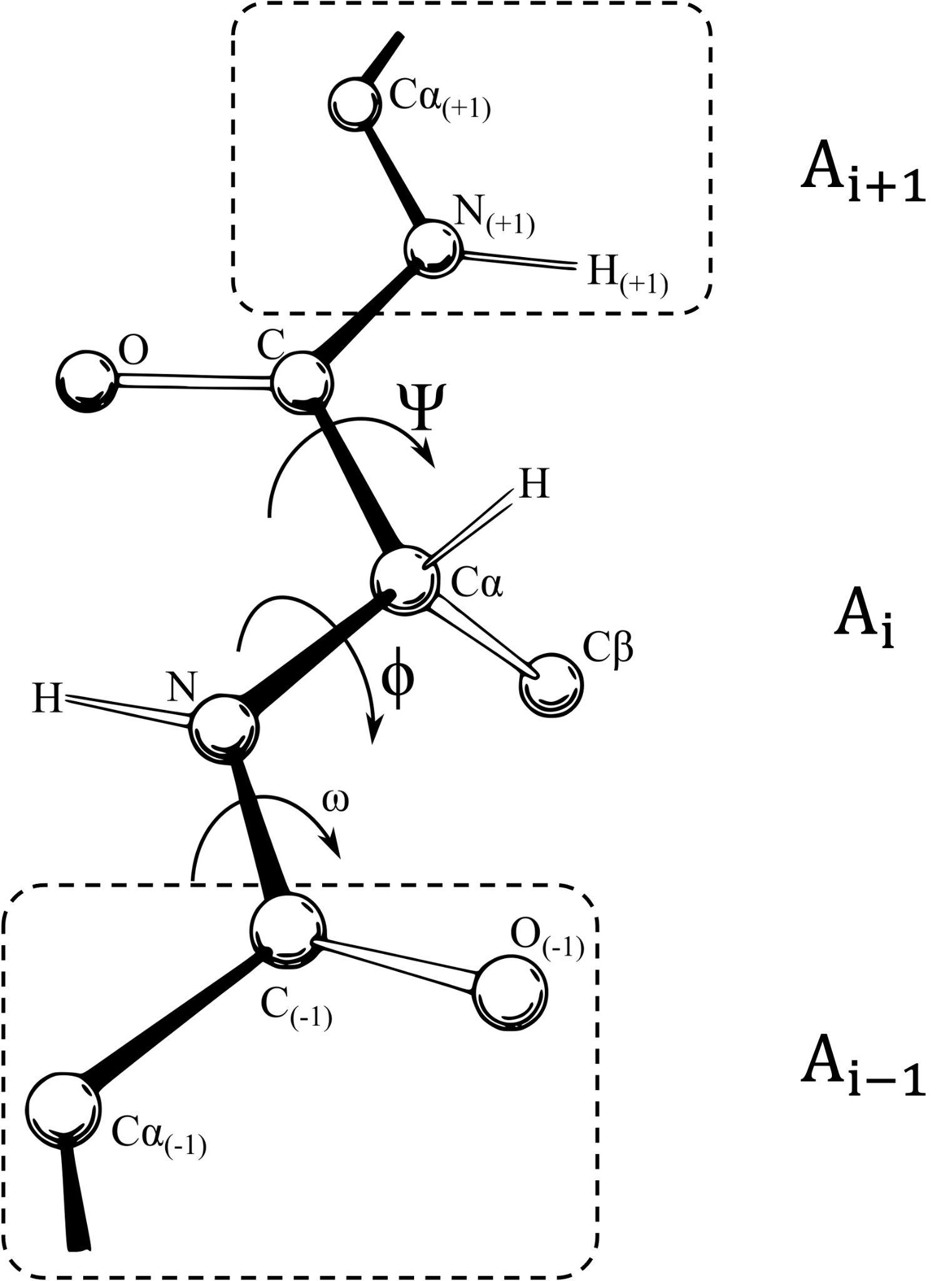
**Advanced Machine Learning (CPSC-8420), S2022**

