This is still far off; but only if we can accurately simulate and rationally manipulate a nervous system can we begin to truly understand it. Once again, Brenner's tiny worm, occupying its unique sweet spot between simplicity and complexity, finds itself on the front line of biology's most challenging problems. ■

Douglas S. Portman *is in the Department of Biomedical Genetics, Del Monte Institute for Neuroscience, University of Rochester, Rochester, New York 14642, USA.*
e-mail: douglas.portman@rochester.edu

1. Cook, S. J. *et al. Nature* **571**, 63–71 (2019).
2. Ankeny, R. *Nature Rev. Genet.* **2**, 474–279 (2001).
3. White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. *Phil. Trans. R. Soc. Lond. B* **314**, 1–340 (1986).
4. Durbin, R. M. *Studies on the Development and Organisation of the Nervous System of* Caenorhabditis elegans. PhD thesis, Univ. Cambridge (1987).
5. Hall, D. H. & Russell, R. L. *J. Neurosci.* **11**, 1–22 (1991).
6. Varshney, L. R., Chen, B. L., Paniagua, E., Hall, D. H. & Chklovskii, D. B. *PLoS Comput. Biol.* **7**, e1001066 (2011).
7. Jarrell, T. A. *et al. Science* **337**, 437–444 (2012).
8. Oren-Suissa, M., Bayer, E. A. & Hobert, O. *Nature* **533**, 206–211 (2016).
9. Bargmann, C. I. & Marder, E. *Nature Methods* **10**, 483–490 (2013).
10. Nguyen, J. P. *et al. Proc. Natl Acad. Sci. USA* **113**, E1074–E1081 (2016).
11. Venkatachalam, V. *et al. Proc. Natl Acad. Sci. USA* **113**, E1082–1088 (2016).
12. Szigeti, B. *et al. Front. Comput. Neurosci.* **8**, 137 (2014).

COMPUTATIONAL MATERIALS SCIENCE

Text mining facilitates materials discovery

Computer algorithms can be used to analyse text to find semantic relationships between words without human input. This method has now been adopted to identify unreported properties of materials in scientific papers. SEE LETTER P.95

OLEXANDR ISAYEV

The total number of materials that can potentially be made — sometimes referred to as materials space — is vast, because there are countless combinations of components and structures from which materials can be fabricated. The accumulation of experimental data that represent pockets of this space has created a foundation for the emerging field of materials informatics, which integrates high-throughput experiments, computations and data-driven methods into a tight feedback loop that enables rational materials design. On page 95, Tshitoyan *et al.*[1] report that knowledge of materials science 'hidden' in the text of published papers can be mined effectively by computers without any guidance from humans.

The discovery of materials that have a particular set of properties has always been a serendipitous process requiring extensive experimentation — a combination of craft and science practised by knowledgeable artisans. However, this trial-and-error approach is expensive and inefficient. There is therefore great interest in using machine learning to make materials discovery more efficient.

Currently, most machine-learning applications aim to find an empirical function that maps input data (for example, parameters that define a material's composition) to a known output (such as measured physical or electronic properties). The empirical function can then be used to predict the property of interest for new input data. This approach is said to be supervised, because the process of learning from the training data is akin to a teacher supervising students by selecting the subjects and facts needed for a particular lesson. A contrasting approach involves using only input data, which have no obvious connection to a specific output. In this case, the goal is to identify intrinsic patterns in the data, which are then used to classify those data. Such an approach is called unsupervised learning, because there are no a priori correct answers and there is no teacher.

Tshitoyan and colleagues collected 3.3 million abstracts from papers published in the fields of materials science, physics and chemistry between 1922 and 2018. These abstracts were processed and curated, for example to remove text that wasn't in English and to exclude abstracts that had unsuitable metadata types, such as 'Erratum' or 'Memorial'. This left 1.5 million abstracts, which were written using a vocabulary of about 500,000 words.

