Simulating cryo-EM maps using a Generative Adversarial Network (cryo-GAN)

Mark Rozanov · Haim J. Wolfson
Blavatnik School of Computer Science, Tel-Aviv University

Abstract

The ability to perform a realistic simulation of a cryo-EM density map (at appropriate resolution) is crucial for the development and analysis of advanced cryo-EM based macromolecular structure modeling algorithms. In particular, such simulation is required for the recent machine learning based algorithms, which have to be trained on very large databases of "realistic" data. Such large databases are currently unavailable, and thus have to be simulated. The currently used simulation methods, such as pdb2mrc from the Chimera package, usually perform averaging on atomic resolution structures from the PDB. It is well known, that the resulting simulative structures (at the appropriate resolution) are significantly more similar to the atomic resolution structures than experimental cryo-EM maps at the same resolution, and thus do not represent the same degree of difficulty for the tested algorithms. Cryo-GAN exploits recent deep learning techniques for creating cryo-EM maps which are indistinguishable from experimental ones. It uses the VAE-GAN architecture which is a compound Generative Adversarial Network (GAN) with Variation AutoEncoder (VAE). Both are proven deep learning techniques for generating 2D and 3D images. The resulting synthetic maps have a high resemblance to the experimental maps. An independent discriminator was trained to distinguish between experimental and synthetic maps. In our computational experiments the discriminator marks all voxels of a map generated by existing simulation (pdb2mrc) as non-real. Less than half of the voxels of a map generated by Cryo-GAN are marked as non-real, meaning the map is indistinguishable from an experimental one.

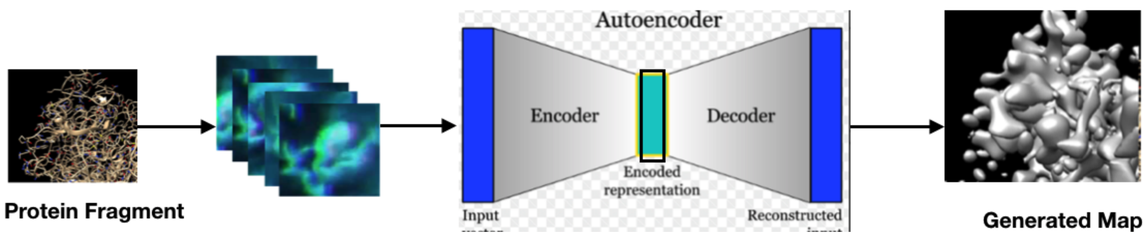
1. Motivation: When a simulation is required

- Create Appropriate Training DataSets for Deep Learning Algorithms: AAnchor [7],[5]. and others.
- Improve the performance of fitting algorithms: Ematch [2] , PowerFit [1], Multifit [8]
- Test the performance of novel modelling algorithms

2. Current Method: pdb2mrc from EMAN [6]

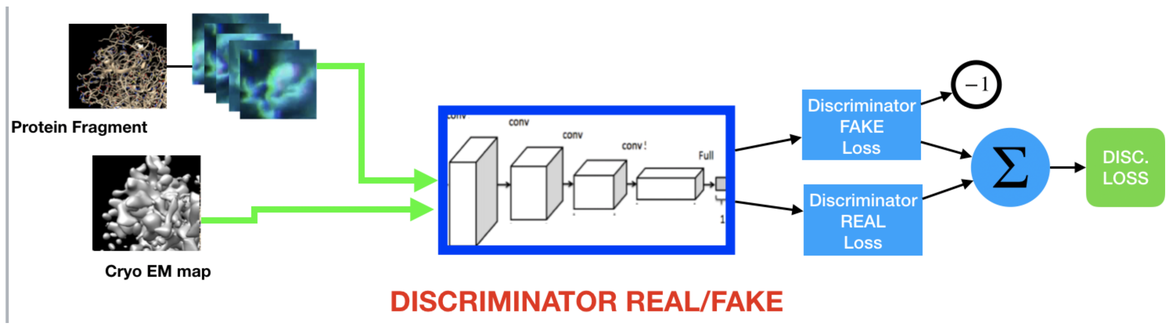
	pdb2mrc	real map
represents an "ideal" universe	✔	✗
contains experimental inaccuracies and real-world noise	✗	✔
affected by physico-chemical properties such as charge, bonds	✗	✔

3. Generating a Realistic Map - use VAE [4] Neural Network



Input: Protein Structure Fragment after Voxelization; **Output:** 3D cryo-EM map

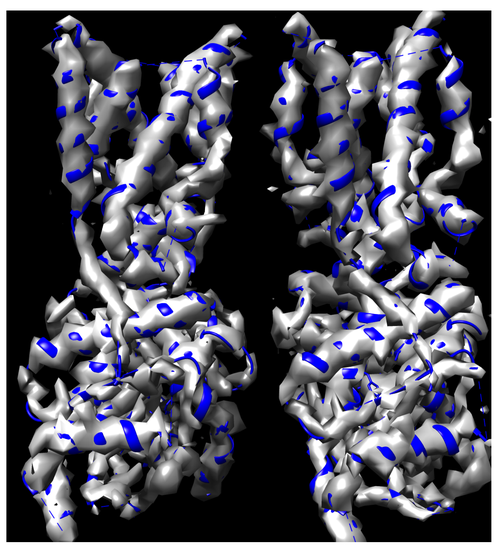
4. Make more Realistic - Adversarial Approach [3]