**Final Project Proposal**

**Gangtong Huang, Zhenzhen Zhang**

**Motivation:**

To analyze the movement of molecules physically, people use molecular dynamics methods to investigate the dynamic evolution of a system. Discrete molecular dynamics (DMD) is a special type of molecular dynamics simulation where pairwise interaction potentials are modeled with discontinuous functions. Same as classical molecular dynamics methods, DMD simulations are also based on the conservation laws of energy, momentum and angular momentum. In actual cases, some detailed structure orientations are strictly defined because of the specific chemical bond lengths and bond angles between bonded atoms, so some additional constraints are needed to confine the collective movement of atom groups so that the dihedral angles formed by these atoms will be restricted in a certain range. For example, as shown in Fig. 1, for each three neighboring amino acids , , and , we should provide constraints of the distance between the backbone alpha carbon atoms and which are correlated to the amino acid sequences and local protein structures. The value of distance solely depends on the backbone dihedral angle of amino acid, . The dihedral angles can be calculated from the protein files including atomic 3D cartesian coordinates which can be downloaded from RCSB online protein data bank (<https://www.rcsb.org/>). Due to the large number of triplet sequences (), the distance data for many sequences is sparsely populated. In 2012, a Bayesian statistical analysis method was used to accurately estimate the relative distribution of these sparsely populated three neighboring amino acids, and the results were assigned as bias potentials as the function of atom distances [1][2]. In the past ten years, with the development of structural biology techniques such as solid-state NMR and cryogenic electron microscopy, more and more protein structures with high resolution have been solved and added to the RCSB online protein data bank. With more entries and developed methods, we are able to recalculate the distance data. The updated data will improve the accuracy of protein structure prediction by DMD simulations, which are extensively used in the study of various biomolecular systems.



**Method:**

The methods involved in this project include SVD and Bayesian statistics. The DMD constraints on the dihedral angles between adjacent amino acid backbone plains are transformed into the constraints between the next-nearest Cα atoms using geometrical relations. The constraints are expressed in the form of potential energy functions based on the conditional probability of the dihedral angles given a 3-amino-acid sequence . The calculation of the conditional probabilities will be mainly based on Bayesian statistics applied to the experimentally measured protein structures in the RCSB Protein Data Bank. In naturally occuring proteins with solved structures, some sequences, such as *TrpTrpTrp*, are very rare, which will increase the error in the estimation of the probabilities. To address this problem, we will make a prior estimation of the as a product , where both and should be considerably larger in the available experimental data. By assuming that has a form of Dirichlet distribution, we will then update the value of with Bayesian-estimated occurrence frequency combined with a correction factor based on experimentally observation frequency to obtain the posterior estimation, which is smoother than the raw counts and will be used to construct the DMD constraints. This process will be enumerated for all possible 20\*20\*20 combinations of to obtain the complete set of constraints that are applicable to future DMD studies.

**Intended experiments:**

1. The construction of a training dataset. In this step, we will download all the available protein structures in the PDB data bank. Some proteins, such as lysozyme, comes with a huge number of highly similar available structures in the PDB data bank, which can cause the over representation of certain combinations . The repetitiveness of structures can be evaluated with respect to amino acid sequence similarity, and we will construct a non-repetitive PDB training dataset. The experimentally observed based on raw counts can then be calculated.
2. The machine learning process. The learning process will be performed on the statistics of for all using methods based on Bayesian estimations and the smoothing method as outlined in the **Method** section.
3. Test dataset and evaluation. The learning outcome will be compared to (and benchmarked against, if time permits) the original version of Ding et. al., 2012.

**Reference**

[1] Shirvanyants, D., Ding, F., Tsao, D., Ramachandran, S., & Dokholyan, N. V. (2012). Discrete molecular dynamics: an efficient and versatile simulation method for fine protein characterization. *The journal of physical chemistry B*, *116*(29), 8375-8382.

[2] Dunbrack Jr, Roland L., and Fred E. Cohen. "Bayesian statistical analysis of protein side‐chain rotamer preferences." *Protein Science* 6.8 (1997): 1661-1681.