Symbolic regression in materials science

Yiqun Wang,* Nicholas Wagner,* and James M. Rondinelli
Department of Materials Science and Engineering, Northwestern University, Illinois 60208, USA

We introduce symbolic regression as an analytic method for use in materials research. First, we briefly describe the current state-of-the-art method, genetic programming-based symbolic regression (GPSR), and recent advances in symbolic regression techniques. Next, we discuss industrial applications of symbolic regression and its potential applications in materials science. Last, we present a GPSR use-case for formulating a transformation kinetics law and show the learning scheme discovers the well-known Johnson-Mehl-Avrami-Kolmogorov (JMAK) form. Finally, we propose that symbolic regression techniques should be considered by materials scientists as an alternative to other machine-learning-based regression models for learning from data.

I. MOTIVATION

A. Era of big data in materials science

Modern scientists perpetuate the scientific process embodied by the works of Tycho Brahe, Johannes Kepler, and Isaac Newton in the heliocentric revolution. Brahe was the observationalist. He took extensive, precise measurements of the position of planets over time. Kepler was the phenomenologist. From Brahes measurements, he derived concise analytical expressions that describe the motion of the solar system in a succinct manner. Last, Newton was the theorist. He realized the motion of objects on earth are described by universal forces. All three scientific modalities are vital in making scientific discoveries: data acquisition (Brahe), data analysis (Kepler), and derivation from first-principles (Newton).

With recent advances in computer science, theoretical modelling, and experimental instrumentation, materials scientists have in many ways created a “mechanical Brahe” and marched into a new era of big data. Datasets of materials information, obtained from advanced characterization techniques,4–6 combinatorial experiments,4–6 high-throughput first-principles simulations,7,8 literature mining,9,10 and other techniques, are created at a faster rate every day with less and less human labor. All of this data enables new opportunities to construct novel laws of phenomenological behavior for systems that previously lacked them.

Inspired by the Materials Genome Initiative (MGI),† the materials community is working collaboratively towards making digital materials data accessible to others. There are multiple existing materials databases,12,13 where tens of thousands of materials information are publicly available. Accessibility to an immense amount of materials data paves way for the next step of “automating Kepler” in the discovery of governing laws in materials processing-structure-properties-performance relationships, which could advance materials discovery, development, and technology innovation.

Since one of the fundamental research objectives of materials science and engineering is to deliver new materials with optimal performance under specified constraints, it is essential to understand how and which features govern the functionality. In other words, which degrees-of-freedom (or parameters) and their corresponding intrinsic relationships (or dependencies) to the material properties should be optimized. However, the multi-scale nature of materials science, e.g. from atomic-scale crystal structure to complex mesoscale domain structures and bulk mechanical properties or from femtosecond laser probes to hour-long recrystallization reactions, makes it particularly challenging to study many hierarchical relationships of different materials families. Given such a high-dimensional parameter space (e.g. chemical composition, crystal structure, external conditions, etc.), materials scientists often explore a finite subspace of all the factors that govern materials properties and performance. In addition, the available data is typically sparsely distributed. Although, access to a large materials database relieves, in part, the limited-data problem, there is an urgent need for a robust data-processing protocol to help discern governing laws in materials science and to deliver designer materials and synthesis/processing procedures.

B. An alternative to machine-learning methods

Much of the burgeoning field of materials informatics focuses on the aforementioned challenges. Machine learning (ML) models are currently the tools of choice for uncovering these physical laws. Although they have shown some promising performance in predicting materials properties,14 typical parameterized machine learning models are not conducive to the next stage of generalizing across domains—the ultimate goal of “automating Newton.”

It is important to note that Newton’s challenge was somewhat made easier, because Kepler’s laws were parsimonious yet predictive. In a modern context, ML models can be predictive but their descriptions are often too verbose (e.g. deep-learning models with thousands of parameters) or mathematically restrictive (e.g. assuming the target variable is a linear combination of input features). Such black-box models have become more prevalent in modern materials science research; however, the interpretability of such models have always been a problem. Although there is a large body of work on data visual-
ization and model understanding to address these issues, those subjects will be out of the scope of this perspective (see Ref. 15 for a review).

In this prospective paper, we focus on an alternative to machine-learning models: symbolic regression. Symbolic regression simultaneously searches for the optimal form of a function and set of parameters to the given problem, and is a powerful regression technique when little if any a-priori knowledge of the data structure/distribution is available.

Figure 1 shows the relative popularity of machine learning and symbolic regression in different research domains. We use data from the “Web of Science Core Collection” database in this analysis. Among all publications whose topics are related to machine learning or symbolic regression, over 50% of the contributions come from the computer science research community, while multidisciplinary engineering is second. Social science and physical science each makes less than 20% of the contribution to the total number of publications. These two techniques are not so popular in materials science research, as the relative contribution is almost negligible compared to other research domains.

It is not surprising to see a dominant contribution from computer science in both the machine learning and symbolic regression communities, since it is where these techniques were born. It is interesting to notice that symbolic regression is significantly more popular than machine learning in social science research. One possible reason for this trend is that social science problems typically do not have a (known) physically motivated governing equation as in many physical sciences, where for example, Newtonian equation-of-motions, Schrodinger equation, etc. can be written formally. Symbolic regression arises naturally as a problem solver since it has the potential to find an appropriate functional form from social science data sets, e.g. questionnaire results, behavior patterns, etc.

