a striking increase in the level of lamin A was observed as a consequence of the treatment. Crucially, compared with the model animals that did not undergo gene editing, those that received the base editor aged with remarkably fewer abnormalities in the usually lifespan-limiting cardiovascular system. These animals also had greater vitality (a better ability to move and a better overall appearance) and a statistically significant lifespan extension.

By directly addressing the root cause of the disease, base editing could offer great advantages over current drug-based therapeutic strategies. Many key questions remain to be answered, however, before people might benefit from the introduction of this technology. For example, what is the optimal distribution of base editor mediated by adeno-associated virus or by other delivery methods? And which organs can be targeted? The adeno-associated virus injection strategy was less efficient in targeting the skin than in targeting other mouse organs.

To what extent can the genetic defect be corrected? High efficiency of editing might be crucial, particularly for efforts to treat other diseases. Previous attempts[11,12] to use gene editing to address the defects underlying Duchenne muscular dystrophy and rare liver diseases met with only limited success. However, Koblan and colleagues' work indicates that correction does not need to reach 100% efficiency to provide positive benefits, opening the possibility of reconsidering this approach for some other diseases, too.

Another important question is whether an immune response might develop that would target components of the editing system. Such a response might result in inefficient treatment if cells harbouring editing components were selectively eliminated[11].

What about long-term considerations? For example, would a single administration of the editors be sufficient? And what would be the best age for treatment to be administered? Progeria is diagnosed relatively early in life compared with many other diseases for which base editing is a possibility. A treatment age of two weeks, for mice, is therefore much lower than the equivalent age, for humans, at which many diseases are diagnosed. Moreover, with animal testing, it is difficult to benchmark the equivalent human age that corresponds to a mouse of a given age. Finally, how would the current drug therapy available for progeria fit with the potential of repair by base editing in a treatment plan?

If base editing is to be used to treat human disease, the safety of such an intervention must be ensured. If it can be, and if this method successfully repairs the progeria-causing alteration in the crucial tissues, such an approach holds tremendous promise as a way of prolonging health, extending lifespan and improving the quality of life of those who have this mutation.

Wilbert P. Vermeij is at the Princess Máxima Center for Pediatric Oncology, Oncode Institute, 3584 CS Utrecht, the Netherlands. **Jan H. J. Hoeijmakers** is in the Molecular Genetics Department, Erasmus MC, 3015 GD Rotterdam, the Netherlands, and also at the Princess Máxima Center for Pediatric Oncology, Oncode Institute, and at the Institute for Genome Stability in Ageing and Disease, CECAD Research Centre, Cologne, Germany.
e-mails: w.p.vermeij@prinsesmaximacentrum.nl; j.hoeijmakers@erasmusmc.nl

1. Koblan, L. W. *et al. Nature* **589**, 608–614 (2021).
2. Niedernhofer, L. J. *et al. Annu. Rev. Biochem.* **87**, 295–322 (2018).
3. Merideth, M. A. *et al. N. Engl. J. Med.* **358**, 592–604 (2008).
4. Eriksson, M. *et al. Nature* **423**, 293–298 (2003).
5. Kumar, A. *et al. Cell* **158**, 633–646 (2014).
6. Nava, M. M. *et al. Cell* **181**, 800–817 (2020).
7. Liu, B. *et al. Nature Med.* **11**, 780–785 (2005).
8. Harhouri, K. *et al. Nucleus* **9**, 265–276 (2018).
9. Beyret, E. *et al. Nature Med.* **25**, 419–422 (2019).
10. Santiago-Fernández, O. *et al. Nature Med.* **25**, 423–426 (2019).
11. Porto, E. M., Komor, A. C., Slaymaker, I. M. & Yeo, G. W. *Nature Rev. Drug. Discov.* **19**, 839–859 (2020).
12. Katrekar, D. *et al. Nature Methods* **16**, 239–242 (2019).

This article was published online on 6 January 2021.

Computational materials science

Machine learning from diverse data sources

Rohit Batra

A strategy for machine learning has been developed that exploits the fact that data are often collected in different ways with varying levels of accuracy. The approach was used to build a model that predicts a key property of materials.

Scientists are always hunting for materials that have superior properties. They therefore continually synthesize, characterize and measure the properties of new materials using a range of experimental techniques. Computational modelling is also used to estimate the properties of materials. However, there is usually a trade-off between the cost of the experiments (or simulations) and the accuracy of the measurements (or estimates), which has limited the number of materials that can be tested rigorously. Writing in *Nature Computational Science*, Chen *et al.*[1] report a machine-learning approach that combines data from multiple sources of measurements and simulations, all of which have different levels of approximation, to learn and predict materials' properties. Their method allows the construction of a more general and accurate model of such properties than was previously possible, thereby facilitating the screening of promising material candidates.

