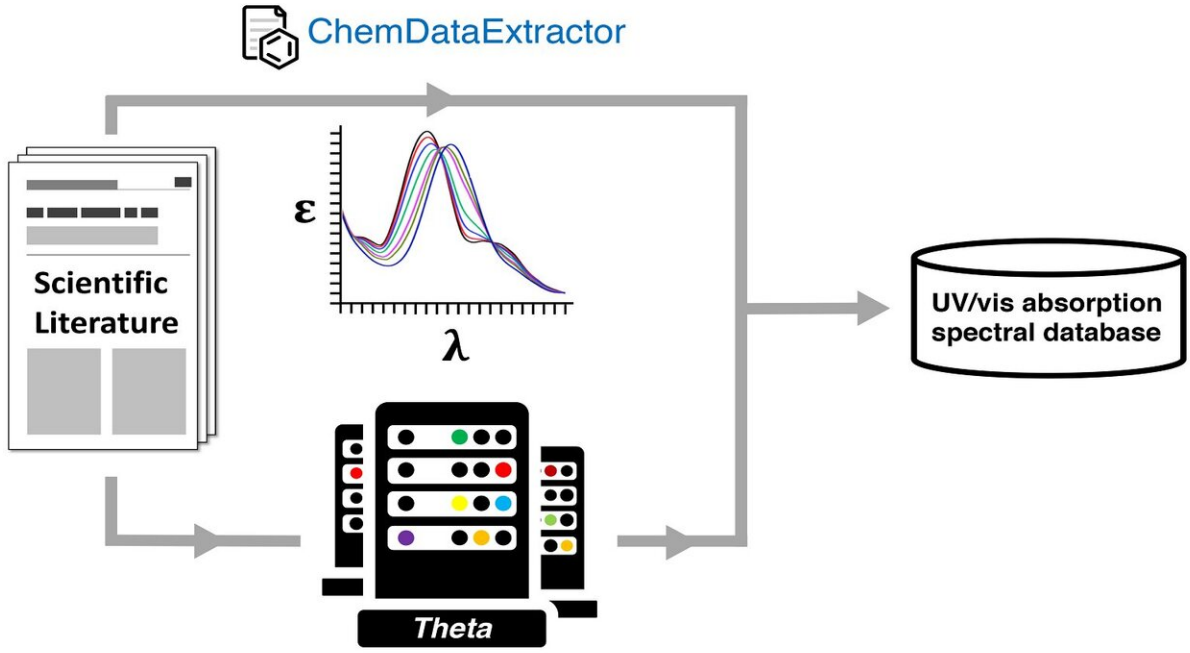
Automatic database creation for materials discovery: Innovation from frustration

September 23 2020, by John Spizzirri



Auto-generating an ultraviolet-visible (UV-vis) absorption spectral database via a dual experimental and computational chemical data pathway using the ALCF's Theta supercomputer. Credit: Jacqueline Cole and Ulrich Mayer / University of Cambridge

A collaboration between the University of Cambridge and Argonne has developed a technique that generates automatic databases to support specific fields of science using AI and high-performance computing.

Searching through reams of scientific literature for bits and bytes of information to support an idea or find the key to solving a specific problem has long been a tedious affair for researchers, even after the dawn of data-driven discovery.

Jacqueline Cole knows the drill, all too well. Head of Molecular Engineering at the University of Cambridge, United Kingdom, she has spent much of her career searching for materials with optical properties that lend themselves to more efficient light collection, like dye molecules that may one day power solar windows.

"I knew that a lot of the information was held in very fragmented form across the literature," she recalls. "But if you collated across thousands and thousands of documents, then you could form your own database."

So Cole and colleagues at Cambridge and the U.S. Department of Energy's (DOE) Argonne National Laboratory did just that, laying out the process in the journal *Scientific Data*.

The paper, says Cole, is a description of how to build a database using natural language processing (NLP) and high-performance computing, much of the latter performed at the Argonne Leadership Computing Facility (ALCF), a DOE Office of Science User Facility.

Among the factors that make the database unique are the scale of the project and the fact that it comprises both experimental and calculated data on both material structures, which describes the atomic or chemical foundation of a thing, and material properties, the functionality provided by those different structures.

