology types, nor indicate optimal processing conditions. Conversely, GPR allowed us to visualize OTA particle morphology as a function of pH and pK_a and delivered a chemically interpretable model. Based on the morphology landscape, our objective was to drive the morphology of particles from one-dimensional (1D) to three-dimensional (3D) shapes. Moreover, by extracting particle yield and volume from each experiment we were able to generate surrogate models for multiple experimental properties at no further cost, allowing us to pursue multi-target tuning of OTA particulate structures. In this manuscript, we present the entire workflow necessary to carry out supervised AI applications on experimental data, with the aim to motivate similar work in the community. Data-efficient AI tools from computer science have the potential to renew experimental practices in chemical engineering and boost the search for advanced sustainable materials.

MATERIALS AND METHODS

Oxidized tannic acid (OTA) particle synthesis. Oxidized tannic acid particles were synthesized using the protocol reported previously [4]. Briefly, aqueous tannic acid solution (2% w/v) was prepared by adding tannic acid powder (1701.20 g/mol, Sigma-Aldrich) into Milli-Q water and rigorously stirring (magnetic bar) until completely dissolved. The pH of the solution was adjusted to a desired pH value with either 1 M KOH, 45 % $(CH_3)_3N$, 1 M NaOH, 0.5 M Na_3PO_4 or 25 % NH_4OH (see **Figure 2a**). All chemicals were reagent-grade and purchased from Sigma-Aldrich. Solutions were covered with perforated Parafilm and were shaken continuously with an orbital shaker for 14 h. All reactions were carried out at room temperature. The grown and precipitated OTA particles were collected and stored at room temperature for further characterization. Despite the simplicity in particle synthesis, multiple experiments were needed to accurately define the conditions that resulted in the given morphology. This required arduous experimentation as well as time since each setup gave specific morphology, depending on the reaction conditions.

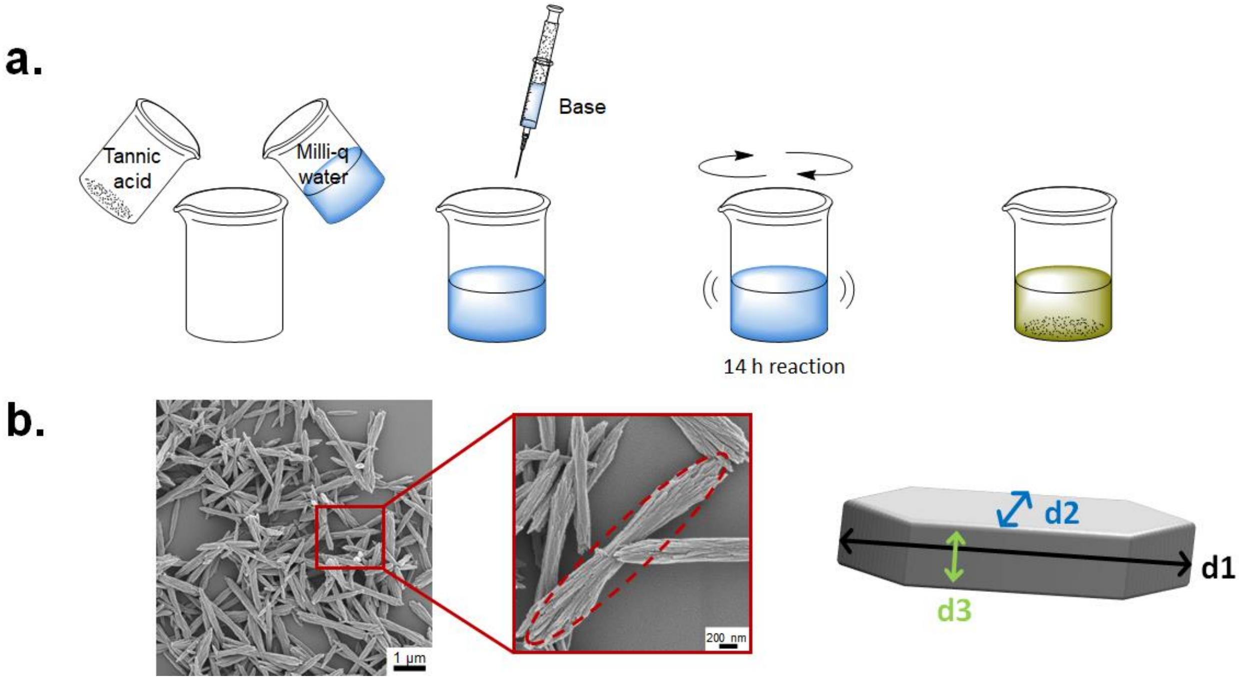

Figure 2. Schematic illustration of the experimental protocol used for data acquisition: a. OTA colloidal particle synthesis; b. SEM image analysis and particle dimension according to characteristic lengths d_1, d_2 and d_3.

Scanning electron microscopy image analysis. The synthesized OTA particles were imaged using a field-emission SEM (Sigma VP, Zeiss, Germany) with Schottky emitter at 1.5 kV without stage bias. For this purpose, aqueous suspensions of the OTA particles were cast onto pre-cleaned silicon wafers, dried in ambient laboratory conditions and sputter-coated with 4 nm Pd/Au. All imaging was performed on the same day with the OTA suspensions freshly prepared. Collected SEM images were then analyzed using ImageJ software [36] to measure the dimensions of the particles (**Figure 2b**). We measured the length, width and height of OTA particles as d_1, d_2, and d_3, such that $d_1 > d_2 > d_3$. Measurements were made for at least 10 different particles visible in the SEM image. The average values are reported here as the best estimate of particle dimensions. Standard deviations were recorded to estimate the experimental uncertainty on particle dimensions. All data points, error analysis and the SEM images are presented in the Supplementary Material (SM) document.

Gaussian Process Regression (GPR) algorithm. GPR is a kernel-based algorithm for supervised regression that relies on Gaussian Process (GP) models to represent black box functions.[23] Given data and the GP prior, Bayes' rule is applied to compute the GP posterior. The GP posterior mean serves as the surrogate model, the statistically most likely form of the unknown function. The GP posterior variance supplies a local measure of confidence in the model, typically rising in regions of search space where data is scarce and decreasing in well-explored regions.