Materials scientists commonly supplement their own 'chemical intuition' with predictions from machine-learning models, to decide which experiments to conduct next[2,3]. For example, artificial intelligence has been used to identify candidate compounds that act as superconductors at high temperatures[4], electrolyte materials that conduct electric currents using lithium ions[5], and electrically insulating polymers that can withstand large electric fields without breaking down[6]. Artificial intelligence has also been used to work out ways of synthesizing materials — that is, to suggest which reagents, catalysts and experimental conditions to use[7].

Most of these studies involve supervised learning, in which a machine is exposed to a large volume of historical data about the chemical composition or atomic structure of materials, and their associated properties, to build a model that can predict the properties of other materials. More importantly, almost all of these studies use models built on data obtained from a single, consistent source. Such models are referred to as single-fidelity models.

However, for most real-world applications, measurements of materials' properties have varying levels of fidelity, depending on the resources available. For instance, the most accurate (high-fidelity) measurements of properties of crystalline materials are made using single crystals, which can be laborious to prepare. Approximate (low-fidelity) measurements are therefore often made using easily synthesizable polycrystalline samples. Similarly, a hierarchy of increasingly accurate, but progressively more expensive, computational modelling schemes are used to calculate materials' properties — high-fidelity modelling is therefore often restricted by its cost.

Overall, this variation in measurement and modelling techniques leads to a heterogeneous data structure — low-fidelity measurements are plentiful, whereas high-fidelity data are sparse. Notably, each type of fidelity has its own advantages: low-fidelity

data cover a diverse range of compounds, whereas high-fidelity data are more accurate. Single-fidelity models trained using only low- or high-fidelity data therefore miss out on high accuracy or generality, respectively (Fig. 1).

Chen *et al.* now report an adaptation of an artificial neural network (a brain-inspired computer system) that they call a multi-fidelity graph network. This can learn about materials' properties using data acquired from different modelling and experimental techniques. As a proof of principle, the authors trained their graph network to learn about bandgaps — a property that controls several electrical and optical properties of solid materials, such as their conductivity and colour. They used bandgap data from five sources: four data sets were the results of different types of quantum-mechanical calculation, and the fifth source was experimental data. The data set that had the lowest fidelity level contained about 50,000 data points, roughly 100 times more than the number in each of the other data sets; this is typical of the heterogeneity of available data in materials science.

The authors' graph network takes a materials graph — a mathematical representation of the structure of a material, consisting of nodes that represent atoms and edges that represent bonds — as an input. It then performs a series of mathematical (convolution) operations to exchange information between its nodes and edges. This produces an output vector known as a latent representation, which is passed on to, and further manipulated by, another artificial neural network to predict the property of interest (in this case, the bandgap). One or more historical data sets are used first to simultaneously train the materials graph and the second artificial neural network, priming them to make predictions.

Machine-learning techniques based on graph networks are among the top-performing methods for single-fidelity learning of materials properties, and do not require a feature-engineering step (in which a material's composition and/or atomic structure is converted into a string of numbers in a machine-readable format), as is necessary for other machine-learning algorithms[8]. To adapt their graph network for multi-fidelity learning, the authors introduced a new variable, in addition to those used to represent graph nodes and edges, that accounts for the fidelity level of a data point. The authors' graph network therefore exchanges information between the atom nodes, the bond edges and the data-fidelity level represented by the new variable. This means that their approach is applicable to any number of fidelity levels.

A comparison of prediction errors clearly demonstrates the benefit of the multi-fidelity approach. For example, models that had four levels of fidelity reduced errors in predictions of bandgaps by 22–45%, compared with single-fidelity models. Similarly, multi-fidelity models involving two, three or five levels of fidelity performed better than did single-fidelity models.

