UC San Diego
UC San Diego News Center

June 13, 2018 | By Warren Froelich

SDSC Comet and Machine Learning Simulates H2O with “Unprecedented Accuracy”

Pioneering Technique will be able to Predict Properties of Other Molecules and Materials

Water, H_2O.

In its molecular form, it’s about as simple as it gets: two atoms of hydrogen and one of oxygen, joined by a chemical bond. Yet, few question that this seemingly simple molecule may be the most important substance on Earth, vital for life and critical for the planet’s geology and climate.

As testimony to its value, for years scientists have tried to create the most accurate molecular model of water, using increasingly powerful computers to simulate its structure in all forms, from the smallest droplets of liquid to its solidified structure of ice. Nevertheless, many questions about some of the anomalous properties of water have remained challenging to answer.

Now, a team led by researchers at UC San Diego’s Department of Chemistry and Biochemistry and the San Diego Supercomputer Center (SDSC), has used machine learning techniques to develop models for simulations of water with “unprecedented accuracy.”

Their pioneering work, published online in April in *The Journal of Chemical Physics*, demonstrates how popular machine learning techniques can be used to construct predictive molecular models, in this case of water but applicable also to other “generic” molecules, based on quantum mechanical reference data. Molecular simulations using modern high-performance computing systems are key to the rational design of novel materials with applications ranging from fuel cells to water purification systems, atmospheric climate models and computational drug design.

“Although computer simulations have become a powerful tool for the modeling of water and for molecular sciences in general, they are still limited by a tradeoff between the accuracy of the molecular models and the associated computational cost,” said Francesco Paesani, professor of chemistry and biochemistry at UC San Diego and the study's principal investigator.

"Now that we've proved this concept with a model of water using machine learning techniques, we are currently extending this novel approach to generic molecules," he added, "meaning that scientists will be able to predict the properties of molecules and materials with unprecedented accuracy."

The new study builds on the highly accurate and successful "MB-pol many-body potential" for water developed in Paesani's lab, which recently has emerged as an accurate molecular model for water simulations from the gas to liquid to solid phases. "This is a new methodology that could revolutionize computational chemistry," said SDSC Director Michael Norman.

As reported in the paper, the researchers investigated the performance of three machine learning techniques – permutationally invariant polynomials, neural networks, and Gaussian approximation potentials – in representing many-body interactions in water. Machine learning typically involves 'training' a computer or robot on millions of actions so that the computer learns how to derive insight and meaning from the data as time advances.

In the quantum world, all three methods have been consistently equivalent in reproducing large datasets involving the interaction of multiple particles – "many body" phenomena such as two-body and three-body energies – as well as water cluster interaction energies, all with great accuracy.

"We have demonstrated that these different machine learning techniques can effectively be employed to encode the highly complex quantum mechanical many-body interactions that arise when molecules interact," said Thuong Nguyen, lead author of the study and a research scholar at UC San Diego when the research was conducted.

As for future efforts, these findings are not only important because the models are highly accurate, but also because it means researchers can choose the algorithms that best map to the available hardware, according to Andreas W. Goetz, a research scientist who directed the work at SDSC.

"Modern many-core processors, for instance, are well-suited to evaluate the complex expressions of the permutationally invariant polynomials, while massively parallel graphics processing units (GPUs) perform exceptionally well for neural networks," Goetz said.

The development of complex neural networks with associated optimization processes was performed on SDSC's *Comet* supercomputer GPU resources and *Maverick*, based at the Texas Advanced Computing Center (TACC), with allocations provided by the eXtreme Science and Engineering Discovery Environment (XSEDE).

Also participating in the study were researchers at the École Polytechnique Fédérale de Lausanne in Switzerland, Cambridge University in England, and the University of Göttingen in Germany.

This research was supported by the National Science Foundation through grant # ACI-1642336.

MEDIA CONTACT

Jan Zverina, 858-534-5111, jzverina@sdsc.edu