THE MATERIAL CODE

Machine-learning techniques could revolutionize how materials science is done.

BY NICOLA NOSENGO

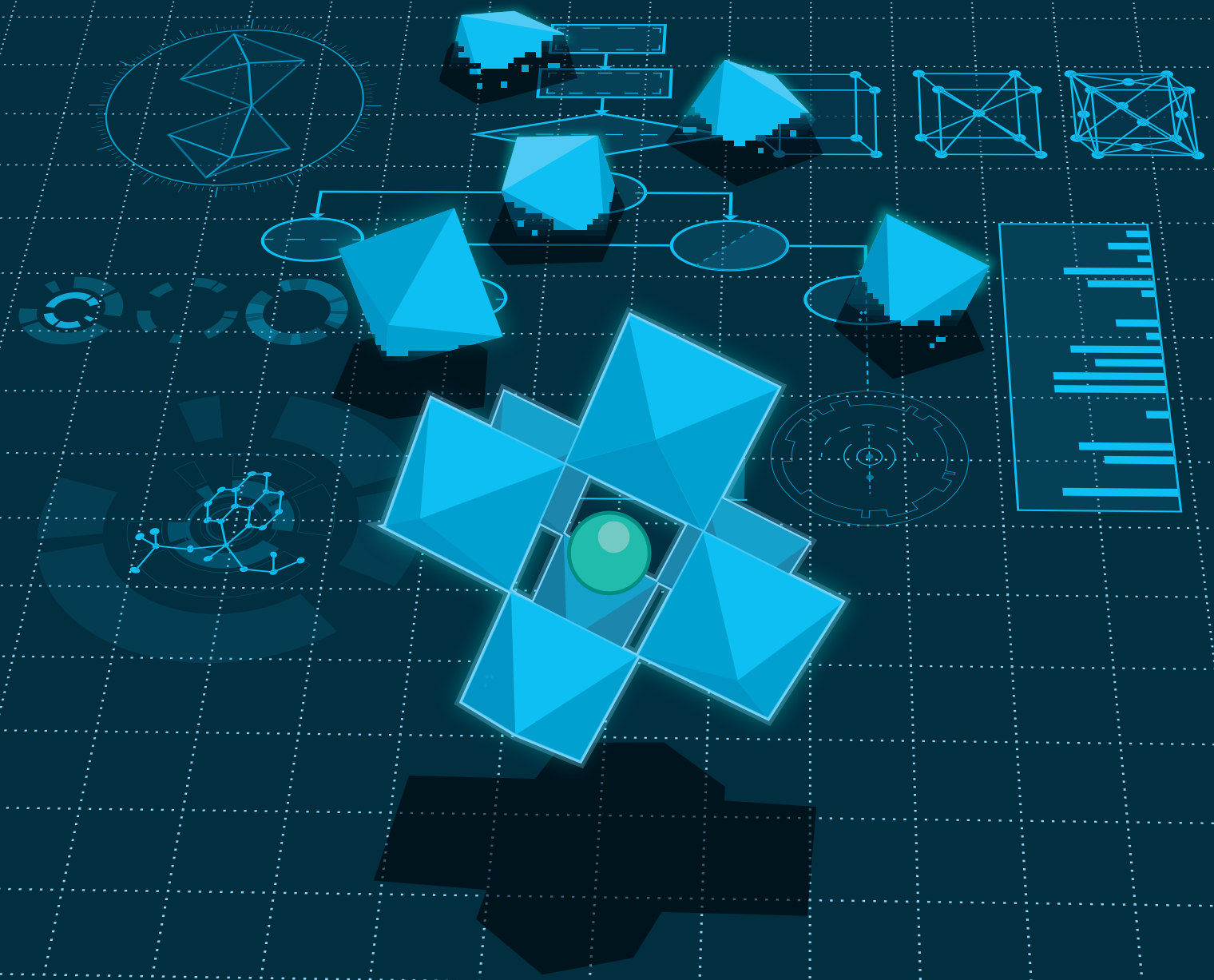


ILLUSTRATION BY NIK SPENCER/NATURE

It's a strong contender for the geekiest video ever made: a close-up of a smartphone with line upon line of numbers and symbols scrolling down the screen. But when visitors stop by Nicola Marzari's office, which overlooks Lake Geneva, he can hardly wait to show it off. "It's from 2010," he says, "and this is my cellphone calculating the electronic structure of silicon in real time!"

Even back then, explains Marzari, a physicist at the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland, his now-ancient handset took just 40 seconds to carry out quantum-mechanical calculations that once took many hours on a supercomputer — a feat that not only shows how far such computational methods have come in the past decade or so, but also demonstrates their potential for transforming the way materials science is done in the future.

