Dalton Transactions

Navigating the Design Space of Inorganic Materials Synthesis using Statistical Methods and Machine Learning

Journal:	*Dalton Transactions*
Manuscript ID	DT-FRO-06-2020-002028.R1
Article Type:	Frontier
Date Submitted by the Author:	26-Jul-2020
Complete List of Authors:	Braham, Erick; Texas A&M University, Department of Chemistry Davidson, Rachel; Texas A&M University, Department of Chemistry Al-Hashimi, Mohammed; Texas A&M University at Qatar, Arroyave, Raymundo; Texas A&M University System, Mechanical Engineering Banerjee, Sarbajit; Texas A&M University, Department of Chemistry

Navigating the Design Space of Inorganic Materials Synthesis using Statistical Methods and Machine Learning

Erick J. Braham,[1,2†] Rachel D. Davidson,[1,2 †] Mohammed Al-Hashimi,[3] Raymundo Arroyave,[2*] and Sarbajit Banerjee[1,2*]

[1]Department of Chemistry, Texas A&M University, College Station, TX 77843, USA; banerjee@chem.tamu.edu

[2]Department of Material Science and Engineering, Texas A&M University, College Station, TX 77843, USA; rarroyave@tamu.edu

[3]Department of Chemistry, Texas A&M University at Qatar, P.O. Box 23874, Doha, Qatar

[†] these authors contributed equally to this work

Abstract:

Data-driven approaches have brought about a revolution in manufacturing; however, their application to the deterministic navigation of reaction trajectories to stabilize crystalline solids with precise composition, atomic connectivity, microstructural dimensionality, and surface structure remains much more challenging. The design of synthetic methodologies for the preparation of inorganic materials is oftentimes inefficient in terms of exploration of potentially vast design spaces spanning multiple process variables, reaction sequences, as well as structural parameters and reactivities of precursors and structure-directing agents. Reported synthetic methods are further limited in terms of the insight they provide into underlying chemical and physical principles. The recent surge in interest in accelerating the discovery of new materials can be considered as an opportunity to re-evaluate our approach to materials synthesis, and for considering new frameworks for exploration that are systematic and strategic in approach. Herein, we outline with the help of several illustrative examples, the challenges, opportunities, and limitations of data-driven synthesis design. The account collates discussion of design-of-experiments sampling methods, machine learning modeling, and active learning to develop experimental workflows that accelerate the experimental navigation of synthetic landscapes.

Introduction:

Data-driven approaches have brought about a revolution in manufacturing, enabling levels of customization and control that were unimaginable with traditional mass-manufacturing.[1,2] Advances in digital manufacturing and nanoscale fabrication have further paved the way to the utilization of a much-expanded palette of materials in technological applications. Powerful as they are, digital manufacturing approaches remain constrained in their ability to structure matter at nanoscale dimensions. An important, yet unresolved, challenge lies at the interface of data science and materials synthesis. From the perspective of inorganic materials chemistry, a fundamental obstacle to the precise structuring of matter that remains to be resolved is to control reaction trajectories to stabilize crystalline solids with precise composition, atomic connectivity (crystallization of a specific polymorph), microstructural dimensionality (particle size, shape, layer thickness, or grain size), and surface structure (texture or surface crystallographic facets). Typically, materials syntheses are developed almost entirely in an empirical manner, based on fragmented knowledge of the underlying sequences of chemical reactions, heterogeneous and homogeneous equilibria, and their coupling with mesoscale mass transport and energy transfer phenomena. Much of current research practice comprises Edisonian trial-and-error methods involving changing a single synthetic variable and observing the response. Such methods are not just inherently inefficient in their exploration of potentially vast design spaces (spanning multiple process variables, reaction sequences, as well as structural parameters and reactivities of precursors and capping ligands) but furthermore do not provide a satisfactory understanding of the underlying chemical and physical principles, ultimately stymying the application of modern process design tools. In this frontier article, we discuss the potential for data science to enable the more efficient navigation of materials design spaces.

The functionality of materials derive from complex convolutions of composition and (atomistic as well as mesoscale) structure, which in turn are determined by their processing history. The design of materials for a specific application requires unraveling the interplay between physical principles that underpin materials function; weighing trade-offs across frontiers of candidate solutions to identify optimal solutions that satisfy multiple constraints; and mapping efficient pathways from starting precursors to arrive at the target material composition and structure. Navigation of synthetic design spaces is challenging because often the structures that are of greatest use are metastable in nature, resident within shallow wells on rugged energy landscapes (**Fig. 1A**).[3–5] Considerable effort has focused on the application of data science methods to accelerate the investigation of structure—property relationships based on mining of crystallographic databases and first-principles calculations of known and putative structures in search of specific function.[6–10] For instance, machine learning of experimental and computational data has enabled high-accuracy predictions of bandgap and crystal structure.[11–15] However, explorations of reaction trajectories and mapping of response surfaces of materials synthesis spaces with a view towards learning process-structure and process-property relationships are much less common.[4,16–20]

