UC Santa Barbara
UC Santa Barbara Previously Published Works

Title

Perspective: Interactive material property databases through aggregation of literature data

Permalink

Journal

ISSN

Authors

Seshadri, Ram
Sparks, Taylor D

Publication Date

2016-05-01

DOI

Peer reviewed

Commentary: Interactive materials properties databases through aggregation of literature data

Ram Seshadri[1, a)] and Taylor D. Sparks[2, b)]

[1)] *Materials Research Laboratory, Materials Department,*
and Department of Chemistry and Biochemistry,
University of California, Santa Barbara, California, 93106,
USA

[2)] *Department of Materials Science and Engineering,*
University of Utah, Salt Lake City, Utah, 84112, USA

(Dated: 22 March 2016)

Searchable, interactive, databases of materials properties, particularly those relating to functional materials (magnetics, thermoelectrics, photovoltaics, . . .) are curiously missing from discussions of machine-learning, and other data-driven methods for advancing new materials discovery. Here we discuss the manual aggregation of experimental data from the published literature for the creation of interactive databases that allow the original experimental data as well additional metadata to be visualized in an interactive manner. The databases described involve materials for thermoelectric energy conversion, and for the electrodes of Li-ion batteries. The data can be subject to machine-learning, accelerating the discovery of new materials.

[a)]Electronic mail:ramseshadri@ucsb.edu

[b)]Electronic mail:sparks@eng.utah.edu

I. INTRODUCTION

Over the last decade there has been an increased effort to integrate and enhance collaboration between computational and experimental materials science. For example, Integrated Computational Materials Engineering (ICME) creates virtual testing and design packages across multiple length and time scales with an emphasis on process engineering and material variation.[1,2] Application of ICME has resulted in a 25% reduction in development time for cast aluminum engines, and other lightweight vehicle components at Ford Motor Company with a savings of USD$100 million.[3,4]

The Materials Genome Initiative (MGI), launched in 2011 by U.S. President Obama, can be considered the next evolution of ICME by extending this approach towards the accelerated discovery of new materials.[5] The goal is to "discover, develop, manufacture, and deploy advanced materials at least twice as fast as is possible today, at a fraction of the cost." This will require a collaborative approach between theory, simulation, modeling, and experimentation leveraging both existing and computed data to create iterative screening methodologies to reduce the number of costly and time consuming physical experiments. A prerequisite to this ambitious goal is (i) a theoretical understanding of the physical mechanisms and structure-property relationships governing materials properties, (ii) computational and modeling tools that are sufficiently rapid and calibrated to simulate virtual experiments, and (iii) experimental databases for training the computational tools and validating the screening methodologies thereafter.

In recent years there have been significant advances in identifying materials descriptors and structure-property relationships in a wide variety of materials systems for many different applications.[6–9] Likewise, the tools and software packages, such as density functional theory (DFT),[10,11] for simulated experiments have improved by considering simplified models and potentials,[7,12–14] as well as preliminary screening techniques[15–18] making high-throughput computational modeling more viable.

Notwithstanding these developments, one of the greatest remaining challenges is the creation of materials property databases that meet the requirements necessary to advance the Materials Genome Initiative, and that are based on experimental data. As mentioned above, these databases can serve to train the computational tools. However, in addition, there are many materials properties that are still beyond the reach of computation. As

an example, while DFT calculations are relatively powerful at estimating the operating potential of electrode materials in Li-ion batteries, the ability to predict capacity retention after extensive cycling — a complex convolution of the ionic and electronic conductivity of the material, its interaction with the electrolyte, and its size and morphology — is a distant dream. Experimental inputs are critically needed in situation such as these.

Database creation is difficult for several reasons. For one, the data necessary for optimization of complex, multi-dimensional systems is not readily available and means for automated data integration from primary literature sources do not yet exist.[19] Additionally, there are unresolved questions regarding how to recognize and assign value to emerging data sets.[20] Finally, the database needs to be open-access in a way that can be indexed and searched, and yet it must have safeguards for industrially proprietary data to ensure industrial partnerships necessary for technology transfer.[4,21]

In this commentary, we describe some of the advantages and shortcomings of current approaches for building interactive databases of materials properties and outline some of the best practices, opportunities, and remaining challenges that lie ahead in this emerging field of materials informatics.