We also report the trend in the number of publications (in a natural logarithm scale) in the following research domains [Figure 1(inset)]: machine learning, application of machine learning in materials science, and symbolic regression. All three domains exhibit a rapid (almost exponential) growth rate, whereas the application of machine learning in materials science has a lower starting point but relatively larger slope than the other two domains. A plateau also seems to appear at the end of machine learning and symbolic regression trend lines, while application of machine learning in materials science shows a steady growth trend towards recent years. The trend we find in machine learning and symbolic regression is a vivid example where research becomes saturated in some domains after decades of study and it becomes more challenging to keep the fast growth rate. However, it might be a golden era to apply these well-established methods to challenges in other research fields, such as materials science.

The trend of symbolic regression applications in materials science is not shown here since the base number is too small; nonetheless, it also reveals a potential research area that has been previously underappreciated. For materials science problems, one is often also presented with the problem of unknown relationships among many variables. Symbolic regression presents an opportunity then to help in the formulation of structure-property relationships derived from these variables. In this prospective paper, we encourage materials scientists and engineers to utilize symbolic regression techniques in solving their domain problems. To facilitate a better understanding of the utility and application of symbolic regression, we next introduce the genetic programming-based symbolic regression (GPSR) method and describe current research frontiers in symbolic regression. Next, we discuss several industrial applications of symbolic regression and propose potential applications in materials science. In addition, we present how GPSR can learn the Johnson-Mehl-Avrami-Kolmogorov (or Avrami) equation to describe recrystallization kinetics. Last, we conclude with some open challenges in materials research that may benefit from symbolic-regression methods.

II. SYMBOLIC REGRESSION AND CURRENT STATE-OF-THE-ART METHODS

A. Genetic programming-based symbolic regression (GPSR)

Symbolic regression is a method of finding a suitable mathematical model to describe observed data. In conventional regression techniques, one optimizes parameters for a particular model provided as a starting point to the algorithm. For instance, a linear regression model
is based on the assumption that the relationship of the dependent variables and regressor is linear; an artificial neural network (ANN) is a nonlinear model which relies on a predefined network infrastructure such as neuron connections and activation function (e.g. sigmoid, softmax function).

In symbolic regression, however, no such a-priori assumptions on the specific form of the function is required. Instead, one provides a mathematical expression space containing candidate function building blocks, e.g. mathematical operators, state variables, constants, analytic functions, and then symbolic regression searches through the space spanned by these primitive building blocks to find the most appropriate solution. In other words, both model structures and model parameters are optimized in symbolic regression. Since there is no need for a predefined function form, optimization algorithms used in symbolic regression are different from conventional analytical/numerical optimization methods (e.g. conjugate gradient, Newton-Raphson method). In this section, we briefly introduce one of the most prevalent methods used in symbolic regression by means of genetic programming.

Genetic programming (GP) was developed by J.R. Koza\(^\text{19}\) as a specific implementation of genetic algorithms (GA),\(^\text{20}\) which are often utilized in the materials community for atomic structure prediction.\(^\text{21-23}\) The idea is to evolve the solution of a given problem following Darwin’s theory of evolution and to find the fittest solution after a number of generations. Instead of using strings of binary digits to represent chromosomes as in GA, solutions in GP are represented as tree-structured chromosomes with nodes and terminals. Figure 2a shows a chromosome example of the mathematical function \(1 + \exp(-x_1)\).
the natural selection process then comes into play. The “fitness” of each individual solution in the initial population is then evaluated by comparing their function output with the true value from the data set. This fitness value describes how well the program performs in terms of solving the problem. The common error metrics used include mean squared error (MSE), root-mean squared error (RMSE), etc. Then GP evolves the current generation by randomly applying genetic operations to individuals, e.g. crossover and mutation. One or more individual from the current generation will be selected as parent(s) based on the fitness score, typically the higher the score, the larger probability to be selected for reproduction. Such a selection rule agrees with the “survival-of-the-fittest” rule since good features are more likely to be inherited by the next generation, which is the essential step towards the optimal solution.

The genetic crossover operation takes two winners of the selection process as parents to breed their offspring. For instance, the two structures in Figure 2a and b are taken as parents. The crossover operator then randomly takes a subtree from parent (b) and substitutes another random subtree in parent (a) with that from (b). One possible offspring from the crossover operator is illustrated in Figure 2c. Crossover is usually the dominant operation in the recombination process. Figure 2d is an example of an offspring from the mutation operation. The mutation operator only takes one parent structure, and randomly substitutes a subtree with another randomly generated structure; in case (d), the constant 1 is mutated to \((x_7 - 0.5)\). Although this operation is more aggressive compared to the crossover operation, since it adds randomness to the system, it is important to have a finite chance of mutation to introduce new variations, e.g. new constants and new features, and avoid being trapped in local minima.

The third category of genetic operations is reproduction, which duplicates the selected program and directly inserts its offspring to the next generation. It guarantees that some of the current generation will be preserved by the next generation, and partially protects the similarity between two generations. Detailed definitions and implementations of each genetic operation can vary from case to case, but the main features should be the similar to what we described here.