Instead of continuing to develop new materials the old-fashioned way — stumbling across them by luck, then painstakingly measuring their properties in the laboratory — Marzari and like-minded researchers are using computer modelling and machine-learning techniques to generate libraries of candidate materials by the tens of thousands. Even data from failed experiments can provide useful input[1]. Many of these candidates are completely hypothetical, but engineers are already beginning to shortlist those that are worth synthesizing and testing for specific applications by searching through their predicted properties — for example, how well they will work as a conductor or an insulator, whether they will act as a magnet, and how much heat and pressure they can withstand.

The hope is that this approach will provide a huge leap in the speed and efficiency of materials discovery, says Gerbrand Ceder, a materials scientist at the University of California, Berkeley, and a pioneer in this field. "We probably know about 1% of the properties of existing materials," he says, pointing to the example of lithium iron phosphate: a compound that was first synthesized[2] in the 1930s, but was not recognized[3] as a promising replacement material for current-generation lithium-ion batteries until 1996. "No one had bothered to measure its voltage before," says Ceder.

At least three major materials databases already exist around the world, each encompassing tens or hundreds of thousands of compounds. Marzari's Lausanne-based Materials Cloud project is scheduled to launch later this year. And the wider community is beginning to take notice. "We are now seeing a real convergence of what experimentalists want and what theorists can deliver," says Neil Alford, a materials scientist who serves as vice-dean for research at Imperial College London, but who has no affiliation with any of the database projects.

As even the proponents are quick to point out, however, the journey from computer predictions to real-world technologies is not an easy one. The existing databases are far from including all known materials, let alone all possible ones. The data-driven discovery works well for some materials, but not for others. And even after an interesting material is singled out on a computer, synthesizing it in a laboratory can still take years. "We often know better what we should be making than how to make it," says Ceder.

Still, researchers in this field are confident that there is a trove of compounds waiting to be discovered, which could kick-start innovations in electronics, energy, robotics, health care and transportation. "Our community is putting together a lot of different parts of the puzzle," says Giulia Galli, a computational materials scientist at the University of Chicago in Illinois. "And when they all click into place, materials prediction will become a reality."

"WE ARE NOW SEEING A REAL CONVERGENCE OF WHAT EXPERIMENTALISTS WANT AND WHAT THEORISTS CAN DELIVER."

GENETIC INSPIRATION

The idea for this high-throughput, data-driven approach to materials discovery hit Ceder in the early 2000s, when he was at the Massachusetts Institute of Technology (MIT) in Cambridge and found himself inspired by the nearly completed Human Genome Project. "By itself, the human genome was not a recipe for new treatments," he says, "but it gave medicine amazing amounts of basic, quantitative information to start from." Could materials scientists learn some lessons from geneticists, he wondered. Could they identify a 'materials genome' that encodes the properties of various compounds in the same way that biological information is encoded in DNA base pairs?

If so, he reasoned, that encoding must lie in the atoms and electrons that make up a given material, and in their crystal structure: the way they are arranged in space. In 2003, Ceder and his team first showed[4] how a database of quantum-mechanics calculations could help to predict the most likely crystal structure of a metal alloy — a key step for anyone in the business of inventing new materials.