II. APPROACHES FOR AGGREGATING DATA FROM LITERATURE

Some of the best examples of open access databases come from the field of crystallography. For example, the widely used Inorganic Crystal Structure Database (ICSD), hosted by FIZ Karlsruhe contains over 180,000 entries on the crystal structures of minerals, metals, and other extended solid inorganic compounds.[22] Other popular structure databases include the Cambridge Structural Database created by the Cambridge Crystallography Data Centre which publishes primarily small molecule organic and metal-organic crystal structures with over 800,000 entries,[23] the Crystallography Open Database (COD) an open access collection of structures with 120,000 entries and limited search infrastructure,[24,25] Pearson's Crystal Structure Database with 274,000 entries,[26] and the Protein Databank for proteins, nucleic acids, and complex assemblies.[27]

Undoubtedly, the success of these crystal structure databases is in part due to early adoption of a standardized format for reporting structures known as the `.cif` (crystallographic information file). This format was proposed by the International Union of Crystallography

(IUCr) in 1991 and updated in 2002.[28,29] Although a few exceptions such as the protein data bank (`.PDB`) and macromolecular (`.mcif`) exist, the original `.cif` file format is supported and often required by journal publishers and thus has found widespread use. A machine-readable, standardized format has been the key for translating information in scholarly publications into new tools of scientific discovery. For example, the International Centre for Diffraction Data merges structural information from `.cif` files into their widely used Powder Diffraction File (PDF) with the Joint Committee on Powder Diffraction Standards (JCPDS). These structural databases are considered among the most frequently and broadly used tools in materials science and have led to exciting discoveries in many fields.[30,31]

Unfortunately, comparable databases for functional materials properties have not been readily available, notwithstanding the enormous number of materials research publications generated each year. For example, no repository exists, even for something as simple as the magnetic or ferroic ordering temperatures of inorganic compounds. In the past, the resources that came closest to what is required is either the CRC Handbook of Chemistry and Physics Online hosted at hbcpnetbase.com or what used to be the Landolt-Börnstein tables hosted by Springer at materials.springer.com, although both of these resources are behind a pay-wall.

In the absence of a single unifying and searchable materials properties database, what does exist is a series of diverse, incomplete, and strongly heterogeneous databases that contain materials properties for very specific applications. In addition, in recent years the ICSD has served as the basis for a number of open access publicly available computationally-generated databases of materials structures and some simulated properties. For example, the Materials Project (`materialsproject.org`) has the ambitious goal of computing materials properties for all known materials using supercomputing clusters.[18] The database currently hosts information from well over 60,000 compounds with many band structures, and some limited elastic tensor, piezoelectric tensor, and Li ion intercalation/conversion properties.[19] Another DFT-based, computationally-generated database of materials properties is Aflow (`aflowlib.org`) with contributing consortium members at a dozen universities worldwide.[32] Aflow has an emphasis on high-throughput computation and accordingly hosts data from over 800,000 compounds with over 72 million calculated materials properties. There is also the Open Quantum Materials Database (OQMD, `oqmd.org`) out of Northwestern University. The OQMD features DFT-calculated thermodynamic and structural properties such
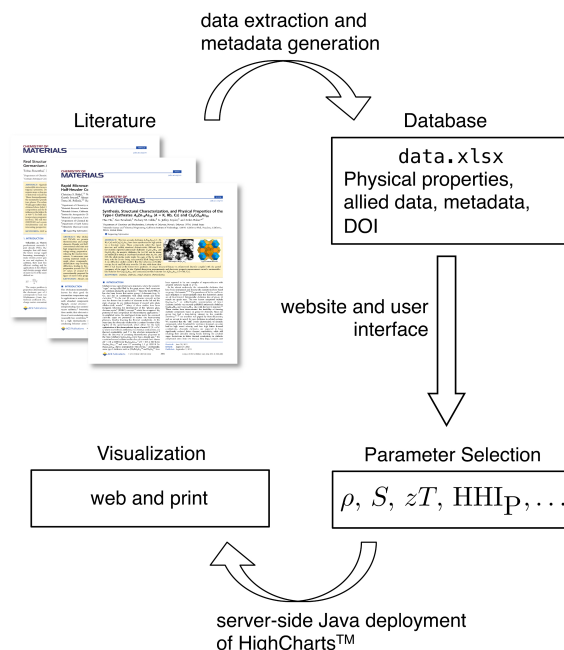
FIG. 1. One possible process for interactive database generation, illustrated for the exmple of thermoelectric materials. Reproduced with permission from Gaultois *et al.* from Ref. 35. Copyright American Chemical Society.

