ISSN: 2692-5397

Modern Concepts in Material Science

DOI: 10.33552/MCMS.2021.03.000570

Iris Publishers

Mini Review

Automated Modelling of Material Chemistry Applying High-Throughput Formulation and Machine Learning

Yannik Schuchmann[1], Christian Schmitz[1], Lasse Wagner[2], Carmen Stoffelen-Janßen[2] and Jost Göttert[2]*

[1]Institute for Coatings and Surface Chemistry, University of Applied Sciences Niederrhein, Germany

[2]Institute for Surface Technology, University of Applied Sciences Niederrhein, Germany

*Corresponding author: Jost Göttert, Institute for Surface Technology, University of Applied Sciences Niederrhein, Adlerstraße 32, 47798 Krefeld, Germany.

Received Date: January 05, 2021

Published Date: January 19, 2021

Introduction

Chemical material development belongs to the fields chemical synthesis of raw materials and formulation of products like plastics, adhesives, coatings, composites, etc. The improvement of material behavior is based on wide range of chemical structures fulfilling the conditions of the application.

Thus, various combinations of reactants or mixtures of chemical substances are possible.

High-throughput technology has been evolved from simple pipetting tasks for example in pharmaceutical research [1] to more complex workflows for material sample preparation and characterization [2]. This accelerated the research and development of new materials by automated repetitive cycles and allowed a deep analysis of the multi-dimensional experimental space based on variation of the raw material synthesis and formulation.

On the other hand, the large sample numbers challenge the human operator with the analysis by extracting complex relationships between all ingredients. Design of experiment (DoE) for planning samples within the experimental space and statistical techniques like analysis of variance (ANOVA) and polynomial regression are commonly used in such cases [3-4]. Nowadays, other modelling methods from the discipline computer science for simulation are rarely applied in material formulation.

Machine learning (ML) includes concepts known for their widespread use in technical applications such as picture recognition or social media but gained more and more attention in material science during the last decade [5]. These techniques allow a more sophisticated modelling of the experimental space reducing the number of samples in comparison to the commonly applied sampling strategies known in DoE with exponentially rising number of samples required. Furthermore, sampling technique using ML plans the location for additional experiments by an algorithm based on the data situation.

In this review we describe a combination of high-throughput equipment and ML algorithms being setup at the University of Applied Sciences Niederrhein [https://ihit.online/de/]. It is an autarchic unit autonomously executing sample preparation and characterization of an user-defined task for coatings material development until a certain optimization or predictability of the target properties is succeeded.

High-Throughput Formulation

The preparation of coating samples (Figure 1) involves the automated dispensing of liquids such as resins or solvents and pigment or filler powders depending on the specification of each sample. The homogenization of the material is realized by dual-axis centrifugal mixing that also allows milling of pigmented systems with beads. The liquid coating is applied on the substrate by spray application or draw down and hardened at room or elevated temperature. Optical properties like color, gloss, surface profile are measured with several sensors [6-8] for surface detection. The mechanical resistance is recognized by automated cutting tests and image detection.

The overall system layout allows individual sampling of coatings demanded by the input of the material recipe. In the end the output of the properties for each sample is automatically provided in a database.

ML Analysis and Sampling

The input can be related to the output properties by modelling the given data. The model itself can be used finding optimal conditions for the application of the material or simulating the material behavior depending on the chemical compounds being incorporated. Well-known techniques fit polynomial equations for the modelling including a certain number of model parameters according to the model being chosen [3]. In general, the user always defines the polynomial degree and to what extend interferences between input parameters need to be considered.

An autonomous analysis for modelling data requires an algorithm finding the best fitting solution. Rather than applying a parametric model by polynomial regression, a non-parametric model by Gaussian process regression can be quite helpful. The target function of the output (eq. 1) is defined by the Gaussian process (GP) with the mean of a normal distribution of functions (m(x)) and the covariance function (k(x,x')) [9].

$$f(x)=\mathcal{GP}\left(m(x),k(x,x')\right) \qquad (1)$$

Figure 2 visualizes the model function for an output according to one input factor based on the experimental data as the observations. The model is able to predict the output and its confidence interval within the bounds given by the input parameters.

