Rationality in Research

1. The Reductionist Trap
2. Fudging Work and Rationalization
3. Generator Systems: Coincident Constraints
4. Amyloid Plaques: Chemical Streetlight, Medical Goodhart
5. Addendum to "Amyloid Plaques: Medical Goodhart, Chemical Streetlight"
6. A Taxonomy of Research
7. How to Find a Problem
8. A Confused Chemist's Review of AlphaFold 2

The Reductionist Trap

This used to be called "Non-Reduction and Nonapples" but thanks to a comment by Slider I've realized this put the emphasis in the wrong place.

Selling nonapples is a great piece about how critiquing a certain technique, or process, or system can often sneak into proposing a non-existent alternative. This new system, as it is overly general, fails to address the problems that the first exists to solve.

I think reductionism vs non-reductionism is sometimes a similar debate, but sometimes not. First let's distinguish between philosophical reductionism, and practical reductionism.

- *Philosophical* reductionism is (sort of) the belief that complex systems act the way they do because of the simpler actions of simpler systems that make them up. I think that this is pretty solidly true in our universe, and I won't discuss it further.
- *Practical* reductionism is the belief that the *best and most effective* way to study complex systems is to study their components individually, then build up to a larger model of the whole system.

(As an aside I actually think a lot of the reductionism vs non-reductionism debate suffers from a motte-and-bailey issue around this)

Microbiological Ecosystems

An example of practical reductionism is the study of microbiological ecosystems: there are lots and lots of different microbes in any given area (such as the sea, the soil, your intestines) which all require different nutrients and produce different sorts of waste. Some photosynthesize, some turn atmospheric nitrogen into a form which is bioavailable, and some break down carbon-containing material. Clearly if we knew everything about every type of microbe present we could understand the ecosystem. Unfortunately, unlike in macroscopic ecosystems, it's difficult to study a single microbe's behaviour while it is inside the ecosystem. They're just too small! So the standard technique is to get a pure culture, containing just one sort of microbe, and study the behaviour of that culture.

Unfortunately, this often doesn't work very well. Microbes in culture can behave differently from how they normally behave in general systems. Lots of microbes are either impossible to form a pure culture of, or require some incredibly specific nutrient only produced by another microbe, that we don't know the identity of. This means the standard "practical reductionist" technique is doomed to failure.

Here people often like to say things like "Reductionism has failed! We need to study whole systems instead of just trying to understand pieces of them!"

I think this kind of fails to capture the benefits of practical reductionism. Practical reductionism solves the problem of "How do we go about studying complex systems?". The answer here is "Break them down into simpler components and study those individually!". Without reductionism, we are left without an obvious solution.

I have avoided using the word "holism" yet. To continue with microbiology, there are a whole set of "holistic" techniques for studying microbiological ecosystems. One example is to mix up all the bacteria and sequence all their genomes at once. This tells you what all of them are doing, as if they were one big organism. In particular it tells you (in theory) all the metabolic pathways that are going on, but doesn't tell you which microbes are doing which metabolism.

This is a way to gain understanding of the system without using the standard "reductionist" techniques, so non-reductionism isn't a nonapple in this case!

Protein-Metal Binding

Another case: the group I am working in was studying how a particular protein binds to a particular metal complex. As part of this, one of the members of the group studied the binding of this metal complex to every single amino acid individually. Turns out that two of the three parts of the protein that bind the metal, will actually not bind the metal when they're not part of the protein! This was mildly interesting but not particularly surprising, as the group already knew that two of the binding parts are bad at binding. Worse than this, very little was learnt from studying the one that did bind!

I'm sure part of the reason for this was that this person was only an undergraduate, and needed something to do which wasn't too difficult, and the overall research wasn't impacted negatively, but I think it speaks to a certain bias.

The bias goes like:

> The protein-metal interaction is made up of amino acid-metal interactions, therefore studying amino acid-metal interactions will help us understand the protein-metal interaction.

Comparing the Two Cases

In the microbiology case: we can study systems *with reductionist, or holistic methods*. Unfortunately, the two methods don't give us the same sort of information, and we can't bridge the gap between the two, even though *we care about both*.

