Title

Research Update: The materials genome initiative: Data sharing and the impact of collaborative ab initio databases

Journal

APL Materials, 4(5)

Authors

Jain, A
Persson, KA
Ceder, G

Publication Date

2016-05-01

DOI

10.1063/1.4944683

Peer reviewed

Research Update: The materials genome initiative: Data sharing and the impact of collaborative ab initio databases

Anubhav Jain, , Kristin A. Persson, and , and Gerbrand Ceder

Citation: APL Materials **4**, 053102 (2016); doi: 10.1063/1.4944683
View online: http://dx.doi.org/10.1063/1.4944683
View Table of Contents: http://aip.scitation.org/toc/apm/4/5
Published by the American Institute of Physics

Research Update: The materials genome initiative: Data sharing and the impact of collaborative *ab initio* databases

Anubhav Jain,[1,a] Kristin A. Persson,[1,2] and Gerbrand Ceder[1,2]
[1]*Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[2]*Department of Materials Science and Engineering, University of California, Berkeley, Berkeley, California 94720, USA*

(Received 21 January 2016; accepted 8 March 2016; published online 24 March 2016)

Materials innovations enable new technological capabilities and drive major societal advancements but have historically required long and costly development cycles. The Materials Genome Initiative (MGI) aims to greatly reduce this time and cost. In this paper, we focus on data reuse in the MGI and, in particular, discuss the impact of three different computational databases based on density functional theory methods to the research community. We also discuss and provide recommendations on technical aspects of data reuse, outline remaining fundamental challenges, and present an outlook on the future of MGI's vision of data sharing.
[http://dx.doi.org/10.1063/1.4944683]

I. INTRODUCTION

Materials innovations are critical to technological progress. Many authors have noted that advanced materials are so vital to society that they serve as eponyms for historical periods such as the Stone Age, the Bronze Age, the Steel Age, the Age of Plastic, and the Silicon Age.[1–3] A recent study by Magee suggests that materials innovation has driven two-thirds of today's advancements in computation (in terms of calculations per second per dollar) and has similarly transformed other technologies such as information transport and energy storage.[4] However, while the benefits of new materials and processes are well established, the difficulties in achieving these breakthroughs and translating them to the commercial market[5] are not widely appreciated. It takes approximately 20 years to commercialize new materials technologies[6] and often even longer to develop those technologies in the first place. The long time scale of materials innovation stifles investment in early stage research because payback is unlikely.[6] Thus, it is natural to wonder whether it is possible to catalyze materials design such that decades of research and development can occur in the span of years.

One contributing factor that stunts materials development is lack of information. For newly hypothesized compounds, the answer to several important questions must typically be guided only by intuition. Can the material be synthesized? What are its electronic properties? What defects are likely to form? The dearth of materials property information can cause researchers to focus on the wrong compounds or persist needlessly in optimizing of materials that, even under ideal conditions, would not be able to meet performance specifications. $LiFePO_4$—a successful lithium ion battery cathode material—serves as a good example of how overlooked opportunities stem from a lack of information. Its crystal structure was first reported[7] in 1938, but $LiFePO_4$'s application to lithium ion batteries was discovered only in 1997 by Padhi *et al.*[8] If the compound's lithium insertion voltage and its exceptional lithium mobility were known earlier, it is likely that $LiFePO_4$'s application to batteries would have arrived one or two decades earlier. If its phase diagram had been previously charted, the synthesis conditions under which this material is made could have been more rapidly optimized.

[a]Author to whom correspondence should be addressed. Electronic mail: ajain@lbl.gov

Today, it is possible to compute many such properties from first-principles using a technique called density functional theory (DFT), which solves the electronic structure of a material by approximating solutions to Schrödinger's equation.[9] New capabilities developed in the last one or two decades make it possible to reimagine the path to materials discovery. As one example of a disruptive change, a single computational materials science research group today might have access to a level of computing power equivalent to personal ownership of the world's top supercomputer from 15 years ago or the sum of the top 500 supercomputers from 20 years ago.[10] New theoretical methods that more accurately predict fundamental materials properties[11,12] have made it possible, in many well-documented cases, to design materials properties in a computer before demonstrating their functionality in the laboratory.[13,14] Similar advancements in experimental capabilities, including combinatorial materials synthesis,[15–17] have made it possible to collect data at unprecedented rates and with much greater fidelity.[18–21] In almost all instances, the situation remains far from ideal. For example, material properties computable with DFT are typically not directly equivalent to the final engineering property of interest. However, several case studies have already shown that some forms of computational materials engineering can save tens of millions of dollars and result in returns on investment of up to 7:1 and with shorter design cycles.[5,22,23]

The Materials Genome Initiative (MGI) recognizes that these advancements, if nurtured, can lead to a discontinuous shift in the time needed to discover and optimize new materials. Its intention is to enable "discovery, development, manufacturing, and deployment of advanced materials at least twice as fast as possible today, at a fraction of the cost."[24] The MGI, which receives major funding from the Department of Energy, Department of Defense, the National Science Foundation, and the National Institute of Standards and Technology, encompasses various individual projects and larger research centers to advance this vision.[25] The four pillars upon which the MGI aims to achieve its goals are (i) to lead a culture shift in materials research that encourages multi-institute collaboration, (ii) to integrate experiment, computation, and theory across length scales and development cycle, (iii) to make digital data accessible and reusable, and (iv) to train a world-class materials workforce.[24]

In this manuscript, we focus on the topic of data reuse, noting that many of the other MGI pillars have been discussed elsewhere.[2,5,26–29] As illustrated in Figure 1, data sharing can drastically shorten the materials research cycle by (i) reducing the burden of data collection for individual research groups and (ii) enabling more efficient development of scientific hypotheses and property prediction models. In this manuscript, we focus on a form of data sharing that is still in its early stages: the use of DFT databases by both theory and experimental groups towards a variety of materials design applications.

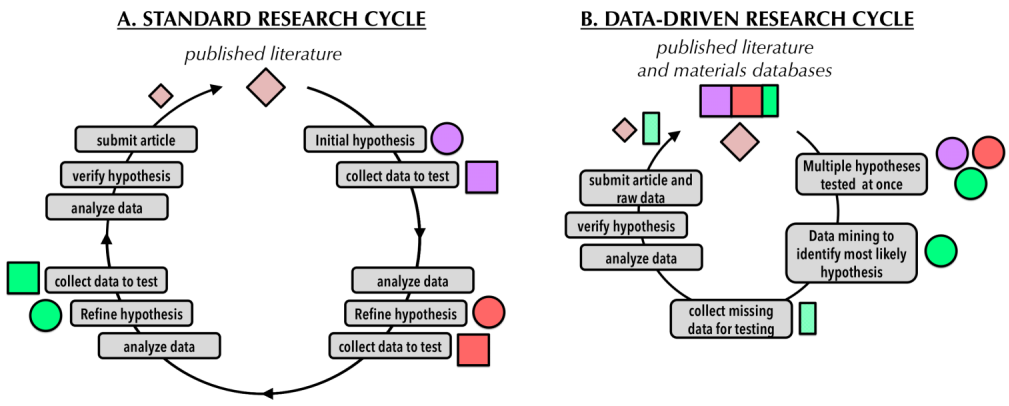

FIG. 1. Schematic to illustrate differences between standard, single-group research (left) and new opportunities afforded by large materials databases (right). The standard research cycle is considerably longer because testing and refining hypotheses typically involves several time-consuming data collection steps. In contrast, the availability of large data sets in the data-driven approach enables multiple hypotheses to be tested simultaneously through novel informatics-based approaches, reducing the burden of data collection.

II. EXAMPLES OF DATA REUSE

Material scientists have been sharing data for several decades. For example, several longstanding crystal structure databases have recently been reviewed by Glasser[30] and their impact on materials science has been reported by Le Page.[31] Experimental researchers routinely rely on diffraction pattern databases such as the Powder Diffraction File from the International Center for Diffraction Data[32] for phase identification of their samples. Thermodynamic databases[33–35] and handbooks[36,37] have been used to build and refine highly successful models such as the Calphad method[27,38,39] for generating phase diagrams. Thus, data sharing in materials science certainly predates the MGI.

