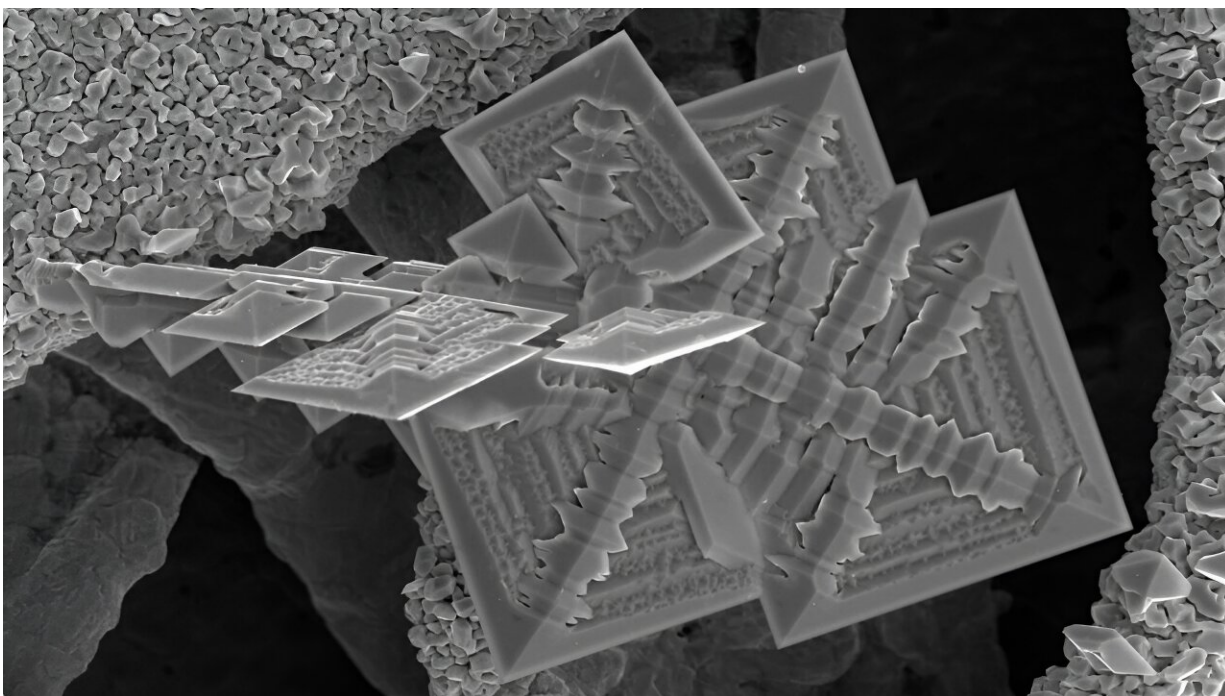# Researchers harness large language models to accelerate materials discovery

January 29 2024, by Molly Sharlach

A scanning electron microscope image of tin crystals, stimulated by electricity and growing on a copper surface. A new method developed by Princeton researchers could speed up the process of designing and testing new crystalline materials. Credit: Lynn Trahey, Argonne National Laboratory

Princeton researchers have created an artificial intelligence (AI) tool to predict the behavior of crystalline materials, a key step in advancing technologies such as batteries and semiconductors. Although computer

simulations are commonly used in crystal design, the new method relies on a large language model, similar to those that power text generators like ChatGPT.

By synthesizing information from text descriptions that include details such as the length and angles of bonds between atoms and measurements of electronic and optical properties, the new method can predict properties of new materials more accurately and thoroughly than existing simulations—and potentially speed up the process of designing and testing new technologies.

The researchers developed a text benchmark consisting of the descriptions of more than 140,000 crystals from the Materials Project, and then used it to train an adapted version of a large language model called T5, originally created by Google Research. They tested the tool's ability to predict the properties of previously studied crystal structures, from ordinary table salt to silicon semiconductors. Now that they've demonstrated its predictive power, they are working to apply the tool to the design of new crystal materials.

The method, presented Nov. 29 at the Materials Research Society's Fall Meeting in Boston, represents a new benchmark that could help accelerate materials discovery for a wide range of applications, according to senior study author Adji Bousso Dieng, an assistant professor of computer science at Princeton.

The paper outlining the method, "LLM-Prop: Predicting Physical And Electronic Properties Of Crystalline Solids From Their Text Descriptions," is now posted to the *arXiv* preprint server.

Existing AI-based tools for crystal property prediction rely on methods called graph neural networks, but these have limited computational power and can't adequately capture the nuances of the geometry and

lengths of bonds between atoms in a crystal, and the electronic and optical properties that result from these structures. Dieng's team is the first to tackle the problem using large language models, she said.

"We have made tremendous advances in computer vision and natural language," said Dieng, "but we are not very advanced yet when it comes to dealing with graphs [in AI]. So, I wanted to move from the graph to actually translating it to a domain where we have great tools already. If we have text, then we can leverage all these powerful [large language models] on that text."

The language model-based approach "gives us a whole new way to look at the problem" of designing materials, said study co-author Craig Arnold, Princeton's Susan Dod Brown Professor of Mechanical and Aerospace Engineering and vice dean for innovation. "It's really about, how do I access all of this knowledge that humanity has developed, and how do I process that knowledge to move forward? It's characteristically different than our current approaches, and I think that's what gives it a lot of power."

For insights into the challenges of crystal design, Dieng and Ph.D. student Andre Niyongabo Rubungo teamed up with Arnold and with Barry Rand, a professor of electrical and computer engineering and the Andlinger Center for Energy and the Environment who focuses on materials for semiconductors and solar energy. Arnold is interested in laser-material interactions, with applications for energy storage.

"The materials in our world are all ones that were developed through scientific hypothesis testing and sometimes luck," said Rand. This process "leads to good outcomes, but it takes time. Through artificial intelligence methods, we could really accelerate that." Furthermore, he said, "it allows us to identify things that probably we as humans wouldn't intuit."

Given a crystal with a particular composition of chemical elements, the team's method can predict properties including the band gap, which relates to the crystal's electronic states and conductivity.

"If you can predict that with high accuracy, when you then go to do the painstaking work of experimentation, you can have more confidence that it's going to yield success," said Rand.

Ph.D. student Rubungo received a best poster award for presenting the work to materials researchers at the fall meeting. Many were surprised by the power of large language models in this context. The field is more accustomed to structured data used as inputs for graph neural networks, but "texts are easier to deal with," said Rubungo. "It's easier to include the information you want in your description, and to modify the tool and remove what you don't want. People were very excited to see that."

As a new tool, he noted, the prediction method has limitations. It uses more computing power and is slower than graph neural networks typically used for this purpose. It could also benefit from expanded [training data](#) to boost its ability to predict properties of novel materials.

Dieng is pursuing collaborations with other materials researchers, and aims to move the work beyond crystals to a broader variety of materials. "This is a nascent area of research, and what advances research is to have a well-established benchmark that's well curated," she said. "We are gathering more data sets into one benchmark that will be hosted at Princeton for researchers to use."