In the protein case: we can study systems *more easily with a reductionist method* than a holistic method. Unfortunately, the two methods don't give us the same sort of information, and we can't bridge the gap between the two, *and we care about the holistic system much more than the reduced systems*.

I think this brings us to a conclusion about bias in studying complex systems:

> Studying simple things is easier than studying complex ones. But just because complex systems are built up of simpler parts, that doesn't mean that studying the simpler parts is the best way to understand the complex system.

And it is particularly tempting to study smaller systems if possible! A simple system is very legible, we can understand it fully. Looking at it *feels* like progress on the bigger problem. A complex system is illegible, we might not make progress for months or days, and our models might not be good at reflecting underlying reality. People who like to understand things can easily be repulsed by this.

This mostly refers to complex systems which already exist in the world. Particularly biological ones. Toy models are very useful for systems we can't study directly (like decision theories and AI). I just think that lots of work is being wasted on studying over-reduced parts of bigger systems.

I suspect the antidote is the question:

> What's the best thing that could come out of this? What results of this investigation would bring me to a greater understanding, and how likely do I think these are?

If you ask this about all your investigations, it should help to clarify *why* a certain object is being studied. If the answer is just that it seems like something that you ought study, you might be reducing more than is helpful.

Fudging Work and Rationalization

If you want to do extraordinary things, you will need to do your best.

In practical sciences, this can come as a form of precision, typical analytical chemistry is done to some degree by hand (!) and involves literally weighing out milligram-precise quantities of substances, and putting them into sub-millilitre-precise quantities of solvents. It is easy to go wrong, or to be lazy. The most accurate flasks look like this:

Lucasbosch, CC BY-SA 3.0, via Wikimedia Commons

Note the single line. If you overshoot the volume, you have to start again. This is a real pain, and can sometimes take tens of minutes. It's also a little embarrassing. But what is overshooting? There's only a certain precision a human can go to, and after all, if you're only half a percent over it can't make much of a difference, can it? These mental patterns are easy to fall into, and they look like *bargaining* with the universe.

The accuracy of your results doesn't depend on how *hard* it was to measure out the right volume. They depend on how accurately you measured that volume. The universe cannot be bargained with.

It's a sort of rationalization, at least that's how it feels to me, when I fall into it. The feeling of rationalization is one of the most important ones to be able to notice as rationalists. I think there are a few reasons for this beyond the standard ones:

I expect to be punished for bad work on some level. When this happens, we bargain with the person doing the punishing, which is reasonable. If someone says "you measured that volume wrong", things like "well it's pretty much close" and "but measuring volumes is really hard!" may well be reasons for them to forgive you. But they are not reasons why the volume is right. When you're doing work like this, it's absurd to feel guilty at an occasional mistake, but unfortunately the guilt-making part of the brain doesn't know this.

I want to protect my opinion of myself as skilled. This has ruined many scientists' careers, when their results are disproven, they turn to fraud.

I want to keep doing things without having to redo the work. Schools and universities teach us that failed work is not something to redo. For the most important work (exams) redoing them is forbidden! In the real world, the more important the work, the more important it is to redo it if it's wrong. Also, redoing the work is lots of effort, and I want to move on to the next bit of the experiment.

As you may have guessed, this applies to many things beyond measuring out a volume. A general term might be "fudging". This is when we thoroughly convince ourselves that we shouldn't have to redo some work, and then slip sideways into the world where the work is actually good. Again, very similar to rationalization, where we convince ourselves that something really should be true, then slip sideways into thinking that it is true.

This is an great sort of rationalization to train yourself on, for a few reasons:

- The costs of doing some slightly subpar work are generally not too large.
- I think it's one of the most common pieces of rationalization around
- There's no actual, permanent status or material loss associated with correcting it, so it's not too painful
- The truth is there in front of you. The truth will also probably be revealed later

If you can learn to recognize this, you'll become better at noticing the tension associated with rationalization in general, which is one of the most important rationalist techniques out there. Fudging work isn't half as dangerous as fudging beliefs.

Generator Systems: Coincident Constraints

The Prototype Plane Perfecter

You are asked with improving the design of a plane, which keeps falling out of the sky when it reaches 80 mph. This makes sense, as it is a prototype, put together with no knowledge of high-speed flight. Unfortunately they always seem to disappear over the horizon just before failure, and crash in such a huge fireball that no evidence as to the cause can be collected.