as $0\,\mathrm{K}$ phase diagrams for over 250,000 compounds. The OQMD emphasizes properties for batteries, Mg alloys, and precipitate strengthening, and new ternary compound discovery.[33] Another important database is the Computational Materials Repository which serves as a software infrastructure for collecting, storing, retrieving, analyzing, and sharing of computationally generated materials properties for disparate applications.[34]

The recent proliferation and growth of these DFT-based computational materials properties databases reflects the broader interest and engagement from the scientific community in pursuing data-driven materials discovery. Notwithstanding the value of computationally calculated properties, these databases are no substitute for comparable databases of experimentally measured materials properties. In fact, DFT in practice can only address a *very small percentage* of data/property needs for the entire materials community. A noteworthy recent database infrastructure created to address this critical niche is led by the start-up company, Citrine Informatics (`citrination.org`).

While the materials research community awaits the continued development of a unified materials property database, such as `citrination.org`, it is nevertheless possible to gener-
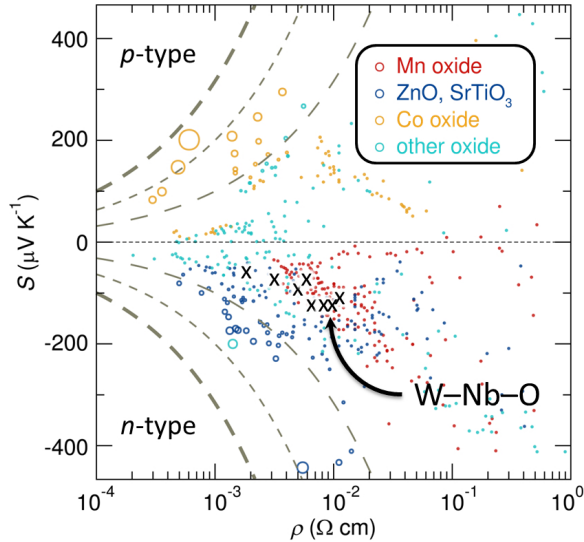
5

FIG. 2. Comparing the thermoelectric performance of early transition metal oxide thermoelectric composites in the W–Nb–O system with the performance of known thermoelectrics (larger circles indicate higher performance). Reproduced with permission from Gaultois *et al.* Ref. 44. Copyright American Institute of Physics.

ate interactive databases for specific applications. Researchers have physically begun to mine data from the literature in order to better understand materials for thermoelectrics,[7,35,36] lithium and lithium-ion batteries,,[37] catalysis,,[21,38] kinetics,[39] and more. The process, shown in Figure 1 for creating interactive databases involves gathering appropriate publications, identifying key data in the publications, and then employing a combination of graduate students and/or post-doctoral fellows, undergraduate interns, and sometimes even high school students to work on data extraction. The process then involves physically entering numbers into a text file or database, frequently after having digitized plots in the publications, using freeware tools such as DATATHIEF.[40] At this stage, metadata such as the crystal structure attributes of the compound being measured, the elemental abundance and availability of the constituents,[35] or the preparation and processing method are entered as well. Finally, the text file is read into web-based visualization suites using software such as HIGHCHARTS,[41] which is freely available for use to academic, not-for-profit entities.