What is novel under the MGI and related programs is (i) the scale at which data sharing is emphasized and (ii) the rapid expansion of theory-driven databases. Data sharing is heavily encouraged and enforced,[40–42] and new experimental databases such as the Structural Materials Data Demonstration Project, Materials Data Facility, and Materials Atlas are being funded. Large Department of Energy research hubs are incorporating combinatorial screening,[43] high-throughput computational screening,[44,45] and data mining[46] into the discovery process. New resources are rapidly coming online: the authors knew of no database of *ab initio* calculations that existed a decade ago, yet a recent review by Lin[47] lists 13 such databases that exist today (many of which contain hundreds of thousands of data points).

Three databases that have received MGI support are the Materials Project (MP),[48] AFLOWlib,[49] and Open Quantum Materials Database (OQMD).[50] The MP is centered at Lawrence Berkeley National Laboratory. Distinguishing features include its large and growing user base (currently over 17 000 registered users), its emphasis on "apps" for exploring the data, and database of elastic[51] and piezoelectric[52] tensor properties. The AFLOWLib consortium is centered at Duke University. A major feature of this resource is its large number of total compounds and its applications towards alloys, spintronics, and scintillation. The OQMD is centered at Northwestern University. Distinguishing features include comprehensive calculations for popular structure types (e.g., Heusler and perovskite compounds) and a focus on new compound discovery.

Unsurprisingly, all three of these databases have been used extensively by their respective development teams. However, although these resources are each only a few years old, there is already an emerging body of literature demonstrating their usage. Next, we present some early examples of research using DFT database resources conducted *without* the involvement of the development team.

A. Applications of computational databases to theoretical studies

A first application of computational databases is to aid and inspire further computational studies. Indeed, database projects such as ESTEST[53] and NoMaD[54] are geared especially to aid in reproducibility, verification, and validation of computational results. Parameters documented and tested by one project[55–57] can, in some cases, develop into a semi-standard methodology for the community. For example, the Hubbard interaction term, pseudopotentials, and reference state energies employed by the MP for DFT calculations are often re-used by the community.[58–64]

The atomistic modeling community has similarly established databases of interatomic potentials such as OpenKIM[1,65] and NIST Interatomic Potentials Repository Project[66] that can be recycled for new applications. Such resources can be complementary, and information from DFT databases can help parameterize interatomic potentials. For example, Pun *et al.*[67] used energies computed by the AFLOWlib and OQMD databases to develop a force field for the Cu-Ta system. Data from the *ab initio* scale can inform materials behavior at higher length scales in other ways as well. For example, Gibson and Schuh[68] employed formation energies from the MP to help establish energy scales for grain boundary cohesion. These kinds of approaches that bridge length scales encompass a key component of the MGI called integrated computational materials engineering (ICME).[22,23,69]

A second use case of DFT databases is as an established reference against which to compare results. For example, a study by Miletic *et al.* used the AFLOWlib database to test their computations of lattice parameters and magnetic moment of YNi_5.[70] Romero *et al.* developed a new structure search algorithm and used all three database resources (MP, AFLOWlib, and OQMD) to verify that none of the new predicted compounds were previously recorded.[71] Sarmiento-Pérez *et al.*,[72] and Jauho

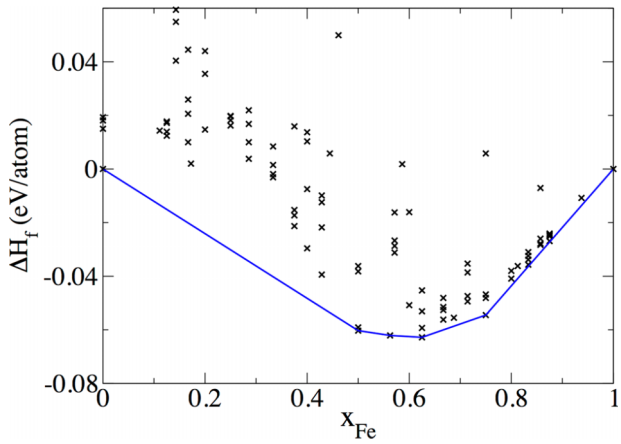

FIG. 2. Example of convex hull stability analysis for Fe-Co alloys, with data points from the AFLOWlib database, used by Troparevsky *et al.*[76] for the purposes of designing high-entropy alloys. Reproduced with permission from Troparevsky *et al.*, JOM **67**, 2350–2363 (2015). Copyright 2015 Springer.

et al.,[73] used extensive sets of the ground state structures from the MP to test new exchange-correlation functionals, and many other examples of such usage can be found.[74,75]

One computational method that has particularly benefited from DFT materials databases is the estimation of the thermodynamic stability of new and hypothetical materials. Estimating this quantity requires calculating all known phases within the chemical space of the target phase and then applying a convex hull analysis to determine the lowest energy combination of phases at the target composition. This process, which was in the past tedious and time-consuming, can today be performed within minimal effort using the data and software tools[77] already provided by larger DFT data providers (Fig. 2). Examples include stability analysis of new photocatalysts[78,79] and other applications[80–84] (assisted by MP data), high entropy alloys[76] (assisted by AFLOWlib data), and dichalcogenides for hydrogen evolution[85] (assisted by OQMD data).

Finally, computational material databases can be used for property prediction and materials screening. For example, Sun *et al.* used the OQMD to help predict lithiation energies of MXene compounds,[86] and Gschwind *et al.* similarly used the MP to predict fluoride ion battery voltages.[87] Hong *et al.* employed MP data to screen new materials as oxygen evolution reaction catalysts.[88] Impressively, some groups have been able to download and use very large data sets in their study. For example, Seko *et al.*[89] extracted 54 779 relaxed compounds as starting points from the MP to virtually screen for low thermal conductivity compounds using anharmonic lattice dynamics calculations. Tada *et al.* used 34 000 compounds from the MP as a basis for screening new materials as two-dimensional electrides.[90] Thus, the usage of computational databases can involve confirming a few data points or screening tens of thousands of compounds.

B. Applications of computational databases to experimental studies

Perhaps an even greater impact to materials science will come from the experimental community's usage of computational databases. A first example application is comparing theoretical and experimental data in order to verify both or fill gaps in information. For example, the MP is often used to look up materials properties such as lattice parameters, XRD peak positions, or even battery conversion voltages.[91–93] The comprehensive nature of these databases can make them very powerful tools when used in concert with experiment. For example, both the AFLOWlib and OQMD contain data on a large number of Heusler-type phases. This helped the research group of Nash perform multiple studies that map out the thermodynamics and phase equilibria of Heusler[94–97] and half-Heusler[98] phases.

Similar to the situation for theorists, one of the most popular uses of computational databases by experimental researchers has been to generate phase diagrams. Multiple studies[80,100–102] have employed the MP Phase Diagram App to establish whether their materials of interest are likely to form under certain conditions. For example Martinolich and Neilson[100] report that $NaFeS_2$ is reported (by

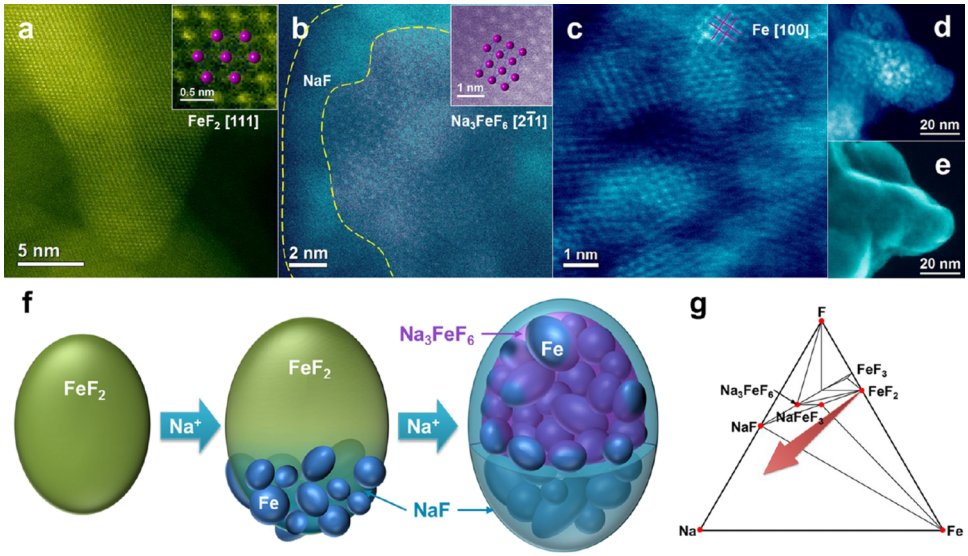

FIG. 3. Information from high resolution STEM images ((a)-(d)) and the computational phase diagram from the Materials Project (g) were employed by He *et al.*[99] to propose a mechanism for sodium incorporation into FeF_2 particles for battery applications. Reprinted with permission from He *et al.*, ACS Nano **8**, 7251–7259 (2014). Copyright 2014 American Chemical Society.