Figure 2 Model describing the observation by the mean (black line) of a distribution of functions and the 95% confidence interval (grey area). Five sample functions of the distribution are shown by the red lines.

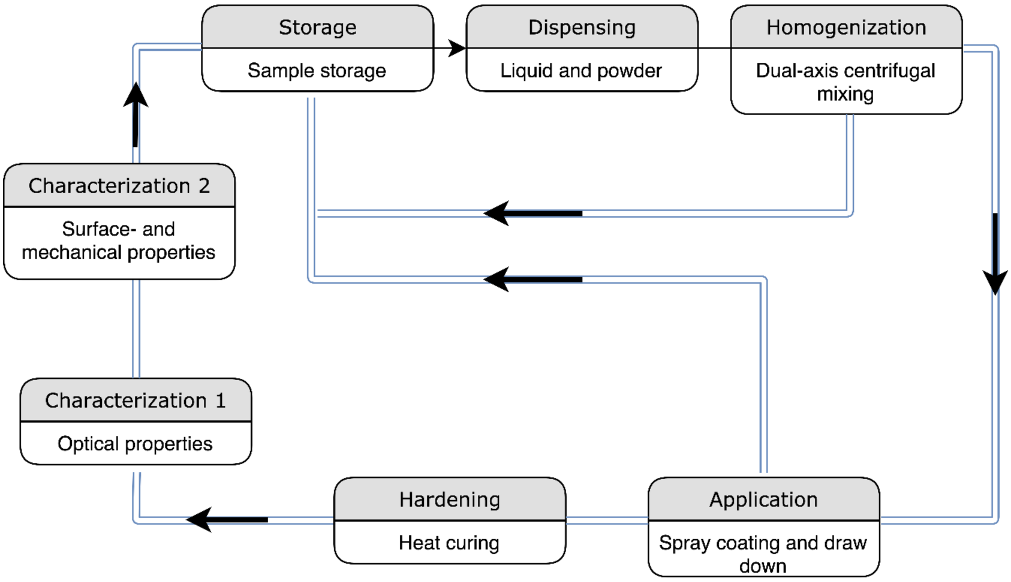

Figure 1: Layout of high-throughput equipment for automated preparation and characterization of hardened coating samples on substrates.

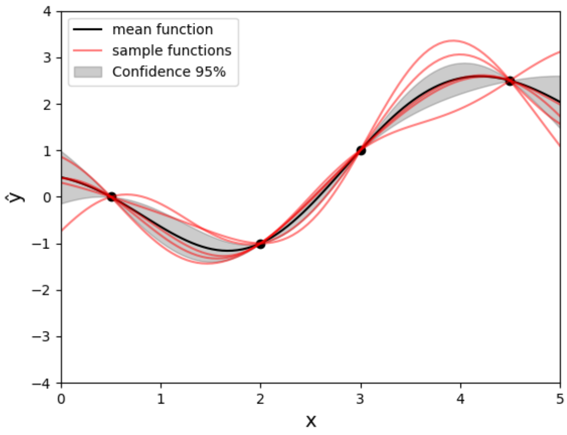

Figure 2: Model describing the observation by the mean (black line) of a distribution of functions and the 95% confidence interval (grey area). Five sample functions of the distribution are shown by the red lines.

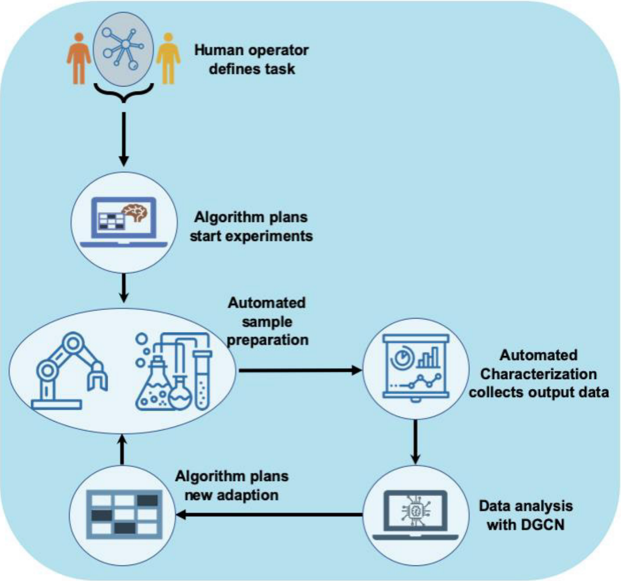
Human operator defines task
Algorithm plans start experiments
Automated sample preparation
Automated Characterization collects output data
Data analysis with DGCN
Algorithm plans new adaption

Figure 3: Combination high-throughput automation of ML algorithm for autarch sample preparation, data analysis and sampling of new experiments.

