Day two - questions and answers

Q&A with AlphaFold Experts	**Oleg Kovalevsky , Augustin Zidek**

Q: I'm still a little unclear on whether NMR structures were used in AlphaFold (2 or 3) training, and in case they were, was it a single structure per ensemble, or multiple structures from an ensemble?
A:They were used in training. We used a single structure per ensemble

Q: Is there a threshold for a deep enough MSA?
A: It depends on the variability within the sequences, usually everything more than a few hundred work well, but around one hundred also may work for some cases

Q: is FAPE no longer part of AF3?
A: It is no longer used. We use frames only for training PAE

Q: For AF3, did the diffusion approach mean lower-resolution structures could be used for training?
A: Cutoff was 9 A resolution structure, fairly high.

Q:If AF3 is less dependent on coevolution since it processes alignments differently: what does it then glean from the alignments? Conservation?
A: Still learns similar info as AF2, likely less to it. Learns important and fundamental input but less sensitive than AF2.

Q: Is AlphaFold scalable (using the terminology of LLMs) ? Does it saturate or improve with scale across network size; data size; "test-time compute", e.g. number of "recycling" steps, diffusion steps ?
A: That is an extensive research question. But we do not believe it to follow scaling laws indefinitely due to the limited structural data available.
Follow-up: did you try to derive any laws/formulas by any axis? E.g. number of parameters

Q: How well suited is AF3 in working with protein data of immunological cells, T-Cells in particular?
A: Similar quality of predictions as with antibodies apparently

Q: Is alpha fold 3 as strongly affected by MSA depth as alpha fold 2?
A: AF3 requires deep enough MSA in general case. E.g. I have an example of an orphan protein with only about ~10 sequences in the MSA and both AF2 and AF3 fail to predict the structure confidently. Designer proteins are an exception. For antigen/antibody interactions MSA obviously can't help but AF3 predicts them with a 60% success rate, which means it has learned something. It works for local interactions like antigen-antibody, but can't help for large orphan proteins yet.
Sergey's ColabFold demonstrated that some MSAs work better than others (jackhmmer MSAs were better so Sergei needed to improve performance of MMSEQ2 to match the results). Explore the tool - need to try and see what's happening! In general, better MSA = better confidence

Q: JSON file in the server- can this be used for local installations?
A: Yes, local AlphaFold3 can automatically import JSON's generated by the Server, but Server can't import JSONs generated by GitHub AF3

Q: Recycling - what happens without it - what happens to the confidence metrics?
A: With AF2 no recycling results in worse structures and metrics, runs faster without.
AF3 default 10 cycles.

Q:What computational requirements are needed?
A: A100 is the tested card comp 6+ 1080 on user side and greater

Q: I'm still trying to get my head around the mini-rollouts. If there is time, I would love to understand this part of the model better!
A: Sorry, please refer to the paper for more details

Q: do you use ground truth (or reasonably accurate) 3D data of small molecules other than PDB? E.g. OpenCrystallography, or something synthetically generated
A: PDB, recall that conformers are generated from the smiles though, which helps generalise

Q: If two proteins do not interact with each other, would the alphafold3 model show that there is no interaction interface between them?
A: It will try to put them together anyway, just confidence scores will be bad (ipTM < 0.6, pTM<0.5, no contacts on PAE plot)

Q: Is AlphaFold used as a "simulator" for inverse design problems? E.g. design a protein or design a small molecule that binds with the given target as shown by AlphaFold.
A: Theoretically yes, but practically it will mean running an astronomical number of folding attempts. There are tools specifically suited for this purpose and AlphaFold is often used to validate the results from those tools (e.g. design a protein with some other tool and then use AlphaFold to check whether it folds as desired)

Q: What is a seed?
A: A random number (integer) used for initialisation. Think about the function as a landscape and a seed is like a point on that landscape when you start your travel to the local minimum (which may be a global minimum as well). On some landscapes you will arrive at the global minimum from any point, on other rough landscapes you may get stuck in local minimums and you need to start your travel from a good point to get to the global one - that's why for the difficult cases trying many seeds may help.

Q: I previously had the impression that AF3 is more "finicky" than AF2 in terms of requiring ligands, ions etc., but after what Oleg said it now sounds like the only detrimental effect of not providing them is getting underestimates of confidence scores, while the actual structure tends to remain the same? Is that more or less true (provided the ligands/ions/PTMs don't alter the structure)?
A: So far it looks like the coordinates are roughly the same, just confidence is lower when the full context is not present. The problem is that if you do the screening, you may not detect the interaction due to low confidence scores.