MP) to be unstable with respect to decomposition into FeS_2, FeS, and Na_2S, which is consistent with their unsuccessful attempts to prepare the ternary phase.

Computed phase diagrams from DFT are also useful for other purposes. For example, He *et al.* investigated the heterogeneous sodiation mechanism of iron fluoride (FeF_2) nanoparticle electrodes by combining *in situ* and *ex situ* microscopy and spectroscopy techniques.[99] The MP Li-Fe-F phase diagram was used to interpret the observed reaction pathways and illustrate the difference between equilibrium reactions and the metastable phases that may form because of kinetic limitations (Figure 3). Another example is from the work of MacEachern *et al.*,[103] who superimposed the results of sputtering experiments onto the MP phase diagram for Fe-Si-Zn. Similarly, Nagase *et al.*[104] employed the computed MP Co-Cu-Zr-B quarternary phase diagram from MP to guide experimental exploration of amorphous phase formation. In another work, the MP phase diagram of Li-Ni-Si was used to highlight the existence of a ternary lithiated phase that could impact the performance of nickel silicides as Li-ion anode materials.[105]

The calculated electronic structure (e.g., the band structure of compounds) is also frequently used in experimental studies despite the known underestimation of band gaps in standard DFT.[107] For example, Fondell *et al.* used the MP-calculated band structure of hematite to explain its limited

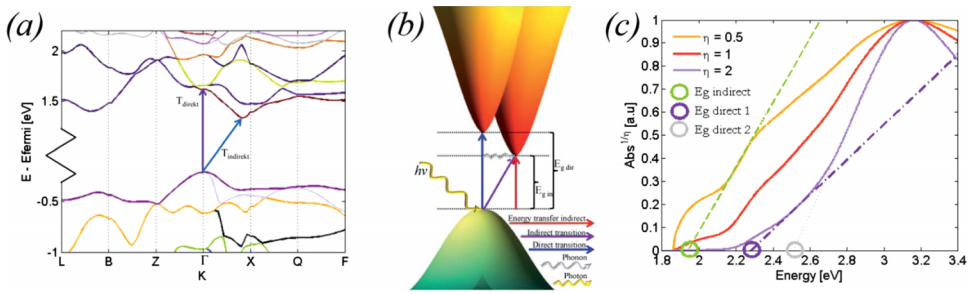

FIG. 4. (a) Calculated band structure of Fe_2O_3 from the Materials Project, (b) a schematic of the indirect nature of the transition (c) and its effect on optical absorption measurements from the work of Fondell *et al.*[106] Reproduced with permission from Fondell *et al.*, J. Mater. Chem. A **2**, 3352 (2014). Copyright 2014 Royal Society of Chemistry.

charge transport in the bulk phase and motivate the need for low-dimensional particles or thin films for solar hydrogen production (Fig. 4).[106]

When considering that the role of density functional theory calculations was limited to the dominion of the theoretical physicist only two to three decades ago, these examples indicate an encouraging trend of DFT calculations becoming practical materials science design tools that are accessible to the broader research community. In alignment with the MGI vision, the examples above highlight a growing trend in which research data (whether computational or experimental) are routinely compared against past results and in which new studies are motivated or explained at least in part by data compiled by computational databases.

III. TECHNICAL ASPECTS OF DATA SHARING

With the benefits of data sharing established, we now concentrate on three technical issues in data sharing: data formats, data dissemination strategies, and data centralization. Our discussion is rooted in our experience in collaboratively building two tools: the Materials Project API (application programming interface) (MAPI),[108] which has been used by over 300 distinct users to download approximately 15×10^6 data points, and the MPContribs framework,[109,110] which allows researchers to contribute external data into the MP.

A. Data formats

Simulation software and experimental instruments typically output their own proprietary data file formats. Professional societies often develop custom data formats, such as the CIF (Crystallographic Information File) standard,[111] which define their own semantics and require building custom parsers to support that standard. Currently, the situation for data formats in materials science resembles a "wild west" scenario that makes it difficult for end users to work with data from multiple sources.

One strategy that can help promote standardization to build materials-centric specifications off of standard data formats that are fully open and cross-platform. For example, tabular data can conform to a comma-separated-values (CSV) format that is easily readable by almost all analysis packages and programming languages. Large, array-based data can be formatted as HDF5[112] or netCDF.[113] For more complex data, e.g., the representation of crystal structures or input parameters and output quantities such as band structures, options include Extensible Markup Language (XML)[114] and Javascript Object Notation (JSON)[115] formats. Successful examples include ChemML[116] and MatML,[117] which build scientific specifications on top of the XML standard. Our preference is for the JSON format, which is smaller in size and simpler to construct. JSON documents can be directly stored in next-generation database technologies such as MongoDB[118] which allow rich search capabilities over these documents. JSON can also be interconverted to the Yet Another Markup Language (YAML) format, which can be easily read and edited in text editors.

B. Data dissemination strategies

Another technical aspect of data reuse concerns the most appropriate way to expose large data sets to the research community. Individual data providers can select from several options for exposing their data, including file download (e.g., as CSV or ZIP archive), a database dump (e.g., a MySQL dump file), or exposing an API over the web. The first two of these methods are relatively straightforward; the final method merits further clarification. An API is a protocol for interacting with a piece of software or data resource. For sharing data over the web, a common pattern is to use a REpresentaitional State Transfer (REST) API.[119] REST APIs are employed by several modern software companies, including Google, Microsoft, Facebook, Twitter, Dropbox, and others. In the most straightforward use case, one can imagine a REST API as mapping web URLs to data in the same way that a folder structure is used to organize files. For example, an API can be set up such that requesting the URL "https://www.materialsproject.org/rest/v2/materials/Fe2O3/vasp" returns a JSON representation of the data on all Fe_2O_3 compounds as computed by the VASP[120] software. However, RESTful APIs are more powerful than file folder trees. For example, the HTTP request used to fetch the URL can additionally incorporate parameters. Such parameters might include an *API key* that manages user access control

TABLE I. A comparison of different methods for exposing data using three different technologies, including positive aspects (pros) and negative aspects (cons) for each selection.

