Generating Protein Structures For Pathway Discovery Using Deep Learning

Konstantia Georgouli[1], Mark Heimann[2], Harsh Bhatia[2], Timothy S. Carpenter[1], Felice C. Lightstone[1], Helgi I. Ingólfsson[1], Peer-Timo Bremer[2]

[1]Physical and Life Sciences, Lawrence Livermore National Laboratory, Livermore, CA, 94550. [2]Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA, 94550.

Abstract

When investigating the intricacies of biological phenomena down to the molecular level, there are fundamental limits in both length- and time-scales that can be probed experimentally. Thus, molecular dynamics (MD) simulations are often used to provide valuable biological insights beyond the scope of experiments resolution. While (approximate) MD simulations normally can provide the necessary spatial resolution, often the time-scales of interest are orders of magnitude too long to be explored even on today's supercomputers. One common problem is that of pathway discovery, where the start and end points of a scientific phenomenon of interest are known or can be estimated but the processes in between are unknown. Given this problem, here, we are using the power of physics-informed deep learning data representation in combination with massive multiscale simulation ensembles to bridge this gap. In particular, we are interested to explore, capture and study the dynamic conformational changes of the RAS-RAF complex along their activation pathway in the context of cancer biology.

Given two simulation ensembles of the start and end conformational states, we train a data representation using deep learning to model all the observed protein conformations in a lower dimensional latent space. The representation is based on a multi-resolution representation of pairwise distances enhanced by a physics-informed loss term using the known MD force field. Using the produced latent space of the representation, we deliberately perform extrapolation between the known ensembles to generate new candidates for future simulations that we automatically execute using **Mu**ltiscale **M**achine-learned **M**odeling **I**nfrastructure (MuMMI) [1, 2]. Ultimately, the new simulations iteratively re-train the system in order to create simulation ensembles that form a densely sampled and simply connected set of conformations in latent space. Assuming sufficient sampling, a simple Dijkstra's traversal in connection with the generative model of the representation provides possible transition paths that can subsequently be validated and optimized using existing techniques. Latent space extrapolation represents a new approach to explore MD time-scales significantly beyond existing capabilities, that is applicable to large protein complexes, and for the first time can provide insights into prolonged signaling events.

Early results on a simplified problem show that the proposed framework can successfully generate simulation ensembles of the pathway that bridges between the two initial ensembles of interest and confirm the value of such a computational approach to establish a detailed understanding of protein dynamics.

References:

[1] Di Natale, F.; Bhatia, H.; Carpenter, T.S.; et al. 2019. A massively parallel infrastructure for adaptive multiscale simulations: modeling RAS initiation pathway for cancer. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-16). Link

[2] Ingólfsson, H.; Neale C.; Carpenter, T.S.; et al. 2022. Machine learning–driven multiscale modeling reveals lipid-dependent dynamics of RAS signaling proteins. PNAS, 119 e2113297119. Link.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-ABS-844988.