PROTEIN FOLDING A MYSTERY BEING UNRAVELLED BY AI

At the moment, we think 2020 will be remembered most of all for the COVID pandemic, which undoubtedly has been an extraordinary historical event. But it may be that an announcement a few weeks ago will be the start of something that changes human lives more than anything that has ever happened before. In the process, it could also be of huge significance for Parkinson’s and other neurological conditions.

Admittedly, this is a huge claim and only time will tell. The event in question? The announcement that a software program developed by DeepMind, the Google-owned artificial intelligence (AI) company, has made a major advance in solving the problem of ‘protein folding’.

Protein folding has long been one of the greatest challenges facing biology. It describes the process in which a protein develops its final, complex three-dimensional (3D) structure, which is determined by a string of building blocks. Proteins are present in all living organisms and consist of long chains of amino acids, linked together like beads on a string. They are absolutely fundamental to the structure of cells and communication between them, as well as regulating all the chemistry in the body.

For a protein to do its job, it must ‘fold’, which describes a process of twisting and bending that transforms it into a complex 3D structure. This structure is fundamental to the protein’s performance of its specific function. If the folding goes wrong – if ‘misfolding occurs’ – the protein won’t form the correct shape, and so it won’t be able to perform its job. And it may cause serious problems, namely diseases like Parkinson's, Alzheimer's and cystic fibrosis.

In Parkinson's, the last decade or two has seen a growing recognition that the formation of the protein alpha-synuclein (α-synuclein) is probably a major factor in the damage done to dopamine-producing neurons, which results in Parkinson's. Misfolding of α-synuclein can cause it to clump together, eventually creating oligomers (molecules comprised of a few similar or identical repeating units) and then thread-like structures called fibrils. These can be damaging and can spread from one neuron to another, where they appear to start the decline and eventual death of the next neuron. We know that people with some inherited forms of Parkinson’s have a mutation in the gene for α-synuclein, which might promote the misfolding of the protein, making it more likely to form oligomers and fibrils.

DeepMind’s program, called AlphaFold2, is being seen as a crucial breakthrough in solving the profound challenge of protein folding. It is based on a computational technique called DeepLearning, which uses the often hidden information contained in vast datasets to solve problems. It has been used widely in fields such as games, speech recognition, autonomous cars, and several areas of science and medicine.

The ability to predict how a protein will fold could revolutionise drug design, and explain the causes of new and old diseases. Today, there are many computer programs that can calculate the structural changes that will occur when small alterations are made to known molecules in a protein. But it has proved impossible to predict how a protein will fold ‘from scratch’ – that is, from the basic string of amino acids. Before DeepLearning, the protein-folding problem seemed impossibly hard, and looked like it would remain unsolved for decades to come.

The folding process happens almost unbelievably quickly, within milliseconds, even though there are a virtually unimaginable number of possible configurations – about 10 to the power of 300 (10^{300}, or 1 followed by 300 zeroes). Compare this with the number of atoms in the universe, estimated at around 10 to the 80! The extraordinary number of 10^{300} helps to illustrate the kind of problem posed by protein folding, even when we know the complete sequence of amino acids that go into making one. Until now, a protein’s structure had to be determined experimentally, which is time-consuming and expensive.

But even though DeepMind’s AlphaFold is a dramatic success story, it is not entirely unexpected. That is because of something that happened in 2017, on a chess board. Then, the best chess performer on the planet was a program called Stockfish-8. It could evaluate 70 million chess positions per second and contained centuries of accumulated human chess strategies, as well as decades of computer experience. It played efficiently and almost brutally, beating all its human challengers. (1)

On December 7, 2017, DeepMind pitted its newly developed chess program AlphaZero, based on DeepLearning, against Stockfish-8. The result was stunning: AlphaZero thrashed it! The two chess engines played 100 games. AlphaZero won 28. How many did Stockfish win? None. Not a single one! There were 72 draws.

This was extraordinary enough, but there is more. AlphaZero carried out 'only' 80,000 calculations per second, about one thousandth of Stockfish-8's 70m. Even more remarkable, AlphaZero **took just four hours to learn chess from scratch by playing against itself a few million times**. This was enough for it to optimise its neural networks as it learned from its experience.

AlphaZero didn't learn anything from humans or chess games played by humans. **It taught itself** and, in the process, derived strategies never seen before. In a commentary in Science magazine, Kasparov himself wrote that by learning from playing itself, AlphaZero developed strategies that "reflect the truth" of chess rather than reflecting "the priorities and prejudices" of the programmers. "It's the embodiment of the cliché 'work smarter, not harder," he said.