You have three hypotheses as to why: the engine is spinning itself apart; the wings are falling off; the instruments are failing. Each of these requires some cost to improve, so your bosses are insistent that you solve the issue, and no more. You make three new planes.

Each of them sets off, and each of them crashes at the exact same distance from the runway (to within random error of course). None of the improvements have increased the top speed.

Naturally, you assume that none of these are constraining it. You look for other reasons the planes might be crashing.

The Mortal Trees

A scattered group of trees wish to live forever. They do not age, so the only threat to them is being toppled by the wind. Each stormy night they all fear terribly for their lives. Their scientist-trees get to work. They find that 80 mph wind is generally fatal.

One school of medical researcher-trees studies trunks. They believe trunk-snapping is the ultimate cause of toppling. With a strong regiment of lignin-enhancers, they manage to fortify themselves physically.

Another school studies roots. They believe that uprooting is the ultimate cause of toppling. With soil rigidification, they ensure their balance is perfect.

The next storm, each waits with baited breath. Both groups find they have lost some of their own, and barely any fewer than the trees in the control group. Those still standing make peace, and accept the inevitability of toppling.

Lessons

In the first story, it is not so clear what's happening. In the second, it is. The difference is that the generator system for prototype plane designs is very different from the generator system for tree designs.

If a plane is put together with no knowledge of high speed flight, then the chances of all of its systems failing at exactly the same speed is very unlikely. In fact this is the case for most systems. In plant growth, one nutrient is usually limiting, the same is

often true for manufacturing. In this case, locating the limiting factor is paramount to improving the system.

The generator system for tree designs is evolution. This is not random. In real life, trees generally do uproot and snap at similar wind speeds. This is because both root stability and trunk strength are metabolically costly, so investing in one being stronger than the other is a poor strategy. A population of trees which uproot at 60 mph but would hypothetically snap at 100 mph will experience two things: if stronger roots are not too costly, genes for these will become more frequent; and genes for weaker trunks will become less frequent. There will be an equilibrium point where winds of a certain strength are so rare that withstanding them is not metabolically costly enough.

Generally, multi-causal models are subject to a significant complexity penalty: if you think 50 different things contributed to falling crime rates in the 90s, you must also explain why all 50 different things happened at the same time. This is not true when the generator system is evolution.

This was written as a direct parallel to ageing. The involution of the thymus gland is basically unrelated to other types of pathological ageing, as far as we can tell. The constraint systems just seem to sort of line up time-wise. The same could be true for other proposed methods of ageing.

Whatever object you're studying, it will have been created by some generator system: evolution, bad engineering, good engineering, free markets, political design. Understanding the generator system gives you good priors. Depending on how cheap experimentation is (does it cost planes worth of money, or does delay cost many human lives).

Amyloid Plaques: Chemical Streetlight, Medical Goodhart

Alzheimer's Disease (AD) is truly, unduly cruel, and truly, unduly common. A huge amount of effort goes into curing it, which I think is a true credit to our civilization. This is in the form of both money, and the efforts of many of the brightest researchers.

But it hasn't worked.

Since AD is characterised by amyloid plaques, the "amyloid hypothesis" that these were the causative agent has been popular for a while. Mutations to genes which encode the amyloid beta protein can cause AD. Putting lots of amyloid into the brain causes brain damage in mice. So for many years, drugs were screened by testing them in mutant mice which were predisposed to AD. If the plaques disappeared, they were considered good candidates.

So why didn't it work?

Lots of things can affect amyloid plaques as it turns out, right up to the latest FDA approved drug, which is just antibodies which target amyloid protein. While this does reduce amyloid, it has no effect on cognitive decline.

Goodhart's law has reared its head: amyloid plaque buildup is a metric for AD progression, but selecting for drugs which reduce it causes the relationship between AD and plaques to fall apart.

Equally, amyloid plaques are very easy to measure in mouse (and human) brains. It can be done by MRI scan, or by dissection. Memory loss and mood changes are harder to measure, and even harder in mice. The methods for measuring amyloid plaques also *feel* better in many ways. There's less variation in potential methods, they can be compared across species, they're qualitative, and they're also more in line with what the average biologist/chemist will be used to.

