eBOOK SERIES

DRUG DISCOVERY WORLD

DDW™

turning science into business

AI in drug discovery: How technology is driving innovation

Sponsored by:

Benchling

Ask new questions, get new answers.

Manage and extract insights from large-scale datasets using the latest AI technology – all in one platform that's compliant with industry-leading security, privacy, and regulatory standards.

Benchling

Your Single Source of Truth for Biotech R&D

DRUG DISCOVERY WORLD
DDW™
turning science into business

edify digital media

Drug Discovery World (ISSN 14694344) ebooks are published by Edify Digital Media Ltd, 1st Floor, 3 More London Riverside, London SE1 2RE, UK.

Main office tel: + 44 (0) 203 840 4720
Website: ddw-online.com

Editor in Chief
Lu Rahman
lu@edifydigitalmedia.com

Deputy Editor
Reece Armstrong
reece@edifydigitalmedia.com

Sales Director
Sarah Orme
sarah@edifydigitalmedia.com

Business Development Manager
Rich York
richard@edifydigitalmedia.com

Publishing Director
Maria Wallace
maria@edifydigitalmedia.com

Production Manager
Emma Bonell
emmabonell@edifydigitalmedia.com

Design
Steven Lillywhite
CRE8 Design Studios Ltd

Follow us on social media
@DDWJournal

Subscription records are maintained at Edify Digital Media Ltd, 1st Floor, 3 More London Riverside, London SE1 2RE, United Kingdom.

Making its mark: AI in drug discovery

By Lu Rahman

The concept of Artificial Intelligence (AI) never fails to stimulate and in drug discovery the possibilities it presents, the potential it holds, are always thought-provoking. The simulation of human intelligence processes by machines finds itself at an exciting moment for use within the drug discovery industry, especially as the world emerges from the Covid-19 pandemic.

We are well aware of the time and cost involved in developing a new drug. From research to clinical trials, FDA assessment and monitoring beyond, the process is by no means quick; anything that helps expediate the route for getting a drug to market is desirable. Using data intelligently, allowing AI into the process means drug developers can make the move from research to a working molecular model, at a much faster rate.

According to MarketsandMarkets the global AI in drug discovery market is projected to reach $1,434 million by 2024 from $259 million in 2019. It says that the growth of the AI in drug discovery market is primarily driven by factors such as “the growing number of cross-industry collaborations and partnerships, the increasing need to control drug discovery and development costs and reduce the overall time taken in this process, the rising adoption of cloud-based applications and services, and the impending patent expiry of blockbuster drugs.”

Despite its potential, this market also experiences challenges – these include a lack of data sets within the drug discovery field, and a shortage of skilled labour. However, important developments are continuing to take place, collaborations are growing and the awarenes of AI and data

benefits within drug discovery is taking on increasing significance.

"Artificial intelligence and machine learning continue to play important roles in drug discovery," said Rupert Vessey, President of Research & Early Development at Bristol Myers Squibb. This comment followed the announcement by AI-driven pharma-tech company Exscientia that Bristol Myers Squibb had chosen to in-license its immune-modulating drug candidate.

When Exscientia announced that the world's first Alzheimer's disease (AD) drug candidate designed by AI was entering Phase I clinical trials, I spoke to CEO, Andrew Hopkins about the company's work and the opportunities for AI designed drugs.

"Implementing AI in drug discovery requires a culture shift with the adoption of new approaches, new processes and ways of doing things," he said.

Examples of the implementation of AI within drug discovery are plentiful. Only recently, biotech company Galapagos announced its intention to in-license drug targets for inflammatory bowel disease (IBD) which were discovered using a biology platform developed by Scipher Medicine. This platform is powered by patient molecular data in autoimmune diseases and artificial intelligence network science-based algorithms. It identifies biologically similar patients and drug targets that are specific to those patients to improve response rates.

"Drug development in autoimmune diseases is entering an era of precision medicine similar to what we have experienced in oncology in the past decade. There is a significant unmet medical need for drugs that target a specific disease population with higher

> **AI and big data is transforming our understanding of human disease and provides a way for us to develop therapeutics at speed**

clinical response rates," said Alif Saleh, Chief Executive Officer, at Scipher Medicine. "Driven by our access to large patient population molecular data, partnerships like the one with Galapagos can bring new and more effective treatments to diseases with currently low drug response rates or limited treatment options."

