

Data-driven discovery of formulas by symbolic regression

Sheng Sun, Runhai Ouyang, Bochao Zhang, and Tong-Yi Zhang

Discovering knowledge from data is a quantum jump from quantity to quality, which is the characteristic and the spirit of the development of science. Symbolic regression (SR) is playing a greater role in the discovery of knowledge from data, specifically in this era of exponential data growth, because SRs are able to discover mathematical formulas from data. These formulas may provide scientifically meaningful models, especially when combined with domain knowledge. This article provides an overview of SR applications in the field of materials science and engineering. Integrating domain knowledge with SR is the key and a crucial approach, which allows gaining knowledge from data quickly, accurately, and scientifically. In the data-driven paradigm, SR allows for uncovering the underlying mechanisms of materials behavior, properties, and functions, in a wide range of areas from basic academic research to industrial applications, including experiments and computations, by providing explicit interpretable models from data, in comparison with other machine-learning "black-box" models. SR will be a powerful tool for rational and automatic materials development.

Introduction

Theories and models in materials science and engineering are usually represented by mathematical formulas, which play vital roles in the understanding of materials behavior and performance as well as in the development of advanced materials and technologies. Before machine learning (ML) became available, theories and models were often, and even now continue to be, developed through intuition, derivation, or summarization from the accumulation of experience and knowledge of materials behavior and performance. This model-developing paradigm, however, requires long times and huge effort, as well as the genius of humans.

The rapid development of artificial intelligence (AI) techniques today provides the opportunity for automatic model construction. The new paradigm of data-driven science using AI techniques such as ML to discover knowledge and science from data generated from both experiments and computational simulations has rapidly expanded in the field of materials science and engineering. This paradigm combined with expert domain knowledge yields state-of-the-art methodologies for model development.[1]

Langley developed an AI program in 1979 with a few simple heuristics to solve a broad range of tasks (e.g., successful rediscovery of the ideal gas law, Kepler's Third Law, Coulomb's Law, Ohm's Law, and Galileo's Law).[2] These models are among the simplest ones. A scientist's dream is to find natural laws in explicit analytic expressions automatically from data. In 2009, Schmidt and Lipson showed that it was possible to distill underlying physical laws of two conservative systems, air-track oscillators and double pendulums, by employing SR based on genetic programming (GP).[3] An air-track oscillator has fine holes in its surface from which a layer of air is pumped to minimize the friction experienced by a car or several cars gliding on the track. Every car is connected to its nearest neighboring cars or the track end by springs. In these two experiments, they recorded the position of cars on the air track and the position of the pendulums over time by using motion-tracking software and then calculated the velocities and accelerations of the oscillators. They successfully rediscovered the underlying equations describing the designed systems: Hamiltonian or Lagrangian equations were recovered if coordinates and velocities were used as data, and the equation of motion was recovered if coordinates, velocities, and accelerations were all used as data. Besides SR, feature (or descriptor) selection using a ML technique termed the least absolute shrinkage and

Sheng Sun, Materials Genome Institute, Shanghai University, China; mgissh@t.shu.edu.cn
Runhai Ouyang, Materials Genome Institute, Shanghai University, China; rouyang@shu.edu.cn
Bochao Zhang, Materials Genome Institute, Shanghai University, China; mgizbc@shu.edu.cn
Tong-Yi Zhang, Materials Genome Institute, Shanghai University, China; zhangty@shu.edu.cn
doi:10.1557/mrs.2019.156

© 2019 Materials Research Society

https://doi.org/10.1557/mrs.2019.156 Published online by Cambridge University Press

selection operator (LASSO), where features are scientifically meaningful properties or mathematic operators and the number of features can be reduced to zero by allowing the total magnitude of the summation over the absolute coefficients in a linear least squares fitting to be smaller than an adjustable parameter, was also shown to have the ability to rediscover physical laws from data numerically generated from targeted equations.[4,5] Similarly, Ouyang et al. developed a data-driven method called "sure independence screening and sparsifying operator" (SISSO) for identifying interpretable physical descriptors for materials properties.[6,7]

Materials informatics is an emerging and rapidly developing area, particularly after the launch of Materials Genome Initiative (MGI) in 2011 (mgi.gov). Ramakrishna et al.[8] defined materials informatics as that which "employs techniques, tools, and theories drawn from the emerging fields such as data science, internet, computer science and engineering, and digital technologies to the materials science and engineering to accelerate materials, products and manufacturing innovations." Agrawal and Choudhary suggested that materials informatics has become the "fourth paradigm" in materials science.[9] Following the guidance of MGI, many free-access and open-to-public materials databases have been built up and materials data are exponentially growing due to the development of high-throughput experiments and high-throughput calculations, as well as data sharing among the materials communities. See Reference 8 for a complete list of available materials databases.

