PHYS.ORG

New chemistry toolkit speeds analyses of molecules in solution

November 29 2022, by Carol Clark

“We've freed the researchers from most of the tedious, manual tasks of data input,” says Emory theoretical chemist Fang Liu (center). Her team members who developed the toolkit include Emory graduate student Ariel Gale (left) and postdoctoral fellow Eugen Husk (right). Not shown is Xiao Huang, who worked on the project as an undergraduate. Credit: Emory University

A new open-source toolkit automates the process of computing molecular properties in the solution phase, clearing new pathways for artificial-intelligence design and discovery in chemistry and beyond. The *Journal of Chemical Physics* published the free, open-source toolkit developed by theoretical chemists at Emory University.

Known as AutoSolvate, the toolkit can speed the creation of large, high-quality datasets needed to make advances in everything from renewable energy to human health.

"By using our automated workflow, researchers can quickly generate 10, or even 100 times, more data compared to the traditional approach," says Fang Liu, Emory assistant professor of chemistry and corresponding author of the paper. "We hope that many researchers will access our toolkit to perform high-throughput simulation and data curation for molecules in solution."

Such datasets, Liu adds, will provide a foundation for applying state-of-the-art machine-learning techniques to drive innovation in a broad range of scientific endeavors.

First author of the paper is Eugen Hruska, a postdoctoral fellow in the Liu lab. Co-authors include Emory Ph.D. candidate Ariel Gale and Xiao Huang, who worked on the paper as an Emory undergraduate and is now a graduate student of chemistry at Duke University.

Exploring the quantum world

A theoretical chemist, Liu leads a team specializing in computational quantum chemistry, including modeling and deciphering molecular properties and reactions in the solution phase.

The world becomes much more complex as it shrinks down to the scale of atoms and small molecules, where quantum mechanics describes the wave-particle duality of energy and matter.

Theoretical chemists use supercomputers to simulate the structures of molecules and the vast array of interactions that can occur during a reaction so that they can make predictions about how a molecule will behave under certain conditions. Understanding these dynamics is key to identifying promising molecules for various applications and for driving reactions efficiently.

Researchers have already generated datasets for the properties of many molecules in the gas phase. Molecular properties in the solution phase, however, remain relatively unexplored in the context of big data and machine learning, despite the fact that most reactions occur in solution.

The problem is that studying a molecule in solution requires much more time and effort.

A complicated process

"In the gas phase, molecules are far from each other," Liu explains, "so when we study a molecule of interest, we don't have to consider its neighbors."

In the solution phase, however, a molecule is closely immersed with many other molecules, making the system much larger. "Imagine a solvent molecule surrounded by layers and layers of water molecules," Liu says. "Depending on its size and structure, a molecule may be covered by tens, or even up to hundreds, of water molecules. In systems of such large size, the computation will be slow and may not even be feasible."

Before running a quantum chemistry program for a molecule in the solution phase it's necessary to first determine the geometry of the molecule and the location and orientation of the surrounding solvent molecules.

"This process is difficult to do," Liu says. "It takes so much time and effort, and it's so complicated, that a researcher can only perform this calculation for a few systems that they care about in one paper," Liu says.

Technical issues can also arise during each step in the process, she adds, leading to errors in the results.

A streamlined solution

Liu and her colleagues replaced the complicated steps required to perform these calculations with their automated system AutoSolvate.

Previously, a computational chemist might have to type hundreds of lines of code into a supercomputer to run a simulation. The command-line interface for AutoSolvate, however, requires just a few lines of code to conduct hundreds of calculations automatically.

"The time for running the simulations may be long, but that's a job for the computer," Liu says. "We've freed the researchers from most of the tedious, manual tasks of data input so that they can focus on analyzing their results and other creative work."

In addition to the command-line interface geared toward more experienced theoretical chemists, AutoSolvate includes an intuitive graphical interface that is suitable for graduate students who are learning to run simulations.

Labs can now efficiently generate many data points for solvated molecules and then use the dataset to build machine-learning models for chemical design and discovery. AutoSolvate also makes it easier to build and share datasets across different research groups.

Setting the stage for machine learning

"During the past 10 years, machine learning has become a popular tool for chemistry but the lack of computational datasets has been a bottleneck," Liu says. "AutoSolvate will allow the research community to curate a huge number of datasets for molecular properties in the solution phase."

Determining the redox potential of a solvent molecule, or the likelihood for an oxidation to occur, is just one example of a key research area that AutoSolvate could help advance. Redox-active molecules hold potential for applications in the development of anticancer drugs and chemical batteries for renewable-energy storage.

"Building up redox-potential datasets will then allow us to use machine learning to look at millions of different compounds to rapidly find the ones with redox potential within the desired range," Liu says.

Instead of a black-box result, such analyses of large datasets can yield interpretable artificial intelligence, or basic rules for molecular models.

"The ultimate goal is to identify rules that can then be applied to solve a broad range of fundamental science problems," Liu says.

More information: Eugen Hruska et al, AutoSolvate: A toolkit for automating quantum chemistry design and discovery of solvated molecules, *The Journal of Chemical Physics* (2022). DOI: 10.1063/5.0084833

Provided by Emory University

Citation: New chemistry toolkit speeds analyses of molecules in solution (2022, November 29) retrieved 7 May 2024 from
https://phys.org/news/2022-11-chemistry-toolkit-analyses-molecules-solution.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.