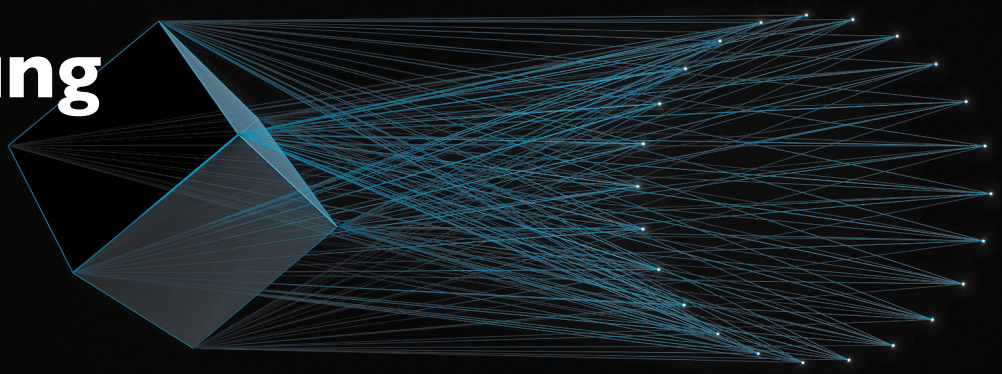
Fast-tracking discovery with data



Andriy Onufriyenko / Moment /Getty

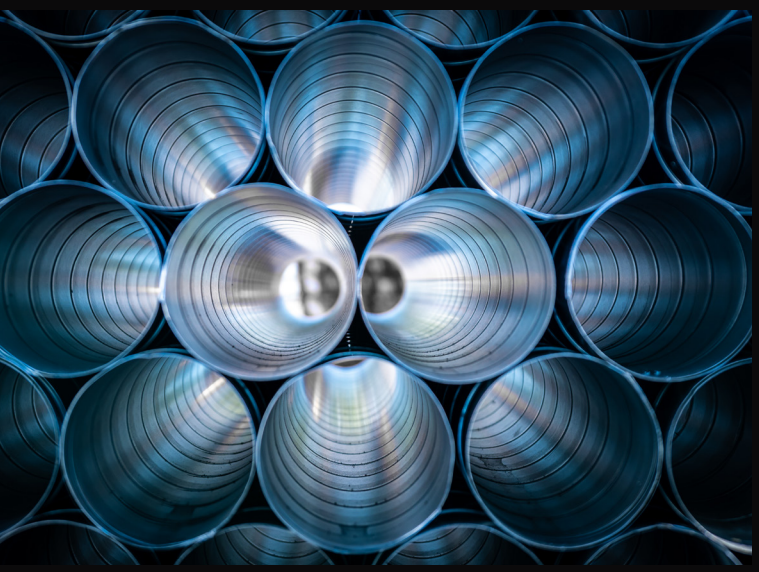
The combination of machine learning with specific domain knowledge is helping scientists to reveal new materials.

"There has been an explosion of research in the past five years to discover new materials, or reveal unseen properties, by using Materials Informatics (MI)," says Tong-Yi Zhang, a materials scientist, a member of the Chinese Academy of Sciences, and dean of the Materials Genome Institute (MGI) of Shanghai University.

Materials informatics can greatly improve understanding and accelerate the discovery of advanced materials by bringing together data science, artificial intelligence (AI), machine learning, deep learning algorithms, computational techniques, and domain knowledge.

Zhang says the institute aims to become a leading research centre for MI and materials genome engineering research. Materials genome engineering, analogous to the human genome, investigates the underlying properties of materials. The institute also plans to become an interdisciplinary hub, integrating computational simulations, experiments, theory and data to engineer new materials and applications.

Reinhard Krull / EyeEm /Getty



Machine learning and other computational techniques are accelerating materials discovery.

Domain knowledge to handle small data sets

MI relies on experimental data, but there is usually not enough available about any given material, due to the high cost and time needed for repeated experiments. Sometimes small data sets yield consistent results that point to useful patterns and conclusions, but at other times small data sets are unreliable, with complex, sparse, and noisy results, because of the greatly variable experimental factors, ranging from the sample size of the tested material, to the testing environment, and the relative fragility of many materials.

A possible solution, according to Zhang, is the integration of machine learning (ML) with domain knowledge, specific to a discipline or field. Zhang put this principle into practice in a study coupling theory and ML on concrete samples to identify their size-dependent properties. He analysed a small data set of the concrete using statistical conclusions and established knowledge. This method enabled Zhang to extract from the data useful results about strength and fracture resistance of the concrete, he says.

In another investigation, Zhang studies how the understanding of materials can be advanced by symbolic regression, a machine learning method that can turn data into mathematical formulae. When integrated with domain knowledge, scientists can gain knowledge from data quickly and accurately, and create explicit interpretable models, uncovering the underlying mechanism of materials properties and performance.

The team has also developed a range of MI software that uses domain knowledge to guide interpretive machine learning. Termin, a formula-finding software performing symbolic regression, can automatically search and suggest the best formula for given data. TCLR, or Tree Classifier for Linear Regression, is a decision tree-based algorithm that looks for correlations between data sets.

Guided by Zhang's domain knowledge emphasis, researchers from MGI have developed AI-guided, high-throughput computational and experimental alloy-design methods which enable alloy development with higher efficiency and lower cost. They have also screened the MGI database to find promising thermoelectric materials that have been successfully synthesized in experiments. ■