Engineering 6 (2020) 609–611

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Views & Comments

On the Data-Driven Materials Innovation Infrastructure

Hong Wang [a], X.-D. Xiang [b], Lanting Zhang [a]

[a] Materials Genome Initiative Center & School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] Department of Materials Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China

In recent years, material genome has been a hot topic in the field of material science. The emergence of the term "material genome" was largely inspired by the successful Human Genome Project. Traditionally, the discovery and development of new materials and new processes depend on scientific intuition and a lengthy trial-and-error process. For years, material scientists have been longing to find some sort of basic building blocks whose structure and defects may determine the properties of materials, similar to the genome in the field of biology. By understanding these building blocks, they hope to be able to design materials on demand, so as to accelerate discovery and development, and reduce costs. Since the launch of Material Genome Initiative [1,2] in the United States in 2011, other major economies such as the European Union [3,4], Japan [5], and China have set up similar scientific programs at the national level. However, despite a wide range attempts, it has been difficult to define what the "material genome" really is. The current consensus is that the term "material genome engineering" (MGE) is used only as a proxy for predictive material research and development [6]. By integrating high-throughput experimentation, high-throughput computation, and material informatics, the relationships between composition, process, structure, and performance—which form the basis of material design—can be established in a faster, more efficient, and less costly fashion.

The working modes of MGE can be roughly classified as experiment-driven, computation-driven, and data-driven, respectively [7]. The experiment-driven mode is based on high-throughput synthesis and characterization experiments, such as combinatorial material chip technology [8], which enables rapid screening or optimization of materials. The computation-driven mode is to predict materials by computational simulation [9], which greatly reduces the scope of promising candidates for quicker experimental verification. The data-driven mode is to build models using a materials informatics approach—that is, by applying artificial intelligence (AI)-based methods such as machine learning to a large amount of data in order to predict candidate materials [10]. For thousands of years in human history of seeking the truth of nature, science has gone through the paradigm shift from experimental observation, to theoretical deduction, then to computational simulation. Today, taking advantage of unprecedented computing power and the large-scale collection of data, modern science is currently entering the "fourth paradigm" [11], which features intensive data + AI. Therefore, the data-driven mode of MGE is an embodiment of the "fourth paradigm."

It should be noted that the essence of the experiment- and computation-driven modes is either factual judgment or the deduction of known physical laws, neither of which fundamentally changes the current way of thinking. In contrast, the data-driven mode uses AI to reveal the relationships embedded in massive data. This approach adds new dimensions and perspectives to the existing routine of material science research. Clearly such "new tool in the box" is expected to produce subversive results. That being said, it is also important to point out that the data-driven mode will by no means supersede the experiment- and computation-driven modes. Rather, it is a powerful supplement and extension of the traditional cognitive paradigm. In addition, domain knowledge should be carried over into the machine learning regime to provide guidance for and validity to the AI-based models.

Sufficient material data is the basic prerequisite for the implementation of the data-driven mode. Although vast quantities of data have been collected in databases around the world, this is only a drop in the ocean when facing the diversity and complexity of material problems. A simple estimation [7] suggests that over 2 million material systems can be composed out of just four elements, which leads to a total of trillions of multidimensional data points at 1% of the data density. In fact, the full implementation of the data-driven materials research is hindered by a lack of material data. In this data era, the ability to quickly generate a large amount of material data has become essential.

In many aspects, the existing materials research infrastructure is designed and developed to fit the current needs. As a new material research routine, MGE requires a whole new infrastructure to be instituted, which should be designed as data-centric to cover from data generation to data utilization, and thus would comprise facilities of data, high-throughput experimentation, and high-throughput computation. A data facility would include a database in conformity with MGE concepts, a library of software modeling tools based on AI methods, and an integrated data platform with data collection, storage, processing, exchanging, sharing, and e-collaboration capabilities [12]. High-throughput experimental facilities and high-throughput computing facilities are effective ways to rapidly generate a large amount of data, while

https://doi.org/10.1016/j.eng.2020.04.004

simultaneously serving as a basis for the experiment-driven and computation-driven modes. As such, the three technical elements of MGE form a deeply fused and indispensable whole, and act in synergy.