Understanding these, we can see how looking for drugs which decrease amyloid plaques in mice *just really feels like* productive research. We can also understand, now, why it wasn't.

Avoiding Wasted Effort

Pointing out biases is fairly useless. Pointing out specific examples is better. But the best way to help others is to point out how it *feels from the inside* to be making these mistakes.

So what does it feel like to be on the inside of these biases? Unfortunately as someone who has not been intimately involved in AD research I can't say exactly. But as someone involved with research in general I can make a guess:

- Research will feel mostly productive. It may feel like you are becoming how you imagine a researcher to be. Papers will be published. This is because you're in

the streetlight.
- What you won't feel is a sense of building understanding. Learning to notice a lack of understanding is one of the most important skills, and it is sadly not an easy thing to explain.
- Think about the possible results of your experiments. Do you expect something you've not seen before? Or do you expect a result with a clear path to success? Creative work usually passes the first. Well-established and effective protocols pass the second. Mouse AD models do not pass either (anymore).
- A positive experimental result will be much easier than a "true" success. This has the benefit (for researchers) of allowing you to seem successful without actually doing good. The ratio of AD papers to AD cures is 1:0 ("Alzheimer's Disease Treatment" returns 714,000 results in Google Scholar)

Beyond this I do not know. Perhaps it is a nameless virtue. But it might be useful to try to identify more cases. I hereby precommit to posting a follow-up with at least five examples of this within the next seven days.

Addendum to "Amyloid Plaques: Medical Goodhart, Chemical Streetlight"

Last week I committed to posting five examples where I think a group of scientists has gone astray. Goodhart's law says you get what you measure. The streetlight effect makes you look where it's easy to look. When combined, all you get from an entire field of research is a bunch of things which are easy to measure.

I have tried to make these as current as possible, with a special focus on finding avenues of research that the general scientific establishment has yet to abandon. Not all of them are like that though, due to time and effort limitations on myself (also an unexpected release of D&D.Sci). They are also mostly worse examples than my original one, which is to be expected.

I would also be remiss not to mention the comment on my last post by JenniferRM which is an excellent example of good reasoning about this sort of issue.

Ageing and Ageing Markers

Lots of research into slowing ageing is actually taking place! This is good news from the perspective of not dying. Unfortunately, a much smaller minority of it is optimal. Lots of research at the moment is focused on interventions which can extend the lifespan of mice, flies, worms, and yeast by a few tens of percentage points. This has given us insights like the role of mTOR, fasting, and the epigenetic clock. However, any intervention which extends life by only a few tens of percent as a maximum is not targeting the "ultimate" causes of ageing. Lots of research seems to look for small increases in longevity rather than looking for building understanding of ageing itself.

Electrocatalysis of Graphene Compounds

This is a closed-book one. Certain chemical reactions involve the transfer of electrons. Graphene can catalyse these reactions, but lots of graphene derivatives can catalyse it even faster. For electronic reasons, both putting some extra electrons into graphene, and taking some out, make it a better catalyst. Many many papers were published based on this theory, until finally someone made graphene doped with bird guano. This was an excellent catalyst, and managed to put the endless searching for better graphene dopants.

Racial Bias Testing

Racism, like most of the large issues facing society, is very complex. One of the big ideas of (relatively) recent times is that individuals who are not explicitly racist can still be biased. The Implicit Association Test is the one where you do classification simultaneously into good/bad and (for example) French/English. If you're faster at grouping croissants with murder and fish 'n' chips with charity than the other way round, then you might be a Francophobe. It is brilliant, elegant, simple, and also very poorly validated.

Bill Clinton's Nanotech

Bill Clinton spent billions on nanotech in 2000. Sadly (and understandably) his administration were not experts in nanotech. This made it almost impossible for them to judge which directions the research at the time needed to go in. Molecular-scale manufacturing and programmable molecules are still a long way away. Most things which are accepted as "nanotech" are co-opted biological molecules doing a slightly different thing to what they do in nature. Sometimes this really is revolutionary (nanopore gene sequencing comes to mind) but a lot of the time it isn't. Optimizing for things which a bureaucratic institute will think of as nanotech destroyed the possibility for actual nanotech.