Method		Considerations
Direct file download, e.g., as CSV or XML, individually or as a folder tree	PROS	• Straightforward for data providers to set up • Transparent for end users to download and explore
	CONS	• Difficult to release data updates in an organized manner • Difficult to search data sets (folder tree or rudimentary filtering is often the only search method) • Difficult to coordinate different data providers; end user must work separately with each such provider
Database dump, i.e., download a single file that can be reconstructed into a database	PROS	• Straightforward for data providers to set up, provided the database technology is already being used • Simple for end users to download large data sets • Rich search is possible through the database query language
	CONS	• Difficult to expose targeted data updates without re-downloading the entire database • User must become comfortable with the database technology used by the provider • In SQL databases, user must tailor their analysis to the way the provider has structured their data, i.e., the *schema*. If the data provider changes the underlying schema of their database (e.g., to improve search efficiency), users must synchronize their analysis code to the database schema updates • Difficult to coordinate amongst multiple data providers because each provider may employ a different database technology
RESTfulAPI	PROS	• Combines the search benefits of hierarchical organization (URL trees) and database search (through REST functions) • Straightforward to support data updates and multiple data versions • Possible to expose only certain portions of the underlying database and control user access • Flexible to changes in the underlying storage technology; changes to the backend, or several data providers with different backends entirely, can maintain the same API to the end user • Possible to expose not only data but also functions that accept parameters from the user
	CONS	• Difficult for data providers to set up; effort is required to design and expose the API • Usage can be difficult for those that are uncomfortable with programming • In general, less transparent to track down and download all the data than with a database dump file (although API design can mitigate this)

over different parts of the data or a constraint that helps in filtering the search. REST URLs can additionally point not only to data but also to backend functions. For example, the URL "https://www.materialsproject.org/rest/v1/materials/snl/submit" might trigger a backend function that registers a request to compute the desired structure embedded in an Hypertext Transfer Protocol (HTTP) POST parameter. RESTful APIs can be intimidating to novices but can be made more user-friendly by making the URL scheme explorable and through intermediate software layers. We present a comparison between data dissemination methods in Table I.

C. Data centralization

A final technical aspect we discuss is the strategy for combining information from distinct data resources. In particular, data can be stored and managed by a small, large, or intermediate number of

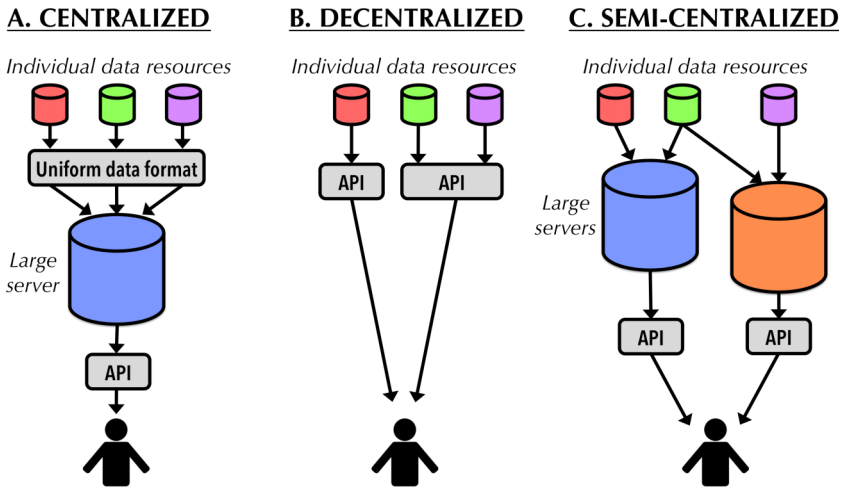

FIG. 5. Potential models for data unification. In the (a) centralized model, data are homogenized and submitted to a single entity that manages data storage and data access through an application programming interface (API). In the (b) decentralized model, each data resource maintains its own data and exposes its own API (either coordinated or uncoordinated with other providers) for the end user. The most likely scenario is the (c) semi-centralized model in which several large providers each handle data contributions for different sub-domains of materials science.

entities, which we broadly classify as "centralized," "decentralized," and "semi-centralized," respectively (Figure 5).

One organization that has made significant progress in establishing a centralized data resource for materials scientists is Citrine Informatics, a company that specializes in applying data mining to materials discovery and optimization. Citrine's data sharing service benefits its core business because its proprietary data mining algorithms become more powerful with access to larger datasets. Data providers and research groups that contribute data to the Citrine database benefit from greater visibility of their work and straightforward compliance with data management requirements. Major benefits of such data centralization by industry include no cost to government funding agencies and the ability to leverage a professional software engineering team to build and support the infrastructure. A potential concern is longevity of the data, which is mitigated through an open API that allows making public copies of the data.

As depicted more generally in Figure 5(a), data contributed to Citrine's platform ("Citrination")[121] are reshaped into their custom Materials Information File (MIF) format,[122] a JSON-based materials data standard. The data are hosted on Citrine's servers and can be accessed by the public either through Citrination's web interface or through a RESTful API. The Citrination search engine today includes almost 250 data sets from various experimental and computational sources. Examples include data on thermoelectric materials, bulk metallic glasses, and combinatorial photocatalysis experiments.

A different organizational structure, depicted in Figure 5(b), is to decentralize data providers and to combine data as needed from multiple data APIs. Advantages of this technique include ease of supporting diverse data, a greater likelihood of maintaining up-to-date search results, and reducing the burden (e.g., storage, bandwidth) on any single data provider. A disadvantage is that each data provider must host and maintain its public API. The decentralized system, which is in some ways analogous to the way that Google indexes web pages hosted by many different servers, requires the development of search engines that are capable of working with multiple data resources. For example, in the chemistry community, the ChemSpider[123] search engine has had considerable success with this strategy.

The most likely scenario is that the community will reach an equilibrium that balances both strategies, as in Figure 5(c). Larger data providers that are continuously generating new data and new features will perhaps maintain their own data service, whereas smaller data outfits that want to submit "one-time" data sets will likely contribute to a larger service.

IV. CHALLENGES AND OUTLOOK

Many of the issues specific to sharing data in materials have been outlined previously.[124,125] Here, we focus on three key challenges that require innovative solutions: parsing unstructured data, combining multiple data sources, and communicating the nuances of different data sets.

Most materials measurements are not available in a structured format. Instead, they are isolated in laboratory notebooks or embedded within a chart, table, or text of a written document. For example, routes for synthesizing materials are almost exclusively reported as text without a formal data structure. Extracting such data into structured knowledge, e.g., using natural language processing, represents a major challenge for materials research.[126]

A second challenge is in combining information from multiple data sets. To illustrate this problem, consider the case of the experimental thermoelectrics data set from the work of Gaultois *et al.*[127] and a second computational data set from the work of Gorai *et al.*,[128] both of which can be downloaded from each project's respective web site or from the Citrination platform. Ideally, one would like to make a comparison, e.g., to assess the accuracy of the computational data with respect to the experimental measurements. Such a study requires matching common entries between databases. As a first step, one can match the entries in each database which pertain to the same composition. However, ensuring that the crystal structures also match is usually much more challenging. In the example of the two thermoelectrics data sets, once can make a match based on Inorganic Crystal Structure Database (ICSD) identification number. In other situations, little to no crystal structure information may be available in one or both data sets. Next, one must align the physical conditions of each entry; in this instance, the computational data are reported at 0 K, whereas the experimental measurements are reported at multiple temperatures between 300 K and 1000 K. Thus, a researcher must select the 300 K experimental data as the most relevant, keeping in mind that the remaining difference in temperature represents a potential source of error. Finally, other material parameters must be considered: for example, the microstructure and defect concentrations of the experimental data can be very different than the pure, single crystal model of many computations. Much of the time, such detailed information on the material is not reported or even known. The situation becomes more troublesome as data sets grow larger because it becomes increasingly time-consuming to employ manual analysis to aid in the process. Better strategies and additional metadata on each measurement are needed to confidently perform analyses across data sets.

A third major challenge is in understanding and communicating the level of accuracy that can be expected from different characterization methods. Interpreting the nuances and evaluating the potential errors of a measurement or simulation technique is typically a subject reserved for domain experts. For example, we have observed that the error bar of "standard" DFT methods in predicting the reaction enthalpies *between oxides* is approximately 24 meV/atom.[129] However, the error bar is much higher (172 meV/atom) for reactions that involve both metals and oxides. More complicated still, by applying a correction strategy, once can reduce the latter error bar to approximately 45 meV/atom,[130] although different correction strategies have been proposed by different research groups.[131–133] Issues of this kind remain very difficult to communicate to non-specialists but are very important when advocating data reuse.

