Making AI Work for Drug Discovery: A Joined-up Approach

AI/data-driven drug discovery is starting to evolve at an encouraging rate now. But the technology's positive impact will depend on how well the technology, and the insights it elicits, are embedded into R&D, says Biorelate's Dr. Ben Sidders.

Gradually, data-driven, AI-enabled drug discovery is becoming a reality, beginning to fulfil the technology's promise and demonstrating its potential more tangibly. The broader signs are encouraging, too – such as the growing profile of 'AI-first' companies such as Recursion and Insilico Medicine, and the observation that many traditional pharmaceutical companies are now embracing AI across their businesses.

Where previously the perceived value of AI in drug discovery and development had failed to live up to the technology's hype, targeted solutions are now emerging which are making a positive impact on aspects of R&D. In turn, these applications are providing some valuable lessons and feedback about how to successfully embed AI within R&D operations.

In target discovery, knowledge graphs are now proving adept at integrating a vast number of data sources into a query-able structure, forming the basis for informed and relatively unbiased target prioritisation decisions and chemistry, where transformers are accelerating small molecule design and synthesis.

Challenges remain, however, predicting synergistic drug combinations has been the topic of extensive research, with only limited success and almost no translational relevance. Nor are we any nearer to being able to predict the effect of a drug on a given patient without first running a clinical trial.

The overriding realisation is that AI's role in life sciences R&D is directly dependent on how decisively, and how well, they integrate the technology – and the insights it surfaces – within the wider R&D operation. Achieving this, in turn, will require a structured approach to AI-enabled R&D transformation, spanning four parallel priority areas: data, model, culture, and validation. Here's how that breaks down.

Data

AI has found most success where the data set is large, complete and in many cases has been generated specifically to solve the problem at hand. The UNI foundation model for computational pathology, for instance, was trained on >100 million images from 20 tissue types.

In contrast one of the largest datasets available to train models for drug combination synergy prediction has 910 combinations of 118 drugs – many orders of magnitude smaller.

Significantly, much of our biomedical knowledge is locked away in unstructured data sources such as the literature. This problem is further exacerbated when we look at data from clinical trial cohorts, which is often sparse, and inconsistent in what is measured. For example, one trial might collect demographics and data for a specific blood-based biomarker; another might also collect genomic data. Then there are differences in the analysis pipelines applied to all these data. Re-processing and harmonising all of these data types is highly labour intensive, and often only the start of the process.

The underlying issue, is that Pharma's data, particularly that from clinical trials, was not generated for AI. To exploit data in a meaningful way using AI, companies must develop a data strategy, be willing to fund and generate data on clinical cohorts if possible, and adopt approaches to maximise the value of unstructured data.

Model

While AI models excel at classification and predictive problems, if AI is to revolutionise drug discovery it must incorporate causality. Predicting that a drug might work in a new indication is valuable, but it is not the same as explaining why the drug will work in that indication. To support internal and regulatory decision-making it is essential to have explainable biology that supports a mechanistic understanding of the particular drug or biology.

The integration of prior knowledge and data-driven insights offers a promising solution. AI combined with highly accurate causal relationships can distil both a broader array of targets with strong promise, and a mechanistic understanding of their biological role in disease.

Causal relationships can be mined from the literature and created from experimental data. These relationships, defining the regulatory interactions between two biological entities, can be combined into structural causal models – a framework to represent and analyse the causal relationships between variables. Such models provide a systematic way to model how changes in one variable can lead to changes in another. These could be used during the training process of more expansive foundation models, but also to build specific mechanistic models that further describe the output from an upstream finding.

Validation

The output from all AI solutions should be validated, experimentally if appropriate, with two provisos. First, the R&D function should be set up so that all data feeds back to the AI model. This helps to mitigate some of the challenges described above, while ensuring that the model can be continually improved.

Second, there needs to be a triage-based validation model. While an AI system is able to identify hundreds of targets, the challenge is to stay open to 'left-field' opportunities that AI might highlight.

Orthogonal in silico approaches might be used to go from 1000 to 100 targets, but to go from 100 to 10 the team should adopt the quickest, most high-throughput experiment to yield the next rung of supporting evidence.

Culture

Underlying many of the data, model and validation issues up to now has been the culture of the organisation and its failure to fully adapt to an AI driven way of thinking or working.

While there are increasing efforts to bridge this gap, upskilling or recruiting talent with AI expertise is essential. At the same time data scientists must be educated in the decision-making process of R&D, and understand/develop methods that directly support that. More could also be done to build the understanding that AI will raise the productivity level of all R&D researchers, and is therefore an opportunity and not a threat.

Building the Future Today

Very soon – within a decade, certainly – we can expect every major decision taken along the drug R&D pipeline to be accelerated by unprecedented access to knowledge. But that relies on companies having done the groundwork, to put the right measures in place.

Data scientists, for their part, will need to develop actionable models with causality at their heart. Biologists must determine how to effectively integrate data science into their workflows. And heads of R&D will need to orchestrate more seamless integration and symbiosis between the two sciences.

Only then, will step changes in R&D success be possible.

Dr. Ben Sidders

Dr. Ben Sidders, Chief Scientific Officer at Biorelate, has been working at the forefront of pharma data science for the last two decades. Formerly Executive Director and Head of Early Data Science within Oncology R&D at AstraZeneca, Ben also previously spent eight years at Pfizer, and has extensive experience of many aspects of drug discovery for major pharma.

Email: ben.sidders@biorelate.com
Web: www.biorelate.com