Review: Inferring a Continuous Distribution of Atom Coordinates from Cryo-EM Images using VAEs
https://arxiv.org/pdf/2106.14108.pdf

The major goal of this manuscript is to create a procedure to model the ensemble of conformations present in a cryoEM particle stack. This goal is accomplished by using a variational autoencoder with special attention paid to the ensemble nature of proteins and the low signal-to-noise in cryoEM images by drawing multiple samples from the posterior and summing their corresponding likelihoods in training. Using “simulated” data where the conformations are derived from MD simulation snapshots and the particle projections are generated using reasonable assumptions, the procedure captures conformational heterogeneity that is coarsely similar to the input distribution. The lack of correlation between individual input snapshots and the corresponding individual output conformation point to how the procedure drives sampling to allow for differences in the input and output distributions (rather than just outputting individual snapshots). How this is accomplished, especially without bias, toward the “mode” conformation are not well explained in this manuscript. Part of this may be semantic as the language used is quite different from other papers that model conformational heterogeneity of macromolecules. Similarly, we suspect there is something to learn from examining the details of how the output structures are “wrong” in Figure 8a (the streaks of conformations at 6 and 36; the lack of sampling at distances: 8.5/34).

Overall, this procedure is radically different from others that approach this problem and therefore very exciting, given the promising results on simulated data shown here.

Major Comments

1. How is the base state chosen? Are the differences you observe similar to the differences from 2 very different potential base states?
2. Why invent a pose portion of the prediction process when we have great software packages that do poses very well? Could you start your method using output from another software package and just get conformational output or is there something to be gained by implementing this that goes beyond what is contained in the manuscript (a future direction?)?
3. Why is backbone continuity the only biological/chemical prior? Would you expect others to help or just add cost? How would you predict the backbone continuity loss to behave across a section of protein with much worse resolution compared to the rest of the protein? What about in a region that was not modeled at all in the reference structure or modeled incorrectly?
4. Can you provide more details about the run to run variance you observed? How different are the loss values you are obtaining and how different are the output structures. What are the number of models you would recommend a user to run?
5. What is the radius of convergence for how big a conformational change or incorrect model you can start from/get to?

6. How do the output structures from your analysis behave in conventional validation metrics? For example:Ramachandran plots/statistics, rotamer analysis, and bond lengths/angles. Can you process the particle stack using a conventional workflow to generate a map and calculate real space correlations, EMRinger scores, Qscore, etc.
7. Are there plans to address defocus? This seems like an important limitation of the current method. What would the loss function look like for keeping that value reasonable, if refined?
8. Please clarify the relationship between *z* and the residue separation in Figure 6. Was there one latent variable, and the distance correlates with that *z*? Or is there a more complex approach?

Minor Comments

1. What resolution range do you expect your method to be useful for? How much varied resolution across the structure is acceptable with good outcomes?
2. Please provide PDB(s) of your output structures.
3. Please provide images of the broken H bonds in the beta sheets as mentioned in supplementary video 1.
4. Please provide images of the degenerate sheet-like structures. Is this the same as the beta sheets mentioned?
5. Please highlight the PDB models that have these degenerate sheet-like structures.

Signed:

Stephanie Wankowicz (UCSF)
Iris Young (UCSF)
Daniel Asarnow (UCSF)
James Fraser (UCSF)