For good measure, another piece of Deep-Learning software created by DeepMind, called AlphaGo, has also demolished the world's best Go players. Go is seen as in some ways even more complex than chess. It has both a larger board with more scope for play, games are longer, and on average there are far more alternatives to consider per move. The number of legal board positions in Go has been calculated to be approximately 2.1×10^{170}.

Another victory for AlphaFold took place in a competition that is effectively the Olympics for molecular modelling, called CASP (Critical Assessment of Structure Prediction). This takes place every two years, when the world's top computational chemists test the abilities of their programs to predict the folding of proteins. Teams are given the linear sequence of amino acids for about 100 proteins for which the 3D shape is known but hasn't yet been published. They then have to compute how these sequences would fold.

In 2018 AlphaFold, a complete beginner in the event, beat all the traditional programs – just. Two years later, its successor program, Alphafold2, won the 2020 competition by a healthy margin. It easily beat its rivals, and its predictions were comparable to the existing experimental results determined through established techniques. A US Professor of Chemistry, Marc Zimmer, has said he expects AlphaFold2 and its progeny will become the methods of choice to determine protein structures.

One of the reasons for AlphaFold2's success is that it could use the Protein Database, which has over 170,000 experimentally determined 3D structures, to train itself to calculate the correctly folded structures of proteins. The potential impact of AlphaFold can be appreciated if one compares the number of all published protein structures – around 170,000 – with the 180 million DNA and protein sequences deposited in the Universal Protein Database. AlphaFold will be a powerful assistant in sorting through these millions of DNA sequences in the search for new proteins with unique structures and functions. Who knows what it may discover?

In fact, what AlphaFold will achieve in the future is not the only thing we don't know. A truly mysterious and intriguing feature of these DeepLearning programs is that **we don't know exactly what the AlphaFold2 algorithm is doing!** We just know that it works. It uses certain correlations, but no one has told it to do this, or indeed to do anything specific. It trains itself. So besides helping to predict the structures of important proteins, if we can get to understand more about AlphaFold's 'thinking', it could give us new insights into the fundamental mechanism underlying the protein folding process.

One of the most common fears expressed about AI is that it will lead to large-scale unemployment. AlphaFold still needs further development, but once it has matured, computational chemists will be involved in improving the program, and trying to understand the underlying correlations used. Then we will be in a position to apply the software to important problems such as the protein misfolding associated with conditions like Parkinson's, Alzheimer's, cystic fibrosis and Huntington's disease.

But there is another, more profound fear connected with AI. The first paragraph of this article said these recent events could be "the start of something that changes human lives more than anything that has ever happened before." The fear is that we are unleashing something that is in a real sense 'beyond us'. We have taken the first steps in creating ***something that can teach itself***, and can thereby ***create a successor that is more powerful than itself***. And we don't fully know how it has achieved this! Immediately, it becomes obvious why this could be such an extraordinary event: we may reach a point, possibly quite soon, where a piece of software designs a better, more powerful version of itself – indeed, it seems as though that is already happening with the DeepLearning programs. There is a term for this, coined by the writer Vernor Vinge, which is the title of a book by one of the world's leading computer engineers and thinkers, Ray Kurzweil: the Singularity(2). If Program A designs a more powerful version of itself, Program B, it will not be long before Program B repeats the process, probably more quickly. And so on – we have a runaway effect, the Singularity.

We have an example of such a runaway event that should give us pause for thought, where a chain reaction passes beyond a critical point to a literally earth-shattering result: the atomic bomb. It is some 75 years since we saw that chain reaction happen, with its catastrophic effects. Who's to say we aren't on the brink of something similar – except this time it will be our intellectual abilities that are dwarfed, with consequences difficult even to imagine? You have been warned!

1. (Bear in mind that it was only in 1997 that the world's best player, Gary Kasparov, lost to IBM's Deep Blue computer, the first time a computer had ever beaten the best human. Until then, many had said such a thing would never happen.)

2. The word singularity came into regular use decades ago in maths and astronomy. It can describe a region where the density reaches infinity – for example, when a sufficiently massive entity collapses on itself. When this happens – or perhaps if – the mathematical rules of physics cease to apply. In 1993,

Vinge applied it to technology, saying that in 30 years, we would have the means to create superhuman intelligence and shortly after, the human era would end.