Technology is continually developing. R&D cloud platform provider Benchling recently released a new AI programme to help scientists predict 3D structures of novel proteins. The company launched a beta version of its AlphaFold programme – an AI platform developed by DeepMind that can predict the 3D structure of a protein from an amino acid sequence. Those accessing the beta will be able to select any amino acid sequence stored in Benchling's R&D cloud, request a 3D structure for it, and visualise the results in Benchling's platform.

"Our team gets excited about two things: science and bringing software to science," said Ashu Singhal, President and Co-Founder of Benchling. "By making AlphaFold available to the biotech industry at the click of a button, scientists will be able to seamlessly experiment with this exciting advancement and find new ways to leverage AlphaFold output in their research. While the use cases for AlphaFold are still being explored and proven, Benchling's goal with its beta feature is to support its community."

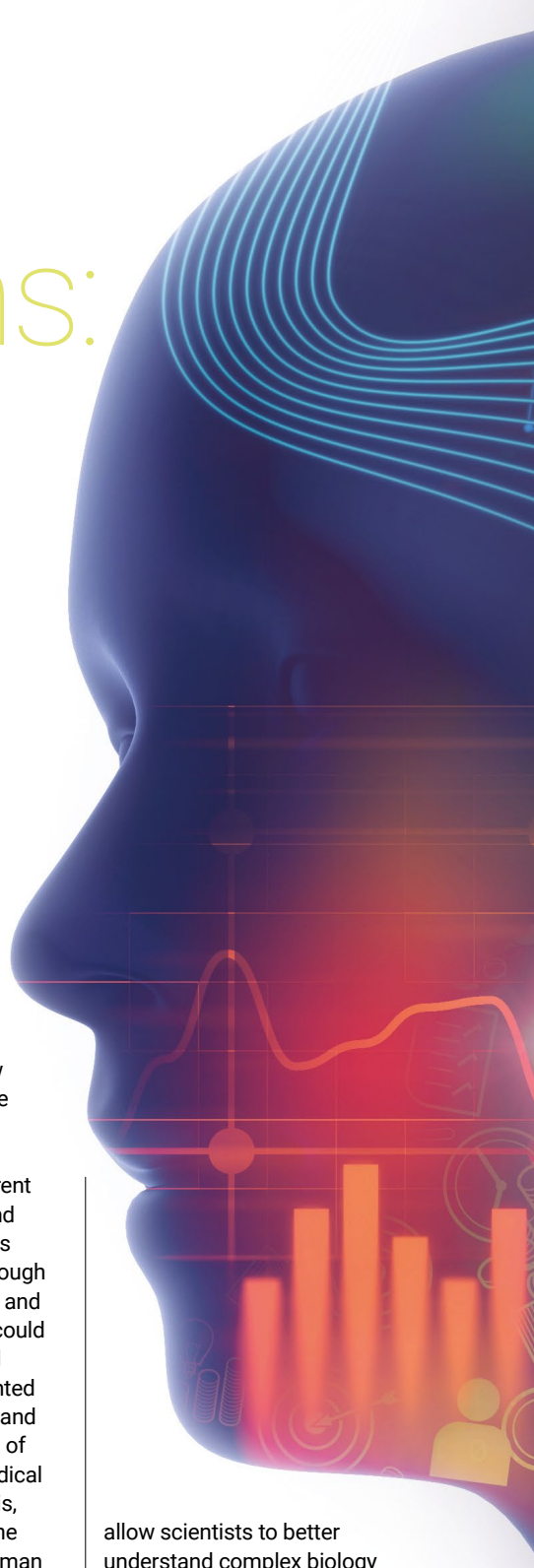
Meanwhile, AstraZeneca has identified an additional target for its idiopathic pulmonary fibrosis (IPF) pipeline using an AI system designed by drug discovery company BenevolentAI.

It is now the third target from that collaboration that has been identified using BenevolentAI's platform. AstraZeneca is using the company's Benevolent Platform across two disease areas – IPF and chronic kidney disease – to identify potential targets. The latest discovery builds upon the recent extension of the collaboration with AstraZeneca to include two new disease areas, systemic lupus erythematosus and heart failure, signed in January 2022.

