ProGAE: A Geometric Generative Model for Disentangling Protein Conformational Space

N. Joseph Tatro[1], Payel Das[2], Pin-Yu Chen[2],
Vigil Chenthamarakshan[2], Rongjie Lai[1]

Rensselaer Polytechnic Institute[1]

IBM Research[2]

ICLR 2021 Workshop
Deep Learning for Simulation (simDL)

Introduction

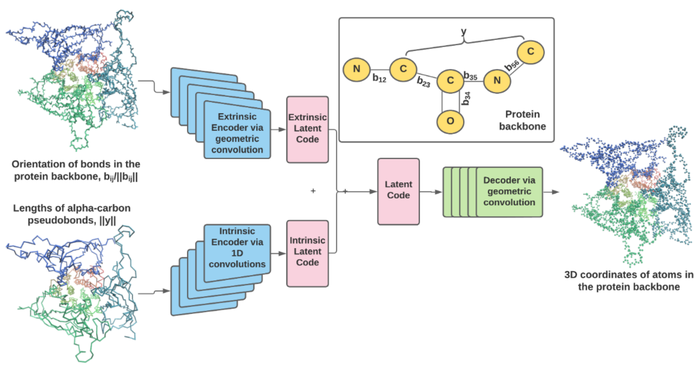
Discussing the generation of protein conformations

- 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

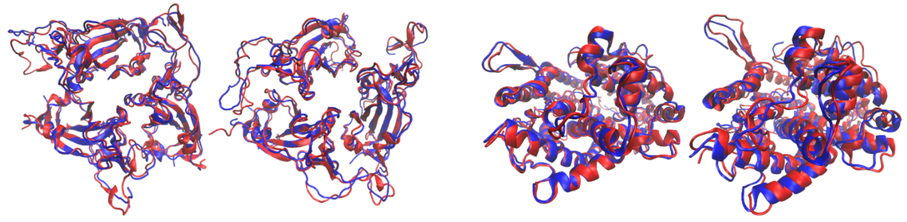
- The input of ProGAE includes:
  - An **intrinsic** signal; the length of pseudbonds in the protein trace
  - An **extrinsic** signal; the orientations of bonds in the protein backbone
- These input signals are separately encoded:
  - To create two distinct latent spaces for greater generative control
- 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



A diagram of the ProGAE architecture, displaying the separate encoding of the intrinsic and extrinsic signals, and the following joint decoding to backbone atom coordinates.

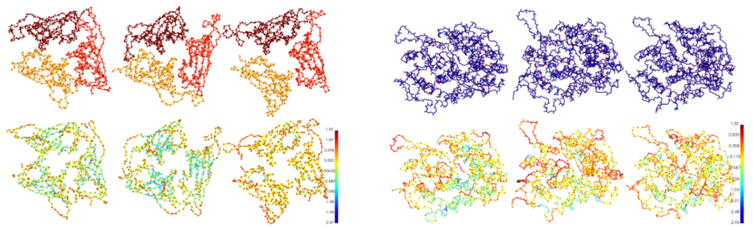
Results

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



ProGAE reconstructions of S protein (left) and hACE2 data (right). Blue and red structures correspond to the reconstructed and ground truth structures, respectively.

- 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

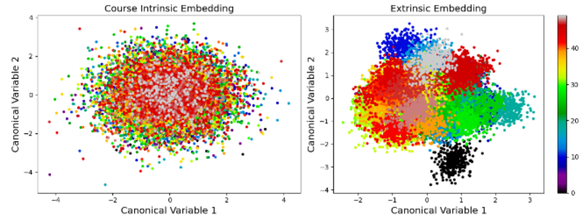


(a) 3 RBDs of S protein (b) Ectodomain of Human ACE2

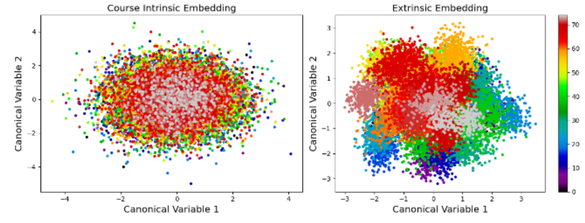
Reconstructions of proteins using ProGAE. The top row displays the ground truth, while the bottom row displays the corresponding generation by the network. Color in the bottom row indicates the log of atom-wise L2 error.

Dataset		Molecular weight	Hydrogen bond donor count	Topological polar surface area
S protein	PCA error (σ)	0.78 ± 0.00	0.81 ± 0.01	0.79 ± 0.00
	Latent error (σ)	$\mathbf{0.55 \pm 0.04}$	$\mathbf{0.56 \pm 0.03}$	$\mathbf{0.61 \pm 0.00}$
hACE2	PCA error (σ)	0.71 ± 0.00	0.65 ± 0.00	0.73 ± 0.00
	Latent error (σ)	$\mathbf{0.55 \pm 0.01}$	$\mathbf{0.57 \pm 0.01}$	$\mathbf{0.53 \pm 0.02}$

Results of linear regression on the extrinsic latent space for predicting physical/chemical properties of the drugs that a protein is bound to. Error is normalized for interpretability. For comparison, performance of linear regression on the PCA embeddings of the orientation of the backbone bonds is reported.



(a) S protein



(b) Human ACE2

Projections of the ProGAE intrinsic and extrinsic latent embeddings, with color denoting the drug responded to in simulation. Clearly the extrinsic space is capturing the variation due to ligand binding.

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).

Dataset		C-CA	C-N	C-O	CA-N	CA-CA
S protein	Int.+Ext. (%)	14.41	22.58	27.00	15.24	15.33
	Ext. Only (%)	19.04	27.34	27.66	18.58	20.89
	Diff of Adding Int.	-4.63	-4.76	-0.66	-3.40	-5.56
hACE2	Int.+Ext. (%)	2.45	8.19	12.07	4.62	0.51
	Ext. Only (%)	4.99	9.81	12.34	5.21	1.59
	Diff of Adding Int.	-2.54	-1.62	-0.27	-0.59	-1.08

Percentage of bonds that are 10% shorter than the minimum seen in training data. The difference (Diff) between the intrinsic+extrinsic ProGAE and the extrinsic-only ProGae is reported.

References

Debsindhu Bhowmik, Shang Gao, Michael T Young, and Arvind Ramanathan. *Deep clustering of protein folding simulations*. BMC bioinformatics, 19(18):47–58, 2018.

Venkata K. Ramaswamy, Chris G. Willcocks, and Matteo T. Degiacomi. *Learning protein conformational space by enforcing physics with convolutions and latent interpolations*, 2020.

N. Joseph Tatro, Stefan C. Schonsheck, and Rongjie Lai. *Unsupervised geometric disentanglement for surfaces via cfan-vae*, 2020.