Decision Theory

Here's the most controversial (on LW at least) one, and the one I'm the least confident in being an actual example of this. I worry that a lot of AI researchers spend a lot of time thinking about decision theory, and this whole process is being driven by finding decision theories which solve more and more esoteric problems. Understanding the nature of decision making is important but I don't feel like our lack of understanding sits in the gaps between UDT and TDT.

A Taxonomy of Research

When doing research you have methods of doing things. These differ from field to field according to the sort of thing you're investigating.

You also have overarching principles which direct your attention and resources along certain lines. The latter are much more interesting to talk about.

One Solution vs Many Solutions

For some avenues of research, there is only one solution. Think discovering Newton's Laws, or finding the protein which is responsible for breakdown of proteins in the stomach. Research is therefore a process of ruling out answers until you get to the only one which is correct. This often comes down to finding the best model. Eventually it comes down to finding a model so good it's the best model possible, which is "the truth".

For others, there are many solutions. Think drug discovery, or developing a new technology. Research here is often a process of search for a decent design, followed by gradual improvement. Other times, it's coming up with a new insight as to how existing things can be used. This is more like problem solving.

Known Solution vs Unknown Solution

For some problems, the general form of the solution is likely to be known. If you're working in battery technology, you probably know the next battery will have an anode, cathode, and an electrolyte which can accept and give up electrons. If you're trying to make a new antibiotic you probably know what sort of molecule it will be.

For other problems, you might have no idea. Until they were discovered, nobody knew what the laws of quantum field theory would look like, nobody expected the world to be made out of weird functions in 4d. Research in gene sequencing has come up with lots and lots of wildly different approaches.