AI and ML-based models can be either explicit or inexplicit. An inexplicit AI and ML-based model function is a "blackbox" (e.g., neural network), which usually involves many flexible mathematical functions containing a number of parameters to fit data. Prediction accuracy is the main goal for such "blackbox" models, and the interpretation of the models is secondary. Although such "blackbox" models are widely applied, they are not the best choice if our goal is to pursue the understanding of underlying mechanisms, in addition to accuracy.

Data to knowledge necessitates models with simple, explicit mathematical expressions. SR is one of the most popular AI and ML methods to obtain explicit models from data for (exactly or approximately) describing target phenomena or mechanisms, and thus can be widely used by materials scientists and engineers to gain knowledge from data.[10]

SR has been applied in a wide range of application areas (e.g., in the field of fluid mechanics[11,12] and dynamical systems[13–17]). It is usually performed based on evolutionary computation (EC),[18] such as GP[19] and its variations, including grammar-guided genetic programming (GGGP),[20] grammatical evolution (GE)[21] and Cartesian genetic programming.[22] Inspired by and analogous to Darwin's theory of evolution by natural selection for excellent offspring, the evolution in GP is realized via iterative selection of optimal candidate solutions. Usually, a candidate solution is called an individual, and the population of individuals after each iteration is named a generation. A new generation is evolved from its parent generation by selecting excellent individuals and adding new individuals. In GP, an individual is expressed in the form of a tree where nonterminal nodes are mathematic operators and terminal nodes are variables or constants and called leaves. Populations are continuously iterated via the process of individual evaluation, selection, and breeding by the operations of crossover and mutation, and eventually evolve to optimal expressions. The general ideas of EC, and the details of the GP, GGGP, and GE methods are presented in References18–21 and 23.

There are many publications on SR in the literature. In the following section, we refer to two specific examples of progress in SR research. An overview on SR applications in materials modeling and simulations is also described in a subsequent section. Finally, concluding remarks are provided to highlight our opinions about materials informatics and AI- and ML-based materials discovery and optimization.

Showcase examples of SR applications

Sette and Boullart[24] illustrate how analytic formulas can be discovered via GP. As an example, we show how GP finds a proper mathematical expression to describe or approximate the simple trigonometric function of $g(x) = 25\sin(x/3.0)$. x is sampled as $\{0, 1, \ldots, 9\}$. If $f = 200 - \sum_{i=1}^{9} |g(x_i) - \mathrm{GP}(x_i)|$, then the task of GP is to find a formula $\mathrm{GP}(x)$ to minimize the summation term, or to let f approach 200. Here, the 200 is used because the summation will always yield values between 0 and 200.[24] The operators set $\{+, -, \times, \div\}$ and terminals (variables and constants) set $\{x, [0, \ldots, 9]\}$ were used for the construction of expression trees in GP, which was run for 30 generations, with an initial population size of 100. The number of generations and population size are preset parameters in a GP evolution, and their values depend on the complexity of the task, as well as the available computational resources. The evolved best function, considering the balance between fitting quality and formula complexity, is $-x^2 + \frac{349}{36}x$, which is a good approximation of $g(x) = 25\sin(x/3.0)$ within $x \in [0,9]$ and appears at the 20th generation with the fitted $\hat{f}$ value of 190 and a complexity of 19. Here, the complexity is defined as the number of terminals and operators used in the expression. The original trigonometric function was not discovered by GP because the sin operator is not included in the operator set. Clearly, the efficiency of GP and the performance of the created formulas depend critically on what operator sets in addition to the terminal sets are used.