The newborns are then added to the next generation after each genetic operation, until the new population size reaches the specified set number. Then the new generation goes through the fitness evaluation and natural selection process again until the fitness value reaches a certain criteria or the maximum number of generations is reached. After termination of GP, the surviving individuals are expected to be highly evolved to adapt to the problem-dependent selection rule. More comprehensive descriptions of GP can be found in Koza’s original paper.\(^{19}\)

B. Advances in symbolic regression

Since Koza introduced the idea of GP in 1992, there have been significant efforts made to improve the performance of the original GPSR algorithm. The major problems to overcome in GPSR include:

1. **Non-deterministic optimization.** It is not guaranteed that the performance of the descendent generations will be better than their parents.
2. **Difficulty in finding the proper constants.** Since the way GP generates constants is random, either in the initial input set or those brought into the population by mutations, there is no effective way to obtain the ideal coefficients as in other numerical regression methods.
3. **Limited capability to preserve good components of the equation** due to the fitness evaluation method. The fitness is evaluated based on the complete structure of an individual. Having a good feature in a subbranch does not necessarily lead to better individual performance, thus good equation components may get lost in the next generation.

We summarize some of the most popular alternative methods to conventional GPSR in Table I and discuss their similarities as well as the differences in four aspects, namely program representation, fitness evaluation, optimization method, and the solution form.

Multiple regression genetic programming (MRGP)\(^{24}\) improves the program evaluation process by performing multiple regression on subexpressions of the solution functions. Instead of evaluating the fitness of each individual solution as a whole, MRGP decouples its mathematical expression tree into subtrees. The fitness of the solution is evaluated based on the best linear combination of these subtree structures. A least angle regression (LARS) algorithm is used to solve the linear regression problem here. Such a fitness evaluation scheme places more emphasis on finding good components even though it might only be a partial solution. For instance, the individuals that contain a correct form of a subtree structure of the correct solution (if known) are more likely to survive the natural selection process and pass these good features to the descendents. MRGP essentially decouples the current basis functions to find the best solution in an enlarged space at the vicinity of the original GP space. Indeed, this feature may be well-suited for multi-scale materials problems where modeling of systems across different length/time-scale is desired.\(^{36}\) While some subexpressions capture relationship among variables within each scale, the final symbolic regression solution assembles models across the scale and returns the multi-scale model.

Geometric semantic genetic programming (GSGP)\(^{26}\) evaluates the semantic performance of a computer program instead of the syntax performance as in conventional GPSR. While still using rooted-tree structure to represent
TABLE I. Comparison of genetic programming-based symbolic regression (GPSR) and its alternative methods: multiple regression genetic programming (MRGP), geometric semantic genetic programming (GSGP), Cartesian genetic programming (CGP), genetic programming-based relevance vector machine (GP-RVM), evolutionary polynomial regression (EPR), and fast function extraction (FFX).

| Program representation | Fitness evaluation | Optimization | Solution form |
|------------------------|--------------------|--------------|---------------|
| GPSR                   | rooted-tree        | individual program | genetic evolution | rooted-tree |
| MRGP                   | rooted-tree        | subexpressions | genetic evolution + linear regression | linear combination of subexpressions |
| GSGP                   | rooted-tree        | distance in semantic space | genetic evolution in semantic space | rooted-tree/semantic vector |
| CGP                    | acyclic graph      | individual program | genetic evolution | 2D grid of nodes |
| GP-RVM                 | rooted-tree        | group of GP individual | genetic evolution + RVM | linear combination of GP individuals |
| EPR                    | vector of integers | individual program | genetic evolution + linear regression | polynomial function |
| FFX                    | basis functions    | individual program | pathwise regularized learning | linear combination of basis functions |

computer programs, GSGP focuses on its semantics, i.e. the behavior of a program. Each program has a semantic vector that corresponds to a point in the high-dimensional semantic space. Since the target point is also known in this space, fitness evaluation is rather straightforward as one only needs to measure the distance of the program from the target point in semantic space. Geometric semantic operators are used to perform genetic operators. The offspring of geometric semantic crossover lies between its parents in semantic space, thus it should be no worse than the poor-performing parent. Optimizing program semantics rather than syntax further frees symbolic regression from specific function forms. For instance, \( add(x_1, x_1) \) is equivalent to \( mul(2, x_1) \) in semantic space, but quite different in terms of syntax. It is reasonable to care more about the behavior of the program than what the function form appears to be.

Cartesian genetic programming (CGP)\(^{27}\) has a more sophisticated design than conventional GP. Here, a computer program is represented as a directed acyclic graph, which may be visualized as a two-dimensional grid of nodes. Each node owns a set of genes that determines the input-output and mathematical function that the node performs; the whole set of genes of the computer programs form its genotype. Decoding the genotype leads to the phenotypic function, i.e. the function form of the computer programs. The genotype-phenotype mapping is a unique feature of CGP which makes it closer to the real natural process.

