Incorporating parameter sampling in MELD to improve protein structure determination using semi-reliable data

Bhumika Singh and **Alberto Perez**

Quantum Theory Project, Department of Chemistry, University of Florida, Gainesville, FL

The rapid development in machine-learning algorithms, improved protein-energy functions, and advances in protein conformational sampling methods have dramatically enhanced our ability in protein structure prediction, even without using any structural template. In CASP14, a machine learning-based model, AlphaFold, accurately predicted static structures of monomeric proteins. Despite its success, it faces the limitation of not giving any information about the protein's dynamics crucial to understanding protein function and regulation. Furthermore, experimental characterization studies are required to validate its predicted structure models – but the determination of the structure via NMR, X-ray crystallography, or CryoEM is not always possible. Our group is interested in integrating experimental data that is insufficient on its own to determine protein structures with simulations. We developed the Modeling Employing Limited Data (MELD) as a physics-based Bayesian inference that combines prior belief with semi-reliable experimental data to resolve data ambiguity and produce a statistically consistent ensemble of structures instead of a single structure. One general challenge in using experimental data is estimating the noise and ambiguity in the original data set, which informs simulations of how much data to enforce. This is a critical step in Bayesian inference: trusting too much data will lead to incorporating noise into the predictions, resulting in incorrect models. While at the opposite, extreme simulations will face problems of convergence and efficiency in sampling phase space and ultimately reduce the performance of the method. In our current study, we use an improved version of MELD that performs Bayesian sampling of parameters based on a given force field and data so that the algorithm can determine the accuracy of the semi-reliable dataset "on the fly". We have applied this methodology to 12 protein targets which we compare to previous results.