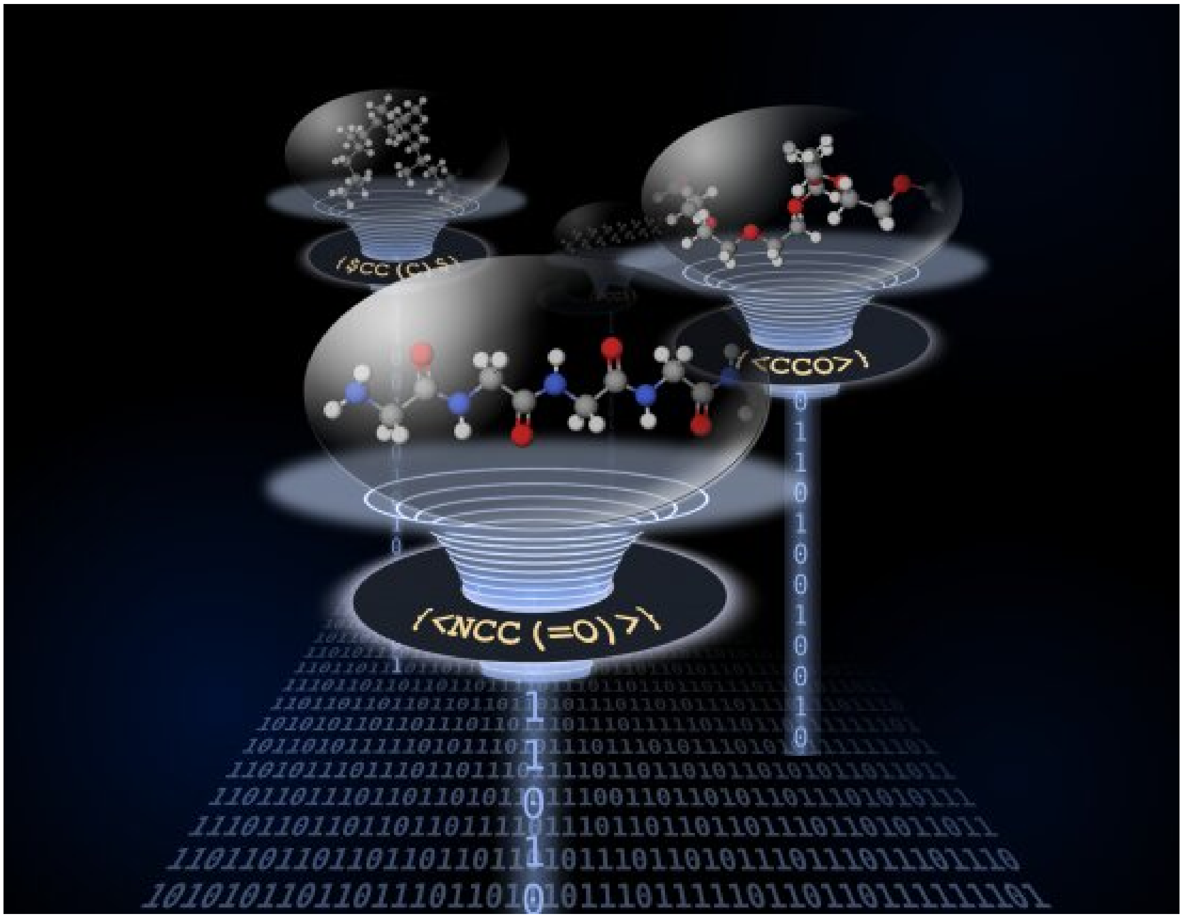
Notation system allows scientists to communicate polymers more easily

September 19 2019, by Melanie Miller Kaufman



In BigSMILES, polymeric fragments are represented by a list of repeating units enclosed by curly brackets. The chemical structures of the repeating units are encoded using normal SMILES syntax, but with additional bonding descriptors that specify how different repeating units are connected to form polymers. This simple design of syntax would enable the encoding of macromolecules over a wide range of chemistries. Credit: Tzyy-Shyang Lin

Having a compact, yet robust, structurally-based identifier or representation system for molecular structures is a key enabling factor for efficient sharing and dissemination of results within the research community. Such systems also lay down the essential foundations for machine learning and other data-driven research. While substantial advances have been made for small molecules, the polymer community has struggled in coming up with an efficient representation system.

For small molecules, the basic premise is that each distinct chemical species corresponds to a well-defined chemical structure. This does not hold for polymers. Polymers are intrinsically stochastic molecules that are often ensembles with a distribution of chemical structures. This difficulty limits the applicability of all deterministic representations developed for small molecules. In a paper published Sept. 12 in *ACS Central Science*, researchers at MIT, Duke University, and Northwestern University report a new representation system that is capable of handling the stochastic nature of polymers, called BigSMILES.

"BigSMILES addresses a significant challenge in the digital representation of polymers," explains Connor Coley Ph.D. '19, co-author of the paper. "Polymers are almost always ensembles of multiple chemical structures, generated through stochastic processes, so we can't use the same strategies for writing down their structures as for small molecules."