"It's probably the first such compilation of a database on such a massive scale, with 5,380 like-for-like pairs of experimental and calculated data," says Cole. "And because it's such a large amount, it serves as a repository

in its own right and really opens the door to predicting new materials."

Many new, large databases are built purely on calculations, an inherent drawback of which is that they are not validated by experimental data. The latter, perhaps most significantly, provides an accurate picture of the material's excited states, which define the dynamic state of electrons and are used to calculate a material's functional properties—optical properties, in this case.

This budding catalog of excited states can then help calculate the properties of materials that have yet to be conceived, further expanding the database.

"Imagine that one wishes to discover a new type of optical material to suit a bespoke functional application, and our database does not contain that particular optical property," explains Cole. "We calculate the optical property of interest from the excited states that are available for each property in our database, and create a material with tailored functions."

The team performed quantum-chemical calculations on each structure for which they had extracted data on optical materials, using the ALCF's Theta supercomputer, thus creating the database of paired experimental and calculated structures and their optical properties.

"One of the biggest challenges was extracting chemical candidates that could serve as dyes for solar cells from 400,000 scientific articles," says Álvaro Vázquez-Mayagoitia, a computational scientist in Argonne's Computational Science division. "We developed a distributed framework to apply artificial intelligence methods, such as those used in natural language processing, on the ALCF's world-class supercomputers."

To automatically extract that information and deposit it in the database,

the team turned to the novel data mining application called ChemDataExtractor. An NLP tool, it was designed to mine text specifically from within chemistry and materials literature, where, Cole says, "the information is strewn across many thousands of papers and is present in highly fragmented and unstructured forms."

Not one for manual article searches, Cole describes the drive to develop the application as innovation from frustration. Initially, she tried more generic NLP packages, but noted that "they don't just fail, they fail spectacularly."

The problem is in the translation, not so much from a human language stance, but from the language of science, although there are some similarities.

A writer, for example, might use a speech recognition program, a form of NLP, to transcribe notes or interviews. The program trains mainly on the writer's voice, picking up patterns and nuances, and begins to transcribe fairly accurately. Now throw in an interview with a subject with a foreign accent and things begin to get wonky.

In Cole's world, the foreign language is science, each domain a different country. Currently, you have to train the program on only one "language," say chemistry, and even then, you have to learn that science's particular dialects.

Inorganic chemists might pose a formula using unfamiliar representations of the well-known chemical element symbols, whereas organic chemists prefer chemical sketches numbered within an illustration box. The information from either typically proves too hard for most mining programs to extract.

"And that's just in a little bit of chemistry," notes Cole. "Because the

way people describe things is so diverse, diversity in domain specificity is absolutely critical."

To that end, the team's database is one of ultraviolet–visible (UV/vis) absorption spectral attributes, which provides an openly available resource for users seeking to find materials with preferred spectral colors.

While the team is using the new database to ferret out organic dyes that might replace traditional metal-organic dyes in solar cells, they have already targeted broader fronts for its use.

Useful as a source of training data for machine-learning methods that predict new optical materials, it can also prove a simple data retrieval option for users of UV/vis absorption spectroscopy, a tool that is widely used across research laboratories around the world as a core technique to characterize new materials.

"The protocols used in this project are already being deployed for similar types of projects," adds Vázquez-Mayagoitia. "For example, the team recently leveraged ChemDataExtractor and ALCF computing resources to produce expansive databases of potential battery chemicals, and magnetic and superconducting compounds."

The optical materials database research appears in the article "Comparative dataset of experimental and computational attributes of UV/vis absorption spectra" in Scientific Data. Additional authors include Edward J. Beard of the University of Cambridge, and Ganesh Sivaraman and Venkatram Vishwanath of Argonne National Laboratory.

A paper detailing their work with magnetic and superconducting materials has been published in *npj Computational Materials*. The battery materials database containing over 290,000 data records has been

published in *Scientific Data*.

More information: Callum J. Court et al. Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning, *npj Computational Materials* (2020). DOI: 10.1038/s41524-020-0287-8

Shu Huang et al. A database of battery materials auto-generated using ChemDataExtractor, *Scientific Data* (2020). DOI: 10.1038/s41597-020-00602-2

Provided by Argonne National Laboratory