For GPR fitting we used an uninformative zero mean GP prior and the radial basis function (RBF) kernel to obtain smooth and continuous landscapes. Data noise was Gaussian-distributed with zero mean. To make the model more robust, we applied inverse gamma priors on the hyperparameters of the kernel, the length scale and variance. During regression, the two hyperparameters were fitted in an automated way by maximizing marginal likelihood: this standard GPR procedure ensures that the results do not depend on manual hyperparameter choices [23].

To compute the surrogate model, we carried out GPR implemented in the Bayesian Optimization Structure Search (BOSS) code. BOSS is an open-source Python code [37,38] for performing GPR and Bayesian optimization (BO) tasks to solve problems in materials science [39–42]. It can read pre-recorded datasets or acquiring data on-the-fly with acquisition functions. BOSS post-processing capabilities allowed us to construct surrogate model landscapes and analyze their features.

RESULTS AND DISCUSSION

We employed 10 experimental data points on crystallized OTA particles collected by Kämäräinen *et al.*[4] to initialize the GPR model. In a departure from earlier work, the prospect of supervised learning required us to carry out experimental data analytics and consider different experimental outcomes, as well as measurement uncertainties. Supervised learning calls for a clear outcome, or label, so samples with ill-defined morphologies were not included into the AI model. Another key part of data digitalization was the conversion of experimental observations into customized descriptors for OTA particle morphology.

We started by analyzing the OTA particle morphology landscapes obtained in the 2-dimensional search space of pH and pK_a for shape predictions. To test the predictive power of the model, we performed 7 more experiments in key regions of the design space. The additional data also served to refine the morphology model. We validated the morphology landscapes against all experimental data collected, including the samples which were not employed in building the model. Lastly, we demonstrated how additional property models for particle yield and volume were built from the same set of experiments and consider multi-objective materials design.

Experimental dataset

The experimental dataset was adapted for GPR supervised learning by presenting each point in [$\boldsymbol{x}$, y] pair format. Here $\boldsymbol{x}$ is the location in the design space of OTA particle processing conditions, and y is the label, the morphology design objective for which we construct the surrogate model. Depending on the number of design parameters, $\boldsymbol{x}$ can be N-dimensional. In this work, $\boldsymbol{x} = (x_1, x_2)$ with x_1 assigned the pH of the solution and x_2 the value of base strength pK_a. We limited the design space of the processing conditions (pH, pK_a) to the range of ([7.0, 12.2], [9.0, 15.5]), thereby avoiding extreme conditions, where experiments may not have been successful.

The morphology of particles was quantified from their measured dimensions (d_1, d_2, d_3). To facilitate comparison between data points, the particle dimension data was scaled by the magnitude of the leading dimension (normalizing the longest dimension to 1.0 for each data point). We defined the morphology label y as:

$$\begin{gathered} y = \frac{d_2}{d_1} + \frac{d_3}{d_1}; \\ d_1 = 1.0 \;\rightarrow\; y = d_2 + d_3 \end{gathered} \tag{1}$$

This label allows us to distinguish between 1D and 3D morphology conditions as follows:

$$y = \begin{cases} 0, & d_1 \gg d_2, d_3; \quad 1D \\ 1, & d_3 \ll d_1, d_2; \quad 2D \\ 2, & d_1 \cong d_2 \cong d_3; \;\; 3D \end{cases} \tag{2}$$

While the 1D-3D signal difference across the realistic particles may be considerably lower than the ideal [0, 2] range, the choice of a physically meaningful property as label y allowed us to formulate interpretable surrogate models and gain immediate insight from GPR applications.

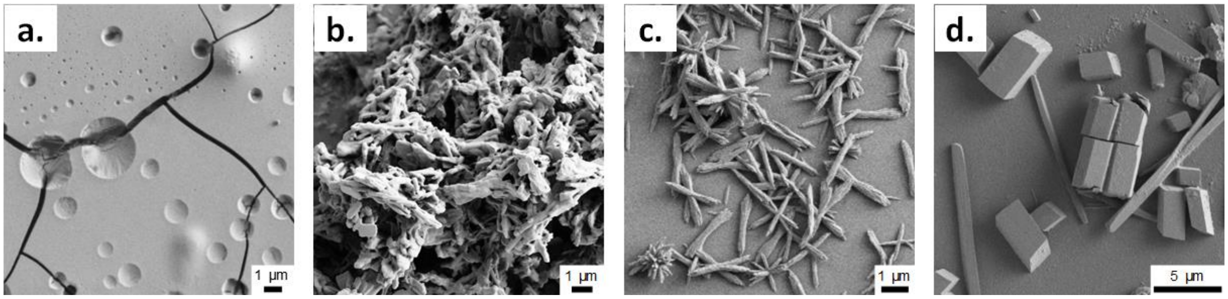

Figure 3. SEM images of precipitated OTA particles: a. no precipitate; b. ill-defined morphologies; c. regular morphology, suitable for parametrization; d. dual morphology.

Next, we review the range of experimental outcomes and discuss their suitability as input for AI application. Unlike in computational research where a numerical result is guaranteed, any experimental data point may result in one of the following outcomes of experimental synthesis, illustrated in **Figure 3**: a. no particle precipitate, b. non-quantifiable, ill-defined particle morphology, c. good quality precipitates with quantifiable dimensions, and d. multi-morphology precipitates. Too many experimental observations in the first two categories would suggest that the chosen design variables are not the key drivers of the chemical synthesis, and that the experimental design space needs further consideration.

In our work, 74% of experiments (17 points) resulted in quantifiable sample morphology. A further 22% (5) data points featuring ill-defined particle morphology could not be employed in building the model, but served to verify the model predictions. In one case, we observed OTA samples that featured two distinct particle morphologies in comparable yields (**Figure 3d**). Such a case indicates a saddle-point in the chemical design space, a two-phase region where both morphologies are in coexistence, and should be approached with caution. Here, we characterized the two morphologies and computed their arithmetic average label y: such treatment reflected the dichotomy in the design space and was supplying this information in the model.