Sette and Boullart[24] also explored the spinning process of yarn in the textile industry and predicted both spinnability and yarn strength deviating from the fiber quality and machine settings. GP was performed on a database of 2160 data points with five machine parameters and five

https://doi.org/10.1557/mrs.2019.156 Published online by Cambridge University Press

continuous fiber characteristics. The operator set included (1) {+, −, ×, ÷, $\sqrt{}$, cos, sin}; (2) a maximum/minimum function; and (3) a step function that groups parameters (0.5, 1.0) to be 1.0 and (0.0, 0.5) to be 0.0. GP evaluations for spinnability and yarn strength yield mathematical expressions that both have prediction accuracy greater than 90%. Partial original features, rather than all of them, appeared in the final expressions, thereby indicating the importance of feature selection.

The dimensional consistency of physical quantities must be maintained in SR for constructing meaningful equations. It was found that the introduction of dimensional analysis (DA) into neural networks could improve prediction performance.[25,26] GGGP provides an easy way by using grammar, which imposes a set of semantic constraints on the available forms of mathematical expressions, to incorporate DA, as well as domain knowledge.[27] Ratle and Sebag[28] argued that the integration of DA in grammar might make this more complex and thus lead to poor performance. They proposed that dedicated design of the population initialization step of GGGP might give admissible individuals (i.e., mathematical expressions) sufficient diversity and limited size, where every individual is dimensionally consistent and its size is also limited to preset maximum tree depth without losing diversity. To prove their arguments, they used the proposed approach to rediscover the Kelvin–Voigt model for two viscoelastic systems: (1) parallel connection of a dashpot η_1 and a spring K_1 (termed Sys1); and (2) series connection of a spring K_2, a dashpot η_2 and Sys1 (termed Sys2). They generated data by assigning random values to materials parameters in the designed model, representing 20 pseudo-materials. By comparing the results from both common grammar (DA-excluded) and DA-included grammar, they showed that integration of DA significantly accelerated the evolution task.

The proposed method was then applied to find an analytic model for force–displacement curves and force–time curves in indentation tests on various materials. Their results indicate that the DA-included algorithm was able to discover accurate expressions for each of the tested samples, whereas only one unified expression was found for all four tested samples by using the common algorithm, thereby demonstrating the significant advantage of using the DA-included algorithm. More examples of SR applications can be found in recent handbooks of GE[29] and GP.[30]

SR applications in materials modeling and simulations

Multiscale simulations of materials properties and phenomena are essential in materials research and design. However, the current widely used density functional theory in materials simulations experiences a large gap in time and length scales with respect to real materials. Advanced methods and strategies are urgently needed to bridge this gap. In this regard, trans-scale (crossing scales) simulations are being developed, where GP is utilized to facilitate multiscale and trans-scale simulations.[31,32] Sastry et al. used GP to symbolically regress mapping of saddle point barriers from only a few calculated points via molecular dynamics, thereby avoiding explicit calculation of all barriers.[31] The GP-regressed barrier function allows one to carry out kinetic Monte Carlo simulations with real-time kinetics based upon realistic atomic interactions. Sastry et al. GP-regressed diffusion energy barriers of a surface atom to its nearby vacancy on the (100) surface of the fcc binary alloy Cu_xCo_{1-x} and predicted the diffusion barriers within 0.1% error by using 3% or less of the total number of barriers with this approach, leading to a dramatic reduction in CPU time on the order of 4~7.

Sastry et al. also conducted SR to discover a constitutive model that describes the relationship between flow stress σ and temperature-dependent strain rate $\dot{\epsilon}$ from experimental data for Al-Zn-Mg-Cu aluminum alloy AA7055, a commercially available aluminum alloy with high strength.[32] The used operator set is $\{+, -, \times, \div, \exp, \sin\}$, and the terminal set is $\{\dot{\epsilon}, T, \exp[1/(RT)], R\}$, where R is the ideal gas constant, T is temperature, and R is a random number. The GP, based on experimental data of compression tests at temperatures ranging from 340°C to 520°C at a strain rate of 10^{-3} s^{-1}, yielded the model:

$$\dot{\epsilon} = c_0 \exp\left(\frac{1}{RT}\right)\left(\frac{\sigma}{\mu}\right)^{4.55},$$

where μ is viscosity and c_0 is a constant. The formula is nearly the same as that proposed by Padilla et al.,[33] who used the following law:

$$\dot{\epsilon} = c_0 \exp\left(\frac{Q}{RT}\right)\left(\frac{\sigma}{\mu}\right)^{4.56},$$

where Q is the activation energy. Furthermore, GP conducted with both low (10^{-3} s^{-1}) and high strain rate (1 s^{-1}) data gave a unified expression:

$$\dot{\epsilon} = \frac{c_0 \exp(1/RT)}{g[\dot{\epsilon}, \exp(1/RT)]}\left(\frac{\sigma}{\mu}\right)^4\left(1-\frac{\sigma}{\mu}\right),$$

with g being a complex expression.[32] This model might imply competition between the fourth-order power law and the fifth-order power law in stress.[32]

The interatomic potential plays a central role in MD and Monte Carlo simulations. Traditionally, potentials are mainly developed based on experience and are thus also called empirical potentials. Recently, data-driven construction of interatomic potentials has become a focus in chemistry and materials science,[34] and the commonly used ML methods include artificial neural networks,[35–39] ridge regression,[40] Gaussian approximation potential,[41] and Gaussian process regression.[42,43] GP is also used to construct interatomic potentials.[44–48]

In general, data-driven potentials work well within the sampling region, whereas extrapolating to "unseen" (without training data) areas is a challenge. On the other hand, empirical

https://doi.org/10.1557/mrs.2019.156 Published online by Cambridge University Press

potentials might be widely applicable but usually suffer from insufficient accuracy. These two aspects may complement each other, and their combination could result in potentials with both improved prediction accuracy and transferability.

Hernandez et al.[48] summarized the frequently appearing function forms in many empirical potentials, such as Lennard-Jones potential, Morse potential, and many-body potentials, and put them into the function set as expert domain knowledge. First, they tested GP to see if it could construct the pairwise Lennard-Jones potential and the many-body Sutton–Chen embedded atomic potential with the training data, which were generated from MD simulations with the previously mentioned potentials under various simulation ensembles and temperatures. Indeed, GP successfully recovered the original potentials, though the parameters were slightly different. After the validation, they constructed GP copper potentials, considering complexity, computational cost, and fitting quality, from data produced from *ab initio* MD simulations under different ensembles or temperatures.[48] Two GP potentials were applied in MD simulations to predict the elastic constants C_{11}, C_{22}, and C_{44} of fcc copper. Their results showed that the two GP potentials have accuracy comparable to those of the established accurate empirical potentials, while the GP potentials bear the advantages of low complexity.[48]

In finite element (FE) analysis, modeling nonuniform, nonlinear, and inelastic materials and structures is challenging. Automatic routes and techniques are required for constructing constitutive models from experimental data directly and applying them to FE analysis. GP is powerful in solving this challenging problem, especially when combining with domain knowledge. Javadi and Reezania[49] proposed an intelligent FE method, in which the constitutive equation of $y = \sum_{j=1}^{n} F(\mathbf{x}, f(\mathbf{x}), a_j) + a_0$ is derived by evolutionary polynomial regression (EPR) from experimental data, where y is the predictor output, $\mathbf{x}$ are input variables, f is some user-provided function, and a_j and a_0 are constants. The function form of F is constructed by evolutionary selection of function blocks provided by the user based on their understandings of the physical behaviors. The authors performed EPR on a data set that included 537 items from triaxial tests of a soil and finally obtained the variation of the next state of deviator stress ($^{i+1}\sigma_d$) as a function of the current one ($^{i}\sigma_d$), as well as the confining pressure (σ_3), the current axial strain ($^{i}\varepsilon_1$), and the axial strain increment ($^{i}(\Delta\varepsilon_1)$). For each load increment, the instant modulus was calculated using

$$E_{\mathrm{EPR}} = \frac{^{i+1}(\sigma_{d1}) - {}^{i}(\sigma_{d1})}{^{i+1}(\varepsilon_1) - {}^{i}(\varepsilon_1)}$$

and substituted into the constitutive equation of $\Delta\sigma = \mathbf{D}\Delta\varepsilon$, where $\mathbf{D}$ is the material stiffness matrix and updated after E_{EPR} was recalculated in each step. The FE analysis in conjunction with EPR gave a reliable result, as compared to the traditional elastic–plastic model and nonlinear elastic model. Recently, they extended the EPR method by employing a multi-objective genetic algorithm to reach a tradeoff result between fitness and function complexity,[50] which means maximization of fitness and minimization of model complexity.