One of the greatest benefits of this type of approach is that enormous amounts of information can be visualized and interpreted using high information density plots. For example, years ago Villars used a series of chemical descriptors such as electronegativity, radii sums,

6

and valence electron counting as axes to generate three-dimensional structural stability diagrams.[42,43] A similarly valuable visualization technique would rely on plotting materials *properties* in addition to chemical descriptors. Information can be encoded not only on the abscissa and ordinate coordinates, but also as the marker size and color creating 4D plots that allow a user to examine large property space at a glance to make decision regarding screening and optimization. An example of this is observed in thermoelectric materials where certain classes of materials have been studied extensively over the years with the community frequently citing that optimization of carrier concentration and microstructure engineering could lead to eventual high performance materials. The insight that can be gained simply by looking at the data, appropriately plotted, cannot be overstated. As one example, the exercise as applied to thermoelectric materials very quickly point out the large regions of parameter space where searching for new high-performing materials would be futile. New materials can also be rapidly displayed in the context of what is known, as demonstrated in Figure 2, showing new classes on thermoelectrics in the space of known ones. To quote the great Yogi Berra: "*you can see a lot just by observing*".

There is great value in constructing *interactive* databases rather than static tables of data. For example, in the thermoelectric and battery databases we have constructed,[45,46] users can select from a list of materials properties or metadata when choosing what parameters to plot for the x and y–axes as well as which parameter to use when scaling the marker size. This ability to select parameters arbitrarily is valuable because it allows users to explore previously unexpected, and unobserved correlations in materials properties. Users are also given options in how to sort the data. In some instances it might be desirable to plot the data sorted as a function of crystal structure or material family, where in other instances it might be more appropriate to plot the data as a function of property measurement condition such as temperature, or C rate (the rate at which the cell is charged/discharged), for example. Once the data has been sorted and assigned marker colors the entire sorted dataset can be toggled on and off via the interactive plot legend. In our recent work we have encoded the plots to allow users to hover over specific data points to expose additional information regarding property measurement and metadata (*i.e.* author, year, synthesis route, processing conditions, comments, temperature etc.)[35,37] Additionally, if a user clicks on a data point it will open a new tab and redirect the user to the primary literature source using the document DOI. Finally, users are given the choice for exporting the plots in either raster
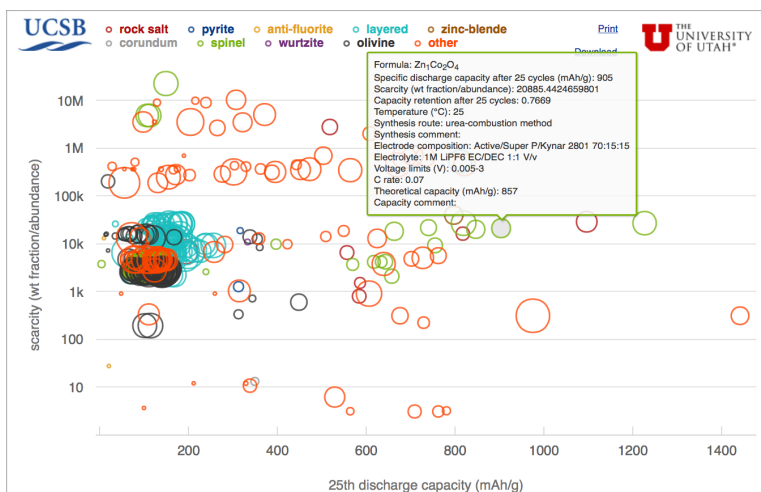
FIG. 3. Screenshot of the battery datamining resource, displaying the constituent element scarcity (inverse crustal abundance) as a function of the reported discharge capacity after the 25th cycle. The symbol size is proportional to the percentage capacity that is retained after 25 cycles (larger is better) and the color indicates the crystal structure type of the active electrode material. Hovering the mouse over data points allows metadata to be read, and clicking on the points takes the reader to the original literature. Website hosted at http://tomcat.eng.utah.edu/sparks/battery.jsp. An equivalent website for thermoelectrics is at http://www.mrl.ucsb.edu:8080/datamine/thermoelectric.jsp.

or vector formats. The interactive nature of the plots, and the ability to use ancillary data that can be added during the aggregation process (elemental scarcity in the example shown here) is displayed in Figure 3.