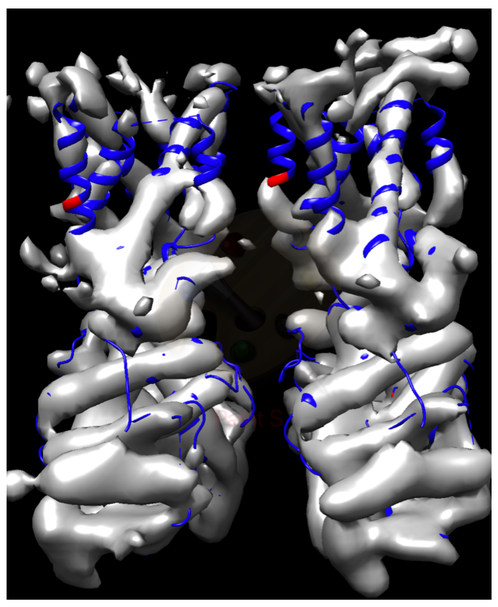
The role of a **discriminator** is to distinguish between real (input) and fake (generated) map.

6. Results: cryoEM map of STING protein (pdb ID 6nt8)

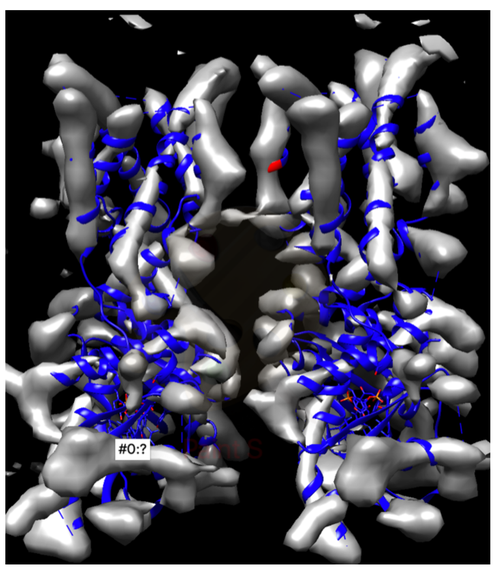
Existing Simulation - pdb2mrc



Real Map - EMD0505



Deep Simulation -cryoGAN

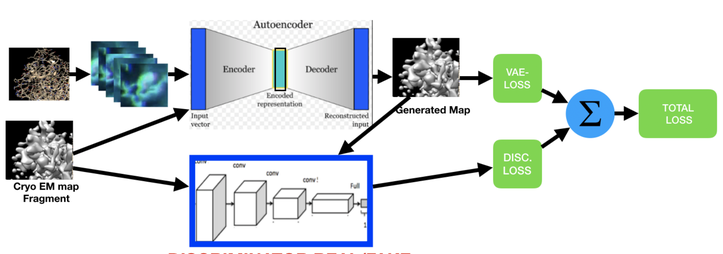


7. Evaluation

- Visual inspection shows that the generated map has the features of an experimental map.
- We trained an additional **discriminator** to distinguish between real and "fake" maps
- The REAL/SYM discriminator labels as real 41000 from 73000 test voxels from map

e-mail: *markroza@tauex.tau.ac.il*

8. VAE-GAN [9]: Training VAE network to confuse the discriminator



References

[1] G. C.P.van Zundert and A. M.J.J. Bonvin. Fast and sensitive rigid-body fitting into cryo-EM density maps with PowerFit. *AIMS Biophysics*, 2015.

[2] O. Dror, K. Lasker, R. Nussinov, and H. Wolfson. EMatch: An efficient method for aligning atomic resolution subunits into intermediate-resolution cryo-EM maps of large macromolecular assemblies. *Acta Crystallogr., Sect D*, D63(1):42—49, 2007.

[3] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[4] Larsen. Autoencoding beyond pixels using a learned similarity metric. 2016.

[5] R. Li, D. Si, T. Zeng, S. Ji, and J. He. Deep convolutional neural networks for detecting secondary structures in protein density maps from cryo-electron microscopy. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2016.

[6] S. J. Ludtke, P. R. Baldwin, and W. Chiu. EMAN : Semiautomated Software for High-Resolution Single-Particle Reconstructions. *Journal ofStructural Biology*, 128:82–97, 1999.

[7] M. Rozanov and H. J. Wolfson. AAnchor: CNN guided detection of anchor amino acids in high resolution cryo-EM density maps. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 88–91, 2018.

[8] E. Tjioe, K. Lasker, B. Webb, H. J. Wolfson, and A. Sali. MultiFit: A web server for fitting multiple protein structures into their electron microscopy density map. *Nucleic Acids Research*, 2011.

[9] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. Technical report.

Acknowledgments

Acknowledgments: This work has been supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv Univesity, by Len Blavatnik and the Blavatnik Family Foundation ,and by the I-CORE program of the Budgeting and Planning Commitee