Experimental uncertainties are common in any practical work, and must be carefully considered. In our efforts, there were uncertainties associated with both OTA sample synthesis and characterization. While we made every effort to fix all aspects of OTA particle synthesis apart from pH and pK_a, unaccounted differences in ambient conditions such as relative humidity could influence the evaporation rate during the experiments, affecting particle yields and morphologies. Changes in impurity content could also affect the observed morphologies. OTA particle dimensions were measured based on visual assignment of particle boundaries: these may introduce minor uncertainties into the mapping from design space to experimental outcome that are difficult to quantify. Irregular particle sizes in our experiments allowed us to perform a statistical analysis of particle dimensions (and thus morphologies). The standard deviations per particle dimension were combined to compute the overall uncertainty Δ on the morphology label y. Since this quantity reflects the knowability of data, it was adopted to represent all sources of experimental error and served as data noise in the GPR surrogate model (see SM for full details). For the precipitate yield, a conservative estimate of 5 % variation was assumed.

Morphology landscapes in the design space

Based on GPR, we computed the initial surrogate model for OTA particle morphology in the 2-dimensional pH-pK_a design space shown in **Figure 4a**. The continuous morphology landscape features areas of interest associated with low y signal (1D) and high y signal (3D) structures. It also indicates that there are regions of design space where no data has been collected and where the model may be less reliable.

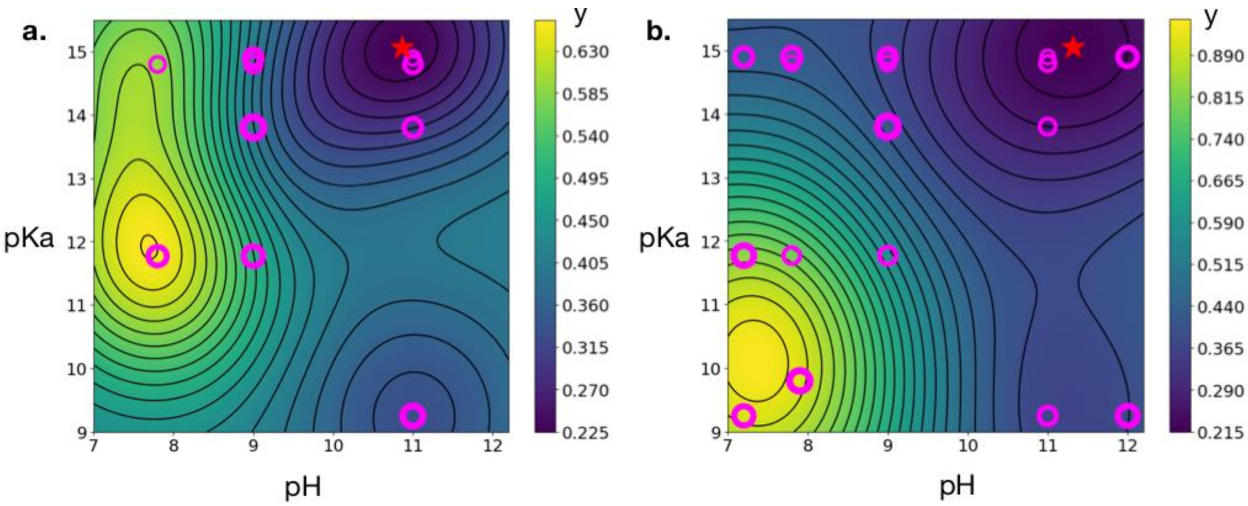

Figure 4. GPR surrogate models for morphology label y in pH-pK_a design space fitted with a. 10 and b. 17 experimental data points. Chart color reflects the value of the morphology label y, with yellow color denoting 3D and dark blue reflecting 1D particle outcome. Magenta circles indicate the loci of the actual experimental data. The red star indicates the processing conditions that produced OTA particles with the most pronounced 1D character (minimum y value in the surrogate model).

The minimum of the surrogate model in **Figure 4a**. suggests that high-pH combined with high-pK_a produced OTA particles with the most strongly pronounced 1D character ($d_1 \gg d_2, d_3$). Conversely, low pH solutions most likely produced 3D particles. To verify these predictions, we sampled further data points at the edges of the design space at pH<7.8 and pH>11, and also at low pK_a values, where data had been sparse. The GPR model that was re-trained with 7 additional experimental points is presented in **Figure 4b**.

The refined surrogate model for OTA particle morphology retains many of the features of the previous GPR fit in **Figure 3a**. The predicted high-pH and high-pK_a conditions for 1D particles remain unchanged. However, the region specific to 3D structures (high y values) is now enhanced, shifting to lower pK_a values. The refined landscape suggests that only low-pH and low-pK_a

processing conditions give rise to 3-dimensional particles. The relatively low value of the morphology signal *y* throughout the design space indicates that many experimental outcomes are 1D-like. Particles that are 2D-like may form only in the region of chemical space that neighbors the 3D structural conditions.

Model validation and predictive power

To extract predictions from the surrogate model, we coarse-grained the landscape into several categories assuming linear progression from 1D to 3D. As illustrated in **Figure 5**, this allows us to define regions of design space where experiments would reliably produce 1D, 2D and 3D OTA particles. We observe that 1D and 3D regions of design space are clear and well separated. The model predicts that solution pH and pK_a are directly correlated: 1D particles are obtained when their values are both high, and 3D when they are both low. In contrast, the 2D particle region spans a limited non-convex area in design space that conforms to the 3D particle region. This implies that 2D particles are difficult to synthesize. The greatest portion of design space was associated with 1D-type structures. The resulting model prediction is that when pH and pK_a are inversely correlated, 1D-like or 1D-2D mixed morphology particles are expected to occur.