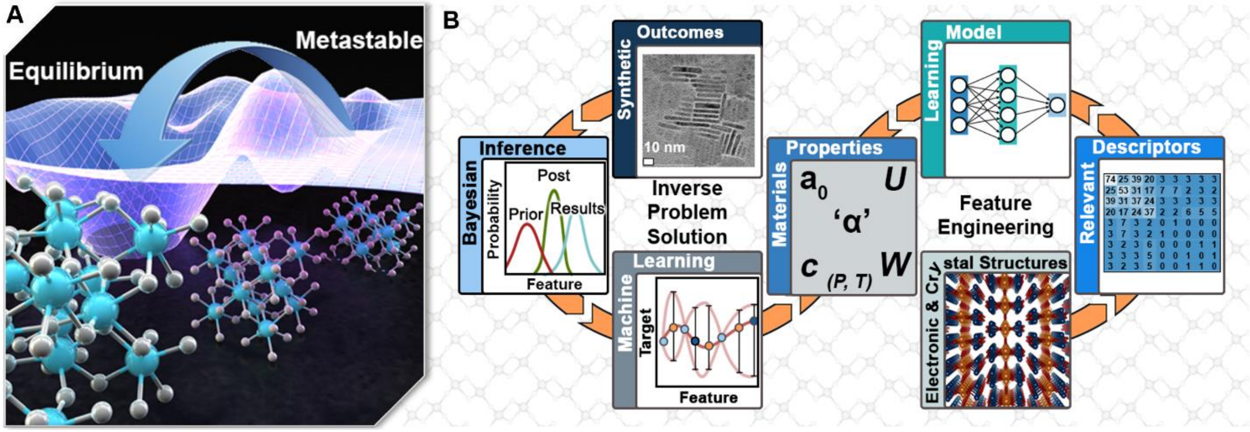


Figure 1: A) Illustration of a rugged energy landscape showing metastable and equilibrium energy wells for polymorphs of HfO_2. B) Schematic illustration of a data-driven approach to feature engineering and inverse synthesis design.

Towards Machine-Learning-Aided Inverse Synthesis Design

Challenges in the application of data science methods to materials synthesis stem from the high dimensionality of problems where *n* synthesis variables create an *n*-dimensional space for exploration, the sparsity and expense of available data, and the non-monotonicity and extreme non-linearity of many thermodynamic functions (evidenced as phase transitions).[21–24] Furthermore, understanding synthesis requires elucidation of process-structure relationships or using processing-function relationships as a proxy. Modeling the former requires the decoding of structure into numerical descriptor(s) or limits the modeling to expressions that accept categorical variables at inputs. The latter, in turn, requires that the entire design space be reasonably represented by a single or small subset of properties. This would typically translate the task of 'learning' a synthesis to the task of optimizing a specific property (e.g., particle size). The strength of coupling between spin, charge, orbital, lattice, and compositional degrees of freedom determine the shape of thermodynamic energy landscapes of periodic solids.[4,24] Strong coupling amongst the degrees of freedom can make it difficult to traverse along pathways to arrive at specific polymorphs.

In conventional high-temperature synthesis, as a system relaxes towards equilibrium, from an initial high-energy state, it scans the landscape for efficient paths to enable dissipation of the available free energy. Conventional metallurgical and ceramic processing provide a large excess of energy, enabling the material to readily find its way towards equilibrium, without being trapped in a metastable state, although there are notable examples, particularly in phase-transforming materials, in which the trapping in metastable states is highly history/processing-dependent.[25] However, given the challenges with ensuring homogeneous energy and mass flows across the system, and the sensitivity of crystallization processes to mesoscale phenomena, such processes can be difficult to control. Solution-phase synthesis, *chimie douce* routes, and templated processes (e.g., molecular beam epitaxy and pulsed laser deposition) can potentially allow for more deterministic navigation of energy landscapes such as to trap the material in a local minimum (**Fig. 1A**). However, the intrinsic path-dependence of these methods increases the dimensionality and complexity of the reaction space.[26–29] Data science methods hold promise for "learning" the design space and enabling the design of synthetic pathways that connect starting precursors and the target structure. In principle, machine learning allows for the possibility of inverse design synthesis (**Fig. 1B**); the generation of models which can take a target structure as the input and predict synthetic routes to generate materials with those properties as outputs.[30–33] Increasingly complex problems have been addressed working towards modeling systems with high costs of experimentation, optimization with multiple target objectives, and providing greater understanding of error in systems with relatively limited amounts of data.

Statistical regression and machine learning methods can aid the mapping of pathways between the target material and precursors based on the fusion of disparate types of data. First, data mined from the literature provide access to specific hyperplanes, in that it is typically data collected through consideration of one-variable-at-a-time (OVAT), which is analogous to examining a singular plane of experiments within a high dimensional reaction space wherein each variable adds another dimension to the space. These hyperplanes oftentimes reflect chemical intuition, serendipity, accessibility of specific precursors, or a combination thereof in terms of experimental design, and can provide valuable inputs to algorithms and provide a means of seeding initial experiments (albeit failed experiments typically go unreported and thus vast sections of the design space are underrepresented in the literature). Codified prior knowledge of thermodynamics and chemical concepts allow for the application of specific constraints (e.g., knowledge of decomposition temperatures or solubility guide precursor selection). Results from first-principles calculations and molecular dynamics simulations, oftentimes coupled with metaheuristic algorithms (simulated annealing or basin hopping),[32,33] algorithms that screen different optimization procedures for their facility with converging at a minimum, can guide exploration of the adjacent phase space to identify potential intermediates that can be exploited as waystations to the target, or conversely, to avoid thermodynamic dead ends. Accurate first-principles descriptions of entire systems and energy landscapes are inaccessible in most cases owing to inadequate energy resolution and high computational costs. The available inputs can then inform the targeted navigation of the synthesis design spaces without having to perform full factorial experiments across multiple dimensions of chemical, process, and temporal variables. Microfluidic platforms,[18] high-throughput robotic arm dispensation systems, entirely mobile robots,[34] and parallelized hydrothermal platforms[35] have emerged as alternatives for rapidly acquiring data to test the validity of data-driven synthesis models.

A major advantage afforded by machine-learning-aided approaches is the ability to progress beyond expensive one-variable-at-a-time (OVAT) sampling methods to more efficiently explore synthetic landscapes, minimization of sampling bias often inherent in human intuition, and a decrease in likelihood of arrival at a local minimum hyperplane. Perhaps more importantly, machine learning algorithms are generalizable and can thus be used as a means of “rule discovery”, thereby unraveling hidden correlations and providing fundamental chemical and physical insight of the underlying reactivity.[36] Here, we will outline the use of design of experiments in exploration of synthetic landscapes and discuss how machine learning algorithms can complement these statistical sampling techniques in order to accelerate materials discovery referencing some illustrative examples from the literature.