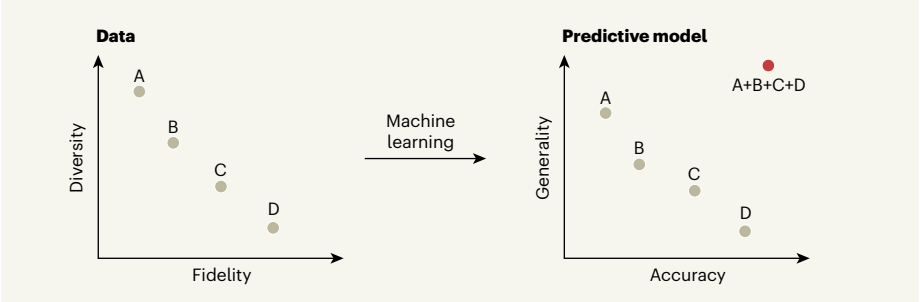

Figure 1 | Multi-fidelity data can improve predictive models developed using machine learning. Accurate (high-fidelity) data about the properties of materials can be difficult or expensive to acquire, and so less-accurate (low-fidelity) data are often obtained instead. Low-fidelity data sets are therefore usually larger than high-fidelity ones, and represent a greater diversity of materials. Machine-learning systems typically use individual data sets to generate predictive models of materials' properties. Diverse low-fidelity data produce general but approximate models, whereas high-fidelity data produce accurate but less-general ones. Here, machine learning based on individual data sets A to D produces four predictive models, the generality and accuracy of which are shown. Chen *et al.*[1] report a machine-learning architecture that can process materials data from multiple sets that have different fidelities, and thereby generates predictive models that are more general and accurate than are those produced using the individual data sets; the red spot indicates the generality and accuracy of a model trained using the combined data sets A to D. Examples are illustrative, and do not depict actual data.

This improvement can be attributed to two key factors. First, the large volume of low-fidelity data represents a more chemically diverse collection of materials than does a single high-fidelity data set; exposure of the graph network to this diversity results in a better and more-general latent representation. Second, there is a high correlation between the low- and high-fidelity bandgap data — many of the bandgaps in the low-fidelity data set are close in value to the equivalent data points in the high-fidelity data sets. This second factor is evident from the higher prediction accuracy that is achieved when using high-fidelity data sets that correlate more closely with the low-fidelity data set.

Chen and colleagues' approach overcomes the limitations of other multi-fidelity approaches[9,10], which are either not easily scalable to large data sets, or cannot handle heterogeneous data or more than two levels of fidelity. The authors' multi-fidelity graph network is therefore a powerful new system for capturing complex relationships between data sets of multiple fidelities. It should be noted, however, that Chen *et al.* did not explore what happens if low- and high-fidelity data points are weighted differently. Such weighting might become necessary when the number of low-fidelity points is so large that it over-represents the full set of multi-fidelity data.

The authors' system is not restricted to materials science, but is generalizable to any problem that can be described using graph structures, such as social networks and knowledge graphs (digital frameworks that represent knowledge as concepts connected by relationships). Furthermore, this approach could, in principle, be used to learn about multiple properties simultaneously (multi-task learning), rather than learning about just one property for which data are available at multiple levels of fidelity.

However, some fundamental questions remain. Are multi-fidelity approaches guaranteed to perform better than single-fidelity models, even when the quality of the low-fidelity data is extremely poor? And what happens when low- and high-fidelity data points are poorly correlated? More research is needed to understand the scenarios for which multi-fidelity learning is most beneficial, balancing prediction accuracy with the cost of acquiring data. In the meantime, the popularity of multi-fidelity methods will surely increase, because they directly exploit the underlying widespread heterogeneity of data in the materials and chemical sciences.

Rohit Batra is at the Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois 60439, USA.
e-mail: rbatra@anl.gov

1. Chen, C., Zuo, Y., Ye, W., Li, X. & Ong, S. P. *Nature Comput. Sci.* **1**, 46–53 (2021).
2. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. *Nature* **559**, 547–555 (2018).
3. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. *npj Comput. Mater.* **3**, 54 (2017).
4. Stanev, V. *et al. npj Comput. Mater.* **4**, 29 (2018).
5. Sendek, A. D. *et al. Chem. Mater.* **31**, 342–352 (2019).
6. Mannodi-Kanakkithodi, A. *et al. Mater. Today* **21**, 785–796 (2018).
7. Huo, H. *et al. npj Comput. Mater.* **5**, 62 (2019).
8. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. *Chem. Mater.* **31**, 3564–35 72 (2019).
9. Rohit, B., Pilania, G., Uberuaga, B. P. & Ramprasad, R. *ACS Appl. Mater. Interf.* **11**, 24906–24918 (2019).
10. Dahl, G. E., Jaitly, N. & Salakhutdinov, R. Preprint at https://arxiv.org/abs/1406.1231 (2014).

This article was published online on 14 January 2021.