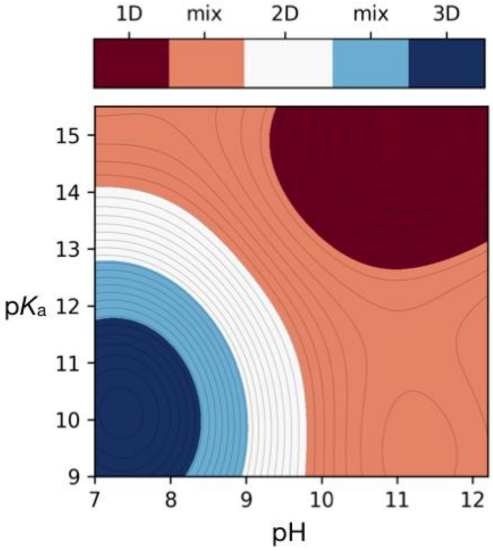

Figure 5. OTA particle morphology prediction by particle dimensionality, indicating mixed 1D-2D/2D-3D regions

In the next step, we validate our model predictions by cross-referencing SEM images of OTA particles with the particle morphology landscape. **Figure 6** portrays the landscape overlaid with SEM image data from the area of design space where the OTA particle synthesis was carried out. Images outlined in red represent cases of non-quantifiable particle dimensions (ill-defined morphology), which were not included in the model construction. The case of dual particle morphologies is indicated in green.

It is immediately clear that the predictions regarding 1D and 3D particle formation were correct. 1D landscape regions are associated with very long needle-like particles (up to 0.1 mm), where the design condition $d_1 \gg d_2, d_3$ is best satisfied. 1D-like regions exhibit a different 1D morphology where the particles are short and matchstick-like. In some cases, the short 1D particles agglomerate into a larger mass where the morphology is not easily identified. These data were not included into the surrogate model, and yet they correlate well with the mixed morphology 1-2D and 2-3D regions of the landscape. The same is true of the dual morphology data points, which correctly occur in the mixed 1D-2D section of the landscape.

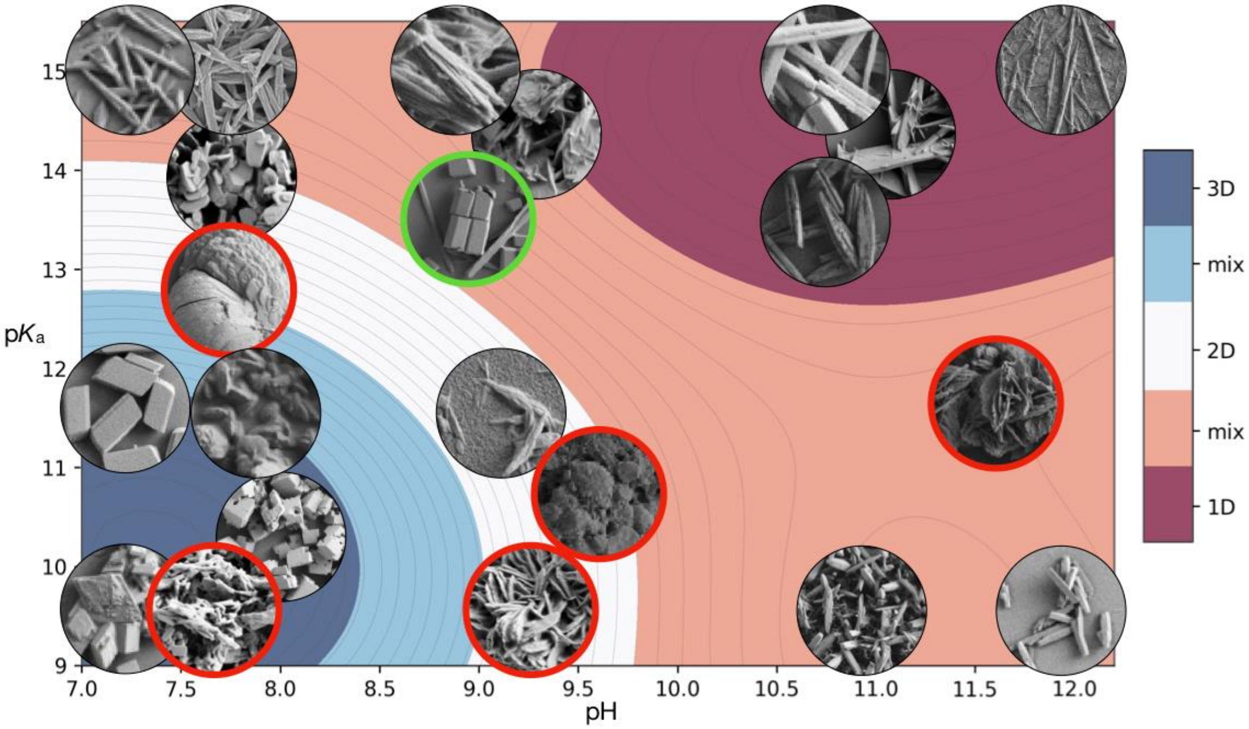

Figure 6. Surrogate model of OTA morphology validated against experimental SEM images. Images with red borders indicate experiments with ill-defined morphology while those in green indicate mixed morphologies.

SEM images reveal few examples of 3D particles obtained in these experiments, about 25% of the total. Even fewer are the 2D particle cases, which present mostly as domino-like platelet structures. As predicted by the surrogate model, 1D particles dominate the design space: short matchstick-like structure are the most common experimental outcome. At intermediate pH and pK_a values, there is a risk of particle aggregation: matchsticks combining into disordered bundles and coral-like growth is observed.

OTA particle yield and functional properties

Having demonstrated that GPR surrogate models for OTA particle morphology have good predictive power, we turn our attention to other experimental information. With each synthesis data point, we recorded the yield of the dried OTA colloidal content. The measurement of particle dimensions further allowed us to analyze and engineer other functional properties such as particle size, volume or surface area. The leading particle length in experiments varied in the range 0.4–130 µm, suggesting that experimental conditions can be used to tailor the particle size to diverse applications. We focus on the ratio of particle surface area to its volume: surface-based chemical processes underpin many technological applications, so maximizing surface area per volume (A/V) complements particle morphology control as an important design objective.