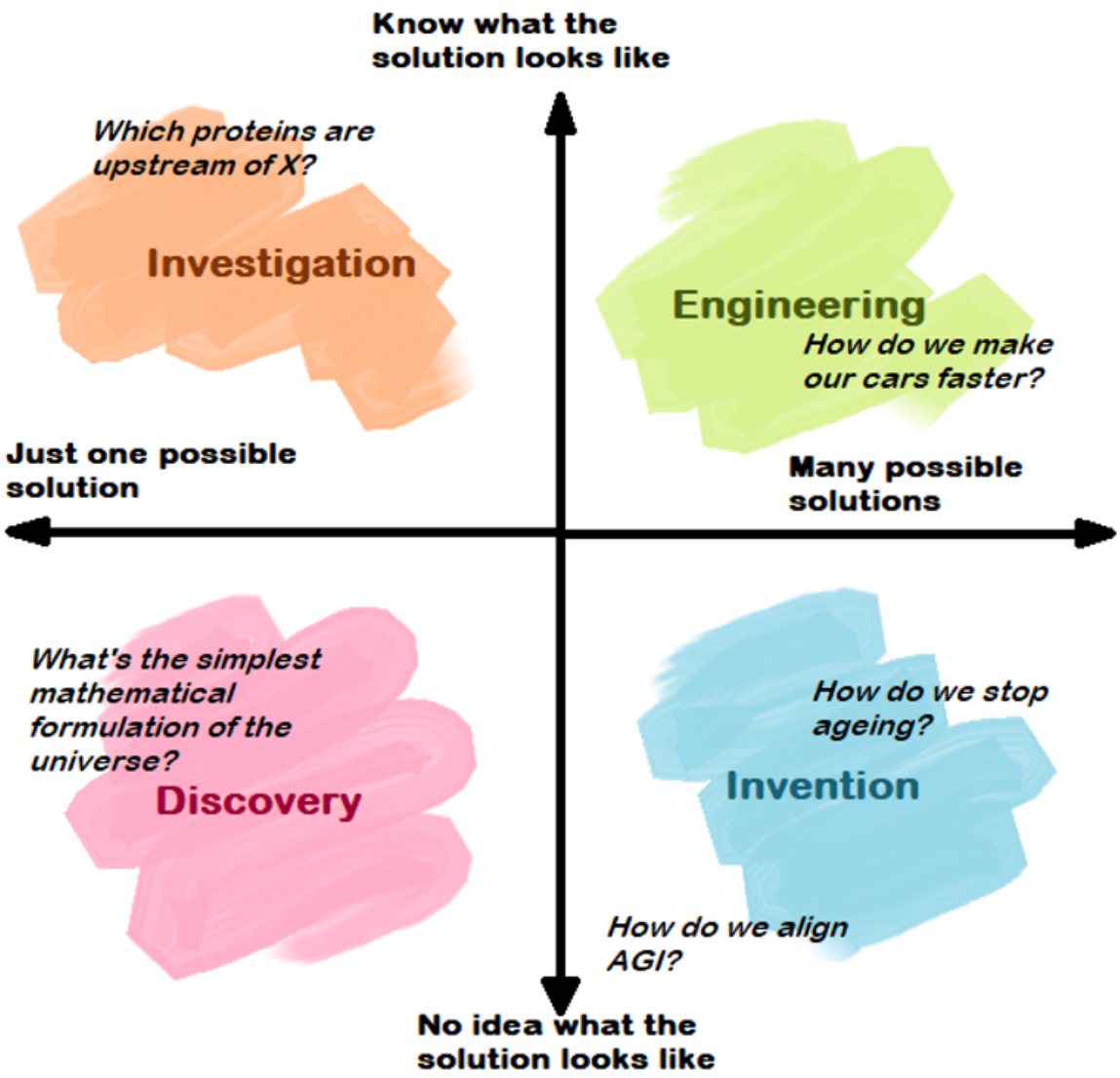
The Grid

Looking for a single answer that you don't know what it looks like is discovery.

Looking for a single answer that you have a pretty good idea of what it will look like is investigation.

Trying to solve a problem with a solution you know the form of is engineering.

Trying to solve a problem with a solution you don't know the form of is invention.



Here's some examples which I've tried to put roughly in the right place

Depending on where you're aiming, you need different tools.

The further down, and also the further right, the more you need to pay attention to unexpectedly unexpected results. These are results which vary orthogonally to the axis of variation you're studying. Being further down and right is also (arguably) where creativity is most important.

Being further left means you need to be more specific. It's about cutting down hypothesis space. This means doing experiments with the aim to gain the very maximum information possible. If you're in the bottom left, this is especially important as your hypothesis space is largest, and the target space is smallest.

In the top left, you can usually stick to established methods. This is often where lots of data can be gathered, and then lots of numerical analysis can be carried out. This is how we get the most detailed models, like understanding not just the structure of a protein, but exactly which order the different parts of it fold in.

The top right has often been banished from scientific research altogether. That's why "engineer" is a whole job. It is, however, often how physical things get designed and made. This is why "engineers" are so well paid.

How to Find a Problem

There are many ingredients for creativity. One is to have a lot of knowledge saved up, so it can combined in novel ways. One is to actually think about things for five minutes. One is to not flinch away from problems when you find them.

But the most important is to have a problem to solve. Or if you're Feynman:

> You have to keep a dozen of your favourite problems constantly present in your mind, although by and large they will lay in a dormant state.

Importance

To have a problem in the back of your mind constantly is somewhat taxing. You need to find a problem which isn't emotionally draining. For this reason the problems need to be important. They don't have to be the most important thing in the world, but they do have to be important enough *to you* to be somewhat self-motivating. If your day-to-day life involves so many interactions with doorbells that you can think about doorbell design all day, then that's enough.

In some ways, I have it easy on this front, which makes it hard for me to give advice. My domain is biological chemistry, so I can do a good job just by pointing myself at the closest ageing-related issue. Ageing is very salient to me, and as I get older I will either succeed or it will become painfully more salient.

Size

The problems need to be of the correct size. I don't hold the problem "cure ageing" in my head all day. That wouldn't work, because anything could in theory be related to "cure ageing" but there's no clear way for my brain to make the right connection. Instead you have to find a correct balance. The problem "replace lost stem cells in human tissue" is about the right level. There are lots of possible answers, but also the line between ideas I come across and that is not too many steps long.

I suspect that the right level is different for different people. This makes my advice somewhat unhelpful. Still, just knowing there is a correct level might be helpful. Perhaps a good idea is to try and keep a few different problems of different levels around, at least until you find the right one.

How to actually find the problem

Hopefully the two issues above have already clued a few readers into my method.