Gandomi[51,52] employed multi-objective GP, as well as multi-gene genetic programming (MGGP), to find models of concrete strength and creep. For MGGP, they used linear combinations of several GP trees to give a prediction, rather than using only one tree in standard GP. They named a single GP tree as a gene. The MGGP is mathematically expressed by:

$$\hat{y}(\mathbf{x}, \mathbf{d}) = d_0 + \sum_{i=1}^{n} d_i G_i(\mathbf{x}),$$

where $G_i(\mathbf{x})$ is the output of each GP tree (or each gene), $\mathbf{d} = \{d_0, d_1, \cdots, d_n\}$ is the gene weight vector, n is the number of the gene, and $\mathbf{x}$ is the vector of the input variables. $\mathbf{d}$ is evaluated by using least squares fitting of the expression to data. Each single GP tree is randomly generated and several GP trees are randomly picked and linearly combined to generate an initial population. Besides crossover and mutation operations in one GP tree as in standard GP, the whole GP trees in MGGP between two expressions are permitted to crossover operations. Their methods presented more accurate predictions for total compliance of concrete than current available models using various statistical indicators.

Versino et al.[53] conducted GP on the deformation behavior of metallic copper to find the expression of strength (represented by flow stress) versus plastic strain (ε^p), plastic strain rate ($\dot{\varepsilon}^p$), and temperature (T): $\sigma = f(\varepsilon^p, \dot{\varepsilon}^p, T)$. Four GP models were derived either by using standard GP or GP subjected to expert knowledge constraints. Their data were extracted from the literature. The four GP models as well as two well-established strength models, the Preston–Tonks–Wallace model[54] and the mechanical threshold stress model,[55] were compared by performing a virtual experiment with a classical large-deformation and high strain-rate test—Taylor anvil impact on a copper cylinder under different temperatures from 295 K to 1235 K and different initial velocities from 48.7 to 394 m/s. Comparing the deformed profile and the final length of the simulated cylinder with those of the experimental one indicates the excellent performance of the GP models.[53]

Concluding remarks

Recent advancements in data-driven methodology, together with fast-growing materials data, have enabled automatic extraction of knowledge from data and accelerated novel materials discovery, which is becoming a new paradigm in materials research and development. Among the cutting-edge data-driven methods, SR provides regression results in analytic formulas or equations, thereby delivering scientific insights. SR is widely applicable across all areas of materials science and engineering to find explicit formulas, hypotheses, and laws for targeted materials properties and functions.

In particular, when SR is coherently combined with domain knowledge, scientifically meaningful models can be obtained

https://doi.org/10.1557/mrs.2019.156 Published online by Cambridge University Press

with much less computation power and time. What distinguishes SR from other ML "blackbox" models lies in the analytic outputs, therefore, the significance of integrating domain knowledge into SR has become especially important. In addition to the fitting quality, the low complexity of outcome formulas enable the formulas to be more robust against data noise. When an anticipated formula has only a few coefficients, SR might not require big data for training. Considering the flexibility of incorporating domain knowledge and the large family of genetic programming algorithms, SR is expected to play increasing roles and has broader applications in data-driven knowledge discovery.

Great challenges remain with SR when domain knowledge is not available or insufficient, where SR outcome models may not be physically meaningful and even hard to interpret. One should keep in mind that material data are not perfect; the data may contain noise, and usually, there aren't enough data points. Since the SR outcome is assessed purely by fitting quality, the SR outcome might have a better fitting quality than the outcome directly from domain knowledge. Under this circumstance, which should we trust, the SR outcome or the outcome directly from domain knowledge? Therefore, it is important in the future to consider how the interpretability of SR outcome models can be improved. At the moment, most terminals in SR are represented by scalar quantities, as overviewed in the present article, and few publications with vector and tensor terminals can be found in the literature. Thus, the development of SR algorithms with mathematical operations on high dimensional terminals might demand more attention from the materials information community.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2017YFB0702101, 2017YFB0701604 and 2018YFB0704400). S.S. and B.Z. also acknowledge the National Natural Science Foundation of China (No. 11672168).