Beyond Guess and Check: Design-of-Experiments and Connections to Machine Learning

Data is at the core of any machine learning model, and for applications in materials science, the lack of data, format of data, and quality of data can frequently represent a bottleneck to progress. For problems in which data is not available elsewhere, starting with the design-of-experiments (DOE) sampling techniques, rather than OVAT methods, can often afford a more richly diverse dataset that is readily amenable to modeling with machine learning algorithms.[37–39] An abiding challenge is to determine the best approaches to represent chemical structure and composition in a manner amenable to the application of statistical regression tools. A major research question is thus to identify the structural and compositional motifs, processing conditions, and reaction sequences that are most strongly associated with the synthetic outcomes. Such “feature engineering” (**Fig. 1B**) is pivotal to developing a scored experimental design approach that allows for identification of the key descriptors underpinning a specific synthetic output and enables iterative improvement of the synthesis models.

Predating the use of machine learning methods, DOE methodologies have shown considerable value in materials synthesis.[18,40–44] These methods aim to first broadly sample large design spaces with as few experiments as possible utilizing approaches such as full and fractional factorial designs, random sampling, or, more recently, Bayesian optimization in which probability estimates from the model are continuously updated as new data is acquired. These sampling methods are typically coupled with response surface modeling to generate a rough model of the system through use of a simple regressor.[45,46] As a notable example of this approach, Murphy and co-workers[47] explored the seed-mediated silver-assisted growth of gold nanorods using fractional factorial DOE along eight independent experimental parameters. In the synthesis, gold seeds are prepared by combining a solution containing a gold precursor and capping ligand (e.g., cetyltrimethylammonium bromide (CTAB)) with a solution containing the reducing agent. The seed solution is then added to a solution containing a capping ligand, additional gold precursor, a weaker reducing agent, and silver nitrate. Numerous studies had previously evaluated effects of different reaction parameters using traditional OVAT methods and had determined that the concentration of gold seeds, temperature, amount of silver nitrate, and concentration of the ascorbic acid reducing agent were all of relevance to determining the aspect ratio of nanorods.[48–51] Using DOE methods, the authors not only demonstrated all of the trends observed previously with the separate OVAT studies but also determined that the interaction of variables was significant. They demonstrated that while there is a positive correlation between concentration of silver nitrate and the nanorod length, it has no primary effect on the length and instead demonstrates a secondary interaction with the amount of reducing agent. These results provided much needed insight into the true role of silver nitrate and the general mechanisms of anisotropic growth. While it was previously postulated that silver or silver bromide absorbed on the gold surface may serve as a blocking layer resulting from underpotential deposition on certain crystallographic faces, the correlation of $AgNO_3$ concentration with the concentration of reducing agent indicates that it more likely shields charge for negatively charged species headed towards the negatively charged (as a result of Br^- adsorption) nanocrystal surface. The authors postulated that surface adsorption of Br^- ions directs anisotropic growth.

In DOE studies, the initial round of sampling is frequently used as a means of down-selecting to variables with the highest influence on synthetic outcomes through feature selection (**Fig. 1B**), allowing for the possibility of a second round of more dense sampling of the design space of interest. In the steepest ascent approach this involves iterative sampling in the direction which heads towards an optimum in output. While the data is ascending, a first-order model is used which does not account for curvature in the output data. Once near the apex in data, a second order model which accounts for curvature of the data provides a better fit.[45] Mora-Tamez et al.[18] explored the colloidal synthesis of Ni_2P nanoparticles to generate a model predicting particle size. In the first-order design, six possible factors were screened for influence on the size of nanoparticles followed by a second-order design model of the dependence on the strongest influencing factors; triphenylphosphine/nickel ratio and temperature. This model was then corroborated with four additional samples all resulting in particles with excellent agreement to the predicted values well within experimental error. The high monodispersity of nanoparticles and low experimental noise of the chosen synthetic method implemented within a microfluidic platform combined with the lack of complex variable correlation in this study allowed an accurate response surface model to be built from a relatively small (9 sample) DOE chosen dataset. This methodology is useful for analyzing the trends, magnitude of influence, and correlation of a large set of variables.

As it is typically implemented, DOE is best suited for optimization problems. The interpolation offered with response surfaces can be predictive for small design areas with linear or quadratic trends but often is constrained in its ability to analyze design spaces with more complex responses and systems where exploration, rather than optimization, is the focus. When coupled with the capabilities of machine learning algorithms and with the incorporation of features representative of chemical structure and composition, the opportunities for systematic exploration of synthetic landscapes are greatly expanded. **Figure 2** depicts a workflow for machine-learning-aided navigation of synthetic design space.

