Project Proposal: Accelerated Diffusion Models for Protein Structure Generation

Background Diffusion models have emerged as the new state-of-the art family of deep generative models. They have shown great potential in a variety of domains ranging from computer vision, natural language processing, computational chemistry and bioinformatics. Intuitively, diffusion models are of two Markov processes. The forward process that smoothly perturbs data by adding noise and the reverse process that converts noise back to data via iterative denoising steps corresponding to training and inference procedure, respectively. Each denoising step in the reverse process typically requires estimating the score function, which is a gradient pointing to the directions of data with higher likelihood and less noise.

On the other hand, the diffusion model has the inherent drawback of long training and sampling time compared to Generative Adversarial Networks (GANs) and Variational Auto-Encoders(VAEs). Since diffusion models leverage a Markov process to convert data distribution by tiny perturbations, a large number of diffusion steps are required in both training and inference phases. This is particularly a bottleneck for diffusion-based generative models in the field of protein structure generation, considering that a large amount of generated structures are usually needed to be generated given a comparably short period of time. We hence propose to enhance current diffusion-based models for protein structure generation with acceleration algorithms to improve their performances.

Previous Work Many samplers for diffusion models rely on discretizing either the reverse-time SDE present or the probability flow ODE. Since the cost of sampling increases proportionally with the number of discretized time steps, many researchers have focused on developing discretization schemes that reduce the number of time steps while also minimizing discretization errors. Noise-Conditional Score Networks (NCSNs) and Critically-Damped Langevin Diffusion (CLD) both solve the reverse-time SDE with inspirations from Langevin dynamics. DDIM sampling process amounts to a special discretization scheme of the probability flow ODE. And Diffusion Exponential Integrator Sampler and DPM-solver leverage the semi-linear structure of probability flow ODE to develop customized ODE solvers that are more efficient than general-purpose Runge-Kutta methods.

Learning-based sampling is another efficient approach for diffusion models. By using partial steps or training a sampler for the reverse process, this method achieves faster sampling speeds at the expense of slight degradation in sample quality. Unlike learning-free approaches that use handcrafted steps, learning-based sampling typically involves selecting steps by optimizing certain learning objectives. The main methods in this aspect are optimized discretization, truncated diffusion and knowledge distillation.

Our Schedule We will first try to reproduce some baseline models that achieved success in protein structure generation. Next we will apply these acceleration algorithms that are generally tested on computer vision tasks to our protein structure generation models. The main obstacle is that protein structures enjoy SE(3)-equivariance, meaning that a certain structure remains the same during translation and rotation, to which we will particularly pay attention when combining current acceleration algorithms with diffusion-based generative models for protein structure generation.