References

1. E.B. Goldstein, G. Coco, *Front. Environ. Sci.* **3**, 1 (2015).
2. P. Langley, "Rediscovering Physics with BACON.3," *Proc. 6th Int. Jt. Conf. Artif. Intell.—Vol. 1* (Morgan Kaufmann Publishers, 1979), pp. 505–507, http://dl.acm.org/citation.cfm?id=1624861.1624976.
3. M. Schmidt, H. Lipson, *Science* **324**, 81 (2009).
4. H. Schaeffer, S.G. McCalla, *Phys. Rev. E* **96**, 023302 (2017).
5. S.H. Rudy, S.L. Brunton, J.L. Proctor, J.N. Kutz, *Sci. Adv.* **3**, e1602614 (2017).
6. L.M. Ghiringhelli, J. Vybiral, E. Ahmetcik, R. Ouyang, S.V. Levchenko, C. Draxl, M. Scheffler, *New J. Phys.* **19**, 023017 (2017).
7. R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L.M. Ghiringhelli, *Phys. Rev. Mater.* **2**, 083802 (2018).
8. S. Ramakrishna, T.-Y. Zhang, W.-C. Lu, Q. Qian, J.S.C. Low, J.H.R. Yune, D.Z.L. Tan, S. Bressan, S. Sanvito, S.R. Kalidindi, *J. Intell. Manuf.* (2018).
9. A. Agrawal, A. Choudhary, *APL Mater.* **4**, 053208 (2016).
10. Y. Wang, N. Wagner, J.M. Rondinelli, "Symbolic Regression in Materials Science," submitted arXiv:1901.04136 (2019), http://arxiv.org/abs/1901.04136 (accessed March 26, 2019).
11. P. Praks, D. Brkić, *Water* **10**, 1175 (2018).
12. E.B. Goldstein, G. Coco, A.B. Murray, M.O. Green, *Earth Surf. Dyn.* **2**, 67 (2014).
13. M.P. Hinchliffe, M.J. Willis, *Comput. Chem. Eng.* **27**, 1841 (2003).
14. D.L. Ly, H. Lipson, *J. Mach. Learn. Res.* **13**, 3585 (2012).
15. T.W. Cornforth, H. Lipson, *Genet. Program. Evolvable Mach.* **14**, 155 (2013).
16. J. Gout, M. Quade, K. Shafi, R.K. Niven, M. Abel, *Nonlinear Dyn.* **91**, 1001 (2018).
17. M. Quade, M. Abel, K. Shafi, R.K. Niven, B.R. Noack, *Phys. Rev. E* **94**, 012214 (2016).
18. K.A. De Jong, *Evolutionary Computation: A Unified Approach* (MIT Press, Cambridge, MA, 2006).
19. J.R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (MIT Press, Cambridge, MA, 1992).
20. R.I. McKay, N.X. Hoai, P.A. Whigham, Y. Shan, M. O'Neill, *Genet. Program. Evolvable Mach.* **11**, 365 (2010).
21. C. Ryan, J.J. Collins, M.O. Neill, "Grammatical Evolution: Evolving Programs for an Arbitrary Language," *Eur. Conf. Genet. Program.* (Springer, 1998), pp. 83–96.
22. J. Miller, Ed., *Cartesian Genetic Programming* (Springer, Heidelberg, Germany, 2011).
23. R. Poli, W.B. Langdon, N.F. McPhee, J.R. Koza, *A Field Guide to Genetic Programming* (Lulu Press, Morrisville, NC, 2008).
24. S. Sette, L. Boullart, *Eng. Appl. Artif. Intell.* **14**, 727 (2001).
25. D.J. Gunaratnam, T. Degroff, J.S. Gero, *Appl. Soft Comput.* **2**, 283 (2003).
26. I.M. Jamadar, D.P. Vakharia, *Measurement* **94**, 177 (2016).
27. P.J. Angeline, K.E. Kinnear, "On Using Syntactic Constraints with Genetic Programming," in *Advances in Genetic Programming* (MIT Press, Cambridge, MA, 1996), https://ieeexplore.ieee.org/document/6277529 (accessed March 23, 2019).