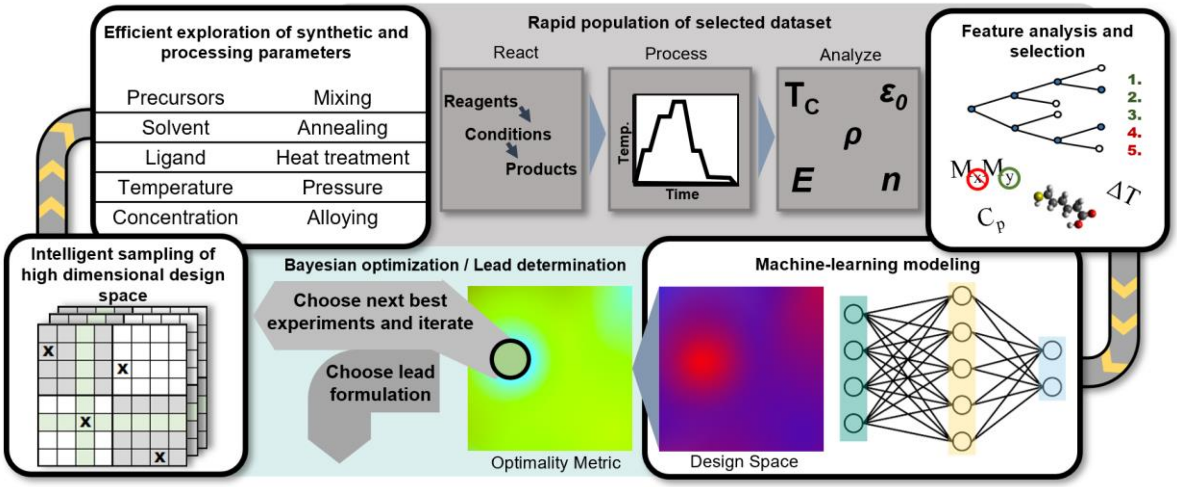

Efficient exploration of synthetic and processing parameters
Precursors
Mixing
Solvent
Annealing
Ligand
Heat treatment
Temperature
Pressure
Concentration
Alloying
Rapid population of selected dataset
React
Process
Analyze
Reagents
Conditions
Products
Temp.
Time
T_C
ε_0
ρ
E
n
Feature analysis and selection
1.
2.
3.
4.
5.
$M_x M_y$
C_p
ΔT
Intelligent sampling of high dimensional design space
Bayesian optimization / Lead determination
Choose next best experiments and iterate
Choose lead formulation
Optimality Metric
Machine-learning modeling
Design Space

Figure 2: Schematic depiction of an example of a machine-learning workflow for the iterative exploration and exploitation of a synthetic design space for inorganic materials.

Efficient Routes to Complex Predictivity:

Canonical machine learning techniques are particularly well suited to large amounts of data with complex prediction or optimization-based goals.[52,53] While the goals of prediction and optimization remain the same for materials synthesis, the cost of data is much higher. While response-surface DOE provides a distinct edge over OVAT sampling and allows for reliable inferences from a small amount of data, it is often unable to handle data that has a more complex response. Machine learning methods have the ability to provide valuable insight and to develop useful models that can be further iteratively improved. However, the sparsity of datasets and the substantial costs of experiments have limited its application in the exploration of synthesis design spaces.

In recent work, we have performed a study using existing OVAT data supplemented with random sampling to build a predictive model of a nanocrystal synthesis.[17] The study used only 74 samples to create a model that was able to predict both the conditions that will lead to quantum confined $CsPbBr_3$ nanoplatelets and their average thickness for a given sample. This relation of three experimental parameters (temperature, ligand choice, and ligand concentration) generated a highly nonlinear response that was mapped within one unit cell layer of accuracy using support vector machine regression (SVM, **Figure 3**). This supervised learning model chooses the fit that minimizes the length of vectors perpendicular to the fit that connect the model to the data. The model was physically interpretable in terms of the competition between enthalpic and entropic considerations and in distinguishing thermodynamic and kinetic regimes under different synthetic conditions. The observed resolution of mapping, ability to handle imperfect data, and insight into a nonlinear reaction space was made possible by using machine learning. Despite the positive use case shown in this work, the number of experiments was high, and parts of the design space may have been under sampled. To increase efficiency and lower variability of the response space, more targeted sampling and iterative design methods are desirable.

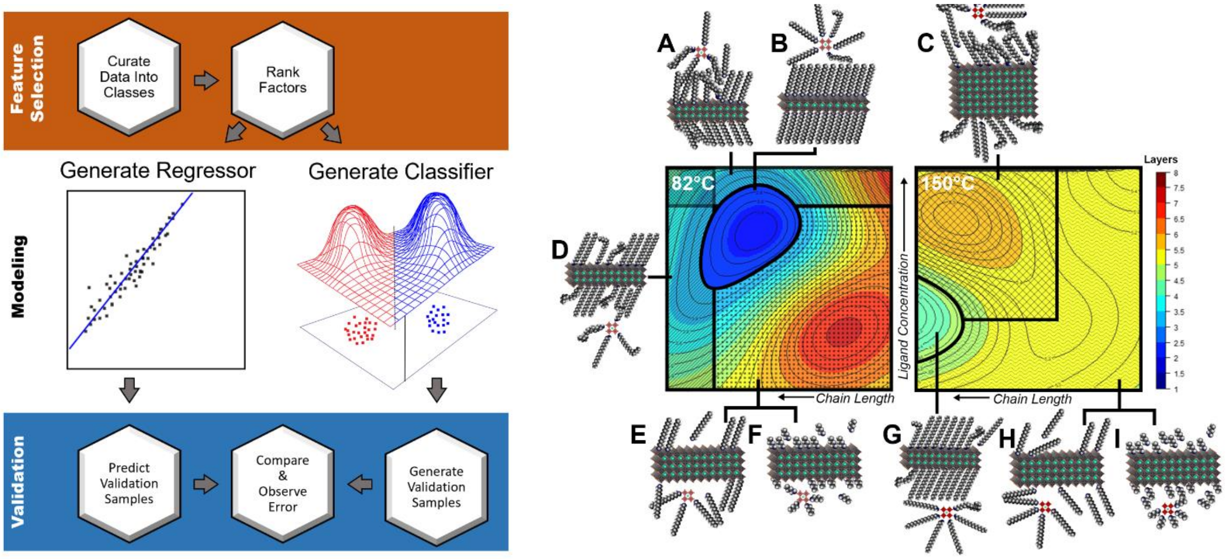


Figure 3: Machine learning flowchart (left) and interpretation of the SVM regression results (right) from the study of $CsPbBr_3$ perovskite nanocrystal growth by Braham et al.[17] Regression heatmaps of the particle thicknesses along the modeled axes of ligand chain length and ligand concentration is shown at two temperatures of 82 and 150°C. Interpretation and illustration of 9 selected growth regimes are depicted in A-I showing findings of chemical significance with a global minimum thickness regime with a close packed monolayer (B), entropy-driven monolayer misalignments for high chain lengths or concentration (A,C,D), incomplete monolayer formation owing to low ligand concentrations or weak intermolecular interactions (E,F,H,I,) and a local minimum at high temperature illustrating the shift in ideal monolayer packing conditions as a function of temperature (G). Reprinted with permission from ref. 15; Copyright 2019, the American Chemical Society.[17]