The authors then analysed the curated text using an unsupervised machine-learning algorithm known as Word2vec (ref. 2), which was developed to enable computers to process text and natural language. Word2vec takes a large body of text and passes it through an artificial neural network (a type of machine-learning algorithm) to map each word in the vocabulary to a numeric vector, each of which typically has several hundred dimensions. The resulting word vectors are called embeddings, and are used to position each word, represented as a data point, in a multidimensional space that represents the vocabulary. Words that share common meanings form clusters within that space. Word2vec can therefore make accurate estimates about the meaning of words, or about the functional relationships between

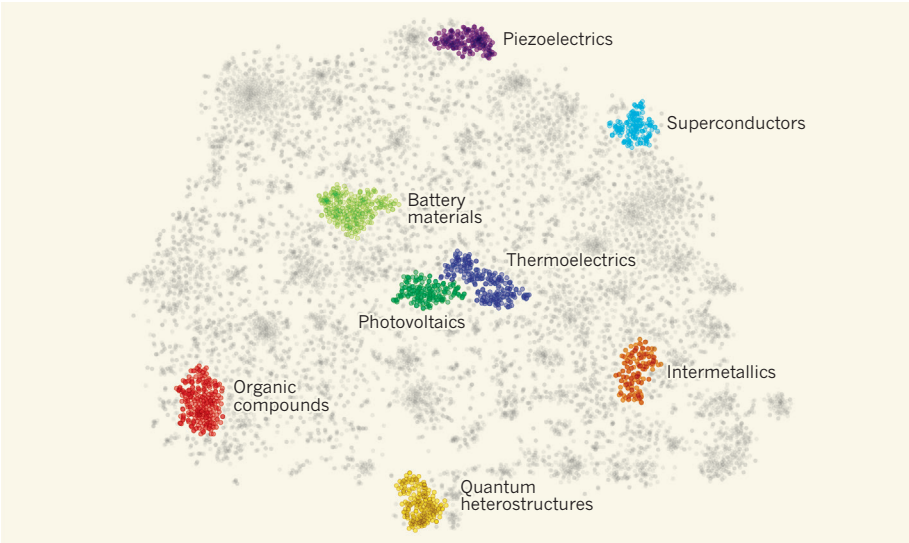

Figure 1 | Clustering of materials from textual analysis of scientific papers. Tshitoyan *et al.*[1] used a machine-learning algorithm to analyse the text in the abstracts of 1.5 million papers to identify relationships between words, including the names of materials. Individual materials were then represented as data points on graphs, and the algorithm clustered the data points together on the basis of the semantic relationships between the words used to describe those compounds. The clusters (coloured) correspond to particular types of material, such as superconductors, battery materials and organic compounds. The authors show that this approach can be used to identify unreported properties of materials mentioned in the scientific literature. (Adapted from Fig. S7a of the paper[1].)

them, on the basis of the patterns of usage of the words in the original text. Importantly, these meanings and relationships are not explicitly encoded by humans, but are learnt in an unsupervised way from the analysed text.

The researchers found that the obtained word embeddings for materials-science terms produced word associations that reflect rules of chemistry, even though the algorithm did not use any specific labels to identify or interpret chemical concepts. When combined using various mathematical operations, the embeddings identified word associations that corresponded to concepts such as 'chemical elements', 'oxides', 'crystal structures', and so on. The embeddings also identified clusters of known materials (Fig. 1) corresponding to categorizations that could be used to classify new materials made in the future.

But Tshitoyan *et al*. went further than just establishing relationships between words — they also demonstrated how their approach could be used for prospective materials discovery. They began by training a machine-learning model to predict the likelihood that a material's name will co-occur with the word 'thermoelectric' in the text (thermoelectric materials are those in which a temperature difference generates a voltage, or vice versa). They then searched the text to find materials that had not been reported to have thermoelectric properties, but whose names have a high semantic relationship with the word 'thermoelectric' — and that might therefore actually be thermoelectric.

The authors validated this approach by training a model using literature published before a particular cutoff year, and then checking to see whether it identified materials that were reported to be thermoelectric in subsequent years. The top 50 materials picked using this method were 8 times as likely to have been studied as a thermoelectric in the 5 years after they had been reported than were randomly chosen materials. Tshitoyan and colleagues' approach therefore demonstrates yet another successful application of text mining, which has now been used in fields ranging from materials science to protein identification[3] and cancer biology[4].