First off, think of something you want, or want to know. Just wanting to "discover things" is not enough. You have to work one level down to achieve that goal. It can even be silly or useless. As long as your emotional brain will let you focus on it.

If you're lucky, that problem will be really big, too big, even. This is good. It means you can just spend a bit more time splitting it into sub-problems. Try and be honest. Imagine that a given sub-problem was solved. Would it really contribute to the bigger issue? If you could solve one sub-problem magically, which would it be?

Once you have picked some then you have to actually do the thing, which is the hard part. As my current research involves a lot of (relatively) boring steps of making and purifying protein, it's easy for me to spend time mulling over the ideas. This might not be the case if you have a different day job or a different brain.

This is all you can really do. You cannot force yourself to have important ideas, only put your brain in the right place to have them. This is a technique which can be practised and improved on. So go and do it.

A Confused Chemist's Review of AlphaFold 2

(This article was originally going to be titled "A Chemist's Review of AlphaFold 2")

Most of the protein chemists I know have a dismissive view of AlphaFold. Common criticisms generally refer to concerns of "pattern matching". I wanted to address these concerns, and have found a couple of concerns of my own.

The main method for assessment of AlphaFold 2 has been the Critical Assessment of Protein Structure (CASP). This is a competition held based on a set of protein structures which have been determined by established experimental methods, but deliberately held back from publishing. Entrant algorithms then attempt to predict the structure based on amino acid sequence alone. AlphaFold 2 did much better than any other entrant in 2020, scoring 244 compared to the second place entrant's 91 by CASP's scoring method.

The first thing that struck me during my investigation is how large AlphaFold is, in terms of disk space. On top of neural network weights, it has a 2.2 TB protein structure database. A model which does *ab initio* calculations i.e. does a simulation of the protein based on physical and chemical principles, will be much smaller. For example Rosetta, a leading *ab initio* software package recommends 1 GB of working memory per processor in use while running, and gives no warnings at all about the file size of the program itself.

DeepMind has an explicit goal of replacing crystallography as a method for determining protein structure. Almost all crystallography is carried out on naturally occurring proteins isolated from organisms under study. This means the proteins are products of evolution, which generally conserves protein structure as a means of conserving function. Predicting the structure of an evolved protein is a subtly different problem to predicting the structure of a sequence of random amino acids. For this purpose AlphaFold 2 is doing an excellent job.

On the other hand, I have a few nagging doubts about how exactly DeepMind are going about solving the protein folding problem. Whether these are a result of my own biases or not is unclear to me. I am certainly sympathetic to the concerns of my peers that something is missing from AlphaFold 2.

Representations

One of the core elements is that the representation(s) of protein structure is fed through the same network(s) multiple times. This is referred to as "recycling" in the paper, and it makes sense. What's interesting is that there are multiple layers which seem to refine the structure in completely different ways.

Some of these updates act on the "pair representation", which is pretty much a bunch of distances between amino acid residues (amino acids in proteins are called residues). Well it's not that, but it's not *not* that. I think it's best thought of as a sort of "affinity" or "interaction" between residues, which is over time refined to be constrained to 3D space.

There is also a separate representation called the "multiple system alignment (MSA) representation" which is not a bunch of distances between residues.

The MSA representation (roughly) starts with finding a bunch of proteins with a similar amino acid sequence to the input sequence. The search space of this is the 2.2 TB of data. Then it comes up with some representation, with each residue of our input protein being assigned some number relating it to a protein which looks like the input protein. To be honest I don't really understand *exactly what* this representation encodes it and I can't find a good explanation. I *think* it somehow encodes two things, although I can't confirm this as I don't know much about the actual data structure involved.

Thing 1 is that the input protein probably has a structure similar to these proteins. This is sort of a reasonable expectation in general, but makes even more sense from an evolutionary perspective. Mutations which disrupt the structure of a protein significantly usually break its function and die out.

Thing 2 is a sort of site correlation. If the proteins all have the same-ish structure, then we can look for correlations between residues. Imagine if we saw that when the 15th residue is positively charged, the 56th one is always negatively charged, and vice versa. This would give us information that they're close to one another.

Evoformer Module

The first bunch of changes to the representations comes from the "evoformer" which seems to be the workhorse of AlphaFold. This is the part that sets it apart from ordinary simulations. A bunch of these models sequentially update the representations.

