Lab leads effort to model proteins tied to cancer - Lab leads effort to model proteins tied to cancer

Lawrence Livermore National Laboratory researchers, along with scientists from Los Alamos National Laboratory, the National Cancer Institute and other institutions, are using machine learning as a virtual magnifying glass to study interesting regions of RAS protein/lipid simulations in higher detail. Credit: Tim Carpenter/LLNL

Computational scientists, biophysicists and statisticians from Lawrence Livermore National Laboratory (LLNL) and Los Alamos National Laboratory (LANL) are leading a massive multi-institutional collaboration that has developed a machine learning-based simulation for next-generation supercomputers capable of modeling protein interactions and mutations that play a role in many forms of cancer.

The research stems from a pilot project in the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program, a collaboration between the Department of Energy's (DOE) Office of Science, the National Nuclear Security Administration (NNSA) and National Cancer Institute (NCI) that is supported in part by the Cancer Moonshot. The work is being published by the 2019 Supercomputing Conference, held Nov. 17-22 in Denver, where it is among the finalists for the conference's Best Paper award.

The paper, which also includes contributions from Oak Ridge National Laboratory (ORNL), the Frederick National Laboratory for Cancer Research (FNLCR) and IBM, describes a predictive and multi-scale approach to model the dynamics of RAS proteins — a family of proteins whose mutations are linked to more than 30 percent of all human cancers — and lipid membranes, as well as the activation of oncogenic signaling through interaction with RAF proteins. NCI established the RAS Initiative in 2013 to explore the biology of mutant (oncogenic) RAS and to ultimately create new therapeutic opportunities for RAS-related cancers.

LLNL computer scientist and lead author Francesco Di Natale, who will present the paper at the conference, said the team took a broad approach to modeling RAS protein interactions. The team began with a macro-model capable of simulating the impact of a lipid membrane on RAS proteins at long timescales and incorporated a machine learning algorithm to determine which lipid "patches" were interesting enough to model in more detail with a molecular-level micromodel. The result is a Massively parallel Multiscale Machine-Learned Modeling Infrastructure (MuMMI) that scales up efficiently on large, heterogenous high performance computing machines like LLNL's Sierra and ORNL's Summit.

performance computing machines like LLNL's Sierra and ORNL's Summit.

While the concept of a multiscale simulation isn't new to molecular dynamics, Di Natale said, introducing machine learning to the previously manual process of selecting patches of interest is a novel approach, saving compute time and money on expensive simulations that may or not provide relevant information.

"The machine learning model lets us remove the human from the loop while still getting really relevant seed data," Di Natale said. "The benefit to having the process automated is that we start seeing multiple RAS proteins in a patch, and it's something that manually would be hard to do. Suddenly, we can see how far apart the RAS proteins naturally orient and then provide results to the community to more accurately model it, because we have all this data. You can start asking questions about what realistically happens rather than guessing at valid parameters."

The mechanism and dynamics of how RAS proteins interact, how they interact with RAF and promote oncogenic signaling and how the lipid (soluble organic compounds that help make up cell membranes) composition affect RAS activity are not well understood. For this paper, the team simulated the interaction between RAS and eight of the most relevant lipids to investigate RAS dynamics and interaction. They simulated a 1-by-1 micrometer membrane patch with 300 different RAS proteins to analyze the membranes in order to generate statistically relevant observations that can be tested experimentally at FNLCR.

"We decided that instead of doing what traditionally has been done with simulations — taking a model membrane with one or two lipids — that we'd try to make it realistic and model a biologically relevant membrane," said LLNL computational biologist Helgi Ingólfsson, a technical lead on the project. "The goal is to characterize RAS aggregation, RAS-protein interactions and RAS-lipid interactions, observing what types of lipids dictate RAS behavior and orient on the membrane. We want to see if we can modulate RAS activity with different types of lipids or some kind of pharmaceutical, not to eliminate RAS activity but modulate it in different ways, like promoting the inactive states."

Scientists from LANL spearheaded the efforts on establishing a gold-standard for this MuMMI framework using atomistic simulations and led the analysis of these large-scale simulations to extract mechanistic understanding of RAS biology in the context of a membrane. LANL expertise in uncertainty quantification was critical for identifying statistically significant biophysical properties within the multiscale framework.

Over the course of the project, the team found that RAS assemblies have a preferred membrane environment, and that using machine learning (ML) to automatically select certain patches for higher-resolution molecular-level simulations has distinct advantages over randomized patch selection, providing wider coverage of the phase space of the RAS lipid environment.

"Exploring the incorporation of AI and ML with high-performance simulation was an

"Exploring the incorporation of AI and ML with high-performance simulation was an integrating theme for the three DOE/NCI pilots that comprise JDACS4C," said Fred Streitz, pilot co-lead and chief scientist in the Artificial Intelligence and Technology Office in the Department of Energy, which coordinates AI projects across the DOE enterprise. "This application really highlights how powerful the two technologies can be, and how much more effective we can be, when working in together."

Frederick National Lab is performing experimental testing to ensure the models are representative of actual biological results. The models will help NCI carry out experiments to test predictions and generate more data that will feed back into the machine learning model, creating a validation loop that will produce a more accurate model, researchers said. The large simulation campaign allows addressing a large range of scientific question without the need for new simulations.

"There we can really see the true potential of having so many different types of simulations, both in so many different environments and also the sheer amount, and ask nearly any question we want," Ingólfsson said. "Now we can answer those questions after the fact because we have enough data to do it."

For the microscale model, the team used a molecular dynamics code adopted for the coarse-grained Martini model. It was adapted for GPUs to run on Sierra, making it likely the only general molecular dynamics code to run completely on GPUs, theresearchers said. The work stretched the limits of the early access Sierra system, as each "patch," representing an area of about 30 by 30 nanometers, contained about 140,000 coarse-grain beads and thousands of individual lipids.

While the system was still in its unclassified environment, the team ran nearly 120,000 simulations on Sierra, taking 5.6 million GPU hours of compute time and generating a massive 320 terabytes of data. The number of simulations was "staggering," researchers said, adding that the largest number of Martini simulations done at one time was only in the thousands prior to this project.

"Operating at this scale demonstrates a lot of the challenges that we're going to have to start discussing," Di Natale said. "You can only store so many hard drives before it becomes both spatially and financially untenable to manage that much data. This is a foray into potential next-generation problems. It's the first evidence of where the field is starting to go, and that in and of itself means more growing pains."

The team is continuing the work on Lassen, Sierra's unclassified companion system and currently the 10th fastest supercomputing system in the world, according to the Top500 list. They are preparing to run the next phase on the world's most powerful supercomputer, Oak Ridge's Summit, later this year, refactoring the code and adding various improvements.

The paper includes 13 other co-authors from LLNL, including Streitz, who is on assignment

The paper includes 13 other co-authors from LLNL, including Streitz, who is on assignment with DOE, and Harsh Bhatia, Peer-Timo Bremer, Tim Carpenter, Gautham Dharuman, Jim Glosli, Felice Lightstone, Tomas Oppelstrup, Tom Scogland, Shiv Sundram, Michael Surh, Yue Yang and Xiaohua Zhang.

Co-authors from other institutions are Dwight Nissley of the NCI RAS Initiative at Frederick National Laboratory, who leads the project for the NCI; Sandrasegaram Gnanakaran and Chris Neale from LANL; Sara Schumacher, Claudia Misale, Lars Schneidenbach, Carlos Costa, Changhoan Kim and Bruce D'Amora of IBM's Thomas J. Watson Research Center; and Liam Stanton of San Jose State University.