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.

Method
Description of the data

Figure 1: In this preliminary study, we bridge between two simulation ensembles of a known problem, how KRAS conformational changes of the start and end conformational states and we denote them as Ensemble a and Ensemble b. (Ensemble a – CRD membrane depth < 5 nm, Ensemble b – CRD membrane depth > 1.5 nm). This plot shows a summary of the data distribution with respect to the CRD distance.

Description of the Deep Learning Model

Given an input data (X, D) where the flattened Cartesian coordinates of the 751 CG beads (n=2253), X, represent a protein conformation, and a set of hierarchical distances (n=15328), D, represent the corresponding hierarchical pairwise distances. Our goal is to learn the mappings (X, D) → z → X and an autoencoder is an appropriate solution for this problem in order to create a meaningful protein latent space.

The following architecture has been defined (MSE loss, Adam optimizer):

During the training, 10% of the data was randomly selected for the validation of the model.

Protein structure generator

(a) before
(b) after

Distribution of CRD distance in real validation structures

Distribution of CRD distance in generated structures

References

Acknowledgements

Conclusions and future work

• We present a deep learning-based approach to solve the problem of pathway discovery.

• Preliminary results show that the proposed framework is capable of producing simulation ensembles of the pathway bridging the two initial ensembles of interest.

• Our approach confirms the value of such a computational approach to establish a detailed understanding of protein dynamics.

• In future, we plan to expand the Multiscale Machine-learning Modeling Infrastructure (MuMMI) [2,3] to support iterative refinement of ML predictions of trajectories.

• We will investigate other sampling techniques in our structure generator such as the Optimal Transport Propagation (OTP). We are going to explore, capture and study the dynamic conformational changes of the KAS-Rho complex along their activation pathway in the context of cancer biology.