GP-RVM\(^{28}\) is an alternative GP method that combines Kaisen programming and a relevance vector machine (RVM) algorithm to solve symbolic regression problems. Kaisen programming (KP) is a collaborative version of genetic programming, where individuals work together with each other to solve the problem. The solution of a Kaisen process is a linear combination of GP individuals, and thus the fitness evaluation is based on a group of individual partial solutions instead of an individual program as a complete solution. RVM is a Bayesian kernel method that could extract important basis functions from the basis set without the prior knowledge to set a threshold and automatically deals with singularity. GP-RVM leverages advantages from both evolutionary algorithm and Bayesian kernel methods: the former mainly explores the parameter space while the latter extracts basis functions to build and solve for the optimal solution function within that space.

Evolutionary polynomial regression (EPR)\(^{29}\) hybridizes the parameter estimation used in conventional numerical regression methods with the evolutionary optimization scheme in GPSR. EPR first explores the function space using genetic algorithms, then performs linear regression (e.g. least squares) to optimize the coefficients of each mathematical building block. Although EPR specifically uses polynomial expansions for the form of the functions, the solution is not necessarily a simple polynomial function since the transformed variables used in the polynomial expansion could be nonlinear functions of independent input variables. Such a hybrid method improves the stochastic GPSR method moving it towards a more deterministic approach although the computational cost may be relatively higher. In fact, the polynomial form of the expressions could make EPR suitable for materials design or multiobjective optimization purposes. Since the analytical gradient and Hessian of the solution can be evaluated, materials scientists may have more insights regarding the system and know what parameters to tune in order to achieve optimal design.

Fast function extraction (FFX)\(^{30}\) is an efficient way to find good basis functions and solve for the best solution within the space it spans. The first step in FFX is to generate a large number of candidate basis functions built from input variables and other predefined variables. The evolutionary optimization scheme is not involved in FFX,
instead, a pathwise regularized learning technique is used to identify the best coefficients and basis functions for the solution. Then, models obtained from the previous step are assessed based on the validation data set as well as their model complexity in order to identify the best solution. FFX is more efficient compared to other GP-based methods due to the deterministic optimization technique. Materials scientists could first use FFX to see whether the input function/variable basis is sufficient for their research problem, before further investigation using symbolic regression methods (either FFX or other variants).

The performance of some of the recently developed symbolic regression techniques has been assessed against popular machine learning methods, and it is reported that symbolic regression performs considerably well compared to state of the art ML algorithms with regards to predictive accuracy. However, the two methods do not simply exist in competition to one another. We also observe a trend of more hybridization between conventional ML algorithms and genetic programming in symbolic regression solvers. These advances have enabled symbolic regression to be used for solving real-world problems, which we will discuss in the following section.

III. APPLICATIONS OF SYMBOLIC REGRESSION

Common motivations underlying the use of GPSR for complex problem solving include when the system in question is not effectively modelled by a linear model. Existing multiple linear regression models are much faster and are already easy to interpret. GPSR is best used for systems with complex interactions between observable variables for which the form of which is not known beforehand—a situation common in materials science and engineering.

In addition, a GPSR approach could be useful for design optimization purposes. Although the evolutionary search process is a black box, the final solution is analytical which potentially contains important information (e.g. gradient, Hessian) about relationships between the design variables and objectives. There is also need for multiobjective optimization such as finding the Pareto optimal combination of model performance and complexity in various domains, where symbolic regression technique has shown to be effective and interpretable. We next describe some applications of symbolic regression in various science and technology domains.

A. Industrial applications

GPSR has been applied to a wide variety of problems in fields outside of materials science and chemistry. Most prominently featured in the popular press was work published by Schmidt and Lipson in Science, which showed GPSR could discover Hamiltonians and Lagrangians for systems of simple harmonic oscillators and double pendulums. Reports of using GPSR for real world systems, however, have been published since Koza’s origination of the idea in the early 1990s and continue today.

Arkov et al. used GPSR to identify equations governing gas turbine engines under multiple optimization conditions. Berardi et al. used GPSR to find easy to interpret models for pipe failures in a UK water distribution system. Bongard and Lipson applied GPSR to generate symbolic equations for nonlinear coupled dynamical systems in mechanics, ecology, and systems biology. The authors also emphasized that their symbolic models are easier to interpret than numerical models, which makes understanding more complex systems easier for future applications.

Cai et al. identified correlation equations from experimental heat transfer measurements using GPSR with a sparsifying constraint. The authors’ predicted correlations had lower percentage error than models developed graphically and numerically, albeit with more formula complexity than those traditional methods. Can and Heavey applied GPSR to develop metamodels for predicting throughput rates in industrial serial production lines. McKay, Willis and Barton developed steady-state models for a vacuum distillation column and a chemical reactor.

La Cava et al. applied GPSR to identify nonlinear governing equations of wind turbines. The Pareto front from their paper is reproduced in Figure 4. The Pareto front illustrates the trade-off between their model complexity as defined by the number and type of operations in the equation and the normalized variance in the prediction.
error. La Cava and other authors also tested modifying standard GPSR with features from epigenetics, such as passive structure, phenotypic plasticity, and inheritable gene regulation. These researchers demonstrated their modifications improved the performance over standard GPSR by finding compact dynamic equations for synthetic data from nonlinear ordinary differential equations as well as real-world systems, e.g. cascaded tanks, a chemical distillation tower, and an industrial wind turbine. GPSR has also been applied to testing the efficient market hypothesis, formulating the synchronization control in oscillator networks, identifying the structure of helicopter engine dynamics, real-time runoff forecasting in France and Singapore, designing circuits, predicting solar power production, finding dynamical equations for metabolic networks, in both cases where a starting model was known and from scratch, modelling global temperature changes, and synthesizing second-order coefficient insensitive digital filter structures.