The GPR surrogate model for OTA particle yield is presented in **Figure 7a**. The irregular features in this landscape suggest that particle yield is strongly correlated with the base employed in the solution, rather than the pK_a value. For example, applying LiOH (pK_a 13.8) to OTA leads to relatively high yields, about 60%, but NaOH (pK_a 14.8) causes the yield to drop below 10%. This observation suggests that particle yield may be better correlated with a different property of the base, such as its size. Solution pH does play a role in the particle yield, with largest yields observed in the pH range 8–11.

The OTA particle A/V landscape, illustrated in **Figure 7b**, presents a central region where the A/V ratio is very high. These mid-range pH and pK_a conditions are associated with 2D particles, where experimental data is scarce. OTA particles synthesized in these conditions tend to produce 2D-like lamellar forms that agglomerate into 3D structures (see Figure 6 for SEM images). It was difficult to measure the shape of these particles, so they were not included into the surrogate model. Nevertheless, such samples clearly had the highest A/V ratio, and this was correctly predicted by the A/V surrogate model despite the paucity of data.

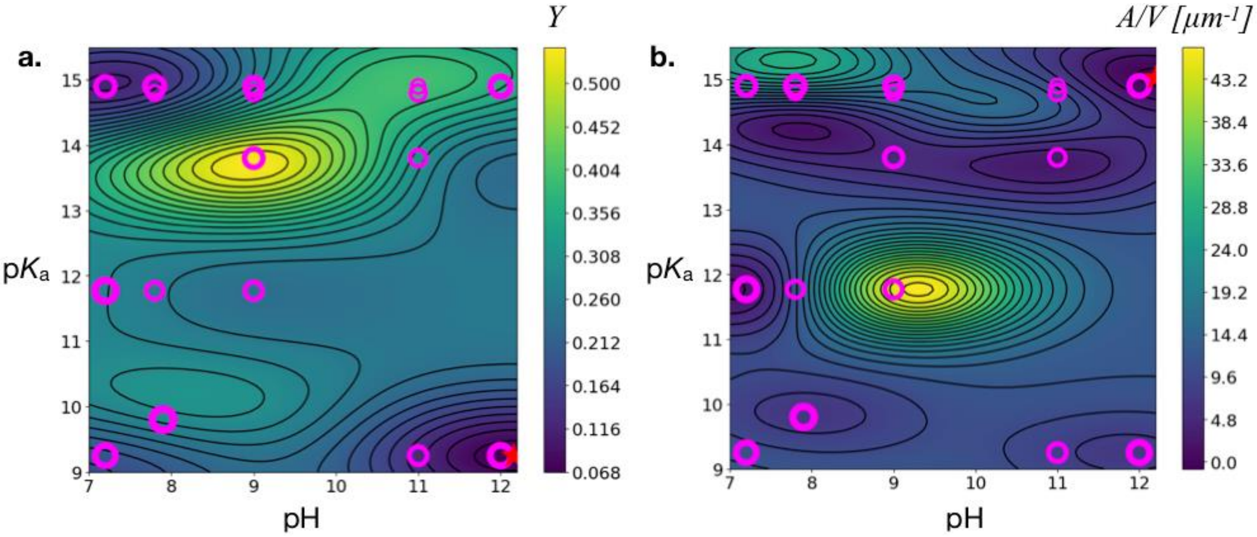

Figure 7. Surrogate models for OTA particle experimental properties, a. particle yield and b. particle A/V ratio, in the design space of pH and pK_a processing conditions. Magenta circles indicate the locations of the experimental data points.

Extracting several surrogate models from the same experimental data (at no additional cost) allows us to cross-reference different properties and infer the conditions that would satisfy several design objectives at once. For example, a high yield of 3D particles can be obtained with NH_4OH in low pH=7 conditions. Highest yield of 1D OTA particles can be achieved with KOH at pH=10–11, which also produces largest particles with most surface area exposed. 1D particles with high

A/V ratio could be produced at very high pK_a, but at relatively low yields. In further work, different label variables can be arithmetically combined into composite labels and landscapes.

Discussion and Outlook

The purpose of this work was to evaluate the predictive power of GPR on a small experimental dataset; therefore, we deliberately constrained the dimensionality of the problem, which also produced interpretable surrogate models. OTA particle morphology is certainly affected by other experimental parameters. Nevertheless, the good predictive power of surrogate models in the relatively simple 2D design space demonstrated that pH and pKa alone are sufficient to control particle morphology, in agreement with the earlier PCA result. Unfortunately, PCA was unable to provide insights into the morphology variation that could be achieved with surrogate models.

The morphology landscape portrays a very clear synthesis trend, but we were unable to interpret it using scientific intuition. The bottom-up OTA particle synthesis is a result of complex self-assembly where OTA particles coordinate into secondary supramolecular structures, which form tertiary nanofilaments and these assemble into quaternary mesoscopic crystals [43]. It is very difficult to develop any inkling about the outcome of such an intricate procedure, nor about how processing conditions might affect it. Instead, the data-driven landscape can guide further research into the chemical processes behind such outcomes and advance fundamental understanding.

Surrogate models are of general value in materials design because they span all design space, are chemically intuitive and interpretable. It is difficult to establish the criteria for quantitative accuracy of surrogate models. Our work shows that qualitative accuracy already translates to good predictive power, marked by the good visual agreement between the morphology landscapes and the SEM images. OTA samples with ill-defined morphology (not included in the GPR) were particularly important in validating the model predictions. The correspondence of these mixed morphology samples with the appropriate regions on the map demonstrates that good quality AI predictions can be achieved in areas where no experiments were previously performed, or included in the model.

The sensitivity of OTA particles to their processing conditions made them an ideal test case for this study, but they remain a challenging material to work with. The composition as well as the molecular structure of tannins are dependent on the source they were extracted from [44,45]. In other words, the plant species and their physiological state dictate the polydispersity and molecular weight, giving rise to inevitable heterogeneity, which complicates the processing and characterization of the materials. The relatively high experimental uncertainties translated into data noise that amounted to 10% of the entire GPR model corrugation. Such noise did not impair the predictive power of the models in this study, but in other work experimental errors could lead to distorted models and less optimal fits.