The data, once available for visualization, can be highly useful for further prediction and identification of trends that might otherwise not be obvious, even to experts who have worked in the field for many years.[48] Additionally, the experimental data itself can serve as a powerful tool and training data set for artificial intelligence and machine-learning algorithms designed to discover new materials. For example, Sparks *et al.* have recently used data mining and machine-learning algorithms to generate ternary diagrams of a predicted probability of thermoelectric performance for hundreds of thousands of compositions and used this tool to explore entirely new compositions and compounds that would not be obvious thermoelectric materials otherwise.[47] An illustration of a previously unexplored composition space of Mn–Ru–Ge, showing the likelihood of low-thermal conductivity materials is displayed in Figure 4.
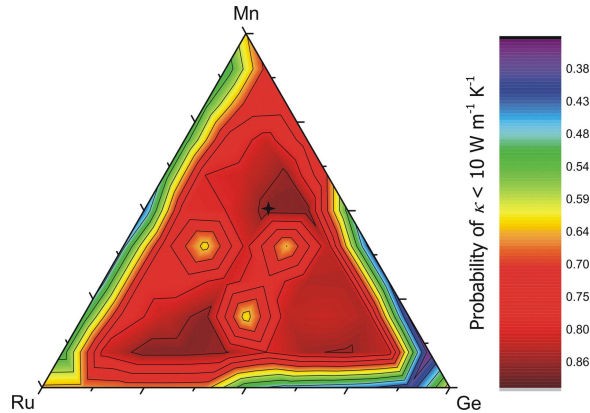
FIG. 4. Machine learning from an experimental database of thermoelectrics, coupled with appropriate DFT calculations, allows low-thermal conductivity phases to the proposed in a previously, relatively unexplored ternary system. Reproduced with permission from Ref. 47. Copyright Elsevier.

Mining data from literature and entering it into a spreadsheet is a time consuming task but not beyond the technical ability of materials researchers. On the other hand, constructing an interactive database with the visualization tools described above requires a software coding background that many materials researchers will not have. One potential solution it so hire third-party software developers to translate the spreadsheet data set into an interactive database, although this can be a somewhat costly and slow process. Alternately, we have developed another solution, a generic Data Visualization Tool,[49] that is both freely available and can be employed almost immediately for any dataset. Users are provided a simple excel spreadsheet and only need to input values for x, y, marker size, and color (how to sort the data). Optionally, the user can provide a DOI and a comment. The appropriately formatted spreadsheet is uploaded, and can be plotted with the option of linear or logarithmic axes.

III. CHALLENGES AND OPPORTUNITIES

Notwithstanding the advances in the databases described thus far, the field has many outstanding questions that will require innovative solutions.

1. How can data acquisition from existing publications be crowd-sourced, or even automated in order to rapidly accelerate the data mining and cataloging process?

2. How can future publications be written in a way to facilitate data mining?

3. What are the prospects of eventually being able to aggregate unpublished data, including failed experiments?

4. What unifying materials property database should be employed and what data infrastructure must it offer?

5. What data should be collected and how can incorrect data be identified for exclusion? What data standards need to be developed? It is important to note that only incorrect data should be excluded. "Bad" or uncorrelated data is actually valuable for developing robust machine learning algorithms that must use all correct, both good and bad, as reference points to make good predictions.

6. How can individual researchers and corporations alike use big data while preventing the loss of intellectual property?

7. How can details such as doping or processing steps be incorporated to fully understand structure-processing-property relationships?

8. How do we overcome the bias in the literature of high performing materials in certain applications?

Some of these challenges have solutions on the horizon, that the authors are aware of. For example, Citrine Informatics (citrination.com) has undertaken the task of developing an open access database that will provide data infrastructure for all materials properties, both calculated and measured experimentally. Likewise, Citrine Informatics have developed a machine vision "figure extraction" tool for automated collection of data from scanned images of existing publications. Funding for MGI related programs has grown and comes from many agencies including the National Science Foundation, Department of Energy, Department of Defense, and National Institute of Standards and Technology, and others.[20] Many funding agencies now require detailed data management plans for collecting and archiving data. Rather than seeing this data management requirement as impractical, or as an onerous burden getting in the way of research, it can be thought of as positive catalyst for accelerating database growth. It may be possible to leverage this new requirement to encourage broader participation in materials properties database growth. Moreover, funding agencies could

play a role in developing or offsetting the cost of existing commercial cloud-based electronic laboratory notebooks where enormous quantities of high-quality yet unpublished work could be eventually data mined.