The existing uses of GPSR within chemistry are more extensive than that for materials science. We refer the reader to the review by Vyas, Goel, and Tambe for further details. Some key studies with relevance to materials science are summarized here. Langdon and Barrett developed a model for oral bioavailability of a small molecule given a few hundred data points from expensive experiments. Their model based on chemical descriptors showed promise for rapid drug screening, but had difficulty generalizing to novel molecules. Vyas et al. also discovered structure-property-relationships for drug absorption using GPSR. Here, they demonstrated $R^2$ values comparable to those achieved with artificial neural networks and support vector regression. Barnapalexis et al. performed a multiobjective optimization using GPSR. They found a function mapping levels of 4 polymers to three different properties of a pharmaceutical release tablet that was more predictive than a shallow neural network. Last, Muzny, Huber, and Kazakov built a correlation model for the viscosity of hydrogen as a function of temperature and pressure.

**B. Opportunities in materials science**

Materials science has many potential areas where GPSR can be applied for the same reasons it find use in other disciplines. Nonlinear systems are abundant in materials science. Changes in materials properties occurring in response to structural, composition, and other external perturbations are frequently nonlinear in proximity to phase transitions or for large stimuli. For instance, changing the concentration of oxygen vacancies in a transition metal compound by an atomic percent can alter its ionic or electronic conductivity by orders of magnitude. The dynamical behavior of materials performance as a function of time is also of broad interest and technological importance, e.g. corrosion of nickel cathodes under different conditions. These are the areas where a dynamical multivariable model would be ideal to understand the correlation among the variables and assist optimization of materials properties, e.g. corrosion resistance.

As we mentioned earlier, materials properties and performance are affected by phenomena that involve multiple length scales. Most theoretical models are formulated to be optimal at a particular length scale. However, recent emphasis has been placed on multiscale and hierarchical modeling in the materials science community, and there is an increased need for effective, descriptive and predicative, multiscale models. Symbolic regression techniques are potential solutions to this challenge by directly searching for the interactions among variables operating and passing between multiple spatial and temporal scales. Another possible approach is to utilize existing simulation methods within each length scale, while using symbolic regression to find the suitable coupling interactions between scales, i.e. connecting models of different scales.

Other applications of GPSR in materials and molecular systems are in areas where supervised machine learning has already demonstrated usefulness in providing new insight or solutions. While machine learning has produced many impressive results, it is commonly understood that ML models exhibit a trade-off between performance on prediction metrics and the ability to explain the predictions of a model due to the complexity of state-of-the-art models like deep neural networks or gradient boosted decision trees. GPSR offers a middle ground with comparable performance but with the added ability to read and directly interpret the output function.

GPSR is also well-suited to the development of new descriptors for materials properties. By combining features in a manner best suited to fitting data, new features are created that can be used as proxies for the property in question. This is also a common application of compressed sensing. Compressed sensing differs from GPSR in that while GPSR uses GP to iteratively evolve a solution, compressed sensing tries to enumerate as many combinations of primary features as possible and then use sparsifying operators to find a small dimensional subset that correlates with the target.

Materials scientists are not just interested in making predictions. They also want to identify the controlling features of a property and what role each feature plays; they want to understand which degrees-of-freedom to design or optimize to achieve targeted properties. Given some desired properties as objectives and the relevant variables, there are a number of numerical algorithms available for performing optimization, but they typically require some a-priori knowledge of the mathematical relationships within the system. The symbolic equations derived from GPSR offer insight into which microscopic or macroscopic knobs to turn for the design of desired functionality, such as corrosion resistance in steels.

Some novel ideas include targeting physical variables for which we do not know the proper mathematical expressions. For instance, the exchange-correlation functional used in density functional theory, or the correlation func-
tion for the viscosity of normal hydrogen.\textsuperscript{57} Contraindi-
cated material property pairs such as ferromagnetism and ferroelectricity or optical transparency and electrical conductivity would be interesting areas to pursue in search for routes to decouple or circumvent perceived coexistence incompatibilities. In these cases, GPSR could provide candidate representations of these terms, where our physical knowledge can be used to filter meaningful results. One advantage of using symbolic regression is that the balance between model accuracy and complexity is tunable with GPSR settings, e.g. stopping criteria, penalty on individual size, etc. This advantage is particularly evident when designing with constraints or optimizing for performance. The recommendations coming from the learned symbolic equations are more actionable than learned functions only optimized for test set accuracy since they are rigorously made to use fewer terms.