The convergence of GP models is an important concern in experimental work where dataset sizes are small. Typically, an iterative convergence procedure is followed. Here, the addition of further 7 data points intended to verify model predictions had a small effect on the qualitative features of the model, so we stopped short of additional experiments. The need for further data can be also evaluated from the values of the GPR posterior variance, which tends to decrease with more data included in the model. We considered the OTA morphology model variance after 10 and 17 experimental points (see **Figure S3**). In this work, the relatively large experimental uncertainties translated into large values of GP variance, which remained unchanged with the addition of more data. This finding indicates that in GPR applications to experimental data, where large noise maintains high variance, GP posterior variance might not be a useful measure of model confidence. However, the variance could be used to guide additional experiments.

In further work, our GPR-based approach could be extended to active learning material design workflows. In Bayesian optimization (BO) [32,33], GPR variance is exploited by acquisition functions to select the sampling location that would most enhance the dataset. Acquisition functions balance data exploration (searching less-visited areas of phase space) with data exploitation (searching near optimum points in phase space) to attain search objectives with relatively few data points. Search objectives can be learning the entire landscape or minimizing and maximizing materials properties across the search space.

By demonstrating that GPR performs well with experimental data related to OTA morphology design, this study opens the route towards BO with experimental data in engineering colloids. Integrating BO into experimental work is challenging [46–48], but there are many benefits [49,50]. With acquisition functions guiding the selection of experiments, good predictive power of machine learning could be achieved with fewer experimental data points, facilitating the study of complex N-dimensional design spaces with more design variables. Moreover, BO allows to drive experimental data collection towards materials with preferred functional properties (morphological, mechanical or chemical) within the search space. The AI-guided search can thus replace trial-and-error experimental approach in materials design.

CONCLUSIONS

Supramolecular OTA constructs present a prospect of novel applications for this versatile and bioactive material. Controlling particle morphology will help us purpose the OTA particulates towards certain functions and application areas. This study combined chemical engineering with GPR supervised machine learning to correlate the processing conditions of OTA colloidal solution with the morphology of the resulting dry OTA particles. The Bayesian surrogate model landscapes revealed the variation of particle morphology in the design space, illustrating the synthesis conditions needed to achieve different particle shapes. The main finding from the OTA morphology landscape is that severe processing conditions (high pH and pK_a) give rise to extended 1D particles with high surface area per volume ratios. Reducing the severity of the solution produces smaller, compact 3D shapes.

Despite the relatively small dataset size and large experimental uncertainty, the data-driven morphology landscape was in good agreement with OTA sample images. It exhibited considerable predictive power on samples that were not originally included in the model, marking the potential for predictive materials design. From the same set of experiments, we built surrogate models for OTA particle shape, yield, and surface-to-volume ratio, and cross-referenced them to demonstrate how multiple design objectives could be satisfied at once.

Mapping processing conditions directly to experimental properties of materials constitutes a practical approach to AI-led chemical engineering, free of human bias. Such procedures could supplant experimental trial-and-error approaches, but also guide further research into the mechanisms of crystallization and self-assembly in complex materials, opening innovative engineering routes towards new phases of matter.

ASSOCIATED CONTENT

SUPPORTING INFORMATION

The following file is available free of charge. Supporting information contains the summary of experimental data points and data analytics, SEM micrographs of all OTA particle samples produced and GPR posterior variance landscapes corresponding to the OTA particle morphology posterior mean.

AUTHOR INFORMATION

Author Contributions

S.J. and T.K. performed all experimental work. M.T. performed all computational work and wrote the manuscript. M.T., O.R. and P.R. conceived the study. All authors participated in refining the manuscript.

Funding Sources

This project received partial funding from the Academy of Finland via the Artificial Intelligence for Microscopic Structure Search (AIMSS) project No. 3316601 and the European Union's Horizon 2020 program under the ERC Advanced Grant Agreement No. 788489, “BioElCell”.

ORCID

Milica Todorović: 0000-0003-0028-0105

Tero Kämäräinen: 0000-0001-8333-4900

Patrick Rinke: 0000-0003-1898-723X

Orlando J. Rojas: 0000-0003-4036-4020

ACKNOWLEDGMENT

This project received partial funding from the Academy of Finland via the Artificial Intelligence for Microscopic Structure Search (AIMSS) project No. 13316601 and the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 788489). We are grateful for the support by the Academic Flagship programme under the FinnCERES Materials Bioeconomy and the FCAI Center for Artificial Intelligence as well as the Canada Excellence Research Chair initiative (OJR). The facilities and technical support provided by Aalto University at OtaNano – Nanomicroscopy Center (Aalto-NMC) are also acknowledged.

REFERENCES

[1] Z. Hu, H.S. Marway, H. Kasem, R. Pelton, E.D. Cranston, Dried and Redispersible Cellulose Nanocrystal Pickering Emulsions, ACS Macro Lett. 5 (2016) 185–189. https://doi.org/10.1021/acsmacrolett.5b00919.

[2] A.E. Hagerman, K.M. Riedl, G.A. Jones, K.N. Sovik, N.T. Ritchard, P.W. Hartzfeld, T.L. Riechel, High Molecular Weight Plant Polyphenolics (Tannins) as Biological Antioxidants, J. Agric. Food Chem. 46 (1998) 1887–1892. https://doi.org/10.1021/jf970975b.

[3] S. Gharehkhani, N. Ghavidel, P. Fatehi, Kraft Lignin-Tannic Acid as a Green Stabilizer for Oil/Water Emulsion, ACS Sustain. Chem. Eng. 7 (2019) 2370–2379. https://doi.org/10.1021/acssuschemeng.8b05193.

[4] T. Kämäräinen, M. Ago, L.G. Greca, B.L. Tardy, M. Müllner, L.S. Johansson, O.J. Rojas, Morphology-Controlled Synthesis of Colloidal Polyphenol Particles from Aqueous Solutions of Tannic Acid, ACS Sustain. Chem. Eng. 7 (2019) 16985–16990. https://doi.org/10.1021/acssuschemeng.9b02378.