In addition to that a large volume of data being needed, the data for MGE should be highly integrated, systematic, consistent, and comprehensive. Ideally, the data should be generated from a "Data Fab" (Fig. 1)—that is, a dedicated data-producing platform that can either be a facility in a centralized location or in the form of a group of virtually linked sites. An experimental "Data Fab" includes setups of systematic and high-throughput synthesis and multi-parameter characterization, which are capable of mass producing data in batches in a standardized manner, just like an industrial production line. It would be advantageous to build such a platform in conjunction with large-scale scientific facilities such as a synchrotron light source, neutron source, and so forth. A computational "Data Fab" can be a facility with a variety of high-throughput computing software and hardware, which is able to generate a large amount of comprehensive material data through batch computing. A "Data Fab" will bring about profound changes in data generation. First, comprehensive data will be intentionally mass produced with a broader goal, rather than being collected from scattered experiments or computation events with a very specific purpose. Second, the "Data Fab" shifts data generation from individual activities to organized efforts. Third, such organized efforts will transform the societal characteristics of data from private property to public resources. As a result, data quality, consistency, and comprehensiveness will be improved, data sharing will become simpler, and the total cost to society will be reduced.

At present, there are a list of databases based on high-throughput computing platforms, or computational "Data Fab," such as the Materials Project [13], Automatic Flow for materials discovery (AFLOW) [14], Open Quantum Materials Database (OQMD) [15], Novel Materials Discovery (NOMAD) [16], and MatCloud [17]. The High-Throughput Experimental Materials Database (HTEM DB) [18] is an open experimental database of inorganic materials synthesized by high-throughput thin-film technology, developed by the National Renewable Energy Laboratory (NREL), United States. It has the basic characteristics of an experimental "Data Fab." The computational and experimental "Data Fab" supported by the National Key Research and Development Program of China is currently underway.

Another important task of MGE is to cultivate an open and collaborative "big science" culture in materials research and development. In order to break down the barrier to data sharing among all researchers, Wilkinson et al. [19] and Mons et al. [20] put forward the findable, accessible, interoperable, reusable (FAIR) data principle for scientific data. Establishing appropriate data standards is an important aspect of ensuring that the data meets the FAIR principles. To this end, the recently released China Standards of Testing and Materials (CSTM) *General rule for materials genome engineering data* [21] is a first attempt to standardize the content of data (many more standards for specific data format still need to be established). Under the *General rule for materials genome engineering data*, data is divided into three classes: sample information, source data (unprocessed data), and processed data (data obtained by analyzing and processing of existing data). Each individual action event (i.e., sample preparation/characterization/calculation/data processing) is defined as a stand-alone entry unit, and is assigned an independent resource identification (such as digital object identifier (DOI) or an identification per Chinese national standard GB/T 32843–2016, etc.). Here, the sample can be either a real object fabricated by experimentation or a virtual object created by computation. Similarly, the source data can be a result of direct measurement, or can be generated by computation/simulation under given conditions. Each data entry should collect as complete a set of metadata related to the action as possible Fig. 1. Listing sample information as a class of data is a unique choice, the greatest advantage of which is to make the samples themselves a part of the resources conforming to FAIR principles, so that samples can be found, shared, and reused, like data.

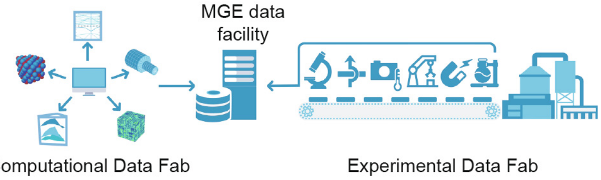

Fig. 1. Conceptualization of Data Fab—a dedicated facility capable of mass producing data in batches in a standardized manner, just like an industrial production line. An experimental Data Fab, as depicted on the right, contains setups of systematic and high-throughput synthesis and characterizations to generate multi-parameter data sets, including mechanical, electrical, optical, thermal, magnetic and acoustic properties and performances, etc. Ideally all property measurements are done on the same sample, preferably simultaneously, sometimes in-situ. A computational Data Fab, as depicted on the left, can be a computing center with a variety of high-throughput computation software and hardware, to generate a large amount of comprehensive data from atomic to macroscopic scales via various methods including Density Functional Theory, Molecular Dynamics, CALPHAD method, phase field simulation, Finite Element Analysis, etc. through batch computing. Data Fab can be a platform either in a centralized location or in the form of a group of virtually linked sites.

In summary, the data-driven mode of MGE brings about a new paradigm in material innovation that is fundamentally different from the current way of thinking and doing, and that needs to be supported by a corresponding whole new infrastructure. The infrastructure will include a data-centric integrated platform consisting of facilities of data, high-throughput experimentation, and high-throughput computation, in order to comprehensively cover data production, storage, analysis, sharing, and collaboration capabilities. Such a platform will generate and utilize an enormous amount of data conforming to the FAIR principles to facilitate the data-driven mode, while simultaneously serving as the basis for exercising the experiment- and computation-driven modes.