An algorithm called Deep Gaussian Covariance Network (DGCN) combines benefits from Gaussian processes and convolutional neural networks and does not require a setting of parameters by the user. The model is locally adjusted to the data set even in case of a small data size.

The DGCN model can be combined with Bayesian optimization that is applied for sampling the next experiments. Operating this sampling method, the location of new experiments collecting information on the analyzed case is planned in separated packages with small sample numbers in contrast to the common DoE that samples all experiments beforehand. The analysis of the systems begins with a number of start samples and the sampling is stepwise proceeded in so called adaptions until the task is finished.

The start samples with no information on the system are planned with Latin hypercube sampling [10,11] achieving a uniform distribution of data points. Latin hypercube sampling randomly chooses input values so that the distribution of data points over the whole experimental space in all dimensions is well-balanced. It avoids overlapping of data points in each dimension and prevents a loss of information if input parameters only show little effect on the output value.

The following adaptions sample new experiments conditioned by the information already given by the collected data. Based on the model and its local uncertainty new test samples within the experimental space will be identified by an acquisition function (AQF) following the highest improvement. The type of acquisition function depends on the task and allows a convergence with less experiments as possible. Expected improvement (eq. 2) [12,13] identifies the global optimum of the problem and eq. 3 finds the most accurate model considering the whole experimental space.

$$AQF(x_*) = (\hat{y}_* - y_{best}) \cdot \Phi(Z) + \hat{s}_* \cdot \phi(Z) \quad (2)$$

$$AQF(x_*) = \hat{s}_* \quad (3)$$

$Z = \frac{\hat{y}_* - y_{best}}{\hat{s}_*}$; $\hat{y}_*$ is the prediction and its standard deviation $\hat{s}_*$ provided by the DGCN model. y_{best} represents the current best output value found so far. Φ(Z) and ϕ(Z) are the cumulative density function and probability density function of a Gaussian distribution. Expected improvement in the literature is often described with an additional factor covering the trade-off between exploitation as the most promising candidate and exploration searching locations with higher uncertainty [14,15].

The optimization is carried out with the AQF instead of the surrogate model itself. The goal is to find the point x_* , that maximizes the AQF. From the eq. 2 two cases can be derived, that meet this requirement:

"AQF" (x_*) is high if $\hat{y}_*$>y_{best}

"AQF" (x_*) is high if $\hat{s}_*$ around x_* is high

In the case of eq. 3 sampling is executed in locations with the highest uncertainty until the overall predictability of the model is acceptable.

Intelligent Automation

The automation of sample preparation including testing and the ML algorithms for automated computer analysis and sampling based on the information collected so far are connected to an autonomous working unit. This unit operates a user-defined case based on a coating recipe with variation of its ingredients until the initial objective is succeeded. The user specifies the task differing between optimization and complete modelling, the bounds of the variations or other constraints and the workflow of the preparation procedure. The ML algorithm connects the elements from DGCN transferring the given data into a virtual model of the material and the AQF suggesting next samples planned in adaptions.

The input parameters itself are chosen by the algorithms processing input and output data collected by the high-throughput equipment. Additionally, the user defines the number of experiments for the start samples and the adaptions. The number of the start samples is usually higher since the first insight of the material behavior needs to be revealed and modelled before the adaptions improve in smaller steps (Figure 3).

The scheme in Figure 3 shows the workflow of the intelligent automation for material development.

On the top the user interferes with the unit defining the task and its conditions of the experimental setup. The algorithm uses this information planning the starting experiments within the boundaries, which are executed on the high-throughput equipment. The data from the starting experiments result in an adaption cycle including sampling of new experiments, execution by the high-throughput equipment, generating the next model with the decision, whether the task is succeeded and sampling new experiments if not.

