I read the article in its entirety.

This was a really cool study, with some revolutionary discoveries and super practical applications to the end result. The researchers involved in this study aimed to solve one specific component of the protein folding problem: structure prediction. This involves predicting the structure of a protein given its sequence of amino acids, and therefore give insight into its function. The methods prior to AlphaFold were unable to accurately predict the structure of a protein with atomic accuracy, especially when there were no homologous structures available for reference. Experimental efforts are costly and require lots of time and resources in order to determine protein structure. However, experimental efforts can be thanked for providing the accurate data for the AlphaFold model to be trained on. At the time of the study, around 100,000 unique proteins had been experimentally determined, which is a just a drop in the bucket to the amount of possible proteins in existence. This computational approach is needed to address this longstanding problem to enable large-scale structural bioinformatics. The techniques the authors used involved creating a neural network incorporating cutting-edge computational techniques and a variety of inputs. The deep learning-based model incorporates both physical and biological data, multi-sequence alignments, and evolutionary constraints. Previous approaches often focused on either the physical interactions or the evolutionary history. The team competed in the fourteenth Critical Assessment of protein Structure Prediction, or CASP14, proving that they could well outperform other methods, and come to shockingly close results to the experimentally tested protein structures. One thing that sets AlphaFold apart is its ability to model both protein backbones and side chains with high fidelity. In order to ensure make use of all the existing data, and ensure that there was no data leakage, they only utilized data from the protein structure database up to a certain date for training, then had the holdout set for testing as all protein structures in the database after that date. Ultimately, they have found a way to predict protein structure to near experimental accuracy.

One interesting and well done aspect of the study was the well-designed and thoroughly thought through neural network architecture. They named this architecture Evoformer, and it was able to process both evolutionary and spatial information through many layers. This design, I’m sure, enabled increased accuracy over other methods because it takes into account multiple representations in order to prepare for structure generation. I also thought that it was super nice that they provided confidence scores so that we can determine the structure’s reliability for future use cases. I believe it was called per-residue confidence scores, or pLDDT. Another aspect of their model that was impressive to me and I believe added to their ability to predict structure with such high precision was their use of a recycling mechanism in the process. They were able to iteratively refine and update their predictions through the process of running information through the model to ensure that predictions were improved upon each iteration. I also think they handled blindness very well.

It was honestly very difficult to find aspects of the paper that could be improved, as it was pretty dense and seemed quite thorough, but I found a few things that they may do in the future to improve the model, as well as future directions. One thing I noticed is they mentioned that the model struggles with proteins that rely heavily on interactions with other proteins, forming multiprotein complexes. This makes sense honestly, as it may require insane amounts of computation to include interactions between several proteins, but it’s a direction that they could look to. They also mentioned that AlphaFold’s accuracy decreases as the amount of sequence data available decreases. The model seems heavily reliant on multi-sequence alignments for inputs. Not sure if this is really a flaw in the study, but rather a direction to go to improve the model overall. Another approach that I might consider for improvement in the future would be including other environmental factors that impact protein folding such as temperature and pH. It is known that protein sometimes folds differently under different conditions, so taking that into account could be beneficial for future directions.

This was my own work.