Frequently, materials scientists look for relationships \((f)\) among multiple variables with the aim to find some closed-form expression such as \(y = f(X)\), where \(y\) is the objective value and \(X\) are a set of variables. Equations in physical science usually contain conservation laws, e.g. energy conservation in the Schrödinger equation or conservation of momentum from Newton’s first law. It has been shown that symbolic regression has the potential to find implicit relationships, despite it still being technically challenging.\textsuperscript{59} The ability to unearth conservation laws with symbolic regression goes beyond the aim of materials property predictions and helps researchers establish insights of the materials systems they study. In fact, we do not necessarily need a rigorous expression of natural laws in every case; sometimes an approximation with a simple yet effective expression serves well for the research purpose.\textsuperscript{70} Symbolic regression could potentially balance the trade-off between model accuracy and simplicity, and might even help scientists discover new equations that redefine our understanding of functional materials in the same way those of Hall and Petch and Harper and Dorn changed our understanding of metal mechanical properties or as more recently Berry phases and topological band theory changed our understanding of electronic structures.

C. Use Case: Discovering the Avrami equation

We now show how to use genetic programming-based symbolic regression to learn the Johnson-Mehl-Avrami-Kolmogorov (JMAK), hereafter, Avrami equation. The Avrami equation quantitatively describes the growth kinetics of phases in materials at constant temperature. Here, we specifically study the recrystallization process of copper, the original experimental data was obtained from Ref. \textsuperscript{71}. We expect the form of the function to be

\[
y = 1 - \exp(-kt^n),
\]

where the phase fraction of transformation \(y\) is a function of time \(t\). The coefficients \(k\) and \(n\) are unknown and change with respect to temperature and other environmental conditions.

We use GPSR as implemented in \texttt{gplearn}\textsuperscript{72} to predict the relationship between fraction transformed \(y\) and time \(t\). The hyper-parameters used in GPSR are listed in Table II. When the population size is divided by the tournament factor, one obtains the number of individuals competing for reproduction each round. The parsimony coefficient regularizes the size of individuals by penalizing over-sized structures. We include operations of addition, subtraction, multiplication, negation, and the natural exponential function into the function set. The power function is not included since it easily causes numerical overflow or invalid operations (e.g. power\((-1,0.5))\). In general, this can be an issue when evaluating power-law dependent phenomena such as electrical transport equations. Working with log transforms of the original variables may be a more stable approach. Crossover operations dominate the genetic operations with a 70\% probability to be applied; the other 30\% chance corresponds to mutation operations (e.g. point mutation, subtree mutation, etc.). Additional details concerning the usage of \texttt{gplearn} are given in its documentation.

Data preprocessing is an essential step in machine learning before feeding data into the solver. Conventional preprocessing methods include shifting data to be zero-centered, and scaling data to unit standard deviation. However, the conventional preprocessing methods are not ideal choices in our case. Zero-centered shifting is not applicable to either the phase fraction transformed \((y)\) or time \((t)\) since we want to obtain the exact function form of the Avrami equation. Furthermore, scaling the time frame would only change the constant \(k\) in the Avrami equation. Here, we scale time from 0 to 10 for all data sets so that the constant \(t\) lies in the range \([-1, 1]\). The \(y\) values remain unchanged as the experimental data range over \([0, 1]\).

We take the experimental copper recrystallization data at temperatures 135\(^\circ\)C, 113\(^\circ\)C, and 102\(^\circ\)C as input and scale the time variable before performing regression. Since the experimental data only contains several points at each temperature (less than 10), we also take data points from interpolated lines and perform symbolic regression on the

| Parameter          | Values                          |
|--------------------|---------------------------------|
| population size    | \{2000, 5000\}                  |
| tournament factor  | \{100, 500\}                    |
| parsimony coefficient | \{0.001, 0.005\}               |
| max generation     | 20                              |
| constant range     | \((-1, 1)\)                     |
| function set       | \{add, sub, mul, neg, exp\}     |
| crossover probability | 0.7                          |
TABLE III. Functions predicted by GPSR and numerical fitting results for the Avrami equation describing copper recrystallization.

| Dataset       | Numerical fitting result | GPSR result | GPSR (transformed y) result |
|---------------|--------------------------|-------------|-----------------------------|
| ideal         | $y = 1 - \exp(-0.6t^2)$  | $y = 0.994 - \exp(-0.58t^2)$ | $1 - y = \exp(-0.58t^2)$  |
| 135°C experimental | $y = 1 - \exp(-0.019t^{1.2})$ | $y = \exp[-\exp(-t + 2.64)]$ | $1 - y = \exp(-0.024t^{1.2})$  |
| 135°C interpolated | $y = 1 - \exp(-0.17t^{3.6})$ | $y = \exp[-\exp(-t + 2.18)]$ | $1 - y = \exp(-0.065t^{3.6}(t - 0.851))$  |
| 113°C experimental | $y = 1 - \exp(-0.018t^{2.8})$ | $y = 0.986 - \exp(-0.051t^2)$ | $1 - y = \exp(-0.014t^{2.8})$  |
| 113°C interpolated | $y = 1 - \exp(-0.042t^{2.5})$ | $y = 0.986 - \exp[-0.108t \times (t - 0.934)]$ | $1 - y = \exp(-0.097(t - 0.764))$  |
| 102°C experimental | $y = 1 - \exp(-0.0054t^{2.9})$ | $y = 0.1t - 0.081$ | $1 - y = \exp(-0.0045t^{2.9})$  |
| 102°C interpolated | $y = 1 - \exp(-0.01t^{3.1})$ | $y = \exp[-(1.11 + 2t) \times \exp(-t + 1.42)]$ | $1 - y = \exp(-0.01t^{3.1})$  |

FIG. 5. GPSR prediction and performance with different data sets. Left panels correspond to the direct experimental data while right panels use interpolated experimental data (a) 135°C experimental data, (b) 135°C extrapolated data, (c) 113°C experimental data, (d) 113°C extrapolated data, (e) 102°C experimental data, and (f) 102°C extrapolated data.

interpolated data for comparison. An ideal dataset is also generated, where $y$ values are directly calculated as $1 - \exp(-0.6t^2)$.