Looking forward, we expect that learning to work with data, and particularly data that are not self-generated, to become an important skill for material scientists. The integration of computer scientists and statisticians into the domain of materials science will also be vital. The path paved by the biological sciences serves as a good model: bioinformatics and biostatistics are today well-recognized fields in their own right, and biology and health-related fields are recognized as being more multidisciplinary than materials science today.[134]

One can also ask whether materials science will be impacted by the revolution in "big data."[135] While big data can be a nebulous term, reports often point to the "three V's"—i.e., the volume, velocity, and variety of the data (and sometimes extended to include veracity and value).[136] Other definitions of big data are that it occurs when there are significant problems in loading, transferring, and processing the data set using conventional hardware and software, or when the data size reaches a scale that enables qualitatively new types of analysis approaches. With respect to volume and velocity (rate at which data is generated), materials science has not hit the "big data" mark as compared with

other fields. For example, the European Bioinformatics Institute stores approximately 20 petabytes of data, and the European Organization for Nuclear Research (CERN) generates approximately 15 petabytes of data and has stored approximately 200 petabytes of results.[137] For comparison, all the raw data stored in the MP amount to approximately 100 terabytes (a factor of 2000 less than CERN) and the final data sets exposed to users amount to approximately 1 terabyte. Data from large experimental facilities such as light sources may change this assessment in the future,[135] as might software that allows running several orders of magnitude more computations on ever-larger computing resources.[138–140] Although materials science data are not particularly large, it is challenging to work with in terms of the variety of data types and the complexity of objects such as periodically repeating crystals or mass spectroscopy data. As for whether data sets will open new qualitative research avenues, the size of materials data sets is likely still too small to apply some machine learning techniques such as "deep learning."[141,142] However, more efficient machine learning algorithms for such kinds of learning are being developed[143] which hint that the coming wave in materials data will indeed open up new types of analysis methods.

Transitioning from data to insight will remain a major research topic. Much recent research has begun applying data mining techniques to materials data sets, but "materials informatics"[144] remains a nascent field of study. To propel this field forward, Materials Hackathons[145] organized by Citrine Informatics and a Materials Data Challenge hosted by NIST[146] have been announced to encourage advancement through competition. Further progress might be stimulated by incorporating data visualization and analysis tools directly into common materials database portals. For example, we envision that in the future, one will be able to log on to the MP, identify a target property, visualize the distribution of that property over known materials, extract relevant descriptors, and build and test a predictive model.

Many challenges lie ahead for uncovering the materials genome, i.e., to understand what factors are responsible for the complex behavior of advanced materials through the use of data-driven methods. However, the recent examples demonstrating that theorists and experimental researchers alike can apply online *ab initio* computation databases towards cutting-edge research problems is an encouraging harbinger of a new and exciting era in collaborative materials discovery.

ACKNOWLEDGMENTS

This work was funded and intellectually led by the Materials Project (DOE Basic Energy Sciences Grant No. EDCBEE). Work at the Lawrence Berkeley National Laboratory was supported by the U.S. Department of Energy Office of Science, Office of Basic Energy Sciences Department under Contract No. DE-AC02-05CH11231. We thank Bryce Meredig for discussions regarding the Citrination search platform.

[1] E. B. Tadmor, R. S. Elliott, S. R. Phillpot, and S. B. Sinnott, "NSF cyberinfrastructures: A new paradigm for advancing materials simulation," Curr. Opin. Solid State Mater. Sci. **17**, 298–304 (2013).
[2] B. G. Sumpter, R. K. Vasudevan, T. Potok, and S. V. Kalinin, "A bridge for accelerating materials by design," npj Comput. Mater. **1**, 15008 (2015).
[3] G. B. Olson, "Designing a new material world," Science **288**, 993–998 (2000).
[4] C. L. Magee, "Towards quantification of the role of materials innovation in overall technological development," Complexity **18**, 10–25 (2012).
[5] J. A. Christodoulou, "Integrated computational materials engineering and materials genome initiative: Accelerating materials innovation," Adv. Mater. Processes **171**(3), 28–31 (March 2013).
[6] T. W. Eagar, "Bringing new materials to market," Technol. Rev. **98**(2), 42–49 (1995).
[7] C. O. Björling and A. Westgren, "Minerals of the Varuträsk pegmatite," Geol. Foeren. Stockholm Foerh. **60**, 67–72 (1938).
[8] A. Padhi, K. Nanjundaswamya, and J. Goodenough, "Phospho-olivines as positive-electrode materials for rechargeable lithium batteries," J. Electrochem. Soc. **144**, 1188–1194 (1997).
[9] G. Ceder, G. Hautier, A. Jain, and S. P. Ong, "Recharging lithium battery research with first-principles methods," MRS Bull. **36**, 185–191 (2011).
[10] H. W. Meuer, E. Strohmaier, J. Dongarra, and H. D. Simon, *The TOP500: History, Trends, and Future Directions in High Performance Computing* (Chapman & Hall/CRC, 2014).
[11] R. O. Jones, "Density functional theory: Its origins, rise to prominence, and future," Rev. Mod. Phys. **87**, 897 (2015).
[12] A. D. Becke, "Perspective: Fifty years of density-functional theory in chemical physics," J. Chem. Phys. **140**, 18A301 (2014).
[13] A. Jain, Y. Shin, and K. A. Persson, "Computational predictions of energy materials using density functional theory," Nat. Rev. Mater. **1**, 15004 (2016).