28. A. Ratle, M. Sebag, *Appl. Soft Comput.* **1**, 105 (2001).
29. C. Ryan, M. O'Neill, J. Collins, Eds., *Handbook of Grammatical Evolution* (Springer International Publishing, Cham, Switzerland, 2018).
30. A.H. Gandomi, A.H. Alavi, C. Ryan, Eds., *Handbook of Genetic Programming Applications* (Springer International Publishing, Cham, 2015).
31. K. Sastry, D.D. Johnson, D.E. Goldberg, P. Bellon, *Phys. Rev. B* **72**, 085438 (2005).
32. K. Sastry, D.D. Johnson, D.E. Goldberg, P. Bellon, *Int. J. Multiscale Comput. Eng.* **2**, 239 (2004).
33. H.A. Padilla, S.F. Harnish, B.E. Gore, A.J. Beaudoin, J.A. Dantzig, I.M. Robertson, H. Weiland, "High Temperature Deformation and Hot Rolling of AA7055," *Metallurgical Modeling for Aluminum Alloys, Proc. Mater. Solutions Conf. 2001: 1st Int. Symp. Metall. Model. Alum. Alloys*, M. Tiryakioglu, L.A. Lalli, Eds. (ASM International, Materials Park, OH, 2003), pp. 1–8.
34. J. Behler, *J. Chem. Phys.* **145**, 170901 (2016).
35. W. Li, Y. Ando, E. Minamitani, S. Watanabe, *J. Chem. Phys.* **147**, 214106 (2017).
36. S.K. Natarajan, J. Behler, *Phys. Chem. Chem. Phys.* **18**, 28704 (2016).
37. P. Wang, Y. Shao, H. Wang, W. Yang, *Extreme Mech. Lett.* **24**, 1 (2018).
38. J. Behler, *Int. J. Quantum Chem.* **115**, 1032 (2015).
39. P.E. Dolgirev, I.A. Kruglov, A.R. Oganov, *AIP Adv.* **6**, 085318 (2016).
40. F. Fracchia, G. Del Frate, G. Mancini, W. Rocchia, V. Barone, *J. Chem. Theory Comput.* **14**, 255 (2017).
41. V.L. Deringer, G. Csányi, *Phys. Rev. B* **95**, 094203 (2017).
42. A. Glielmo, P. Sollich, A. De Vita, *Phys. Rev. B* **95**, 214302 (2017).
43. Z. Li, J.R. Kermode, A. De Vita, *Phys. Rev. Lett.* **114**, 096405 (2015).
44. W.M. Brown, A.P. Thompson, P.A. Schultz, *J. Chem. Phys.* **132**, 024108 (2010).
45. A. Kenoufi, K.T. Kholmurodov, *Biol. Chem. Res.* **2**, 1 (2015).
46. D.E. Makarov, H. Metiu, *J. Chem. Phys.* **108**, 590 (1998).
47. A. Slepoy, M.D. Peters, A.P. Thompson, *J. Comput. Chem.* **28**, 2465 (2007).
48. A. Hernandez, A. Balasubramanian, F. Yuan, S. Mason, T. Mueller, "Fast, Accurate, and Transferable Many-Body Interatomic Potentials by Genetic Programming," submitted arXiv:1904.01095 (2019), http://arxiv.org/abs/1904.01095 (accessed April 10, 2019).
49. A.A. Javadi, M. Rezania, *Adv. Eng. Inform.* **23**, 442 (2009).
50. A. Faramarzi, A.M. Alani, A.A. Javadi, *Comput. Struct.* **137**, 63 (2014).
51. A.H. Gandomi, S. Sajedi, B. Kiani, Q. Huang, *Autom. Constr.* **70**, 89 (2016).
52. A.H. Gandomi, A.H. Alavi, *Neural Comput. Appl.* **21**, 171 (2012).
53. D. Versino, A. Tonda, C.A. Bronkhorst, *Comput. Methods Appl. Mech. Eng.* **318**, 981 (2017).
54. D.L. Preston, D.L. Tonks, D.C. Wallace, *J. Appl. Phys.* **93**, 211 (2003).
55. P.S. Follansbee, U.F. Kocks, *Acta Metall.* **36**, 81 (1988). □