Other challenges will require a new innovations, investments, and a shift in the scientific culture.[20] For example, to accomplish the MGI goals, direct funding for large scale data infrastructure development and data acquisition is warranted. Dr. Francine Berman, chair of the US branch of the Research Data Alliance, stated *"publicly accessible data requires a stable home and someone to pay the mortgage."*[50]

Drawing from the success of the crystallographic databases, an immediate opportunity for innovation is to employ a standardized, machine-readable `.mif` (Materials Information File) format for submitting materials properties. Going forward, it is clear that the impetus for the creation of such databases must be associated with journal-mandated requirements for the deposition of relevant property data, appropriately curated and formatted, in precisely the same manner that is mandated for crystal structure information. We see an opportunity for professional societies to guide the database formation by convening authoritative committees to set standards and best practices that publishers could then adopt.

One innovative idea for materials data aggregation is to rely on crowd-sourcing from volunteers in the materials community at large in much the same way that ACS' Chemical Abstracts Service did before switching to full-time professional abstractors. Incentives for this volunteer work could come from publishers, professional societies, or private organizations perhaps in the form of formal recognition or reduced fees for services such as journal access or conference registration. However, as with other crowd-sourced content generation, such as Wikipedia, careful data curation from experts will be necessary.

IV.   CONCLUDING REMARKS

The manual, human-intensive approach of abstracting and aggregation data from the literature, while appearing at first sight, primitive and somewhat tedious, can be surprisingly powerful, and even somewhat efficient, once the initial framework is established. The creation of interactive databases then allows such data to be used very effectively, and when appropriately visualized, the collected data can be very revealing and insightful. The irony of requiring to digitize data from the literature — data that was originally created in a

digital format — should not be missed. Clearly, better data curation, for unpublished and published data, and new and improved methods of data archival and retrieval are required. Journal publishers in the area, guided by input from professional societies, could help provide the necessary leadership, by requiring and archiving the raw data associated with plots in publications. However, the greatest challenge is to make available to the world, the vast quantities of unpublished data that never leave the realm of laboratory notebooks.

ACKNOWLEDGMENTS

REFERENCES

[1] J. Allison, D. Backman, and L. Christodoulou, J. Miner. Met. Mater. Soc. **58**, 25 (2006).

[2] J. Allison, J. Miner. Met. Mater. Soc. **63**, 15 (2011).

[3] W. J. Joost, J. Miner. Met. Mater. Soc. **64**, 1032 (2012).

[4] P. Patel, MRS Bull. **36**, 964 (2011).

[5] J. P. Holdren *et al.*, National Science and Technology Council OSTP. Washington, USA (2011).

[6] J. C. Tan and A. K. Cheetham, Chem. Soc. Rev. **40**, 1059 (2011).

[7] J. Yan, P. Gorai, B. Ortiz, S. Miller, S. A. Barnett, T. Mason, V. Stevanović, and E. S. Toberer, Ener. Env. Sci. **8**, 983 (2015).

[8] C. E. Wilmer, O. K. Farha, Y.-S. Bae, J. T. Hupp, and R. Q. Snurr, Ener. Env. Sci. **5**, 9849 (2012).

[9] D. Johrendt, J. Mater. Chem. **21**, 13726 (2011).

[10] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).

[11] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).

[12] K. F. Garrity, J. W. Bennett, K. M. Rabe, and D. Vanderbilt, Comp. Mater. Sci. **81**, 446 (2014).

[13] G. K. Madsen, J. Am. Chem. Soc. **128**, 12140 (2006).

[14] W. Setyawan and S. Curtarolo, Comp. Mater. Sci. **49**, 299 (2010).