[14] G. Hautier, A. Jain, and S. P. Ong, "From the computer to the laboratory: Materials discovery and design using first-principles calculations," J. Mater. Sci. **47**, 7317–7340 (2012).
[15] R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, and H. Lam, "Combinatorial and high-throughput screening of materials libraries: Review of state of the art," ACS Comb. Sci. **13**, 579 (2011).
[16] J.-C. Zhao, "High-throughput experimental tools for the materials genome initiative," Chin. Sci. Bull. **59**, 1652–1661 (2014).
[17] J. Hattrick-Simpers, C. Wen, and J. Lauterbach, "The materials super highway: Integrating high-throughput experimentation into mapping the catalysis materials genome," Catal. Lett. **145**, 290–298 (2014).
[18] S. Nemšák, A. Shavorskiy, O. Karslioglu, I. Zegkinoglou, A. Rattanachata, C. S. Conlon, A. Keqi, P. K. Greene, E. C. Burks, F. Salmassi, E. M. Gullikson, S. Yang, K. Liu, H. Bluhm, and C. S. Fadley, "Concentration and chemical-state profiles at heterogeneous interfaces with sub-nm accuracy from standing-wave ambient-pressure photoemission," Nat. Commun. **5**, 5441 (2014).
[19] D. S. Su, B. Zhang, and R. Schlögl, "Electron microscopy of solid catalysts—Transforming from a Challenge to a toolbox," Chem. Rev. **115**, 2818 (2015).
[20] N. Balke, S. Jesse, A. N. Morozovska, E. Eliseev, D. W. Chung, Y. Kim, L. Adamczyk, R. E. García, N. Dudney, and S. V. Kalinin, "Nanoscale mapping of ion diffusion in a lithium-ion battery cathode," Nat. Nanotechnol. **5**, 749–754 (2010).
[21] M. E. Holtz, Y. Yu, D. Gunceler, J. Gao, R. Sundararaman, K. A. Schwarz, T. A. Arias, H. D. Abruña, and D. A. Muller, "Nanoscale imaging of lithium ion distribution during *in situ* operation of battery electrode and electrolyte," Nano Lett. **14**, 1453–1459 (2014).
[22] P. Patel, "Materials genome initiative and energy," MRS Bull. **36**, 964–966 (2011).
[23] J. Allison, "Integrated computational materials engineering: A perspective on progress and future steps," JOM **63**, 15–18 (2011).
[24] National Science and Technology Council, *Materials Genome Initiative Strategic Plan* (National Science and Technology Council, 2014), available at https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/mgi_strategic_plan_-_dec_2014.pdf.
[25] A. White, "The materials genome initiative: One year on," MRS Bull. **37**, 715–716 (2012).
[26] A. A. White, "Universities prepare next-generation workforce to benefit from the materials genome initiative," MRS Bull. **38**, 673–674 (2013).
[27] G. B. Olson and C. J. Kuehmann, "Materials genomics: From CALPHAD to flight," Scr. Mater. **70**, 25–30 (2014).
[28] A. A. White, "Interdisciplinary collaboration, robust funding cited as key to success of materials genome initiative program," MRS Bull. **38**, 894–896 (2013).
[29] A. White, "Workshop makes recommendations to increase diversity in materials science and engineering," MRS Bull. **38**, 120–122 (2013).
[30] L. Glasser, "Crystallographic information resources," J. Chem. Educ. **93**, 542 (2015).
[31] Y. Le Page, "Data mining in and around crystal structure databases," MRS Bull. **31**, 991 (2006).
[32] J. Faber and T. Fawcett, "The powder diffraction file: Present and future," Acta Crystallogr., Sect. B: Struct. Sci. **58**, 325–332 (2002).
[33] C. W. Bale, E. Bélisle, P. Chartrand, S. A. Decterov, G. Eriksson, K. Hack, I.-H. Jung, Y.-B. Kang, J. Melançon, A. D. Pelton, C. Robelin, and S. Petersen, "FactSage thermochemical software and databases—Recent developments," Calphad **33**, 295–311 (2009).
[34] P. J. Linstrom and W. G. Mallard, NIST Chemistry WebBook, NIST Standard Reference Database Number 69, 2013.
[35] M. W. Chase and J. A. N. A. Force, NIST-JANAF thermochemical tables, 1998.
[36] O. Kubaschewski, C. Alcock, and P. Spencer, *Materials Thermochemistry*, 6th ed. (Pergamom Press, Oxford, 1993).
[37] T. B. Massalski and H. Okamoto, *Binary Alloy Phase Diagrams*, 2nd ed. (ASM International, 1990).
[38] C. E. Campbell, U. R. Kattner, and Z. K. Liu, "File and data repositories for next generation CALPHAD," Scr. Mater. **70**, 7–11 (2014).
[39] L. Kaufman and J. Ågren, "CALPHAD, first and second generation—Birth of the materials genome," Scr. Mater. **70**, 3–6 (2014).
[40] A. A. White, "Mandates for public access to publications and data on the horizon for US researchers," MRS Bull. **38**, 531–532 (2013).
[41] J. R. Kitchin, "Data sharing in surface science," Surf. Sci. (published online 2015).
[42] J. R. Kitchin, "Examples of effective data sharing in scientific publishing," ACS Catal. **5**, 3894–3899 (2015).
[43] D. Guevarra, A. Shinde, S. K. Suram, I. D. Sharp, F. Toma, J. A. Haber, and J. Gregoire, "Development of solar fuels photoanodes through combinatorial integration of Ni-La-Co-Ce oxide catalysts on $BiVO_4$," Energy Environ. Sci. **9**, 565 (2015).
[44] L. Cheng, R. S. Assary, X. Qu, A. Jain, S. P. Ong, N. N. Rajput, K. A. Persson, and L. A. Curtiss, "Accelerating electrolyte discovery for energy storage by high throughput screening," J. Phys. Chem. Lett. **6**, 283–291 (2015).
[45] X. Qu, A. Jain, N. N. Rajput, L. Cheng, Y. Zhang, S. P. Ong, M. Brafman, E. Maginn, L. A. Curtiss, and K. A. Persson, "The electrolyte genome project: A big data approach in battery materials discovery," Comput. Mater. Sci. **103**, 56–67 (2015).
[46] S. K. Suram, J. A. Haber, J. Jin, and J. M. Gregoire, "Generating information rich high-throughput experimental materials genomes using functional clustering via multi-tree genetic programming and information theory," ACS Comb. Sci. **17**, 224–233 (2015).
[47] L. Lin, "Materials databases infrastructure constructed by first principles calculations: A review," Mater. Perform. Charact. **4**, MPC20150014 (2015).
[48] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).

[49] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, "AFLOWLIB.ORG: A distributed materials properties repository from high-throughput *ab initio* calculations," Comput. Mater. Sci. **58**, 227–235 (2012).
[50] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD)," JOM **65**, 1501–1509 (2013).
[51] M. De Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. K. Ande, S. Van Der Zwaag, J. J. Plata, C. Toher, S. Curtarolo, G. Ceder, K. A. Persson, and M. Asta, "Charting the complete elastic properties of inorganic crystalline compounds," Sci. Data **2**, 150009 (2015).
[52] M. de Jong, W. Chen, H. Geerlings, M. Asta, and K. A. Persson, "A database to enable discovery and design of piezoelectric materials," Sci. Data **2**, 150053 (2015).
[53] G. Yuan and F. Gygi, "ESTEST: A framework for the validation and verification of electronic structure codes," Comput. Sci. Discovery **3**, 015004 (2010).
[54] See http://nomad-repository.eu/cms/ for the NoMaD repository.
[55] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, G. Ceder, S. Ping Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, "A high-throughput infrastructure for density functional theory calculations," Comput. Mater. Sci. **50**, 2295–2310 (2011).
[56] W. Setyawan and S. Curtarolo, "High-throughput electronic band structure calculations: Challenges and tools," Comput. Mater. Sci. **49**, 299–312 (2010).
[57] C. E. Calderon, J. J. Plata, C. Toher, C. Oses, O. Levy, M. Fornari, A. Natan, M. J. Mehl, G. Hart, M. B. Nardelli, and S. Curtarolo, "The AFLOW standard for high-throughput materials science calculations," Comput. Mater. Sci. **108**, 233–238 (2015).
[58] M. W. Penninger, C. H. Kim, L. T. Thompson, and W. F. Schneider, "DFT analysis of NO oxidation intermediates on undoped and doped $LaCoO_3$ perovskite," J. Phys. Chem. C **119**, 20488–20494 (2015).
[59] Y. Zhu, X. He, and Y. Mo, "First principles study on electrochemical and chemical stability of the solid electrolyte-electrode interfaces in all-solid-state Li-ion batteries," J. Mater. Chem. A **4**, 3253 (2015).
[60] Y. Zhu, X. He, and Y. Mo, "Origin of outstanding stability in the lithium solid electrolyte materials: Insights from thermodynamic analyses based on first principles calculations," ACS Appl. Mater. Interfaces **7**, 23658 (2015).
[61] A. Narayan, A. Bhutani, S. Rubeck, J. N. Eckstein, D. P. Shoemaker, and L. K. Wagner, "Computational and experimental investigation of unreported transition metal selenides and sulphides," e-print arXiv:1512.02214 [cond-mat.mtrl-sci] (2015).
[62] M. Burbano, M. Duttine, O. Borkiewicz, A. Wattiaux, A. Demourgues, M. Salanne, H. Groult, and D. Dambournet, "Anionic ordering and thermal properties of FeF_3 center dot $3H_2O$," Inorg. Chem. **54**, 9619–9625 (2015).
[63] R. Sarmiento-Pérez, T. F. T. Cerqueira, S. Körbel, S. Botti, and M. A. L. Marques, "Prediction of stable nitride perovskites," Chem. Mater. **27**, 5957–5963 (2015).
[64] I. Jandl, H. Ipser, and K. W. Richter, "Thermodynamic modelling of the general NiAs-type structure: A study of first principle energies of formation for binary Ni-containing B8 compounds," Calphad **50**, 174–181 (2015).
[65] E. B. Tadmor, R. S. Elliott, J. P. Sethna, R. E. Miller, and C. A. Becker, "The potential of atomistic simulations and the knowledge base of interatomic models," J. Mater. **63**, 17 (2011).
[66] C. A. Becker, F. Tavazza, Z. T. Trautt, R. A. Buarque, and D. Macedo, "Considerations for choosing and using force fields and interatomic potentials in materials science and engineering," Curr. Opin. Solid State Mater. Sci. **17**, 277–283 (2013).
[67] G. P. Purja Pun, K. A. Darling, L. J. Kecskes, and Y. Mishin, "Angular-dependent interatomic potential for the Cu–Ta system and its application to structural stability of nano-crystalline alloys," Acta Mater. **100**, 377–391 (2015).
[68] M. A. Gibson and C. A. Schuh, "A survey of *ab-initio* calculations shows that segregation-induced grain boundary embrittlement is predicted by bond-breaking arguments," Scr. Mater. **113**, 55–58 (2016).
[69] J. Allison, D. Backman, and L. Christodoulou, "Integrated computational materials engineering: A new paradigm for the global materials profession," JOM **58**, 25–27 (2006).
[70] G. I. Miletic and A. Drašner, "DFT study of the cohesive and structural properties of YNi_5H_x compounds," J. Alloys Compd. **622**, 1041–1048 (2015).
[71] I. Valencia-Jaime, R. Sarmiento-Pérez, S. Botti, M. A. L. Marques, M. Amsler, S. Goedecker, and A. H. Romero, "Novel crystal structures for lithium–silicon alloy predicted by minima hopping method," J. Alloys Compd. **655**, 147–154 (2016).
[72] R. Sarmiento-Pérez, S. Botti, and M. A. L. Marques, "Optimized exchange and correlation semilocal functional for the calculation of energies of formation," J. Chem. Theory Comput. **11**, 3844–3850 (2015).
[73] T. S. Jauho, T. Olsen, T. Bligaard, and K. S. Thygesen, "Improved description of metal oxide stability: Beyond the random phase approximation with renormalized kernels," Phys. Rev. B **92**, 115140 (2015).
[74] V. S. Kandagal, M. D. Bharadwaj, and U. V. Waghmare, "Theoretical prediction of a highly conducting solid electrolyte for sodium batteries: $Na_{10}GeP_2S_{12}$," J. Mater. Chem. A **3**, 12992 (2015).
[75] G. M. Dongho Nguimdo and D. P. Joubert, "A density functional (PBE, PBEsol, HSE06) study of the structural, electronic and optical properties of the ternary compounds $AgAlX_2$ (X = S, Se, Te)," Eur. Phys. J. B **88**, 113 (2015).
[76] M. C. Troparevsky, J. R. Morris, M. Daene, Y. Wang, A. R. Lupini, and G. M. Stocks, "Beyond atomic sizes and Hume-Rothery rules: Understanding and predicting high-entropy alloys," JOM **67**, 2350–2363 (2015).
[77] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python materials genomics (pymatgen): A robust, open-source python library for materials analysis," Comput. Mater. Sci. **68**, 314–319 (2013).
[78] I. E. Castelli, D. D. Landis, K. S. Thygesen, S. Dahl, I. Chorkendorff, T. F. Jaramillo, and K. W. Jacobsen, "New cubic perovskites for single-and two-photon water splitting using the computational materials repository," Energy Environ. Sci. **5**, 9034 (2012).
[79] I. E. Castelli, T. Olsen, S. Datta, D. D. Landis, S. Dahl, K. S. Thygesen, and K. W. Jacobsen, "Computational screening of perovskite metal oxides for optimal solar light capture," Energy Environ. Sci. **5**, 5814 (2012).