Acknowledgement

This work is supported in part by the National Key Research and Development Program of China (2018YFB0703600).

References

[1] Holdren JP. Materials genome initiative for global competitiveness [Internet]. Washington, DC: National Science and Technology Council; 2011 Jun [cited 2018 Mar 31]. Available from: https://www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf.

[2] MGI strategic plan [Internet]. Washington, DC: National Science and Technology Council; [cited 2018 Mar 31]. Available from: https://www.mgi.gov/sites/default/files/documents/mgi_strategic_plan_-_dec_2014.pdf.

[3] Accelerated metallurgy—the accelerated discovery of alloy formulations using combinatorial principles [Internet]. Luxembourg: The Community Research and Development Information Service; c1994–2019 [updated 2019 Aug 2; cited 2019 Mar 5]. Available from: https://cordis.europa.eu/project/id/263206.

[4] Jarvis D, Raabe D, Singer R, van Houtte P, Vahlas C, Alford N, editors. Metallurgy Europe: a renaissance programme for 2012–2022. Strasbourg: EFS; 2012.

[5] JST support program for starting up innovationhub, Information Integration, new scientific approach for materials research [Internet]. Tsukuba: National Institute for Materials Science; c2020 [cited 2019 Mar 5]. Available from: https://www.nims.go.jp/eng/research/MII-I/index.html.

[6] Wang H, Xiang Y, Xiang X, Chen L. Materials genome enables research and development revolution. Sci Technol Rev 2015;33(10):13–9.

[7] Wang H, Xiang X, Zhang L. Data + AI is the core of materials genomic engineering. Sci Technol Rev 2018;36(14):15–21.

[8] Xiang X, Sun X, Briceño G, Lou Y, Wang K, Chang H, et al. A combinatorial approach to materials discovery. Science 1995;268(5218):1738–40.

[9] Ceder G, Persson K. The stuff of dreams. Sci Am 2013;309(6):36–40.

[10] Raccuglia P, Elbert KC, Adler PD, Falk C, Wenny MB, Mollo A, et al. Machine-learning-assisted materials discovery using failed experiments. Nature 2016;533(7601):73–6.

[11] Hey T, Tansley S, Tolle KM, editors. The fourth paradigm: data-intensive scientific discovery. Redmond: Microsoft Research Press; 2009.

[12] Ward C, Brinson LC, Galli G, Kalidindi SR, MehtaA, Meredig B, et al. Building a materials data infrastructure: opening new pathways to discovery and innovation in science and engineering [Internet]. Pittsburgh: The Minerals, Metals & Materials Society; [cited 2019 Mar 5]. Available from: http://www.tms.org/Publications/Studies/Materials_Data_Infrastructure/Materials_Data_Infrastructure.aspx?hkey=d228f86c-e269-49a2-a638-395285b760e4.

[13] Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. The materials project: a materials genome approach to accelerating materials innovation. APL Mater 2013;1(1):011002.

[14] Curtarolo S, Setyawan W, Hart GLW, Jahnatek M, Chepulskii RV, Taylor RH, et al. AFLOW: an automatic framework for high-throughput materials discovery. Comput Mater Sci 2012;58:218–26.

[15] Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, et al. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. Comput Mater 2015;1(1):15010.

[16] Draxl C, Scheffler M. NOMAD: the FAIR concept for big data-driven materials science. MRS Bull 2018;43(9):676–82.

[17] MatCloud [Internet]. Beijing: Materials Genetic Engineering Information Technology Application Laboratory; c2016 [cited 2019 Mar 5]. Available from: http://matcloud.cnic.cn/static/view/about.html.

[18] Zakutayev A, Wunder N, Schwarting M, Perkins JD, White R, Munch K, et al. High throughput experimental materials database (2018) [Internet]. Denver: National Renewable Energy Laboratory; [cited 2019 Mar 5]. Available from: https://htem.nrel.gov/#/about/content.

[19] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3(1):160018.

[20] Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European open science cloud. Inf Serv Use 2017;37(1):49–56.

[21] China Standards of Testing and Materials (CSTM). T/CSTM 00120–2019: general rule for materials genome engineering data. Chinese standard. Beijing: Metallurgical Industry Press; 2019. Chinese.