Data-Driven Molecular Engineering of Solar-Powered Windows

Jacqueline M. Cole
University of Cambridge and Argonne National Laboratory

Editors:
James J. Hack, jhack@ornl.gov; Michael E. Papka, papka@anl.gov

Buildings are the centerpiece of modern living, with more than half of the world's population now living in urban environments. This demographic evolution has led to building use becoming the main drain of our energy resources. According to the US Energy Information Administration, in 2016, 40 percent of the total energy consumption in the United States came from building use. However, we could overcome this energy drain by embedding new environmental technologies into future cities to realize energy-sustainable buildings.

"Smart windows" that generate electricity from sunlight hold exciting prospects for meeting entire cities' building energy demands in a fully sustainable fashion. The dye-sensitized solar cell—a next-generation photovoltaic technology that mimics photosynthesis—is a particularly strong contender for smart windows, given its transparent and low-cost nature.[1] Yet, like so many solar-cell technologies, a lack of suitable materials for these devices is holding up innovation. In particular, the discovery of new types of dye molecules that absorb just a little more light than current dyes could transform the innovation prospects for this solar-powered window technology. A modest boost in photovoltaic performance is all that is required, because price-to-performance governs solar-cell industry economics. Additionally, manufacturing dye-sensitized solar cells is very cheap relative to other solar-cell technologies.

Research at the Argonne Leadership Computing Facility (ALCF) at Argonne National Laboratory, a US Department of Energy Office of Science User Facility, seeks to realize such dye discovery through a new initiative in data science. In 2016, the ALCF Data Science Program (ADSP) provided a significant award of computing time on ALCF resources and personnel support for a project aimed at discovering new dye materials that would be suitable for dye-sensitized solar cells. The central idea behind this project is to develop a new design approach: one that marries the latest technical capabilities in natural language processing, machine learning, and quantum-chemical calculations to the world-leading supercomputing resources available at Argonne. The overarching concept is to search through a representative set of all possible chemical molecules and use artificial intelligence to target the chemicals whose molecules have optical properties that would yield optimum device function in dye-sensitized solar cells.

 Published by the IEEE Computer Society

MINING THE SCIENTIFIC LITERATURE

The first step in this dye discovery process is to produce a large dataset of chemical molecules that includes their molecular structure and optical properties. Such data are available from the scientific literature in a fragmented form. For example, an isolated data entry might arise from an academic research article about the synthesis of a new chemical, and that article might give a general characterization of the chemical's optical absorption properties but not specify a particular optical application. The sheer volume of literature containing such information would require several lifetimes of human effort to manually curate a database that assembled all molecular structures paired with their optical properties.

Fortunately, such human effort can now be avoided thanks to a new text-mining software tool called ChemDataExtractor, which auto-generates materials databases.[2] ChemDataExtractor uses natural language processing to mine user-specified chemical data types (chemical records) from documents; custom-built chemical dictionaries enable high precision and recall in data extraction. The tool also uses unsupervised and supervised learning algorithms, as well as rule-based knowledge classifiers, to improve upon its chemical cognition. For example, the tool uses clustering methods to classify common themes in chemical words by apportioning them the same binary code, such as 10011011101 for words containing the chemical descriptor "nano" (nanoparticles, nanocrystals, nanowires, and so on).

This ADSP project is using ChemDataExtractor to auto-generate a materials database of tens of thousands of chemicals with their paired optical absorption property information, as shown in Figure 1. These property data manifest in the literature as optical absorption spectral data, which feature parameters such as the wavelength of light at which a chemical most absorbs and the intensity of this absorption.