Cao *et al*. have described and demonstrated this synergistic relationship between DOE and ML in a perspective article examining the optimization of the power conversion efficiency for a bulk heterojunction photovoltaic device created via spin casting a mixture of a low-band-gap donor polymer and fullerene as an acceptor with the addition of diiodooctane, which is thought to decrease donor-acceptor phase segregation.[37] These authors specifically considered the influence of the weight percentage of the low-band-gap polymer, total solution concentration, spin-casting speed, and the volume percent of the diiodooctane additive. They sampled the reaction space using a fractional factorial, analyzed variance in the data using ANOVA analysis to understand feature correlations, and fitted the data using an SVM with a radial basis function kernel. The

SVM was then used to generate a visual map of the space, which informed the design of a second round of fractional factorial sampling, eliminating the addition of diiodooctane as a variable and narrowing the range of the other factors to target the area of the space demonstrating the highest power conversion efficiency values. While this second round of sampling narrowed in on the optima, a purely exploitative approach can potentially converge on local rather than global optima. However, this study exemplifies the promise of using a combination of DOE and ML to sample, model, and explore a synthesis space. While ML algorithms can oftentimes uncover hidden correlations among variables and provide some predictivity, a ‘one-shot’ fitting of a model to a space often lacks predictive capability beyond that of interpolation.

Active learning, sometimes referred to as sequential learning, is an iterative process where a utility or acquisition function is applied to the output of an initial surrogate model (typically a ML model) to strategically select a new area of the design space to sample.[54–56] Such an approach, illustrated in **Figure 2**, allows for rapid updating of the model and enables efficient exploration of the synthetic design space. Active learning approaches often leverage exploration strategies from global optimization methods with a Bayesian-optimization-based approach being most popular.[57] In this particular iterative approach, the acquisition functions are based on Bayes Theorem and leverage information previously observed to find a posterior distribution using scores from the surrogate model. The acquisition function then chooses the most valuable experiment to perform next, balancing a preference towards choosing samples that would either be the most helpful to improve the extent to which the model captures the dataset (exploration) or move towards a predicted maximum or minimum of the surrogate function (exploitation). This typically limits use of the Bayesian optimization strategy to problems using regression-based models. Just as no single ML model works well to fit every dataset, various active learning workflows are better suited for different problems.[56] Utility or acquisition functions vary in the degree to which they favor exploration or exploitation of the data, allowing for users to focus more on creating either the most accurate design space model or finding an optimal solution. Xue *et al.* demonstrated the efficacy of this approach in the systematic exploration of the synthesis of a Ti_{50} ($Ni_{50\text{-}x\text{-}y\text{-}z}Cu_xFe_yPd_z$) shape memory alloy.[58] In order to optimize the transformation temperature, the authors synthesized an initial set of 53 alloys, and applied a polynomial model to serve as the surrogate model coupled with iterative sampling using expected improvement as the acquisition function. The model effectively identified samples with increased transformation temperatures and captured the influence of atomic size on local strain and influence of bond strength on the transformation temperature. While this process can in principle be used to explore a wide variety of synthetic landscapes, examples involving active learning of experimental synthesis spaces are still limited. The iterative framework adds another layer of constraints on the framing of the problem, as the data must be suitably modeled by a regressor in order to reasonably predict the subsequent samples to be measured. Strategies for Bayesian sequential learning are currently being developed to overcome challenges.[54,59–62] For example, Wang *et al*. developed a Bayesian optimization approach to enable nested-batch sampling.[59] In this method the algorithm predicts the most beneficial batch of samples to run next, rather than a ranked list of single experiments. It additionally allows for the user to rank variables to avoid variance within a given batch of variables that would not be feasible within a batchwise process. This addresses two problems which are unique to synthesis; some variables are more expensive to vary than others and sampling two drastically different samples may be much more expensive than sampling two similar samples. For example, it is often simple to vary concentration, as a single stock solution can be made and then diluted to different extents. However, varying the reaction temperature may be limited by the number of independent thermal profiles accessible within a single autoclave, or varying the solvent may negate the opportunity to work with a stock solution. To consider this constraint, the acquisition function estimates the value of information for each batch of samples that could be generated.

Most Bayesian optimization efforts have focused on single-objective optimization. However, in materials chemistry, multiple objectives (along a Pareto frontier) must be optimized at once. For instance, minimizing defect density and positioning dopant atoms within a particle while controlling particle size. This problem can be framed as identifying the optimal sequence of observations (via experiments or simulations) that is most efficient at identifying the Pareto frontier of candidate solutions, which is a graphical representation of the tradeoffs between two output parameters. In common Bayesian optimization methodologies, it is assumed that the search for the global optimum of the descriptive function is sequential, evaluating the function one step at a time, regardless of the number of objectives to optimize. This means that even in multi-objective Bayesian optimization it is necessary to quantify the utility of a potential experiment as a scalar quantity. A powerful scalar utility metric used in multi-objective optimization is the so-called Expected Hyper-Volume Improvement (EHVI).[63] Similar to the utility functions used in single-objective Bayesian optimization, EHVI is constructed by balancing the exploration and exploitation of the design space in order to efficiently locate the Pareto frontier. Recently, we have developed a multi-objective (up to three objectives) optimal materials discovery framework[64] and demonstrated its efficacy by identifying regions in the microstructural space that yielded optimal performance in a precipitation-strengthened NiTi-based shape memory alloy. The alloy composition as well as microstructural features (specifically, the precipitate volume fraction) of precipitation-hardened nickel—titanium alloys have been optimized within a pre-defined budget of experimental steps.[64] This approach demonstrates the promise of multi-objective Bayesian optimization methods to develop optimal sequence of experiments allowing for simultaneous control of different synthetic outcomes.

An alternative approach to overcome the constraints of a Bayesian-based active learning approach was demonstrated by Moosavi and co-workers, who decoupled the tasks of learning and exploring the space by using a ML algorithm to understand correlations among variables and gain insight into the reaction mechanisms, while using a metaheuristic global optimization strategy to iteratively explore the space in their search for a metal-organic framework (HKUST-1) with the highest surface area.[65] In doing so, they relax the need for a good initial fit to a surrogate model and gain greater flexibility in choice of ML model, as the model does not need to work well within an optimization workflow. Looking broadly for opportunities to leverage tools in adjacent areas of work will likely be key in increasing the use and diversity of spaces explored experimentally using iterative methods.