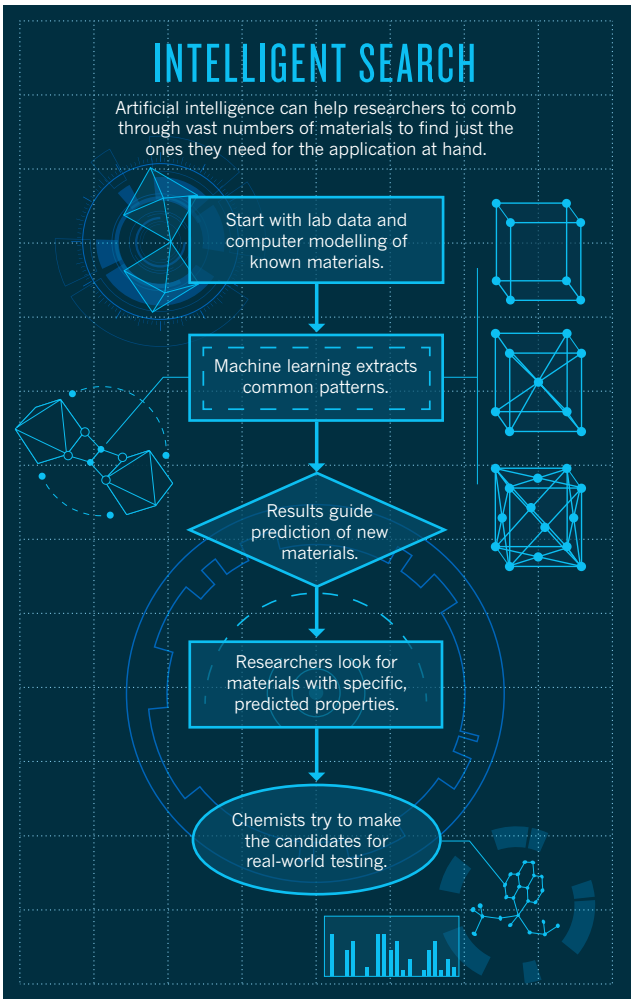
In the past, these calculations had been long and difficult, even for supercomputers. The machine had to go through an inordinate amount of trial and error to find the 'ground state': the crystal structure and electron configuration in which the energy was at a minimum and all the forces were in equilibrium. But in their 2003 paper[4], Ceder's team described a shortcut. The researchers calculated the energies of common crystal structures for a small library of binary alloys — mixes of two different metals — and then designed a machine-learning algorithm that could extract patterns from the library and guess the most likely ground state for a new alloy. The algorithm worked well, slashing the computer time required for the calculations (see 'Intelligent search').

"That paper introduced the idea of a public library of materials properties, and of using data mining to fill the missing parts," says Stefano Curtarolo, who that same year left Ceder's group to start his own laboratory at Duke University in Durham, North Carolina. The idea then gave birth to two separate projects. In 2006, Ceder started the Materials Genome Project at MIT, using improved versions of the algorithm to predict lithium-based materials for electric-car batteries. By 2010, the project had grown to include around 20,000 predicted compounds. "We started from existing materials and modified their crystal structure — changing one element here or another one there and calculating what happens," says Kristin Persson, a former member of Ceder's team who continued to collaborate on the project after she moved to the Lawrence Berkeley National Laboratory in California in 2008.

At Duke, meanwhile, Curtarolo set up the Center for Materials Genomics, which focused on research on metal alloys. Teaming up with researchers from Brigham Young University in Provo, Utah, and Israel's Negev Nuclear Research Center, he gradually expanded the 2003 algorithm and library into AFLOW, a system that can perform

calculations on known crystal structures and predict new ones automatically[5].

Researchers from outside the original group were getting interested in high-throughput computations as well. One such researcher was chemical engineer Jens Nørskov, who started using them to study catalysts for breaking down water into hydrogen and oxygen[6] while he was at the Technical University of Denmark in Lyngby, and later expanded the work as director of the SUNCAT Center for the computational study of catalysis at Stanford University in California. Another was Marzari, who was part of a large team developing Quantum Espresso: a program for quantum-mechanics calculations that was launched[7] in 2009. That is the code running on his mobile phone in the video.



NIK SPENCER/NATURE

MATERIALS GENOMICS

Still, computational materials science did not become mainstream until June 2011, when the White House announced the multimillion-dollar Materials Genome Initiative (MGI). "When people at the White House became familiar with Ceder's work they got very excited," says James Warren, a materials scientist at the US National Institute of Standards and Technology and executive secretary of the MGI. "There was a general awareness that computer simulations had got to the point where they could have a real impact on innovation and manufacturing," he says — not to mention the 'genomics' name, "which was evocative of something grand."

Since 2011, the initiative has invested more than US$250 million into software tools, standardized methods to collect and report experimental data, centres for computational materials science at major universities and partnerships between universities and the business sector for research on specific applications. But it is unclear how far this largesse has actually advanced the science. "The initiative brought a lot of good things, but also some re-branding," says Ceder. "Some groups started calling their research genomics this and genomics that, even though it had little to do with it."

One thing the MGI definitely did do, however, was to help Ceder and others realize their vision of an online database of materials properties. In late 2011, Ceder and Persson relaunched their Materials Genome Project as the Materials Project — having been asked by the White House to give up the 'genome' label to avoid confusion with the national effort. The following year, Curtarolo posted his own database, called AFLOWlib, based on the software he had developed at Duke[8]. And in 2013, Chris Wolverton, a materials researcher at Northwestern University in Evanston, Illinois, launched the Open Quantum Materials Database (OQMD)[9]. "We borrowed the general idea from the Materials Project and AFLOWlib," says Wolverton, "but our software and data are homegrown."

All three of these databases share a core of around 50,000 known materials taken from a widely used experimental library, the Inorganic Crystal Structure Database. These are solids that have been created at least once in a laboratory and described in a paper, but whose electronic or magnetic properties may have never been fully tested; they are the starting point from which new materials can be derived.

