Promises and perils of computational materials databases

Over the past decade, the materials science community has fostered the development of materials databases from high-performance computation. While these databases have achieved great success, there are still several challenges to be addressed for the community to realize the full potential of the materials-by-design era.

M. K. Horton, S. Dwaraknath and K. A. Persson

The ultimate goal of a materials database is to assist in the design of materials. These are materials that will form the basis of the modern world. However, finding the right material for the job can be decades-long, painstaking work. Materials databases, and the methods they rely on, form the foundation of new ways of finding materials to enable a transition from trial-and-error discovery to materials-by-design. This is achieved by allowing a scientist to search the database for their ideal material by desired property, and also by providing a bird's eye perspective that can allow new insights into the physics that governs these materials, and therefore can improve the process by which completely new materials are designed.

The ability to realize these computational databases is relatively new. Historically, the calculation of a single material from first principles might take months or years. However, the inevitable march of high-performance computing, along with key efforts in infrastructure to allow the automation of calculations, has made computing properties of vast quantities of materials feasible in a way that was simply not previously possible. Given the youth of these databases, best practices are still being established, and there is much to learn about both the opportunities they offer and the unique challenges they must confront.

The current success of materials databases

Born out of the Materials Genome Initiative, and now approaching its tenth birthday, the Materials Project[1] is an example of a widely used computational materials database that has been enormously successful. Founded with the goal of creating an open, web-based resource of computed properties of materials, it now has a database of millions of properties of over 100,000 crystalline materials, along with a website and tools that help filter and analyze the data it provides. At the time of writing, the Materials Project has over 150,000 registered users from across academia, education, government and industry, and offers not only properties of materials like thermodynamic stability but also piezoelectric, dielectric and elastic tensors, magnetic orderings, phonon dispersion and more. Crucially, use of data from the Materials Project has led to the real-world synthesis and characterization of many materials, including those for carbon capture[2], phosphors[3], photocatalysts[4], magnetocalorics[5] and thermoelectrics[6], among other applications.

This success is shared by several other excellent computational databases, including, but not limited to, the Open Quantum Materials Database[7], Materials Cloud[8], NOMAD Laboratory[9], AFLOW[10], JARVIS[11], NRELMatDB[12] and numerous additional efforts dedicated to specific classes of materials, such as two-dimensional materials[13], topological materials[14] and organic crystals[15]. Clearly, computational materials databases are here to stay. The questions then become: what is the role of these databases in the future of our community? What new tools and techniques are necessary for them to improve? How do we ensure that they are useful?

Challenges for the years ahead

To answer these questions, we review several vital challenges for computational materials databases.

Literacy of data and methods. An effective materials database needs to make information as easy to access and as understandable as possible, especially given that their audience will necessarily include people without training in computational methods, and the burden is rightly on the database builders to do this. Computational methods are not a cure-all, and there are both physical and practical limitations to the accuracy of their predictions that need to be communicated. For instance, the absence of an error bar in a first-principles calculation can give a false sense of certainty to the unwary, leading to profound misunderstandings. The danger here is twofold: the data are misinterpreted in a way that hinders understanding or progress, or discarded because their true values are obscured. An example of this might be a typical low-cost calculation of the electronic structure of a material, where the magnitude of the bandgap is systematically underestimated, but which generally gives the right character and nature of the electronic transitions present and therefore still provides substantial utility in understanding the underlying physics. Nevertheless, the magnitude of the bandgap itself is presented without an error bar, since the prediction is exact for that level of theory. This is compounded by the fact that databases often have to trade accuracy of prediction with computational speed to achieve necessary scale, so they cannot use the latest, most-accurate methods.