[5] T.S. Sileika, D.G. Barrett, R. Zhang, K.H.A. Lau, P.B. Messersmith, Colorless multifunctional coatings inspired by polyphenols found in tea, chocolate, and wine, Angew. Chemie - Int. Ed. 52 (2013) 10766–10770. https://doi.org/10.1002/anie.201304922.

[6] Z. Hu, R.M. Berry, R. Pelton, E.D. Cranston, One-Pot Water-Based Hydrophobic Surface Modification of Cellulose Nanocrystals Using Plant Polyphenols, ACS Sustain. Chem. Eng. 5 (2017) 5018–5026. https://doi.org/10.1021/acssuschemeng.7b00415.

[7] V. Tulyathan, R.B. Boulton, V.L. Singleton, Oxygen Uptake by Gallic Acid as a Model for Similar Reactions in Wines, J. Agric. Food Chem. 37 (1989) 844–849. https://doi.org/10.1021/jf00088a002.

[8] A. Dutta, S.K. Dolui, Tannic acid assisted one step synthesis route for stable colloidal dispersion of nickel nanostructures, Appl. Surf. Sci. 257 (2011) 6889–6896. https://doi.org/10.1016/j.apsusc.2011.03.025.

[9] J. Scoccia, M.D. Perretti, D.M. Monzón, F.P. Crisóstomo, V.S. Martín, R. Carrillo, Sustainable oxidations with air mediated by gallic acid: Potential applicability in the reutilization of grape pomace, Green Chem. 18 (2016) 2647–2650. https://doi.org/10.1039/c5gc02966j.

[10] S.K. Bhangu, R. Singla, E. Colombo, M. Ashokkumar, F. Cavalieri, Sono-transformation of tannic acid into biofunctional ellagic acid micro/nanocrystals with distinct morphologies, Green Chem. 20 (2018) 816–821. https://doi.org/10.1039/c7gc03163g.

[11] K.T. Chung, T.Y. Wong, C.I. Wei, Y.W. Huang, Y. Lin, Tannins and human health: A review, Crit. Rev. Food Sci. Nutr. 38 (1998) 421–464. https://doi.org/10.1080/10408699891274273.

[12] B. Badhani, N. Sharma, R. Kakkar, Gallic acid: A versatile antioxidant with promising therapeutic and industrial applications, RSC Adv. 5 (2015) 27540–27557. https://doi.org/10.1039/c5ra01911g.

[13] S.-C. Wang, Y. Chen, Y.-C. Wang, W.-J. Wang, C.-S. Yang, C.-L. Tsai, M.-H. Hou, H.-F. Chen, Y.-C. Shen, M.-C. Hung, Tannic acid suppresses SARS-CoV-2 as a dual inhibitor of the viral main protease and the cellular TMPRSS2 protease., Am. J. Cancer Res. 10 (2020) 4538–4546.

[14] H. Ejima, J.J. Richardson, F. Caruso, Metal-phenolic networks as a versatile platform to engineer nanomaterials and biointerfaces, Nano Today. 12 (2017) 136–148. https://doi.org/10.1016/j.nantod.2016.12.012.

[15] V.N. Manoharan, Colloidal matter: Packing, geometry, and entropy, Science (80-.). 349 (2015). https://doi.org/10.1126/science.1253751.

[16] J. Lin, H. Chen, W. Xu, Geometrical percolation threshold of congruent cuboidlike particles in overlapping particle systems, Phys. Rev. E. 98 (2018). https://doi.org/10.1103/PhysRevE.98.012134.

[17] T. Moberg, K. Sahlin, K. Yao, S. Geng, G. Westman, Q. Zhou, K. Oksman, M. Rigdahl, Rheological properties of nanocellulose suspensions: effects of fibril/particle dimensions and surface characteristics, Cellulose. 24 (2017) 2499–2510. https://doi.org/10.1007/s10570-017-1283-0.

[18] A. Albanese, P.S. Tang, W.C.W. Chan, The effect of nanoparticle size, shape, and surface chemistry on biological systems, Annu. Rev. Biomed. Eng. 14 (2012) 1–16. https://doi.org/10.1146/annurev-bioeng-071811-150124.

[19] Y. Xu, M. Cao, Q. Zhang, Recent advances and perspective on heterogeneous catalysis using metals and oxide nanocrystals, Mater. Chem. Front. 5 (2021) 151–222. https://doi.org/10.1039/d0qm00549e.

[20] Q. Zhang, G. Cao, Hierarchically structured photoelectrodes for dye-sensitized solar cells, J. Mater. Chem. 21 (2011) 6769–6774. https://doi.org/10.1039/c0jm04345a.

[21] M. Chen, Y. Zhang, L. Xing, Y. Liao, Y. Qiu, S. Yang, W. Li, Morphology-Conserved Transformations of Metal-Based Precursors to Hierarchically Porous Micro-/Nanostructures for Electrochemical Energy Conversion and Storage, Adv. Mater. 29 (2017) 1607015. https://doi.org/10.1002/adma.201607015.

[22] J.A. Champion, Y.K. Katare, S. Mitragotri, Particle shape: A new design parameter for micro- and nanoscale drug delivery carriers, J. Control. Release. 121 (2007) 3–9. https://doi.org/10.1016/j.jconrel.2007.03.022.

[23] M. Seeger, Gaussian processes for machine learning., Int. J. Neural Syst. 14 (2004) 69–106. https://doi.org/10.1142/S0129065704001899.

[24] R. Batra, L. Song, R. Ramprasad, Emerging materials intelligence ecosystems propelled by machine learning, Nat. Rev. Mater. 44 (2020) 1–24.

[25] X. Wang, N. Rai, B. Merchel Piovesan Pereira, A. Eetemadi, I. Tagkopoulos, Accelerated knowledge discovery from omics data by optimal experimental design , Nat. Commun. 11 (2020) 611.