Professor Maria Belvisi, SVP and Head of Research and Early Development, Respiratory and Immunology at AstraZeneca said: "IPF is a devastating disease with median survival of around three years and there is a serious need for better treatment options. At AstraZeneca, we are continuously striving to improve our R&D productivity in order to quickly bring potentially innovative treatments to complex diseases such as IPF. Our partnership with BenevolentAI furthers our commitment and we are proud to deliver a second novel target for IPF to our portfolio."

Not only is AI and big data transforming our understating of human disease, it also provides a way forward for us to develop and design therapeutics at speed, and with less cost. Innovation is key in all industries and the drug discovery sector, this is where AI comes into its own. As it continues to offer a faster route to market and a decrease in the financial outlay required, this technology holds a key role in drug discovery now and in the future.

Intelligent designs: How pharma researchers are using AI



Here, DDW's **Reece Armstrong** explores how AI is being utilised in the lab and what the technology can really do within pharma.

Artificial intelligence (AI), a branch of computer science encompassing machine learning and deep learning, and which technologies such as automation, natural language processing (NLP) and image recognition all fall into, offers significant opportunity for the life sciences industry to change how it operates. The drug discovery and development process is costly and arduous, offering no certainty of success which often rests on whether a company has chosen the correct drug molecule and target. AI has long been touted as a kind of new frontier in life sciences, something that when applied to its full potential, could fundamentally change the way pharmaceutical companies develop medicines.

We're still in the early days of the sector though, with the use of AI growing, but still sparse and largely represented by biotechs and technology companies looking to collaborate with pharma. With its use scattered across the industry, just how are researchers utilising AI to improve drug discovery?

The challenge of data

The pharmaceutical industry is built on data. Every successful therapy brings with it the necessary supporting data that confirms its efficacy and place on the market. In the drug discovery stage, researchers are burdened with the task of combing through vast amounts of datasets in order to find potential targets for their molecule. Technologies such as high-throughput screening (HTS), which can analyse vast amounts of chemical and biological compounds in order to expedite target identification, still produce hundreds of gigabytes worth of data, which researchers need to sift through in order to understand their compound.

Mark Davies, SVP of Informatics and Data at BenevolentAI describes the challenge scientists face.

"Under the current model of discovering new medicines, scientists have to manually explore vast quantities of information. They must navigate different dimensions of disease and dysregulated mechanisms and pathways, reason through potential safety concerns and assess approaches that could lead to success in clinical trials. Even the most talented scientists cannot access and analyse the sheer volume of available data and biomedical literature needed to do this, all whilst grappling with the inherent complexity of human biology," Davies says.

According to Davies, one of the benefits AI brings to researchers is its ability to help them tackle the explosion of data that technologies produce throughout drug discovery.

"AI approaches focused on information extraction and inference allows us to directly tackle this data volume challenge by pulling out the potential relationships between those entities, and in doing so, create new connections that allow scientists to better understand complex biology and point to new treatment approaches. As a result, AI can help address major drug development pain points, reducing costs and time to market," he says.

Tom Sharrock, lead of AI Engagement at Lifebit agrees: "AI enables researchers to analyse and understand volumes of data on an order of magnitude greater than the same researchers would be capable of on their own. This enables conclusions to


+11,00.00

be drawn on available data faster with improved certainty, which in turn can improve experimentation and drawing conclusions from a hypothesis."

Having this improved certainty as Sharrock describes is vital for drug discovery, especially when success rates are so low in the industry and costs are so high. With the average cost of bringing a new drug to market estimated between $985 million – $1.3 billion[1], it makes sense that companies will be looking for solutions to both reduce costs and improve their chances of success.

For Sanjay Saraf, Head of Data and Analytics Product Management at Benchling, one of the main benefits that AI and informatics is bringing to researchers is "the increased speed of therapeutic production."

"AI and informatics play a large role in driving efficiencies for researchers and lab technicians with faster decision making, as the data can reduce the number of unnecessary experiments carried out."

"Another is enabling better scientific choices and being able to discover novel therapeutics that wouldn't have been hypothesised previously. Finally, it allows for improved visibility into investment choices. Scientists no longer 'throw something at the wall and hope it sticks.' Instead, researchers can make calculated investment decisions with a rough likelihood of success," Saraf adds.

Speaking of choices, the decision a company makes in terms of what indication its molecule will focus on can be make or break.

BenevolentAI's Mark Davies says that "one of the main reasons why drugs fail is that they actually don't work in the patients they are tested in."