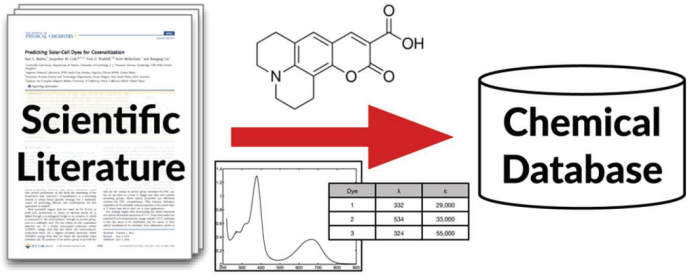

Figure 1. Hundreds of thousands of documents from the scientific literature are fed into the text-mining software tool ChemDataExtractor to automatically assemble a chemical database. This database contains tens of thousands of unstructured data of molecular structures with their paired optical property information. Reprinted with permission from M. C. Swain and J. M. Cole.[2] Copyright 2016 American Chemical Society.

MINING THE RESULTING DATABASE

The second step in this dye discovery process is to mine the materials database by embedding algorithms into the search engine. These algorithms use encoded forms of structure-property relationships that are good for solar-cell dyes, explaining the underlying physics and chemistry in a way that a computer can read. These relationships are established using case studies on known solar-cell dyes, which come from a range of materials characterization efforts to support this dye discovery process. For example, the case study on a dye molecule that is known to function well in dye-sensitized solar cells has recently been studied at the ALCF to develop new molecular design rules to aid the predictive power of its dye discovery program.[3]

The research team used the quantum-chemical calculation software package NWChem on its newly released supercomputer, a Cray XC40 named Theta, to simulate the molecular structure of solar-cell working electrode device interfaces that feature this known high-performance dye.[4]

These computer simulations were complemented by synchrotron-based experiments that revealed the molecular structure of this dye at exceptionally high resolution. The combined results of theory and experiment enabled the formation of new structure-property relationships for solar-cell dyes to add to the growing knowledge base of molecular design rules that will help ongoing dye material prediction efforts (see Figure 2).[5]

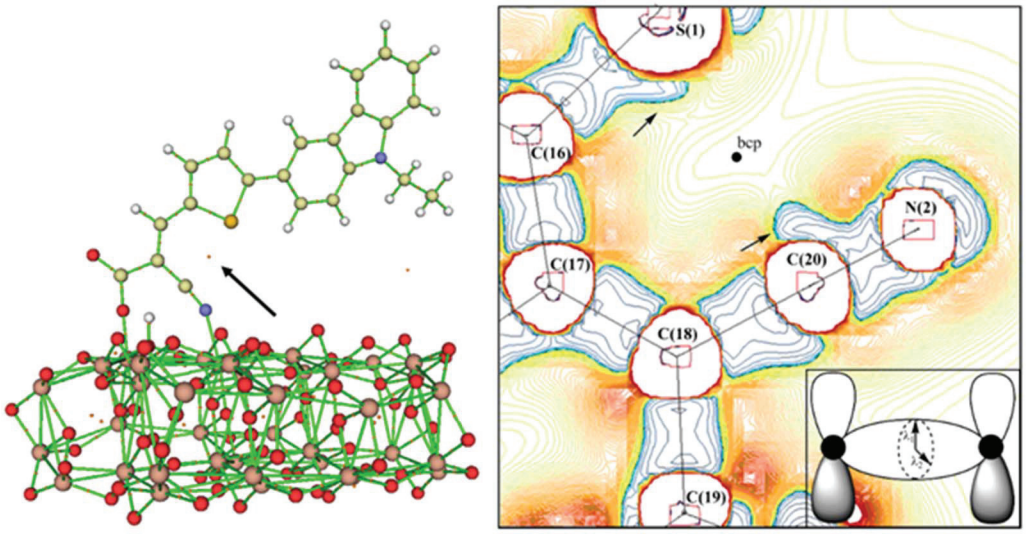

Figure 2. (left) Results from quantum-chemical calculations that employ density functional theory on Theta to produce the electronic structure of the working electrode of a dye-sensitized solar cell. This comprises the MK-44 dye molecule adsorbed onto the surface of titanium dioxide, as modeled using a $(TiO_2)_{38}$ cluster; the arrow highlights the small orange circle between two atoms in the dye molecule, which denotes a "bond critical point" that proves the discovery of a new type of chemical bond between sulfur and carbon atoms. (right) The Laplacian map of a plane of the MK-44 molecule from experimental data that shows this bond critical point of this newly discovered chemical bond at high atomic resolution. This new discovery about chemical bonding helped formulate new structure-property relationships that can be encoded into molecular design rules to aid dye discovery through this data science program. Reprinted with permission from J. M. Cole et al.[3] Copyright 2017 American Chemical Society.