Another issue is that systematic parameter sets are required to run hundreds of thousands of calculations, but these parameter sets might not be ideal in all cases. An example here might be the relative stability of polymorphs of the same material, which are calculated using different choices of exchange-correlation functional, the 'magic' approximation that bridges the divide between independent-particle mean-field methods and quantum reality. Choice of functional has an important effect on which polymorphs are predicted stable, but no choice is completely congruent with observed reality. Or, similarly, predictions are dependent on the degree of self-interaction corrected for using Hubbard methods. Again, which choice is correct? The Materials Project confronted this issue by establishing a scheme to mix calculation results so that the best choice can be made on a per-material basis. Still, this approach is a blunt tool for a subtle problem. The problem remains that there is rarely

a one-size-fits-all solution, and making practical compromises and educating the audience becomes essential.

Confronting bias in machine learning. Data challenges become exacerbated when employing machine learning to investigate big data sets. Any machine learning model begins with data sanitization and, since the data from these databases are largely already sanitized, they make an ideal starting point. The danger here is that ultimately any model trained on these data will contain within it the fundamental limitations of the data themselves. For example, the bandgaps previously mentioned provide a convenient target since computational materials databases can contain an order of magnitude more values than the few thousand known to experiment. But what is being trained here is a model for the underlying simulation technique and not of physical reality itself. Additionally, a tendency towards technologically relevant or naturally abundant materials in the databases might starve the machine learning model of the data that it needs to be appropriately generalizable, and this is often overlooked in the analysis. Another bias not present in nature is that towards crystals with fewer elements and smaller unit cells. While developing these models needn't be the mandate of a computational materials database, it is important to develop methods that remove or, at least, reduce these biases to enable effective use of the data for those who are performing machine learning tasks.

Linking experiment with computation. Experiment serves as an anchor to ensure that computation is relevant. Without it, computation is unmoored, forever simulating ever-more-idealized systems. However, the variables that are easy to modify in an experiment, such as changing temperature, applying a field, or adding a minute addition of another element, can often be very difficult to approach computationally. Also, from the computational side, experimental datasets are often not presented in a way that makes them easy to compare to predictions. To take a concrete example, there is no standardized experimental method to measure and report defect formation energies, which has forced the first-principles defects community to use higher-order methods to establish a 'ground truth', but there is no guarantee these higher-order methods are accurate either. And if they are not, how will computational materials databases build datasets that the materials science community can actually use to engineer or understand defects in real-world materials?

It is also important to calculate properties that are useful to scientists. Computation can often readily provide a specific number, such as a prediction of formation enthalpy, while failing to pay attention to the underlying, motivating question, such as 'can I make this?' In this case, methods[16,17] have been developed to try to answer whether a material might be synthesizable or, at least, to establish when a material will not be synthesizable[18]. However, these methods are expensive to apply at scale, and even when they provide a tenuous answer, they do not share the means by which the synthesis might be performed. Developing and applying new methods that try to address the underlying questions, rather than just calculating those properties that are easy to calculate, is crucial for materials databases to remain relevant for practical applications. And dialogue, trust and mutual learning between both computational and experimental communities will be essential to make this happen.

Availability and reproducibility of data. The final challenge is in the accessibility of the database itself. Science has to be reproducible to be useful, but many computational results simply are not. In scientific databases, one proposal has been an aspirational set of standards, 'FAIR', for datasets that are 'findable, accessible, interoperable and reusable'[19], to ensure that the datasets live on and that we avoid a scientific digital dark age. However, many current databases fall short of these aspirations. Old data can become lost, corrupted or lack the metadata to query it effectively. Data can change as new, better techniques become available, and publications can refer to older data that are now inaccessible. To help address this, the Materials Project was the first within our community to provide a modern application programming interface (API) to help improve data access and allow scientists retrieve mass data on demand and, more recently, a consortium of materials databases has come together to provide a community-standard API[20], which will be an essential next step to allow interoperability between databases.

