Redefining medicines discovery and accelerating breakthroughs

Dr Gemma Holliday, Lead Cheminformatician at Medicines Discovery Catapult, shares her observations on how data is helping to unlock and accelerate the pursuit of new opportunities on the horizon. Medicines Discovery Catapult (MDC) is a government-funded organisation involved in reshaping the UK's medicines discovery industry.

Part of a network of Catapults established by Innovate UK, MDC is an independent organisation that industrialises and drives the adoption of new tools and technologies for the sector. It pioneers next generation approaches to the discovery and proof of well-targeted medicines, diagnostics, and biomarkers. As a champion of open innovation and collaboration, it provides access to skills, technology, informatics, and research assets that are transforming the development of future medicines.

What key trends look set to revolutionise drug discovery in the coming years?

Drug discovery generates huge quantities of complex biological, chemical, clinical, and safety information which means data science is becoming increasingly critical to making the best decisions on which drugs to optimise or progress. In the not-too-distant future Complex Cell Models (CCMs) could make it possible to predict a drug's effects in clinical trials, while 3D biological models could ultimately replace traditional 2D cell cultures and animal testing in pre-clinical research.

Deep learning models and Artificial Intelligence (AI) techniques are increasingly being applied to drug discovery across a variety of fields. These include advanced image analysis, molecular structure, and function predictions, as well as the automated generation of innovative chemical entities with bespoke properties.

As AI capabilities become embedded into the research process, explainable AI is emerging as a key technique that will be vital for delivering transparent insights into how AI powered models work and why the findings of these increasingly sophisticated algorithms should be trusted. In other words, we need to be able to demonstrate this is what we predicted, this is how we got there, and here is the evidence as to why these findings can be trusted.

This will prove critical if the drug discovery ecosystem is to successfully integrate data, computation, and experimental biology to deliver more effective drugs faster and with a higher success rate.

Does that mean the open sharing of data will become increasingly important for propelling the next decade of AI-assisted discovery?

Absolutely. AI can sift through millions of molecules to search for candidate drugs, opening the doors to a much faster path to new treatments. The more high quality data you have access to, the better. In my field, broad and high quality data sets that have better coverage and aren't sparse or biased means better algorithms, better predictions, better clinical designs.

When you consider the ever increasing volume of data we need to curate, concentrate and use in medicines discovery, it's clear that collaboration and open data sharing is becoming increasingly important to pharmaceutical companies large and small. Typically, drug discovery encompasses more than 100 disciplines – so this clearly needs to be a team sport.

The more people share – including clinical research that didn't deliver against its target objectives – the better. For me, this 'negative' trial data is worth its weight in gold. After all, how can you build without knowing what has been tried before – what didn't work and why. While this data might not have furthered one particular drug discovery path, it's findings might provide the vital 'aha' input that enables a completely different medical discovery.

Ultimately, the industry-wide sharing of negative trial data would help significantly reduce phase II and phase III attrition rates and prevent the industry from potentially wasting vast sums of money on compounds for which the molecular target has essentially been invalidated. These resources could then be focused on other subjects that offer more potential – which in turn would improve drug discovery productivity.

It's not just negative trial data that's important, of course. The earlier discovery phases matter too.

It's an issue the industry has been grappling with in recent times. But the collaborative sharing of open data sets that occurred in response to the COVID-19 pandemic demonstrates how drug discovery could be catalysed to deliver huge benefits for public health.

So – will data husbandry will become increasingly critical for drug discovery?

Yes. In this field, data increasingly drives everything we do and transposing that data into knowledge is the name of the game. I myself am passionate about the challenge of surfacing value-added learnings from negative data and building expertise on top of this data that eliminates human bias.

When you consider just how much data exists in the drug discovery ecosystem today and how much is at risk of being potentially lost, it makes you appreciate the scale of the task that lies ahead. If we stopped creating wet lab data today, it would take over 30 years to annotate it all – and no human could ever hope to review all of that without the assistance of AI-powered algorithms.

In the academic community, data sharing has taken off in recent years. But when it comes to maintaining and mapping all that data for the benefit of all, many databases are still being maintained by a small number of academic groups. The question is, who will continue to maintain these repositories when key personnel retire?

Does the industry need some form of federated database and management infrastructure? I don't know what the long term answer is. However, today's lack of coordination between different data domains means we're at risk of losing valuable datasets that could potentially fuel our future drug discovery capabilities.