Where the three databases differ is in the hypothetical materials they include. The Materials Project has relatively few, starting with some 15,000 computed structures derived from Ceder's and Persson's research on lithium batteries. "We only include them in the database if we're confident the calculations are accurate, and if there is a reasonable chance that they can be made," says Persson, who is now director of the Materials Project and has a joint affiliation with the University of California, Berkeley. Another 130,000 or so entries are structures predicted by the Nanoporous Materials Genome Center at the University of Minnesota in Minneapolis. The latter focuses on zeolites and metal–organic frameworks: sponge-like materials with regularly repeating holes in their crystal structures that can trap gas molecules and could be used to store methane or carbon dioxide.

AFLOWlib is the largest database, featuring more than a million different materials and about 100 million calculated properties. That's because it also includes hundreds of thousands of hypothetical materials, many of which would exist for only a fraction of a second in the real world, says Curtarolo. "But it pays off when you want to predict how a material can actually be manufactured," he says. For example, he is using data from AFLOWlib to study why some alloys can form metallic glass — a peculiar form of metal with a disordered microscopic structure that gives it special electric and magnetic properties. It turns out that the difference between good glass formers and bad ones depends on the number and energies of unstable crystal structures that 'compete' with the ground state while the alloy cools down[10].

Wolverton's OQMD includes around 400,000 hypothetical materials, calculated by taking a list of crystal structures commonly observed in nature and 'decorating' them with elements chosen from almost every part of the periodic table[9]. It has a particularly wide coverage of perovskites — crystals that often display attractive properties such as superconductivity and that are being developed for use in solar cells as microelectronics. As the name suggests, this project is the most open of the three: users can download the entire database, not just individual search results, onto their computer.

All of these databases are works in progress, and their curators still spend a good share of their time adding more compounds and refining the calculations — which, they admit, are far from perfect. The codes tend to be quite good at predicting whether a crystal is stable or not, but less good at predicting how it absorbs light or conducts electricity — to the point of sometimes making a semiconductor look like a

metal. Marzari notes that even for battery materials, an area in which computational materials science is having its best success stories, standard calculations still have an average error of half a volt, which makes a lot of difference in terms of performance. "The truth is, some errors come with the theory itself: we may never be able to correct them," says Curtarolo.

Each group is developing its own techniques to adjust the calculations and make up for these systematic errors. But in the meantime they are already doing science with the data — and so are users from other groups. The Materials Project has identified several promising cathodes that may work better than existing ones in lithium batteries[11], as well as metal oxides that could improve the efficiency with which solar cells capture sunlight and turn it into energy[12]. And earlier this year, researchers from Trinity College Dublin used the AFLOWlib database to predict 20 Heusler alloys, a class of magnets that can be used for sensors or computer memories, and managed to synthesize two of them, confirming that their magnetic properties are very close to the predictions (see go.nature.com/v7djio).

"THE TRUTH IS, SOME ERRORS COME WITH THE THEORY ITSELF: WE MAY NEVER BE ABLE TO CORRECT THEM."

EUROPEAN EXPANSION

Materials genomics has also crossed over to Europe — although usually by other names. Switzerland, for example, has created MARVEL, a network of institutes for computational materials science with the EPFL as its lead and Marzari as director. Using a new computational platform[13], he is creating a database called Materials Cloud that he is using to search for 'two-dimensional' materials, such as graphene, that are made from just a single layer of atoms or molecules. Such materials could be used in applications ranging from nanoscale electronics to biomedical devices. To find good candidates, Marzari is subjecting more than 150,000 known materials to what he calls 'computational peeling': calculating how much energy it would take to separate a single layer from the surface of an ordinary crystal. By the time the database is ready for public release later this year, he expects that preliminary runs will have yielded some 1,500 potential two-dimensional structures that can then be tested in experiments.

A few kilometres away in Sion, high in the Swiss Alps, computational chemist Berend Smit has set up another EPFL centre that develops algorithms for predicting hundreds of thousands of nanoporous zeolites and metal–organic frameworks. Other algorithms — including one that scans for certain pore shapes using techniques derived from facial-recognition software — then seek out the best candidates for absorbing carbon dioxide from the flues of fossil-fuel power plants[14].

Smit's work also shows that materials genomics can bring bad news. Many researchers had hoped to use nanoporous materials to build car tanks that could store more methane in less space. But after screening more than 650,000 computed materials, Smit's group concluded that most of the best ones have already been made[15]. New ones could bring only minor improvements, and energy targets currently set by US agencies — which bet on major technological improvements in methane storage — may be unrealistic.