[26] R. Yuan, Y. Tian, D. Xue, D. Xue, Y. Zhou, X. Ding, J. Sun, T. Lookman, Accelerated Search for BaTiO 3-Based Ceramics with Large Energy Storage at Low Fields Using Machine Learning and Experimental Design, Adv. Sci. 6 (2019) 1901395.

[27] Z. Ren, S. Tian, T. Heumueller, E. Birgersson, F. Lin, A. Aberle, S. Sun, I.M. Peters, R. Stangl, C.J. Brabec, T. Buonassisi, F. Oviedo, H. Xue, M. Thway, K. Zhang, N. Li, J.D. Perea, M. Layurova, Y. Wang, Physics-guided characterization and optimization of solar cells using surrogate machine learning model, in: 2019 IEEE 46th Photovolt. Spec. Conf., IEEE, 2019: pp. 3054–3058.

[28] P. V Balachandran, B. Kowalski, A. Sehirlioglu, T. Lookman, Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning, Nat. Commun. 9 (2018) 1–9.

[29] F. Häse, L.M. Roch, C. Kreisbeck, A. Aspuru-Guzik, PHOENICS: A universal deep Bayesian optimizer, ACS Cent. Sci. 4 (2018) 1134–1145.

[30] L. Himanen, A. Geurts, A.S. Foster, P. Rinke, Data-Driven Materials Science: Status, Challenges, and Perspectives, Adv. Sci. 6 (2019) 1900808. https://doi.org/10.1002/advs.201900808.

[31] P.I. Frazier, J. Wang, Bayesian optimization for materials design, in: T. Lookman, F.J. Alexander, K. Rajan (Eds.), Inf. Sci. Mater. Discov. Des., Springer International Publishing, Cham, 2015: pp. 45–75. https://doi.org/10.1007/978-3-319-23871-5_3.

[32] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian optimization of machine learning algorithms, Adv. Neural Inf. Process. Syst. 4 (2012) 2951–2959.

[33] E. Brochu, V.M. Cora, 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.LG]. (2010). http://arxiv.org/abs/1012.2599.

[34] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space , London, Edinburgh, Dublin Philos. Mag. J. Sci. 2 (1901) 559–572. https://doi.org/10.1080/14786440109462720.

[35] H. Hotelling, Relations Between Two Sets of Variates, Biometrika. 28 (1936) 321.

[36] C.A. Schneider, W.S. Rasband, K.W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis, Nat. Methods. 9 (2012) 671–675. https://doi.org/10.1038/nmeth.2089.

[37] Bayesian Optimization Structure Search (BOSS) code, (2020). https://cest-group.gitlab.io/boss/index.html. Accessed January 21, 2021.

[38] GPy by SheffieldML, (n.d.). http://sheffieldml.github.io/GPy/. Accessed January 21, 2021.

[39] L. Fang, E. Makkonen, M. Todorović, P. Rinke, X. Chen, Efficient Amino Acid Conformer Search with Bayesian Optimization, J. Chem. Theory Comput. (2021).

https://doi.org/10.1021/acs.jctc.0c00648.

[40] M. Todorović, M.U. Gutmann, J. Corander, P. Rinke, Bayesian inference of atomistic structure in functional materials, Npj Comput. Mater. 5 (2019) 35.

[41] J. Järvi, P. Rinke, M. Todorović, Detecting stable adsorbates of (1S)-camphor on Cu(111) with Bayesian optimization, Beilstein J. Nanotechnol. 11 (2020) 1577–1589.

[42] A.T. Egger, L. Hörmann, A. Jeindl, M. Scherbela, V. Obersteiner, M. Todorović, P. Rinke, O.T. Hofmann, Charge Transfer into Organic Thin Films: A Deeper Insight through Machine-Learning-Assisted Structure Search, Adv. Sci. 7 (2020) 2000992.

[43] I.M. and J.G. X. Qiu, X. Wang, Y. He, J. Liang, K. Liang, B. L. Tardy, J. J. Richardson, M. Hu, H. Wu, Y. Zhang, O. J. Rojas, Superstructured mesocrystals through multiple inherent molecular interactions, Science (80-.). (2021).

[44] L. Mouls, J.P. Mazauric, N. Sommerer, H. Fulcrand, G. Mazerolles, Comprehensive study of condensed tannins by ESI mass spectrometry: Average degree of polymerisation and polymer distribution determination from mass spectra, Anal. Bioanal. Chem. 400 (2011) 613–623. https://doi.org/10.1007/s00216-011-4751-7.

[45] L. Mouls, V. Hugouvieux, J.P. Mazauric, N. Sommerer, G. Mazerolles, H. Fulcrand, How to gain insight into the polydispersity of tannins: A combined MS and LC study, Food Chem. 165 (2014) 348–353. https://doi.org/10.1016/j.foodchem.2014.05.121.

[46] A.E. Gongora, B. Xu, W. Perry, C. Okoye, P. Riley, K.G. Reyes, E.F. Morgan, K.A. Brown, A Bayesian experimental autonomous researcher for mechanical design, Sci. Adv. 6 (2020).

[47] L.M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L.P.E. Yunker, J.E. Hein, A. Aspuru-Guzik, ChemOS: Orchestrating autonomous experimentation, Sci. Robot. 3 (2018). https://doi.org/10.1126/scirobotics.aat5559.

[48] R. Kurchin, G. Romano, T. Buonassisi, Bayesim: A tool for adaptive grid model fitting with Bayesian inference, Comput. Phys. Commun. 239 (2019) 161–165.

[49] M.M. Flores-Leonar, L.M. Mejía-Mendoza, A. Aguilar-Granda, B. Sanchez-Lengeling, H. Tribukait, C. Amador-Bedolla, A. Aspuru-Guzik, Materials Acceleration Platforms: On the way to autonomous experimentation, Curr. Opin. Green Sustain. Chem. 25 (2020) 100370. https://doi.org/10.1016/j.cogsc.2020.100370.

[50] R. Shimizu, S. Kobayashi, Y. Watanabe, Y. Ando, T. Hitosugi, Autonomous materials synthesis by machine learning and robotics, APL Mater. 8 (2020) 111110.