[15] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nature Mater. **12**, 191 (2013).

[16] S. Wang, Z. Wang, W. Setyawan, N. Mingo, and S. Curtarolo, Phys. Rev. X **1**, 021012 (2011).

[17] J. Carrete, W. Li, N. Mingo, S. Wang, and S. Curtarolo, Phys. Rev. X **4**, 011019 (2014).

[18] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, Comp. Mater. Sci. **50**, 2295 (2011).

[19] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, APL Mater. **1**, 011002 (2013).

[20] A. White, MRS Bull. **37**, 715 (2012).

[21] J. K. Nørskov and T. Bligaard, Angew. Chem. Intl. Edn. **52**, 776 (2013).

[22] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, Acta Crystallogr B **58**, 364 (2002).

[23] F. H. Allen, Acta Crystallogr B **58**, 380 (2002).

[24] R. T. Downs and M. Hall-Wallace, Amer. Mineral. **88**, 247 (2003).

[25] S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, and A. Le Bail, J. Appl. Crystallogr. **42**, 726 (2009).

[26] P. Villars, *Pearson's crystal data: Crystal structure database for inorganic compounds* (ASM International, 2007).

[27] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucl. Acid Res. **28**, 235 (2000).

[28] S. R. Hall, F. H. Allen, and I. D. Brown, Acta Crystallogr. A **47**, 655 (1991).

[29] I. D. Brown and B. McMahon, Acta Crystallogr. B **58**, 317 (2002).

[30] G. R. Desiraju, Angew. Chem. Intl. Edn. **34**, 2311 (1995).

[31] A. C. Anderson, Chem. Biol. **10**, 787 (2003).

[32] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, *et al.*, Comp. Mater. Sci. **58**, 227 (2012).

[33] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, J. Miner. Met. Mater. Soc. **65**, 1501 (2013).

[34] D. D. Landis, J. S. Hummelshoj, S. Nestorov, J. Greeley, M. Dulak, T. Bligaard, J. K. Norskov, and K. W. Jacobsen, Comp. Sci. Eng. **14**, 51 (2012).

[35]M. W. Gaultois, T. D. Sparks, C. K. Borg, R. Seshadri, W. D. Bonificio, and D. R. Clarke, Chem. Mater. **25**, 2911 (2013).

[36]P. Gorai, D. Gao, B. Ortiz, S. Miller, S. A. Barnett, T. Mason, Q. Lv, V. Stevanović, and E. S. Toberer, Comp. Mater. Sci. **112**, 368 (2016).

[37]L. Ghadbeigi, J. K. Harada, B. R. Lettiere, and T. D. Sparks, Energ. Environ. Sci. (2015).

[38]"Catapp," http://suncat.slac.stanford.edu/catapp/.

[39]W. Mallard, F. Westley, J. Herron, R. Hampson, and D. Frizzell, *NIST chemical kinetics database*, Vol. 126 (National Institute of Standards and Technology, 1992).

[40]B. Tummers, "Datathief III," http://datathief.org (2006).

[41]"Highcharts," http://highcharts.com.

[42]J. Villars, J. Less Common Met. **92**, 215 (1983).

[43]J. Villars, J. Less Common Met. **99**, 33 (1984).

[44]M. W. Gaultois, J. E. Douglas, T. D. Sparks, and R. Seshadri, AIP Adv. **5**, 097144 (2015).

[45]http://www.mrl.ucsb.edu:8080/datamine/thermoelectric.jsp.

[46]http://tomcat.eng.utah.edu/sparks/battery.jsp.

[47]T. D. Sparks, M. W. Gaultois, A. Oliynyk, J. Brgoch, and B. Meredig, Scripta Mater. **111**, 10 (2016).

[48]M. W. Gaultois and T. D. Sparks, Appl. Phys. Lett. **104**, 113906 (2014).

[49]http://www.mrl.ucsb.edu:8080/data/visualization.jsp.

[50]J. Markoff, "How to share scientific data," http://www.nytimes.com/2013/08/13/science/how-to-share-scientific-data.html (2013).