Looking forward, one promising area of development to improve database accessibility is the construction of ontologies for materials science, such as the European Materials and Modelling Ontology (EMMO). Ontologies have been successful in bioinformatics to provide cross-database connectivity by establishing clear definitions of what individual data entries represent. For materials science, ontologies could differentiate similar concepts without destroying their connectivity. For instance, both an experimental and computational bandgap would be part of an ontology. Bandgaps from specific measurement techniques or computational methods would create sub-terms that enable understanding of each term. By linking to a canonical definition, data from multiple databases can be mixed more easily. Most importantly, ontologies provide a machine-understandable construct of how data are connected. This is critical to enable the complex search that scientists truly desire when looking at a large corpus of materials data, but involves not only a technological solution but also a willingness by the community to adopt such systems.

Concluding remarks

What do these challenges ultimately mean? It means we have work to do. It means that our field is a vibrant one, with much shared opportunity. In all this, it's also notable what challenges are absent: the good news is that computers and algorithms will become ever faster, and computational limitations will fall away. What has taken the past decade to create in the Materials Project will one day be achievable in weeks, and this will be a good thing since it will open up new avenues for research. And let us not lose sight of the end goal: a resource with all materials real and imagined, a Library of Babel, or Hitchhiker's Guide to crystals unknown. What a wonderful resource that will be. ❐

M. K. Horton[1,2], S. Dwaraknath[1] and K. A. Persson[2,3] ✉

[1]*Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.*
[2]*Department of Materials Science and Engineering, University of California, Berkeley, CA, USA.*
[3]*Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.*
✉*e-mail: kapersson@lbl.gov*

Published online: 14 January 2021
https://doi.org/10.1038/s43588-020-00016-5

References
1. Jain, A. et al. in *Handbook of Materials Modeling* (eds Andreoni, W. & Yip, S.) 1–34 (Springer, 2018).
2. Dunstan, M. T. et al. *Energy Environ. Sci.* **9**, 1346–1360 (2016).
3. Li, S. et al. *Chem. Mater.* **31**, 6286–6294 (2019).
4. Yan, Q. et al. *Proc. Natl Acad. Sci. USA* **114**, 3040–3043 (2017).
5. Cooley, J. A. et al. *Chem. Mater.* **32**, 1243–1249 (2020).
6. Zhu, H. et al. *J. Mater. Chem. C* **3**, 10554–10565 (2015).
7. Saal, J. E. et al. *JOM* **65**, 1501–1509 (2013).
8. Leopold, T. et al. *Sci. Data* **7**, 299 (2020).
9. Draxl, C. & Scheffler, M. *J. Phys. Mater.* **2**, 036001 (2019).
10. Curtarolo, S. et al. *Comput. Mater. Sci.* **58**, 218–226 (2012).
11. Choudhary, K. et al. *npj Comput. Mater.* **6**, 173 (2020).
12. Stevanović, V., Lany, S., Zhang, X. & Zunger, A. *Phys. Rev. B* **85**, 115104 (2012).
13. Zhou, J. et al. *Sci. Data* **6**, 86 (2019).
14. Vergniory, M. G. et al. *Nature* **566**, 480–485 (2019).

15. Borysov, S. S., R. Geilhufe, R. M. & Balatsky, A. V. *PLoS ONE* **12**, e0171501 (2017).
16. Aykol, M. et al. *Nat. Commun.* **10**, 2018 (2019).
17. Stevanović, V. *Phys. Rev. Lett.* **116**, 075503 (2016).
18. Aykol, M., Dwaraknath, S. S., Sun, W. & Persson, K. A. *Sci. Adv.* **4**, eaaq0148 (2018).
19. Wilkinson, M. D. et al. *Sci. Data* **3**, 160018 (2016).
20. Andersen, C. et al. *The OPTIMADE Specification* (Zenodo, 2020); https://doi.org/10.5281/zenodo.4195050

Acknowledgements
M.K.H., S.D. and K.A.P. acknowledge support by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under contract no. DE-AC02-05-CH11231 (Materials Project program KC23MP).

Competing interests
The authors declare no competing interests.