TYING IT TOGETHER

Machine learning aids the dye discovery process at various stages. For example, consider the case of missing data. The process relies on optical property information being available in the academic literature, but sometimes only partial information about a given chemical is available. In such cases, the missing data stand to prevent that chemical from being included in the database. Yet, machine learning can come to the rescue. A machine-learning algorithm can use the information that is already in the database as training data and exploit chemical similarity as the learning heuristic to populate missing data for the affected chemicals.

The ALCF supercomputers are also auto-generating a vast array of computational data on the chemical molecules in the database to complement the largely experimental data that ChemDataExtractor is collating from the literature. High-throughput density functional theory and time-dependent density functional theory are the basis of the quantum chemical calculations that form these computational data. Such a computer-intensive task necessitates the supercomputing resources of Theta and Mira, ALCF's IBM Blue Gene/Q.

This combination of computational and experimental data will create an ideal data source for mining. A short list of predicted dye materials will go forward for experimental validation; these dyes will be synthesized, fabricated in dye-sensitized solar-cell devices, and tested for photovoltaic performance. Results from this dye discovery process—good or bad—will be fed back into the cognition part of the workflow as a positive feedback loop, such that subsequent cycles of

dye discovery can proceed with, hopefully, better and better accuracy in materials prediction. The ultimate goal is to use data science to engineer new molecules that are tailored to a given device application.

ACKNOWLEDGMENTS

This research used resources of the ALCF, a US Department of Energy Office of Science User Facility supported under contract DE-AC02-06CH11357. The author is the recipient of the Royal Commission for the Great Exhibition of 1851 2014 Design Fellowship that co-sponsors this research, and she acknowledges Álvaro Vázquez-Mayagoitia of Argonne National Laboratory as the co-investigator of this ADSP project.

REFERENCES

1. B. O'Regan and M. Grätzel, "A Low-Cost, High-Efficiency Solar Cell Based on Dye-Sensitized Colloidal TiO2 Films," *Nature*, vol. 353, no. 6346, 1991, pp. 737–740.
2. M.C. Swain and J.M. Cole, "ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature," *J. Chemical Information and Modeling*, vol. 56, no. 10, 2016, pp. 1894–1904.
3. J.M. Cole et al., "Discovery of S···C≡N Iintramolecular Bonding in a Thiophenylcyanoacrylate-Based Dye: Realizing Charge Transfer Pathways and Dye…TiO2 anchoring Characteristics for Dye-Sensitized Solar Cells," *ACS Applied Materials & Interfaces*, vol. 9, no. 31, 2017, pp. 25952–25961.
4. M. Valiev et al., "NWChem: A Comprehensive and Scalable Open-Source Solution for Large Scale Molecular Simulations," *Computer Physics Communications*, vol. 181, no. 9, 2010, pp. 1477–1489.
5. J.M. Cole et al., "Data Mining with Molecular Design Rules Identifies New Class of Dyes for Dye-Sensitised Solar Cells," *Physical Chemistry Chemical Physics*, vol. 16, no. 48, 2014, pp. 26684–26690.

ABOUT THE AUTHOR

Jacqueline M. Cole is head of the Molecular Engineering group at the University of Cambridge, a joint initiative between the Cavendish Laboratory (Physics) and the Department of Chemical Engineering and Biotechnology, in partnership with the ISIS Facility, STFC Rutherford Appleton Laboratory. She currently holds the Royal Commission for the Great Exhibition of 1851 2014 Design Fellowship with Argonne National Laboratory. Cole received a PhD in physics from the University of Cambridge and a PhD in chemistry from the University of Durham. Contact her at jmc61@cam.ac.uk.

This article originally appeared in
Computing in Science & Engineering, *vol. 20, no. 1, 2018.*