The best individual after 20 generations of evolution is collected for each hyper-parameter setting. Finally, we manually pick the optimal individual with closest function form and constants to the Avrami equation within each data set. Our parameter fitting and GPSR evolution results are shown in Table III and Figure 5.

**Gplearn** successfully recovers the relationship between time ($t$) and fraction transformed ($y$) in most cases. With the ideal input data, **Gplearn** evolves almost a perfect form of the Avrami function as well as the constants (see the first row of Table III). The performance on the raw experimental and interpolated data are generally worse than the ideal case but still reasonable. In most cases the exponential function form and polynomial function of $t$ are both recovered. One source of error is the lack of the power function in function set, which was intentionally omitted as non-integer powers of variables cannot be correctly represented here. Rather, polynomial functions are used as an approximation. With the limited choice of mathematical functions, GP produces results that are very close semantically but exhibit different syntax.

We see in multiple cases GPSR produces functions of the form $\exp[-\exp(\Theta)]$, where the expected function form is $1 - \exp(\Theta)$. Here we introduce a mathematical trick to show their mathematical equivalence. Since

$$\lim_{x \to \infty} e^x \sim 1 + x,$$

$e^x$ and $1 + x$ are called equivalent infinitesimals; thus, $\lim_{x \to \infty} e^{-x} \sim 1 - x$ is also true when $x$ is small. In our case, the exponent $\exp(\Theta)$ quickly reaches zero when it becomes larger; therefore, the equivalent infinitesimal relationship holds and the predicted function form is equivalent to the expected one appearing in the Avrami expression. Based on this example, we encourage professional materials researchers, who are novice data scientists, to perform careful analysis of GPSR results when making the final interpretation of the obtained model.

In real-world applications where the actual function form in unknown, the exact syntax of GPSR results may not matter. There are potentially many cases where GPSR would produce semantically similar/equivalent functions that vary in their syntax, i.e. different function form. Considering the trade-off between model accuracy and complexity, in some cases it might be a virtue to find a simple approximated solution instead of using rigorous but complex relationships among multiple variables.

The GPSR result from the 102°C experimental data set (Figure 5e), turns out to be a linear relationship and deviates significantly from the original data. The poor result may be caused by the relatively small slope of the original data and the insufficient number of available data points. Taking the model performance and complexity into consideration, we find that the linear relationship survives due to its simple form. With interpolated data as input, functions evolved from GPSR agree better with the Avrami equation, as shown in Figure 5f.

Interestingly, the performance of GPSR can also be very close semantically but exhibit different syntax.
Symbolic regression has shown competitive performance to other machine learning-based regression models in various research domains. While there are some shortcomings of the current state-of-the-art GPSR, e.g. high computational cost, non-deterministic optimization, there are numerous active research efforts focusing on improving the performance of symbolic regression to expand its use in real-world applications. The ability of symbolic regression to distill natural laws from data sets with high-dimensional parameter space makes it an ideal technique for materials science research, since these researchers typically face sparse data sets with multiple variables. Freed from having a fixed form of equations, symbolic regression can potentially reveal the significant interactions among physical variables. We recommend more materials scientists utilize symbolic regression techniques in their own research domain.

**ACKNOWLEDGMENTS**

Y.W. acknowledges partial support from the Predictive Science and Engineering Design (PS&ED) program at Northwestern University. All authors acknowledge support from the National Science Foundation (NSF) through the Designing Materials to Revolutionize and Engineer our Future (DMREF) program under award number DMR-1729303.
identification of wind turbine models using evolutionary multiobjective optimization,” Renewable Energy 87, 892–902 (2016).

Shu-Heng Chen and Chia-Hsuan Yeh, “Toward a computable approach to the efficient market hypothesis: An application of genetic programming,” Journal of Economic Dynamics and Control 21, 1043–1063 (1997).

Gary J. Gray, David J. Murray-Smith, Yun Li, Ken C. Sharman, and Thomas Weinbrenner, “Nonlinear model structure identification using genetic programming,” Control Engineering Practice 6, 1341–1352 (1998).

Soon Thiam Khu, Shie Yui Liong, Vladan Babovic, Henrik Madsen, and Nitin Muttil, “Genetic programming and its application in real-time runoff forecasting,” Journal of the American Water Resources Association 37, 439–451 (2001).

Shie-Yui Liong, Tirtha Raj Gautam, Soon Thiam Khu, Vladan Babovic, Maarten Keijzer, and Nitin Muttil, “GENETIC PROGRAMMING: A NEW PARADIGM IN RAINFALL RUNOFF MODELING,” Journal of the American Water Resources Association 38, 705–718 (2002).