Q: Also, if you can comment on this: are you planning to build an "optimiser" similar to AlphaFill for adding ions/ligands etc.?
A: Not sure what does this mean; AF3 can model proteins with any ions or ligands - please use AF Server or local AF3 installation from GitHub to model complexes

Q: Is AF3 likely to be better at side chains? (I know it's difficult to say without independent evaluation, but can you speculate based on the architecture differences from AF2?)
A: Probably depends on the case

Q: Does alphafoldserver.com use a limited set of sequence databases to improve speed? Does it e.g. exclude BFD? I've noticed that the set of sequence databases varies between the AlphaFold DeepMind papers (AF2, AF2 human proteome and AF3).
A: It uses 'small BFD' - exactly as described in the AF3 paper (Server uses the same set of the databases as indicated in the AF3 paper and that GitHub version is using). Server is quick due to massive parallelisation of the MSA building and inference.

Q: When given a sequence that is in the PDB (i.e. in the training data), how likely are AF2/AF3 to "recall" it correctly? Can we assume that PDB structures are predicted correctly and that there is no benefit to using the PDB structures directly over AlphaFold's?
A: We recommend using PDB structures instead of AF predictions when PDB structure is available; one can complement PDB structures with predictions though for the analysis (e.g. for full-length protein when only a domain is available in the PDB, protein-protein complexes, etc.)

Q: What were the "interesting results" that people were seeing with high numbers of recycling (10,000)? Were the structures better? Did they converge on some sort of minimum?
A: I think there is a mess-up between the number of recycled and the number of seeds. 10k seeds were used with AF2 to get improved antigen-antibody predictions in some published works

Q: Are there specific parameters that indicate if an interaction with a ligand is more likely to be real?
A: confidence scores. Check details of the model, run validation software, check clashes, bond angles, distances, if everything is fine. Look at the predicted contacts between ligands and proteins.

Q: I thought AF2 reverts to CPU if no GPU is present? This could be an option if no GPUs are available to a user, couldn't it? I think it takes several hours for a small protein? I suppose this question is quite relevant for some users. The better option is most likely alphafoldserver.com, though?
A: AF is coded for GPU so it doesn't go to CPU. With modifications, it can be run on CPU. AFserver is cloud based. Can use a virtual machine on cloud, Amazon AWS, Google Cloud to run it. Can be set up on a cloud.

Q: Do you plan to use large scale molecular dynamics simulations to generate data for the next versions of AlphaFold?!
A: We can't speak about the future developments until the official announcement.

Q: Would it be beneficial to routinely run relaxation after AF3, even if it is a slow/expensive step?
A: You can check predictions for geometry violations, if there are violations, minimisation can try to fix them.

Q: What have been some of your favourite moments in working on AlphaFold? Or some of your less favourite ones? What is it like to be working on it?
A: When I was able to confidently model the protein that I have failed to crystallise during my PhD. My least favourite moment was when we could not model one really huge structure for our academic partners due to memory limitations.

Q: A few days ago, we reached out to request the parameter models, but we have not yet received a response. We were wondering if this delay is typical and would appreciate any clarification on the process or criteria for their release.
A: It takes a couple of days. Please write to the alphafold@deepmind.com

Q: I am very new to alphafold, if the sequence have ‘N” region how it will handle that areas? Is there maximum N acceptance count? How would be the results accuracy vary in Alphafold2 vs 3 sequences with N regions?
A:
Unofficial answer: At least AlphaFold 2 only accepts the standard 20 amino acids (it doesn’t understand any other characters) and will crash otherwise (I’m not sure at which stage). You could try to remove N regions, or you could try replacing them with flexible linker sequences (e.g. GGGGS repeated 3 times)?
Slightly more official answer: you can also try alanines (A) instead of (N) - glycine is a very special amino acid and you will likely get disorder prediction; alanines are more compatible with forming a structure.

Q: Do you consider AlphaMissense to be a fairly definitive solution for point mutation effects, as AlphaFold is for folding, or is there still potential for improvement? Basically: how good is AlphaMissense?
A: Benchmarks and comparisons are provided in the paper 🙂

Q: How to interpret the PAE plots of two proteins in a complex?
A: The similar way as you do for the single chain - if proteins are interacting, you expect to see “green” (or blue in ColabFold colouring) contacts with low PAE between interacting proteins, or at least between the interacting domains of those proteins.

Q: When running AlphaFold 2 with PDB100 for template searching, it should in principle always retrieve a protein’s actual PDB structure(s), which should then allow it to produce structures that are as good as what is present in the PDB, therefore obviating the need for using PDB structures in addition AlphaFold ones, shouldn’t it? I’m trying to say: if you use PDB100, do you still need to look at PDB structures separately from AlphaFold? Or does AlphaFold tend to ignore the templates whenever the MSA is quite good?
A: AlphaFold ignores templates if MSA is deep, so if you want to do template-based modelling, you may wish to provide a custom MSA with reduced depth or use ColabFold option to limit MSA depth. Still, the PDB deposition may contain a lot of additional information, like binding of the ligands or crystallisation additives, additional conformational states etc. I recommend looking at both PDB entries and the predicted structure, to get maximal information about the protein of interest.