As intriguing as these examples are, there are still many hurdles to overcome before materials genomics can live up to its promises. One of the largest is that computer simulations still give few clues on how an interesting material can be made in a lab — let alone mass produced. "We come up with interesting ideas for new compounds all the time," says Ceder. "Sometimes it takes two weeks to make it. Other times we still can't make it after six months, and we don't know whether we haven't done the right thing, or it just can't be made."

Both Ceder and Curtarolo are trying to develop machine-learning algorithms to extract rules from known manufacturing processes to guide the synthesis of compounds.

Another limitation is that materials genomics has been hitherto applied almost exclusively to what engineers call functional materials — compounds that can perform a task such as absorbing light in a solar cell or letting electrical current pass in transistor. But the technique does not lend itself well to studying structural materials, such as steel, that are needed to build, for example, aircraft wings, bridges or engines. This is because mechanical properties such as a material's springiness and hardness depend on how it is processed — something that quantum-mechanical codes by themselves can not describe.

Even in the case of functional materials, current computer codes work well only for perfect crystal structures — which are only a small part of the materials realm. "The most interesting materials of the future will probably be assembled at the microscopic level in creative ways," says Galli. They may be assemblies of nanoparticles, crystals with strategically placed defects in their structures, or heterogenous materials made by intertwining different compounds and phases. To predict such materials, says Galli, "you need to calculate many properties at once and how the system will evolve in time and at specific temperatures". There are methods to do that, she says, "but they are still too computationally expensive to be used in high-throughput studies".

In the short term, more data exchange with experiments can give computations a reality check and help to refine them. To that end, Ceder is working with a group at MIT on software that reads papers in experimental materials science and automatically extracts information on crystal structures in a standard format. "We plan to begin adding these data to the Materials Project in a few months," he says.

And in the long run, some help will come from Moore's law: as computational power continues to increase, some techniques that are out still of reach for current computers may soon become viable.

"We've moved away from the artisanal era of computational materials science, and into the industrial phase," says Marzari. "We can now create assembly chains of simulations, put them to work, and explore problems in totally new ways." No computationally predicted material is on the market just yet. "But let's talk again in ten years," says Galli, "and I think there will be many." ■

Nicola Nosengo *is a freelance writer based in Rome.*

1. Raccuglia, P. *et al. Nature* **533,** 73–76 (2016).
2. Björling, C. O. & Westgren, A. *Geol. Fören. Stock. För.* **60,** 67–72 (1938).
3. Padhi, A. K., Nanjundaswamy, K. S. & Goodenough, J. B. *J. Electrochem. Soc.* **144,** 1188–1194 (1997).
4. Curtarolo, S., Morgan, D., Persson, K., Rodgers, J. & Ceder, G. *Phys. Rev. Lett.* **91,** 135503 (2003).
5. Curtarolo, S. *et al. Comput. Mater. Sci.* **58,** 218–226 (2012).
6. Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I. & Nørskov, J. K. *Nature Mater.* **5,** 909–913 (2006).
7. Giannozzi, P. *et al. J. Phys. Condens. Matter* **21,** 395502 (2009).
8. Curtarolo, S. *et al. Comput. Mater. Sci.* **58,** 227–235 (2012).
9. Kirklin, S. *et al. npj Comput. Mater.* **1,** 15010 (2015).
10. Perim, E. *et al.* Preprint at http://arxiv.org/abs/1601.08233 (2016).
11. Jain, A., Shin, Y. & Persson, K. A. *Nature Rev. Mater.* **1,** 15004 (2016).
12. Castelli, I. E. *et al. Adv. Energy Mater.* **5,** 1400915 (2015).
13. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. *Comput. Mater. Sci.* **111,** 218-230 (2016).
14. Lin, L.-C. *et al. Nature Mater.* **11,** 633–641 (2012).
15. Simon, C. M. *et al. Energy Environ. Sci.* **8,** 1190–1199 (2015).

CORRECTIONS

The News Feature ‘The material code’ (*Nature* **533,** 22–25; 2016) omitted Gerbrand Ceder’s first name. In addition, it wrongly implied that the phrase ‘materials genome’ was invented solely by Gerbrand Ceder. The phrase was independently invented and copyrighted by Zi-Kui Liu of Pennsylvania State University.

CLARIFICATION

The News Feature ‘The material code’ (*Nature* **533,** 22–25; 2016) did not make it clear that the director of the Materials Genome Project is Kristin Persson, and that she has an affiliation with the University of California, Berkeley.