Markus Quade, Markus Abel, Kamran Shafi, Robert K. Niven, and Bernd R. Noack, “Prediction of dynamical systems by symbolic regression,” Physical Review E 94, 012214 (2016), arXiv:1602.04648.

Michael D. Schmidt, Ravishankar R. Vallabhaivasula, Jerry W. Jenkins, Jonathan E. Hood, Abhishek S. Soni, John P. Wilkso, and Hod Lipson, “Automated refinement and inference of analytical models for metabolic networks,” Physical Biology 8, 055011 (2011).

Karolina Stanislawksa, Krzysztof Krawiec, and Zbigniew W. Kundzewicz, “Modeling global temperature changes with genetic programming,” Computers & Mathematics with Applications 64, 3717–3728 (2012).

Kazuyoshi Uesaka and Masayuki Kawamata, “Synthesis of low-sensitivity second-order digital filters using genetic programming with automatically defined functions,” IEEE Signal Processing Letters 7, 83–85 (2000).

Renu Vyas, Purva Goel, and Sanjeev S. Tambe, “Genetic programming applications in chemical sciences and engineering,” in Handbook of Genetic Programming Applications (Springer International Publishing, Cham, 2015) pp. 99–140.

W B Langdon and S J Barrett, “Genetic Programming in Data Mining for Drug Discovery,” in Evolutionary Computation in Data Mining, Vol. 163 (Springer-Verlag, Berlin/Heidelberg, 2005) pp. 211–235.

Renu Vyas, Purva Goel, M. Karthikeyan, S.S. Tambe, and B.D. Kulkarni, “Pharmacokinetic Modeling of Caco-2 Cell Permeability Using Genetic Programming (GP) Method,” Letters in Drug Design & Discovery 11, 1112–1118 (2014).

P. Barmpalexis, K. Kachrimanis, A. Tsakonas, and E. Georgrakis, “Symbolic regression via genetic programming in the optimization of a controlled release pharmaceutical formulation,” Chemometrics and Intelligent Laboratory Systems 107, 75–82 (2011).

Chris D. Muzny, Marcia L. Huber, and Andrei F. Kazakov, “Correlation for the Viscosity of Normal Hydrogen Obtained from Symbolic Regression,” Journal of Chemical & Engineering Data 58, 969–979 (2013).

Alexey A Markov, Mikhail V Patrakiev, Vladislav V Khariton, Yevheniy V Pivak, Ilya A Leonidov, and Viktor L Kozhevnikov, “Oxygen Nonstoichiometry and Ionic Conductivity of Sr 3 Fe 2 - x Sc x O 7 - δ,” Chemistry of Materials 19, 3980–3987 (2007).

Akio Nakamura and J Bruce Wagner, Defect Structure, Ionic Conductivity, and Diffusion in Yttria Stabilized Zirconia and Related Oxide Electrolytes with Fluorite Structure, Tech. Rep.

L Daiza, C M Rangel, J Baranda, M T Casais, M. J. Martinez, and J A Alonso, “Modified nickel oxides as cathode materials for MCFC,” Journal of Power Sources 86, 329–333 (2000).

Modeling Across Scales: A Roadmapping Study for Connecting Materials Models and Simulations Across Length and Time Scales, Tech. Rep. (2015).

Aref Yadollahi, Nima Shamsaei, Scott M. Thompson, and Denver W. Seely, “Effects of process time interval and heat treatment on the mechanical and microstructural properties of direct laser deposited 316L stainless steel,” Materials Science and Engineering A 644, 171–183 (2015).

Logan Ward and Chris Wolverton, “Atomistic calculations and materials informatics: A review,” Curr. Opin. Solid State Mater. Sci. 21, 167–176 (2017).

Luca M. Ghiringhelli, Jan Vyibral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler, “Big data of materials science: Critical role of the descriptor,” Physical Review Letters 114 (2015), 10.1103/PhysRevLett.114.105503, arXiv:1411.7437.

Luca M Ghiringhelli, Jan Vyibral, Emre Ahmetcik, Runhai Ouyang, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler, and Luca M Ghiringhelli, “SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates,” Physical Review Materials 2, 083802 (2018), arXiv:1710.03319.

Garret N Vanderplaats, Numerical optimization techniques for engineering design (Vanderplaats Research and Development, Incorporated, 2001).

M Shimada, H Kokawa, Z J Wang, Y S Sato, and I Karibe, “Optimization of grain boundary character distribution for intergranular corrosion resistant 304 stainless steel by twin-induced grain boundary engineering,” in Acta Materialia, Vol. 50 (2002) pp. 2331–2341.

Michael Schmidt and Hod Lipson, “Symbolic regression of implicit equations,” Genetic Programming Theory and Practice 7, 73–85 (2009).

Ulf von Barth and Lars Hedin, “A local exchange-correlation potential for the spin polarized case. i,” Journal of Physics C: Solid State Physics 5, 1629 (1972).

B. F. Decker and D. Harker, “Activation energy for recrystallization in rolled copper,” JOM 2, 887–890 (1950).

Trevor Stephens, “gplearn,” (2016).