The first few transformations are the two representations interacting to exchange information. This makes sense as something to do and I'm not particularly sure I can interpret it any more than "neural network magic". The MSA representation isn't modified any further and is passed forwards to the next evoformer run.

The pair representation is updated based on some outer product with the MSA representation then continues.

The next stages are a few constraints relating to 3D Euclidean space being enforced on the pair representation. Again not much commentary here. All this stuff is applied 48 times in sequence, but there are no shared weights between the iterations of the evoformer. Only the overall structure is the same.

Structure Module and AMBER

This section explicitly considers the atoms in 3D space. One of the things it does is use a a "residue gas" model which treats each residue as a free floating molecule. This is an interesting way of doing things. This allows all parts of the protein to be updated at once without dealing with loops in the structure. Then a later module applies a constraint that they have to be joined into a chain.

They also use the AMBER force-field (which is a simulation of the atoms based on chemical principles) to "relax" the protein sequence at some points. This does not improve accuracy by the atom-to-atom distance measures, but it does remove

physically impossible occlusions of atoms. The authors describe these as "distracting" but strongly imply that the AMBER part isn't very important.

Attention

I think this is what gives AlphaFold a lot of its edge, and unfortunately I don't understand it all that well. It's very similar to human attention in that the network first does some computations to decide where to look, then does more computations to make changes in that region. This is much better (I think) than an "mechanical" model which simulates the protein atom by atom, and devotes equal amounts of computation to each step.

Thoughts and Conclusions

The second placing team in CASP14 was the Baker group, who also used an approach based on neural networks and protein databases. Knowing this, it doesn't surprise me much that DeepMind were able to outperform them, given the differences in resources (both human and technical) available to them. Perhaps this is a corollary of the "bitter lesson": perhaps computation-specialized groups will always eventually outperform domain-specialized groups.

I do have two concerns though:

My first concern is that I strongly suspect that the database-heavy approach introduces certain biases, in the no-free-lunch sense of the word. The selection of proteins available is selected for in two ways: first by evolution, and secondly by ease of analysis.

Evolved proteins are not subject to random changes, only changes which allow the organism to survive. Mutations which significantly destabilize or change the structure of the protein are unlikely to be advantageous and outcompete the existing protein. CASP14 seems to be the prime source of validation for the model. This consists entirely of evolved proteins, so does not provide any evidence of performance against non-evolved (i.e. engineered) proteins. This strongly limits the usage of AlphaFold for protein engineering and design.

Secondly, not all proteins can be crystallized easily, or even at all. Also, some proteins only take on defined structure (and can only be crystallized) when bound to another molecule. DeepMind are working on functional predictions of small molecules binding to proteins, but including bound non-protein molecules in their structural predictions is outside the current scope of AlphaFold.

Both of these cases are much rarer than the typical case of protein crystallography, which is the main aim of AlphaFold. For most protein researchers, particularly medical researchers, I suspect that the trade-off of using the database approach is worth it.

My second concern is more nebulous, and relates to their usage of AMBER. This feels like they're outsourcing their low-level physical modelling.

This sort of thing is actually quite common when analysing crystallography results, which often require multiple rounds of refinement to get from crystallographic data to a sensible (read: physically possible, without atoms overlapping each other) structure.

However the "first-draft" output of a crystallographic data is basically just a fit to the output of a 3D Fourier transform on some potentially noisy x-ray scattering data.

This somehow still feels different to Alphafold. If lots of their neural network layers are outputting "impossible" structures with atoms overlapping then it suggests those layers have failed to learn something about the physical properties of the world. I'm not sure whether or not this will turn out to be important in the end. It may be that I just have some irrational unease about models which learn in a very different way to human minds, learning complicated rules before simple ones.

AlphaFold will only grow in accuracy and scope. I suspect it will eventually overcome its current limitations.

Sources and Thanks:

Many thanks to the LessWrong mod team, and their feedback system. This could have not been written without the feedback I received on an early draft of the article.

The actual paper: https://www.nature.com/articles/s41586-021-03819-2

CASP14: https://predictioncenter.org/casp14/

OPIG who have a much more in-depth analysis of the mechanics of AlphaFold 2: https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/