BITS :: Call for Abstracts 2023 - Oral communication

Type	Oral communication
Session	Structural Bioinformatics
Title	Direct generation of protein structural ensembles via deep generative modeling
All Authors	Giacomo Janson (1), Gilberto Valdes-Garcia (1), Lim Heo (1) & Michael Feig (1)

Affiliation

(1) Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA.

Motivation

Since the activity and regulation of a protein is determined by the dynamical properties of its 3D structure, Structural Biology is increasingly shifting towards the sequence -> structure -> dynamics -> function paradigm. In light of this, the characterization of structural ensembles of proteins is crucial. Computational methods for sampling conformational spaces, such as molecular dynamics (MD), provide an essential help. Unfortunately, the high-dimensionality and large kinetic barriers of protein energy landscapes result in significant computational costs which limit the complexity of systems that can be simulated. Therefore, novel strategies for accelerating the generation of biologically relevant structural ensembles are greatly needed. Motivated by this, our research group has endeavored in developing machine learning (ML) methods, leveraging recent advances in deep generative models (DGMs), with the goal of modeling structural ensembles at reduced computational costs. In this presentation, our recent progresses in this field will be highlighted.

Methods

DGMs are unsupervised ML methods that use neural networks to learn probability distributions with the goal of efficiently sampling from them. We adopt this class of models and train them on molecular simulation data to generate physically-realistic protein conformations for constructing full conformational ensembles. As a proof-of-principle, we used a generative adversarial network (a type of DGM) based on a transformer neural network and trained it on datasets of coarse-grained simulations of intrinsically disordered proteins (IDPs) and peptides. These molecules were chosen as testbeds because of their highly dynamical nature, which we aimed to capture via a DGM. Our model, called idpGAN, was trained on data from numerous protein sequences with the goal of learning rules relating primary sequence to structural dynamics. Once trained, the model can be used to efficiently generate structures for novel, arbitrary sequences, circumventing the need to run expensive simulations.

Results

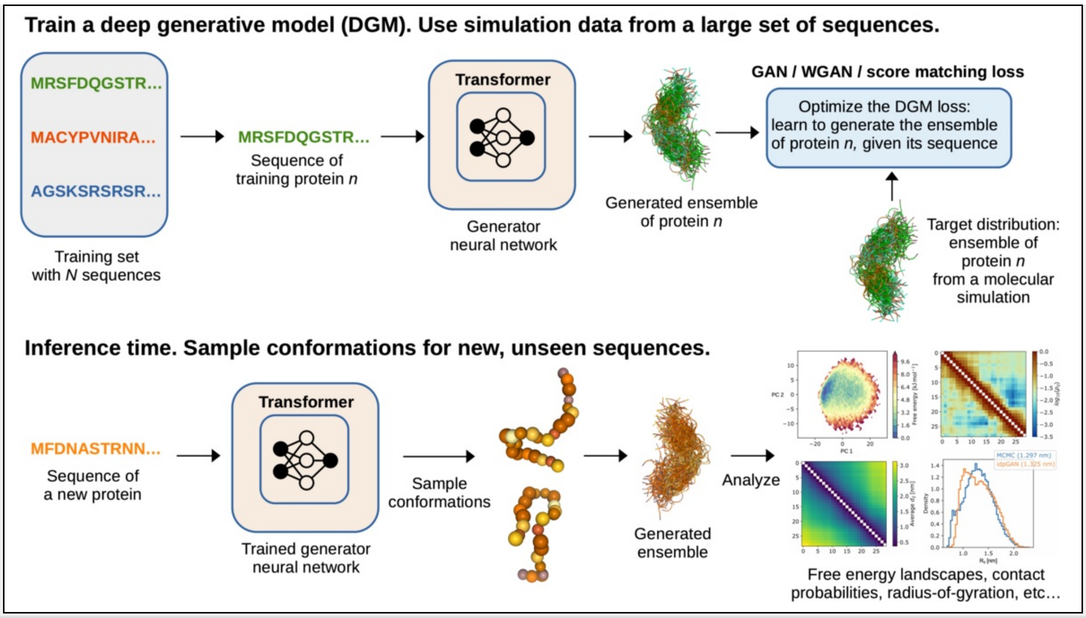
We rigorously evaluated idpGAN and demonstrate that it can predict sequence-dependent ensembles for sequences not present in the training set, showing that transferability can be achieved beyond training data. While other DGMs have been previously applied on simulation data of single protein systems, to our knowledge, idpGAN is the first ever DGM to be applied on simulations of multiple protein sequences with the goal of learning a general model. Our results show that DGMs are a promising strategy to model conformational ensembles of proteins at greatly reduced computational costs. Results will also be shown for the application of diffusion models (another type of DGM) to model challenging protein systems at increasing accuracy. In the final part of the presentation, the strengths and weaknesses of DGM-based approaches will be compared to classical physics-inspired simulation methods and the promises and challenges that await future research in this rapidly-evolving field will be highlighted.

Info

-

filename	figure.png

Figure


Train a deep generative model (DGM). Use simulation data from a large set of sequences.
MRSFDQGSTR...
MACYPVNIRA...
AGSKSRSRSR...
Training set with N sequences
MRSFDQGSTR...
Sequence of training protein n
Transformer
Generator neural network
Generated ensemble of protein n
GAN / WGAN / score matching loss
Optimize the DGM loss: learn to generate the ensemble of protein n, given its sequence
Target distribution: ensemble of protein n from a molecular simulation
Inference time. Sample conformations for new, unseen sequences.
MFDNASTRNN...
Sequence of a new protein
Transformer
Trained generator neural network
Sample conformations
Generated ensemble
Analyze
MCMC (1.297 nm)
idpGAN (1.325 nm)
Free energy landscapes, contact probabilities, radius-of-gyration, etc...

Availability	-

Corresponding Author

Name, Surname	Giacomo, Janson
Email	jansongi@msu.edu
Submitted on	20.04.2023