[80] T. Krishnamoorthy, H. Ding, C. Yan, W. L. Leong, T. Baikie, Z. Zhang, S. Li, M. Asta, N. Mathews, and S. G. Mhaisalkar, "Lead-free germanium iodide perovskite materials for photovoltaic application," J. Mater. Chem. A **3**, 23829–23832 (2015).
[81] Z. R. Liu and D. Y. Li, "Stability and formation of long period stacking order structure in Mg-based ternary alloys," Comput. Mater. Sci. **103**, 90–96 (2015).
[82] R. Sarmiento-Pérez, T. F. T. Cerqueira, I. Valencia-Jaime, M. Amsler, S. Goedecker, A. H. Romero, S. Botti, and M. A. L. Marques, "Novel phases of lithium-aluminum binaries from first-principles structural search," J. Chem. Phys. **142**, 024710 (2015).
[83] Z.-H. Cai, P. Narang, H. A. Atwater, S. Chen, C.-G. Duan, Z.-Q. Zhu, and J.-H. Chu, "Cation-mutation design of quaternary nitride semiconductors lattice-matched to GaN," Chem. Mater. **27**, 7757 (2015).
[84] K. Choudhary, T. Liang, K. Mathew, B. Revard, A. Chernatynskiy, S. R. Phillpot, R. G. Hennig, and S. B. Sinnott, "Dynamical properties of AlN nanostructures and heterogeneous interfaces predicted using COMB potentials," Comput. Mater. Sci. **113**, 80–87 (2016).
[85] M. Pandey, A. Vojvodic, K. S. Thygesen, and K. W. Jacobsen, "Two-dimensional metal dichalcogenides and oxides for hydrogen evolution: A computational screening approach," J. Phys. Chem. Lett. **9**, 1577–1585 (2015).
[86] D. Sun, Q. Hu, J. Chen, X. Zhang, L. Wang, Q. Wu, and A. Zhou, "Structural transformation of MXene (V_2C, Cr_2C, and Ta_2C) with O groups during lithiation: A first principles investigation," ACS Appl. Mater. Interfaces **8**, 74 (2015).
[87] F. Gschwind, G. Rodriguez-Garcia, D. J. S. Sandbeck, A. Gross, M. Weil, M. Fichtner, and N. Hörmann, "Fluoride ion batteries: Theoretical performance, safety, toxicity, and a combinatorial screening of new electrodes," J. Fluorine Chem. **182**, 76–90 (2016).
[88] W. T. Hong, R. E. Welsch, and Y. Shao-Horn, "Descriptors of oxygen-evolution activity for oxides: A statistical evaluation," J. Phys. Chem. C **120**, 78–86 (2016).
[89] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka, "Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization," Phys. Rev. Lett. **115**, 205901 (2015).
[90] T. Tada, S. Takemoto, S. Matsuishi, and H. Hosono, "High-throughput *ab initio* screening for two-dimensional electride materials," Inorg. Chem. **53**, 10347 (2014).
[91] M. J. Young, M. Neuber, A. C. Cavanagh, H. Sun, C. B. Musgrave, and S. M. George, "Sodium charge storage in thin films of MnO_2 derived by electrochemical oxidation of MnO atomic layer deposition films," J. Electrochem. Soc. **162**, A2753–A2761 (2015).
[92] J. C. Weber, P. T. Blanchard, A. W. Sanders, J. C. Gertsch, S. M. George, S. Berweger, A. Imtiaz, K. J. Coakley, T. M. Wallis, K. A. Bertness, P. Kabos, N. A. Sanford, and V. M. Bright, "GaN nanowire coated with atomic layer deposition of tungsten: A probe for near-field scanning microwave microscopy," Nanotechnology **25**, 415502 (2014).
[93] T. T. Tran and M. N. Obrovac, "Alloy negative electrodes for high energy density metal-ion cells," J. Electrochem. Soc. **158**, A1411 (2011).
[94] M. Yin, J. Hasier, and P. Nash, "A review of phase equilibria in Heusler alloy systems containing Fe, Co or Ni," J. Mater. Sci. **51**, 50–70 (2015).
[95] M. Yin, P. Nash, W. Chen, and S. Chen, "Standard enthalpies of formation of selected Ni_2YZ Heusler compounds," J. Alloys Compd. **660**, 258–265 (2016).
[96] M. Yin and P. Nash, "Intermetallics enthalpies of formation of selected Pd_2YZ Heusler compounds," Intermetallics **58**, 15–19 (2015).
[97] M. Yin and P. Nash, "Standard enthalpies of formation of selected Ru_2YZ Heusler compounds," J. Alloys Compd. **634**, 70–74 (2015).
[98] M. Yin and P. Nash, "Standard enthalpies of formation of selected XYZ half-Heusler compounds," J. Chem. Thermodyn. **91**, 1–7 (2015).
[99] K. He, Y. Zhou, P. Gao, L. Wang, N. Pereira, G. G. Amatucci, K. W. Nam, X. Q. Yang, Y. Zhu, F. Wang, and D. Su, "Sodiation via heterogeneous disproportionation in FeF_2 electrodes for sodium-ion batteries," ACS Nano **8**, 7251–7259 (2014).
[100] A. J. Martinolich and J. R. Neilson, "Pyrite formation via kinetic intermediates through low-temperature solid-state metathesis," J. Am. Chem. Soc. **136**, 15654–15659 (2014).
[101] R. D. Bayliss, S. N. Cook, D. O. Scanlon, S. Fearn, J. Cabana, C. Greaves, J. A. Kilner, and S. J. Skinner, "Understanding the defect chemistry of alkali metal strontium silicate solid solutions: Insights from experiment and theory," J. Mater. Chem. A **2**, 17919–17924 (2014).
[102] B. Rousseau, V. Timoshevskii, N. Mousseau, M. Côté, and K. Zaghib, "A novel intercalation cathode material for sodium-based batteries," Electrochem. Commun. **52**, 9–12 (2015).
[103] L. MacEachern, R. A. Dunlap, and M. N. Obrovac, "A combinatorial investigation of Fe-Si-Zn thin film negative electrodes for Li-ion batteries," J. Electrochem. Soc. **162**, A229–A234 (2014).
[104] T. Nagase and Y. Umakoshi, "Amorphous phase formation in Co–Cu–Zr–B-based immiscible alloys," J. Alloys Compd. **649**, 1174–1181 (2015).
[105] Z. Du, T. D. Hatchard, R. A. Dunlap, and M. N. Obrovac, "Combinatorial investigations of Ni-Si negative electrode materials for Li-ion batteries," J. Electrochem. Soc. **162**, A1858–A1863 (2015).
[106] M. Fondell, T. J. Jacobsson, M. Boman, and T. Edvinsson, "Optical quantum confinement in low dimensional hematite," J. Mater. Chem. A **2**, 3352 (2014).
[107] A. J. Cohen, P. Mori-Sánchez, and W. Yang, "Insights into current limitations of density functional theory," Science **321**, 792–794 (2008).
[108] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K. A. Persson, "The materials application programming interface (API): A simple, flexible and efficient API for materials data based on representational state transfer (REST) principles," Comput. Mater. Sci. **97**, 209–215 (2015).
[109] P. Huck, D. Gunter, S. Cholia, D. Winston, A. T. N'Diaye, and K. Persson, "User applications driven by the community contribution framework MPContribs in the materials project," Concurr. Comput. Pract. Exp. (published online 2015).