The consistency of microstructure in an alloy is an excellent case study in not only synthetic optimization but building fundamental scientific understanding of process—structure relationships though statistical learning. The influence of microstructure on key materials properties such as tensile strength and cycling fatigue creates a crucial need for a clear understanding of processing-structure-property relationship that go beyond empirical constructs. Recent studies by Elwany and co-workers have applied statistical learning methods to reveal processing—microstructure/mesostructure relationships in additive manufacturing techniques, specifically laser powder-bed fusion.[66,67] Using a Ni-Nb alloy as a model system, this work utilizes machine learning and a set of materials properties/processing features (melt pool depth, diffusivity of liquid, the Gibbs-Thomson coefficient, and the equilibrium partition coefficient) to model an experimental "base truth" dataset of microsegregation in the grain structure of the alloy.[67] The statistical model showed a level of predictivity to the experimentally compiled dataset but was not able to reach the level of multiphysics phase-field simulations. The approach demonstrates that with a sufficient dataset and by investigating alternative machine learning or featurization methods, design principles underpinning processing-structure relationships can be revealed.

Looking Forward

The salient features of OVAT, DOE, ML, and active learning sampling and modeling methods are summarized in **Figure 4**. OVAT sampling is limited largely by the sampling bias of the scientist and in its limitations in demonstrating correlation among variables. While DOE is an excellent and efficient qualitative strategy for unraveling the impacts of different variables on a synthetic outcome, the standard response surface methodology falls short in fitting highly non-linear responses. Iterative processes in DOE are generally exploitative (rather than exploratory) in nature. DOE is also incompatible with working from found/sparse data. When designing a new modeling or optimization experiment, DOE sampling methods represent an excellent choice for building an initial dataset that can be used in conjunction with ML.

When refining and improving a ML model, emerging active learning techniques provide excellent targeted sampling to achieve a certain goal or for multi-objective optimization, given that the function most appropriate for the goal/s and modeling method is used. The combination of DOE, ML, and active learning would allow for a more robust and efficient path to navigating the design space of materials synthesis. The sparsity of data and the relative cost of synthesis and characterization represent significant barriers. Greater flexibility in automated synthesis platforms, either based on microfluidic systems or robotic arms performing multiplexed synthesis, hold promise for resolving these bottlenecks.[10,34] DOE techniques, ML, and automation of synthesis together represent a promising toolbox for accelerating materials synthesis, providing foundational understanding of the underlying chemical reactivity, and for extracting design principles in order to precisely control reaction trajectories.

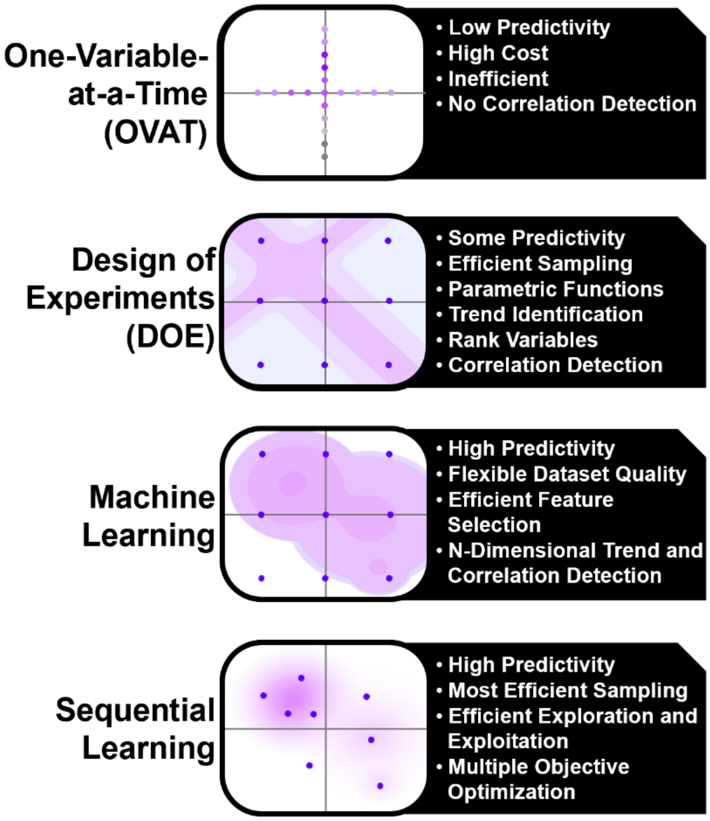

One-Variable-at-a-Time (OVAT)
• Low Predictivity
• High Cost
• Inefficient
• No Correlation Detection
Design of Experiments (DOE)
• Some Predictivity
• Efficient Sampling
• Parametric Functions
• Trend Identification
• Rank Variables
• Correlation Detection
Machine Learning
• High Predictivity
• Flexible Dataset Quality
• Efficient Feature Selection
• N-Dimensional Trend and Correlation Detection
Sequential Learning
• High Predictivity
• Most Efficient Sampling
• Efficient Exploration and Exploitation
• Multiple Objective Optimization

Figure 4: Schematic contrasting the sampling and modeling features contained in OVAT, DOE, ML, and sequential/active learning.

Acknowledgements

This work was supported by the NSF under DMR 1627197. The research was funded in part by award #A-1978-20190330 from the Welch Foundation. E.J.B. and R.D.D. acknowledge support from the Data-Enabled Discovery and Design of Energy Materials (D3EM) program funded through NSF Award No. DGE-1545403. The authors also gratefully acknowledge the support of this work from the Qatar National Research Found (QNRF) and the National Priorities Research Program, project number NPRP10-0111-170152.

References:

1 G. Daehn, G., and Spanos, *Metamorphic Manufacturing: Shaping the Future of On-Demand Components*, Pittsburgh, PA, 2019.

2 A. Bajpayee, M. Farahbakhsh, U. Zakira, A. Pandey, L. A. Ennab, Z. Rybkowski, M. K. Dixit, P. A. Schwab, N. Kalantar, B. Birgisson and S. Banerjee, *Front. Mater.*, 2020, **7**, 52.