The result from sample execution and analysis is a coatings recipe achieving optimal material properties. The more detailed analysis decreasing the overall uncertainty results in a model, which is the virtual copy describing the relationship between input and output parameters. This model can be applied for simulation of material properties being asked for a certain recipe or predicting the right formulation for on-demand material properties.

General Summary

The combination of ML algorithm handling the data regarding an automated analysis that suits nearly all cases to a certain extend and automated sample preparation is one realization of artificial intelligence for material design. In comparison to the traditional way of creating new formulations for materials this tool accelerates the development and is able to find improved solutions due to a better interpretation of the multi-dimensional character of the problem.

Acknowledgement

The authors thank the European Union, the MWIDE NRW, the Ministerie van Economische Zaken en Klimaat and the provinces of Limburg, Gelderland, Noord-Brabant and Overijssel for funding the D-NL-HIT project of the Interreg program. C.S. also thanks the county of North Rhine Westphalia for financial support in the program Karriereweg FH-Professur.

Conflict of Interest

No conflict of interest.

References

1. DA Wells, TL Lloyd (2002) Automation of sample preparation for pharmaceutical and clinical analysis. Comprehensive Analytical Chemistry 37: 837-868.
2. DP Tabor, LM Roch, SK Saikin, Kreisbeck C, Sheberla D, et al. (2018) Accelerating the discovery of materials for clean energy in the era of smart automation. Nat Rev Mater 3: 5-20.
3. MJ Anderson, PJ Whitcomb (1996) Optimization of Paint Formulations Made Easy with Computer-Aided DOE for Mixtures. Journal of Coatings Technology 996: 71-75.
4. MJ Anderson, PJ Whitcomb (1999) Designing Experiments that Combine Mixture Components with Process Factors. Paint & Coatings Industry 68-72.
5. Q Wei, RG Melko, JZY Chen (2017) Identifying polymer states by machine learning. Phys Rev E 95, arXiv:1701.04390.
6. DIN EN ISO 2813:2015-02 - Paints and varnishes - Determination of gloss value at 20°, 60° and 85° (ISO 2813:2014). German Institute for Standardisation, Germany.
7. DIN EN ISO 11664-5:2017-01 Colorimetry - Part 5: CIE 1976 L*u*v* Colour space and u', v' uniform chromaticity scale diagram (ISO/CIE 11664-5:2016). German version EN ISO 11664-5:2016, Germany.
8. DIN EN ISO 25178-1:2016-12 Geometrical product specifications (GPS) - Surface texture: Areal - Part 1: Indication of surface texture (ISO 25178-1:2016). German version EN ISO 25178-1:2016, Germany.
9. CE Rasmussen (2003) Gaussian processes in machine learning. Summer School on Machine Learning, Springer, Berlin, Heidelberg, pp. 63-71.
10. P Paxton, PJ Curran, KA Bollen, J Kirby, F Chen (2001) Monte Carlo experiments: Design and implementation. Structural Equation. Modeling 8(2): 287-312.
11. A Olsson, Anders, G Sandberg, O Dahlblom (2003) On Latin hypercube sampling for structural reliability analysis. Structural safety 25(1): 47-68.
12. T Wagner, M Emmerich, A Deutz, W Ponweiser (2010) On Expected-Improvement Criteria for Model-based Multi-objective Optimization. International Conference on Parallel Problem Solving from Nature, Springer, Berlin, Heidelberg, pp 718-727.
13. E Vazquez, J Bect (2010) Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. Journal of Statistical Planning and Inference 140(11): 3088-3095
14. K Cremanns, D Roos (2017) Deep Gaussian Covariance Network. Machine Learning, arXiv:1710.06202v2.
15. C Schmitz, K Cremanns, G Bissadi (2020) Application of Machine learning algorithms for use in material chemistry. Krefeld.

Citation: Yannik Schuchmann, Christian Schmitz, Lasse Wagner, Carmen Stoffelen-Janßen and Jost Göttert . Automated Modelling of Material Chemistry Applying High-Throughput Formulation and Machine Learning. Mod Concept Material Sci. 3(4): 2021. MCMS. MS.ID.000570. **DOI:** 10.33552/MCMS.2021.03.000570.