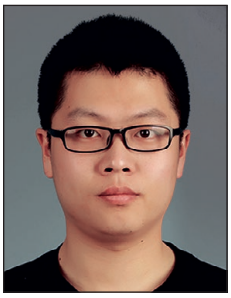


Sheng Sun is an associate professor at the Materials Genome Institute at Shanghai University, China. He received his BS degree in aircraft design and his MS degree in nanomechanics from the Nanjing University of Aeronautics and Astronautics, China, in 2005 and 2008, respectively. He earned his PhD degree in bioengineering from The Hong Kong University of Science and Technology in 2011. His current research focuses on materials informatics by combining multiscale simulations, mathematical modeling, and machine learning. Sun can be reached by email at mgissh@t.shu.edu.cn.

https://doi.org/10.1557/mrs.2019.156 Published online by Cambridge University Press



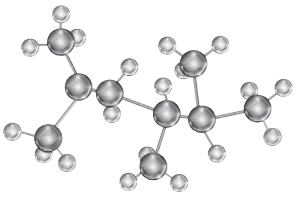
Runhai Ouyang joined Shanghai University, China, in 2019 as an associate professor at the Materials Genome Institute. He obtained his PhD degree in physical chemistry in 2014 from the Dalian Institute of Chemical Physics, Chinese Academy of Sciences, China. He conducted postdoctoral research at The University of Sydney, Australia, at the University of California, Riverside, and at the Fritz Haber Institute of the Max Planck Society, Germany, He was a co-creator of the SISSO (sure independence screening and sparsifying operator) method. Ouyang can be reached by email at rouyang@shu.edu.cn.



Bochao Zhang is pursuing his MS degree at the Materials Genome Institute at Shanghai University, China. He received his BS degree in mining engineering from the China University of Mining and Technology, China, in 2017. His research focuses on high-throughput materials modeling of electrode/electrolyte interfaces. Zhang can be reached by email at mgizbc@shu.edu.cn.



Tong-Yi Zhang is the founding dean of the Materials Genome Institute at Shanghai University, China. He and was previously at The Hong Kong University of Science and Technology. He is a member of the Chinese Academy of Sciences, China, and the Hong Kong Academy of Engineering Sciences, Hong Kong, and a Fellow of the Hong Kong Institution of Engineers, ASM International, and the International Fracture Congress. His awards include two Second Prizes of the State Natural Science Award of China and the Ho Leung Ho Lee Prize for Scientific and Technological Progress. His current research focuses on materials informatics and mechanoinformatics. Zhang can be reached by email at zhangty@shu.edu.cn.



IWGO 2019

3RD INTERNATIONAL WORKSHOP ON GALLIUM OXIDE AND RELATED MATERIALS

AUGUST 12–15, 2019 • COLUMBUS, OHIO, USA • THE OHIO STATE UNIVERSITY

PREREGISTER BY JULY 29 AND SAVE!

The 3rd International Workshop on Gallium Oxide and Related Materials (IWGO-3) is the premier international forum for high impact research on gallium oxide and related materials. The 2019 Workshop, scheduled for August 12–15 on the campus of The Ohio State University (OSU) in Columbus, Ohio, will be held in a single-session format to allow for maximum interaction between researchers of various technical disciplines. It will also include a special session for the U.S. Workshop on Gallium Oxide (GOX) highlighting U.S.-funded basic research efforts. Mark your calendar today and plan to attend.

SCIENTIFIC PROGRAM

- Theory, Modeling and Simulation
- Bulk Growth
- Epitaxial Growth
- Electronic Transport and Breakdown Phenomena
- Dielectric Interfaces
- Advanced Characterization Techniques
- Electronic and Photonic Devices and Applications
- Material and Device Processing and Fabrication Techniques
- Heterogeneous Material Integration

For the most up-to-date information on IWGO-3, visit MRS.ORG/IWGO-3.

GENERAL CO-CHAIRS

Debdeep Jena
Cornell University

James Speck
University of California, Santa Barbara

TECHNICAL PROGRAM CO-CHAIRS

Gregg Jessen
Air Force Research Laboratory

Siddharth Rajan
The Ohio State University

LOCAL ARRANGEMENTS CHAIR

Steve Ringel
The Ohio State University

PUBLICITY CHAIR

Andrew Green
Air Force Research Laboratory

IWGO-3 is managed by MRS CONFERENCE SERVICES Because the Experience Matters
mrs.org/conference-services



https://doi.org/10.1557/mrs.2019.156 Published online by Cambridge University Press