Co-authors are Coley; associate professor of chemical engineering Bradley D. Olsen at MIT; Warren K. Lewis Professor of Chemical Engineering Klavs F. Jensen at MIT; assistant professor of chemistry Julia A. Kalow at Northwestern University; associate professor of chemistry Jeremiah A. Johnson at MIT; William T. Miller Professor of Chemistry Stephen L. Craig at Duke University; graduate student Eliot Woods at Northwestern University; graduate student Zi Wang at Duke University; graduate student Wencong Wang at MIT; graduate student

Haley K. Beech at MIT; visiting researcher Hidenobu Mochigase at MIT; and graduate student Tzyy-Shyang Lin at MIT.

There are several line notations to communicate molecular structure, with simplified molecular-input line-entry system (SMILES) being the most popular. SMILES is generally considered the most human-readable variant, with by far the widest software support. In practice, SMILES provides a simple set of representations that are suitable as labels for chemical data and as a memory-compact identifier for data exchange between researchers. As a text-based system, SMILES is also a natural fit to many text-based machine learning algorithms. These characteristics have made SMILES a perfect tool for translating chemistry knowledge into a machine-friendly form, and it has been successfully applied for small molecule property prediction and computer-aided synthesis planning.

Polymers, however, have resisted description by this and other structural languages. This is because most structural languages such as SMILES have been designed to describe molecules or chemical fragments that are well-defined atomistic graphs. Since polymers are stochastic molecules, they do not have unique SMILES representations. This lack of a unified naming or identifier convention for polymer materials is one of the major hurdles slowing down the development of the polymer informatics field. While pioneering efforts on polymer informatics, such as the Polymer Genome Project, have demonstrated the usefulness of SMILES extensions in polymer informatics, the fast development of new chemistry and the rapid development of materials informatics and data-driven research make the need for a universally applicable naming convention for polymers important.

"Machine learning presents an enormous opportunity to accelerate chemical development and discovery," says Lin He, acting deputy division director for the National Science Foundation (NSF) Division of

Chemistry. "This expanded tool to label structures, specifically devised to address the unique challenges inherent to polymers, greatly enhances the searchability of chemical structural data, and brings us one step closer to harnessing the data revolution."

The researchers have created a new structurally-based construct as an addition to the highly successful SMILES representation that can treat the random nature of polymer materials. Since polymers are high molar mass molecules, this construct is named BigSMILES. In BigSMILES, polymeric fragments are represented by a list of repeating units enclosed by curly brackets. The chemical structures of the repeating units are encoded using normal SMILES syntax, but with additional bonding descriptors that specify how different repeating units are connected to form polymers. This simple design of syntax would enable the encoding of macromolecules over a wide range of different chemistries, including homopolymer, random copolymers and block copolymers, and a variety of molecular connectivity, ranging from linear polymers to ring polymers to even branched polymers. As in SMILES, BigSMILES representations are compact, self-contained text strings.

"Standardizing the digital representation of polymeric structures with BigSMILES will encourage the sharing and aggregation of polymer data, improving model quality over time and reinforcing the benefits of its use," says Jason Clark, the materials lead in Open Innovation for Renewable Chemicals and Materials at Braskem, who was not associated with the research. "BigSMILES is a significant contribution to the field in that it addresses the need for a flexible system to represent complex polymer structures digitally."

Clark adds, "The challenges faced by the plastics industry in the context of the circular economy begins with the source of raw materials and continues all the way through end-of-life management. Addressing these challenges requires the innovative design of polymer-based materials, which has traditionally suffered from lengthy development cycles. Advances in artificial intelligence and machine learning have shown promise to accelerate the development cycle for applications utilizing metal alloys and small organic molecules, motivating the plastics industry to seek a parallel approach." BigSMILES digital representations facilitate the evaluation of structure-performance relationships by application of data science methods, he says, ultimately accelerating the convergence to the polymer structures or compositions that will help enable the circular economy.

"A multitude of complicated polymer structures can be constructed through the composition of three new basic operators and original SMILES symbols," says Olsen, "Entire fields of chemistry, materials science, and engineering, including polymer science, biomaterials, materials chemistry, and much of biochemistry, are based upon macromolecules which have stochastic structures. This can basically be thought of as a new language for how to write the structure of large molecules."

"One of the things I'm excited about is how the data entry might eventually be tied directly to the synthetic methods used to make a particular polymer," says Craig, "Because of that, there is an opportunity to actually capture and process more information about the molecules than is typically available from standard characterizations. If this can be done, it will enable all sorts of discoveries."

More information: Tzyy-Shyang Lin et al. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules, *ACS Central Science* (2019). DOI: 10.1021/acscentsci.9b00476

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT *research, innovation and teaching.*

Provided by Massachusetts Institute of Technology

Citation: Notation system allows scientists to communicate polymers more easily (2019, September 19) retrieved 10 April 2024 from https://phys.org/news/2019-09-notation-scientists-polymers-easily.html