3 D. P. Shoemaker, Y.-J. Hu, D. Y. Chung, G. J. Halder, P. J. Chupas, L. Soderholm, J. F. Mitchell and M. G. Kanatzidis, *Proc. Natl. Acad. Sci.*, 2014, **111**, 10922–10927.

4 A. Parija, G. R. Waetzig, J. L. Andrews and S. Banerjee, *J. Phys. Chem. C*, 2018, **122**, 25709–25728.

5 W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson and G. Ceder, *Sci. Adv.*, 2016, **2**, e1600225.

6 J. J. de Pablo, N. E. Jackson, M. A. Webb, L.-Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, J. Analytis, I. Dabo, D. M. DeLongchamp, G. A. Fiete, G. M. Grason, G. Hautier, Y. Mo, K. Rajan, E. J. Reed, E. Rodriguez, V. Stevanovic, J. Suntivich, K. Thornton and J.-C. Zhao, *npj Comput. Mater.*, 2019, **5**, 41.

7 A. O. Oliynyk and A. Mar, *Acc. Chem. Res.*, 2018, **51**, 59–68.

8 A. O. Oliynyk, L. A. Adutwum, B. W. Rudyk, H. Pisavadia, S. Lotfi, V. Hlukhyy, J. J. Harynuk, A. Mar and J. Brgoch, *J. Am. Chem. Soc.*, 2017, **139**, 17870–17881.

9 D. L. McDowell and S. R. Kalidindi, *MRS Bull.*, 2016, **41**, 326–337.

10 D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson and A. Aspuru-Guzik, *Nat. Rev. Mater.*, 2018, **3**, 5–20.

11 Y. Zhuo, A. Mansouri Tehrani and J. Brgoch, *J. Phys. Chem. Lett.*, 2018, **9**, 1668–1673.

12 A. O. Oliynyk, L. A. Adutwum, J. J. Harynuk and A. Mar, *Chem. Mater.*, 2016, **28**, 6672–6681.

13 G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis and T. Lookman, *Sci. Rep.*, 2016, **6**, 19375.

14 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.

15 M. A. F. Afzal, C. Cheng and J. Hachmann, *J. Chem. Phys.*, 2018, **148**, 241712.

16 W. Sun, D. A. Kitchaev, D. Kramer and G. Ceder, *Nat. Commun.*, 2019, **10**, 1–9.

17 E. J. Braham, J. Cho, K. M. Forlano, D. F. Watson, R. Arròyave and S. Banerjee, *Chem. Mater.*, 2019, **31**, 3281–3292.

18 L. Mora-Tamez, G. Barim, C. Downes, E. M. Williamson, S. E. Habas and R. L. Brutchey, *Chem. Mater.*, 2019, **31**, 1552–1560.

19 N. D. Burrows, S. Harvey, F. A. Idesis and C. J. Murphy, *Langmuir*, 2017, **33**, 1891–1907.

20 J. Kirman, A. Johnston, D. A. Kuntz, M. Askerka, Y. Gao, P. Todorović, D. Ma, G. G. Privé and E. H. Sargent, *Matter*, 2020, **2**, 938–947.

21 B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta and L. Ward, *Mol. Syst. Des. Eng.*, 2018, **3**, 819–825.

22 P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73–76.

23 E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder and E. Olivetti, *Chem. Mater.*, 2017, **29**, 9436–9444.

24 J. L. Andrews, D. A. Santos, M. Meyyappan, R. S. Williams and S. Banerjee, *Trends Chem.*, 2019, 1, 711–726.

25 A. Talapatra, R. Arróyave, P. Entel, I. Valencia-Jaime and A. H. Romero, *Phys. Rev. B - Condens. Matter Mater. Phys.*, 2015, **92**, 054107.

26 J. R. Chamorro and T. M. McQueen, *Acc. Chem. Res.*, 2018, **51**, 2918–2925.

27 J. Gopalakrishnan, *Chem. Mater.*, 1995, **7**, 1265–1275.

28 N. T. K. Thanh, N. Maclean and S. Mahiddine, *Chem. Rev.*, 2014, **114**, 7610–7630.

29 J. Livage, M. Henry and C. Sanchez, *Prog. Solid State Chem.*, 1988, **18**, 259–341.

30 A. Jain, J. A. Bollinger and T. M. Truskett, *AIChE J.*, 2014, **60**, 2732–2740.

31 A. Talapatra, S. Boluki, T. Duong, X. Qian, E. Dougherty and R. Arróyave, *Phys. Rev. Mater.*, 2018, **2**, 113803.

32 D. C. Lonie and E. Zurek, *Comput. Phys. Commun.*, 2011, **182**, 372–387.

33 A. Shamp, T. Terpstra, T. Bi, Z. Falls, P. Avery and E. Zurek, *J. Am. Chem. Soc.*, 2016, **138**, 1884–1892.

34 B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick and A. I. Cooper, *Nature*, 2020, **583**, 237–241.

35 D. Y. Shahriari, A. Barnabè, T. O. Mason and K. R. Poeppelmeier, *Inorg. Chem.*, 2001, **40**, 5734–5735.

36 S. M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K. C. Stylianou and B. Smit, *Nat. Commun.*, 2019, **10**, 539.

37 B. Cao, L. A. Adutwum, A. O. Oliynyk, E. J. Luber, B. C. Olsen, A. Mar and J. M. Buriak, *ACS Nano*, 2018, **12**, 34.

38 T. Lundstedt, E. Seifert, L. Abramo, B. Thelin, Å. Nyström, J. Pettersen and R. Bergman, *Chemom. Intell. Lab. Syst.*, 1998, **42**, 3–40.

39 R. G. Brereton, *Applied Chemometrics for Scientists*, John Wiley & Sons, Ltd, Chichester, UK, 2007.

40 A. J. Rondinone, A. C. S. Samia and Z. J. Zhang, *J. Phys. Chem. B*, 2000, **104**, 7919–7922.

41 T. Taghvaee, S. Donthula, P. M. Rewatkar, H. Majedi Far, C. Sotiriou-Leventis and N. Leventis, *ACS Nano*, 2019, **13**, 3677–3690.