"This might be because the wrong target is picked, which will lead to the compound failing in the clinic. Improving success rates by selecting the right target will have a huge impact on the development of new medicines for patients and we are already starting to witness the impact of AI-driven approaches in this field of drug discovery," he says.

"This is because AI is really good at spotting hidden patterns that are otherwise buried in vast quantities of data. Equally, AI might also uncover information that tells scientists that a target is not worth pursuing."

> **The pharmaceutical industry is built on data.**

Investment

For a growing number of years now there has been significant interest in the application of AI within pharma. Investment levels, alongside mergers and

acquisitions have been steadily rising, pointing to the ongoing interest in the sector. Indeed, the market for AI applications in drug development Is estimated to rise to $5.1 billion by 2025, according to data from Emersion Insights[2].

Mark Davies sees a "significant uptick in interest from pharma and biotechs," which is being driven by a need to "ensure the quickest speed to market at the lowest possible cost."

In turn, companies utilising AI in successful ways should make a return on investment, according to Davies, "as the likelihood of the correct target being discovered increases dramatically."

"Suboptimal targets are a primary driver behind the failure of a drug at the clinical stage, and the use of AI should result in higher success rates and a quick return on investment," he adds.

In terms of resources, Professor Roland Wiest & Dr Richard McKinley, lecturers in CAS AI in Medical Imaging at the Swiss Institute for Translational and Entrepreneurial Medicine, started out with a very modest set-up. When asked about the levels of investment required for AI, Wiest & McKinley say it depends entirely on the "kind of question being answered" and "how many people are working in the lab on AI."

"For imaging we started in 2016 with a single gaming laptop, which was sufficient

to make substantial headway in the problem of segmenting multiple sclerosis lesions. We have subsequently made substantial investments, which mean we can run multiple AI projects concurrently, but our infrastructure is of course far below what would be available in industrial labs."

It's important to remember though that AI, machine learning and informatics are all part of the same puzzle when it comes to expediting drug discovery and development. High volumes of data caused by advanced instrumentation, robotic automation, and new scientific techniques, are things that researchers must apply AI to in order to gain deeper understanding.

"Machine Learning cannot be effective without accurate and standardised data," Sanjay Saraf says. The challenge however, according to Saraf, is having data in a central location and in a standardised format, especially when it is being stored is "disparate places."

"For example, some R&D teams don't yet have their systems in the cloud, and as a result they lack the computational horsepower or the skill set of data scientists to work effectively with data.

"To overcome this, many mechanisms are needed – for systemising the data that goes into the system, for identifying which features of the data are important, for training the ML model, for users to provide input and feedback on the output from the model, as well as other monitoring and management tools."

Investing in these types of tools is something that Saraf says "every company looking to leverage AI and Machine Learning will have to make."

"Without making the substantial investments into building both the teams and data architecture to support Machine Learning, businesses will fail to reap the benefits of these technologies," Saraf explains.

For Dr Arne Kusserow of Merck KGaA,(Merck Group), the investment angle boils down to the simple point of you get what you put into it.

"To keep it simple, digitalisation, automatisation, Machine Learning and AI are intended to increase efficiency. If prices can be kept, this in turn leads to higher cross-margins and revenues. The history of industrialisation proves this. It's not a reasonable question of how much money you have to invest for an invention that increases your efficiency, it is relevant if it pays off. If it pays off and others do it, while you don't, your competitors gain efficiency and will outcompete you. It's a matter of survival," Dr Kusserow tells DDW.

> **Under the current model of discovering new medicines, scientists have to manually explore vast quantities of information.**

Covid-19

The incredible speeds at which Covid-19 vaccines were successfully developed and brought to market highlighted to the world and the industry that lengthy development timelines aren't always assured. The pandemic also highlighted the need for life sciences companies to adopt technologies that can help them work more efficiently. Whether this being remote technologies to enable at-home working, cloud platforms for data, or software for clinical trials so they could continue to operate.

BenevolentAI was able to witness this level of fast-tracked development when it used its platform to help discover a potential Covid-19 treatment. The company used its platform to identify an existing rheumatoid arthritis drug owned by Eli Lilly and Company and repurpose it into a treatment to prevent Covid-19 patients progressing to being placed on a ventilator. The drug, which was used in combination with remdesivir, was granted Emergency use Authorisation (EUA) by the FDA within a year of data being published.