[110] P. Huck, A. Jain, D. Gunter, D. Winston, and K. Persson, "A community contribution framework for sharing materials data with materials project," in *IEEE 11th International Conference on e-Science* (IEEE, 2015).
[111] S. R. Hall, F. H. Allen, and I. D. Brown, "The crystallographic information file (CIF): A new standard archive file for crystallography," Acta Crystallogr., Sect. A: Found. Crystallogr. **47**, 655 (1991).
[112] M. Folk, G. Heber, Q. Koziol, E. Pourmal, and D. Robinson, "An overview of the HDF5 technology suite and its applications," in *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases (AD'11)* (ACM, 2011), pp. 36–47.
[113] R. Rew and G. Davis, "NetCDF: An interface for scientific data access," IEEE Comput. Graphics Appl. **10**, 76–82 (1990).
[114] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, and F. Yergeau, "Extensible markup language (XML)," World Wide Web J. **2**, 27–66 (1997).
[115] D. Crockford, "The application/json media type for javascript object notation (json)," RFC 7159, 2006.
[116] P. Murray-Rust, H. S. Rzepa, and M. Wright, "Development of chemical markup language (CML) as a system for handling complex chemical content," New J. Chem. **25**, 618–634 (2001).
[117] J. G. Kaufman and E. F. Begley, "MatML: A data interchange markup language," Adv. Mater. Process. **161**, 35–39 (2003).
[118] See http://www.mongodb.org for MongoDB, I. MongoDB.
[119] R. T. Fielding, *Architectural Styles and the Design of Network-based Software Architectures* (University of California, Irvine, 2000).
[120] G. Kresse and J. Furthmüller, "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," Phys. Rev. B **54**, 11169–11186 (1996).
[121] See http://www.citrination.com Citrine Informatics Citrination materials search platform.
[122] See http://citrine.io/mif for Citrine Informatics MIF Schema.
[123] H. E. Pence and A. Williams, "ChemSpider: An online chemical information resource," J. Chem. Educ. **87**, 1123–1124 (2010).
[124] J. A. Warren and R. F. Boisvert, "Building the materials innovation infrastructure: Data and standards," NIST Report No. NISTIR 7898, 2012.
[125] C. H. Ward, J. A. Warren, and R. J. Hanisch, "Making materials science and engineering data more valuable research products," Integr. Mater. Manuf. Innovation **3**, 22 (2014).
[126] T. N. Bhat, L. M. Bartolo, U. R. Kattner, C. E. Campbell, and J. T. Elliott, "Strategy for extensible, evolving terminology for the materials genome initiative efforts," JOM **67**, 1866–1875 (2015).
[127] M. W. Gaultois, T. D. Sparks, C. K. H. Borg, R. Seshadri, W. D. Bonificio, and D. R. Clarke, "Data-driven review of thermoelectric materials: Performance and resource considerations," Chem. Mater. **25**, 2911 (2013).
[128] P. Gorai, D. Gao, B. Ortiz, S. Miller, S. A. Barnett, T. Mason, Q. Lv, V. Stevanović, and E. S. Toberer, "TE design lab: A virtual laboratory for thermoelectric material design," Comput. Mater. Sci. **112**, 368–376 (2016).
[129] G. Hautier, S. S. P. Ong, A. Jain, C. C. J. Moore, and G. Ceder, "Accuracy of density functional theory in predicting reaction energies from binary to ternary oxides and its implication on phase stability," Phys. Rev. B **75**, 155208 (2011).
[130] A. Jain, G. Hautier, S. P. Ong, C. J. Moore, C. C. Fischer, K. A. Persson, and G. Ceder, "Formation enthalpies by mixing GGA and GGA + U calculations," Phys. Rev. B **84**, 045115 (2011).
[131] S. Grindy, B. Meredig, S. Kirklin, J. E. Saal, and C. Wolverton, "Approaching chemical accuracy with density functional calculations: Diatomic energy corrections," Phys. Rev. B **87**, 075150 (2013).
[132] V. Stevanovi, X. Zhang, and A. Zunger, "Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies (FERE)," Phys. Rev. B **85**, 115104 (2011).
[133] S. Lany, "Semiconductor thermochemistry in density functional calculations," Phys. Rev. B **78**, 1–8 (2008).
[134] R. van Noorden, "Interdisciplinary research by the numbers," Nature **525**, 306 (2015).
[135] A. A. White, "Big data are shaping the future of materials science," MRS Bull. **38**, 594–595 (2013).
[136] S. R. Kalidindi and M. De Graef, "Materials data science: Current status and future outlook," Annu. Rev. Mater. Res. **45**, 171–193 (2015).
[137] V. Marx, "The big challenges of big data," Nature **498**, 255–260 (2013).
[138] M. Wilde, M. Hategan, J. M. Wozniak, B. Clifford, D. S. Katz, and I. Foster, "Swift: A language for distributed parallel scripting," Parallel Comput. **37**, 633–652 (2011).
[139] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, and B. Kozinsky, "AiiDA: Automated interactive infrastructure and database for computational science," Comput. Mater. Sci. **111**, 218 (2016).
[140] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter, and K. A. Persson, "FireWorks: A dynamic workflow system designed for high-throughput applications," Concurr. Comput. Pract. Exp. **27**, 5037 (2015).
[141] Y. Bengio, "Learning deep architectures for AI," Found. Trends Mach. Learn. **2**, 1 (2009).
[142] L. Deng and D. Yu, "Deep learning: Methods and applications," Found. Trends Signal Process. **7**, 197–387 (2013).
[143] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," Science **350**, 1332–1338 (2015).
[144] K. Rajan, "Materials informatics," Mater. Today **8**, 35–48 (2005).
[145] G. Mulholland and B. Meredig, "Hackathon aims to solve materials problems," MRS Bull. **40**, 166–167 (2015).
[146] A. White, "Federal agencies announce materials data challenge," MRS Bull. **40**, 906–907 (2015).