The combination of unsupervised machine learning and text mining for scientific discovery is intriguing, given the burgeoning growth of both supervised and unsupervised methods for natural-language processing in the past few years, and the increasing availability of digitized scientific literature that encompasses more than 100 years of publications. Of course, many challenges remain. Chief among them is the fact that unsupervised methods are typically less accurate than models obtained from supervised learning. Moreover, although word embedding looks promising as a way of identifying materials that have particular properties, it cannot be used to identify materials not described in the literature, whose names are not part of the existing vocabulary. However, such methods could be used to find previously unrecognized properties of existing materials, which could then be repurposed.

The field of materials informatics is emerging in parallel with the growth of materials databases, in much the same way as chemoinformatics arose 20 years ago with the establishment of chemistry databases[5]. Progress is fast, because methods based on data and literature mining are established tools for data scientists working in the chemical and materials sciences[6]. Future studies that use natural-language processing and unsupervised learning in ways similar to those used by Tshitoyan *et al*., or that use both unsupervised and supervised learning, can be expected to increase the impact of data science on materials design and discovery. So, will the next big discovery in superconductors, for example, be made through conventional human intuition or by machine? In all likelihood, it will be a smart combination of both human and machine intelligence. ■

Olexandr Isayev *is in the Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA.*
e-mail: olexandr@olexandrisayev.com

1. Tshitoyan, V. *at al. Nature* **571**, 95–98 (2019).
2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. *Proc. 26th Int. Conf. Neural Information Processing Syst.* go.nature.com/2wvucor (2013).
3. Spangler, S. *et al. Proc. 20th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining* 1877–1886 (ACM, 2014); https://doi.org/10.1145/2623330.2623667
4. Choi, B.-K. *et al. Proc. Natl Acad. Sci. USA* **115**, 10666–10671 (2018).
5. Brown, F. K. *Annu. Rep. Med. Chem.* **33**, 375–384 (1998).
6. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. *Nature* **559**, 547–555 (2018)

NEUROSCIENCE

Star-like cells drive hyperactivity

A molecular dialogue between neurons and star-shaped cells called astrocytes in the striatum of the mouse brain leads to behavioural hyperactivity and inattentiveness that are reminiscent of attention-deficit hyperactivity disorder.

ZHIHUA GAO & HAILAN HU

Astrocytes are star-shaped cells that account for about 40% of cells in the mammalian brain. Initially considered to be the 'glue' that sticks neurons together, astrocytes actually have crucial roles in brain homeostasis and in regulating the formation, maturation, function and elimination of synapses, the connections through which neurons communicate with each other[1–4]. Although much progress has been made in elucidating the roles of astrocytes[1–4], our understanding of how they regulate neural circuits and affect behaviours that are associated with neurological and psychiatric disorders is just emerging[5–9]. Writing in *Cell*, Nagai *et al.*[10] present evidence in mice that selective activation of astrocytes in the striatum, a brain region that integrates signals from many parts of the brain to coordinate voluntary movement[11], drives behavioural changes that resemble the symptoms of attention-deficit hyperactivity disorder (ADHD) in humans through a dialogue with striatal neurons.

ADHD is a prevalent psychiatric and neurodevelopmental disorder that affects approximately 5% of children worldwide, and its major symptoms include excessive activity (or restlessness) and difficulty in sustaining attention[12]. Although dysfunction in the striatum has been implicated in ADHD[13], the underlying mechanisms of how the striatum — and, in particular, striatal astrocytes — might contribute to the disorder, remain elusive. The striatum largely consists of a special type of medium-sized neuron that is inhibitory (that is, it suppresses the activity of connected neurons) and that features many tiny protrusions called spines that receive synaptic inputs from other neurons. When activated, these medium spiny neurons (MSNs) release the inhibitory neurotransmitter molecule GABA (γ-aminobutyric acid) to reduce the activity of other neurons, and together the MSNs control behavioural movement[14].

Because MSNs are intermingled with astrocytes and form close contacts with them[15], Nagai *et al.* set out to examine whether MSN activation might affect the activity of surrounding astrocytes. The authors monitored astrocyte activity by making these cells express a genetically encoded calcium indicator — a protein that fluoresces in response to increases in the concentration of calcium ions (which are involved in cell signalling). They found that, when they stimulated MSNs using an electric current, the calcium-ion signalling in nearby astrocytes increased. This increase depended on the release of GABA from the MSNs, and on the activation of type B GABA receptors ($GABA_B$ receptors),