ProGAE: A Geometric Generative Model for Disentangling Protein Conformational Space

Introduction

Discussing the generation of protein conformations

Recent work has investigated using machine learning to model the conformational space of proteins (Bhowmik et al., 2018; Ramaswamy et al., 2020).

We are interested in achieving greater interpretability of the models used to generate the conformational space.

Namely, we consider more detailed geometric interpretability.

We propose a novel architecture, ProGAE:

Inspired by recent work on unsupervised geometric disentanglement (Tatro et al., 2020)

A geometric autoencoder that directly learns from 3D protein structure

Separately encodes intrinsic and extrinsic geometries for greater latent space interpretability.

Network Architecture

A model for separately encoding intrinsic and extrinsic protein geometry

The input of ProGAE includes:

An intrinsic signal; the length of pseudobonds in the protein trace

An extrinsic signal; the orientations of bonds in the protein backbone

These input signals are separately encoded:

To create two distinct latent spaces for greater generative control

The corresponding ProGAE output, after joint decoding, are the 3D coordinates of the backbone atoms.

The structure of the architecture uses geometric convolutional layers:

Variants of graph convolutions

Allows the architecture to scale for large proteins such as the Sars-Cov-2 S protein simulation from one of our datasets.

Results

Establishing the contributions of intrinsic and extrinsic geometric signals to protein reconstruction

We summarize our results:

ProGAE is able to reconstruct our proteins from datasets to within the experimental resolution associated with simulation.

As the datasets used are simulations of proteins binding to experimental drugs.

The extrinsic latent space can be used to classify the drug the protein is bound to and infer physiochemical properties

The presence of the intrinsic signal improves the quality of bond geometry in the reconstructions.

Acknowledgments

This work was supported by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons)

R. Lai is supported in part by NSF CAREER Award (DMS—1752934).

References