42 H. Maleki, L. Durães and A. Portugal, *J. Phys. Chem. C*, 2015, **119**, 7689–7703.

43 M. A. B. Meador, L. A. Capadona, L. McCorkle, D. S. Papadopoulos and N. Leventis, *Chem. Mater.*, 2007, **19**, 2247–2260.

44 C. Barglik-Chory, C. Remenyi, H. Strohm and G. Müller, *J. Phys. Chem. B*, 2004, **108**, 7637–7640.

45 J. J. Sheng, in *Modern Chemical Enhanced Oil Recovery*, Elsevier, 4th edn., 2011, pp. 1–11.

46 M. Almeida Bezerra, R. E. Santelli, P. Oliveira, L. Silveira Villar, A. Am´, A. Escaleira, M. A. Bezerra, R. E. Santelli, E. P. Oliveira, L. S. Villar and L. A. Escaleira, *Talanta*, 2008, **76**, 965–977.

47 S. E. Lohse, N. D. Burrows, L. Scarabelli, L. M. Liz-Marza and C. J. Murphy, *Chem. Mater*, 2014, **26**, 24.

48 N. R. Jana, *Small*, 2005, **1**, 875–882.

49 F. Hubert, F. Testard, G. Rizza and O. Spalla, *Langmuir*, 2010, **26**, 6887–6891.

50 T. K. Sau and C. J. Murphy, *Langmuir*, 2004, **20**, 6414–6420.

51 F. Hubert, F. Testard and O. Spalla, *Langmuir*, 2008, **24**, 9219–9222.

52 U. S. Shanthamallu, A. Spanias, C. Tepedelenlioglu and M. Stanley, in *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, IEEE, Larnaca, 2017, pp. 1–8.

53 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.

54 R. Dehghannasiri, D. Xue, P. V. Balachandran, M. R. Yousefi, L. A. Dalton, T. Lookman and E. R. Dougherty, *Comput. Mater. Sci.*, 2017, **129**, 311–322.

55 P. V. Balachandran, B. Kowalski, A. Sehirlioglu and T. Lookman, *Nat. Commun.*, 2018, **9**, 1668.

56 B. Settles, *Active Learning Literature Survey*, 2009.

57 T. Lookman, P. V. Balachandran, D. Xue and R. Yuan, *npj Comput. Mater.*, 2019, **5**, 21.

58 D. Xue, D. Xue, R. Yuan, Y. Zhou, P. V. Balachandran, X. Ding, J. Sun and T. Lookman, *Acta Mater.*, 2017, **125**, 532–541.

59 Y. Wang, K. G. Reyes, K. A. Brown, C. A. Mirkin and W. B. Powell, *SIAM J. Sci. Comput.*, 2015, **37**, B361–B381.

60 R. Aggarwal, M. J. Demkowicz and Y. M. Marzouk, in *Springer Series in Materials Science*, 2016, vol. 225, pp. 13–44.

61 T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi and K. Tsuda, *Mater. Discov.*, 2016, **4**, 18–21.

62 J. Ling, M. Hutchinson, E. Antono, S. Paradiso and B. Meredig, *Integr. Mater. Manuf. Innov.*, 2017, **6**, 207–217.

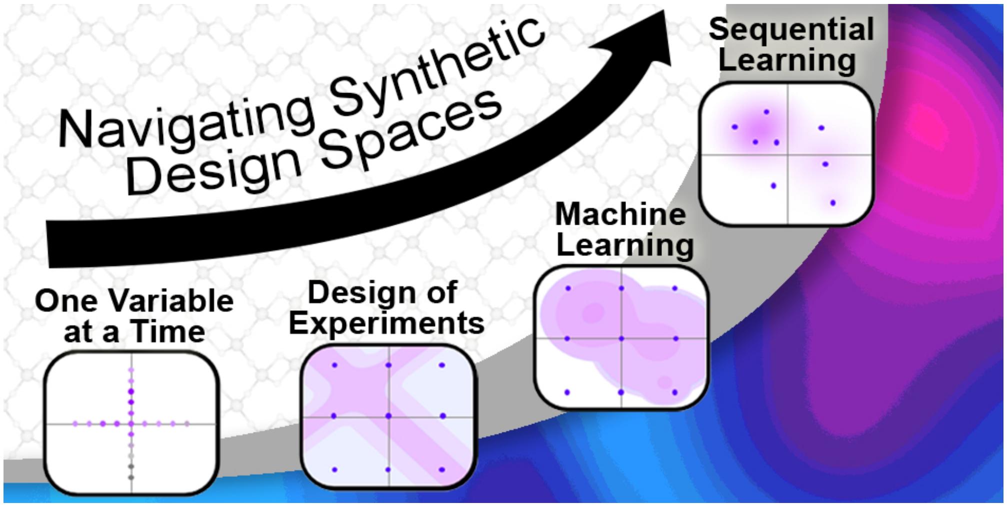
63 M. T. M. Emmerich, A. H. Deutz and J. W. Klinkenberg, in *2011 IEEE Congress of Evolutionary Computation, CEC 2011*, 2011, pp. 2147–2154.

64 A. Solomou, G. Zhao, S. Boluki, J. K. Joy, X. Qian, I. Karaman, R. Arróyave and D. C. Lagoudas, *Mater. Des.*, 2018, **160**, 810–827.

65 S. M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K. C. Stylianou and B. Smit, *Nat. Commun.*, 2019, **10**, 539.

66 J. Mingear, B. Zhang, D. Hartl and A. Elwany, *Addit. Manuf.*, 2019, **27**, 565–575.

67 S. Ghosh, R. Seede, J. James, I. Karaman, A. Elwany, D. Allaire and R. Arroyave, *Philos. Mag. Lett.*, 2020, **100**, 271–282.



333x166mm (72 x 72 DPI)