"I think the pandemic has pushed the industry to recognise that a change is needed in the way drugs are discovered and developed. To have an AI-enabled prediction robustly validated in large, randomised control trials has really helped build trust, confidence and credibility for our tech and approach at BenevolentAI," the company's Mark Davies says.

Dr Arne Kusserow of Merck Group thinks the pandemic has absolutely had an effect on the AI sector.

According to Dr Kusserow, pharmaceutical companies are searching for solutions that enable them to reduce the "time needed for acquiring scientific data", to share with a contract research organisation, and also to reduce the time a "regulating authority needs to understand the data." If found, these solutions can remarkably reduce the time needed to bring a drug successfully through clinical trials and approval.

"This is the lesson learned: it is possible to bring drugs faster to market and it make a huge difference. We see a strongly increased demand for our solutions," Dr Kusserow says.

Benchling's Sanjay Saraf describes how at the start of the pandemic, "the majority of companies were slowed down by their lack of ability to do in-person research, but they quickly pivoted to 'in silico' design approaches for future therapeutics." This digital approach enabled them to work faster and more efficiently than if they were in a lab.

Saraf cites examples of setting up computer models to show how an experiment would happen, rather than setting up and running that same experiment in-person. However, Saraf warns that this digital approach hasn't resulted in a total shift of how pharma companies operate and that "scientists still need to test their therapeutics in the lab."

"While the pandemic accelerated tech investment in biotech and awareness among consumers on the life-changing work pharma companies did in the fight against Covid, it may have increased time to market for most non-Covid related therapeutics," Saraf adds.

The human element

As the technology currently stands, the majority of AI and machine learning technologies cannot operate by themselves, relying on data input or oversight from humans to ensure errors are kept to a minimum. This presents an interesting paradigm where both humans and AI systems are reliant on each other to get the best possible outcomes.

For researchers in the lab, Benchling's Sanjay Saraf says that "while these tools can be used to solve extremely constrained problems, they need to be put into scientific and operational context by humans to be usable in the lab. For most analytical workflows, a human still needs to remain in the driver seat."

"Over the course of a set of experiments, most researchers will rely on some degree of human intuition. That can be as simple as identifying an obvious outlier, or excluding a class of experiments because the mechanism of action has too much overlap with critical proteins.

While AI and Machine Learning as we know them today, can't completely replace human intuition, it can analyse data, build knowledge and apply it to future problems. These technologies can also make new suggestions to either incorporate or ignore. To the degree that we consider these deductions 'obvious', we could say that AI and Machine Learning helps to reduce human error," Saraf adds.

One of the other challenges that these technologies present is the potential of bias creeping into any algorithms or predictive systems that pharmaceutical companies are using. Human bias in inevitable and unfortunately in pharma this has led to a lack of diversity in clinical trials, leading to certain medicines not working as effectively for certain patient populations compared to others.

For Kathy Brunner, CEO of Acumen Analytics, bias presents a bigger to challenge to the market uptake of AI technologies.

"The larger challenge to uptake I would say is algorithm bias which AI has to overcome in the future. Bias can find a way to lurk into algorithms in some ways. AI technology utilises training datasets to make predictions. And these datasets happen to include human-based decisions laced with gender, sex, race, or any other personal information," Brunner says.

Future trends

Over the coming years, more pharmaceutical companies will begin to embrace the possibilities that AI offers, and there's no doubt that more drug candidates discovered through AI will enter clinical trials. But what can we expect going forward?

For Lifebit, AI represents an entire change to how the healthcare industry currently cares for patients.

"Our vision is that when an individual enters the clinic, their genomic sequence, electronic health records (EHR) and diagnostic testing results will be integrated into an AI-powered system that can help diagnose their specific condition. Furthermore, once a diagnosis is complete, the system will also recommend the most suitable personalised treatment for the individual," Tom Sharrock says.

> **Over the course of a set of experiments, most researchers will rely on some degree of human intuition.**

BenevolentAI sees data as being one of the biggest challenges the industry currently faces regarding AI.

"There is a great deal of publicly available data but there is not enough of the right data for more specific applications – i.e. assay or clinical trial data – so within the drug discovery ecosystem, collaborating with the right data providers and generating the right experimental data is key. Another challenge is that the world's biomedical data unfortunately is not consistently standardised to common formats, so I hope that another key trend we will see is the consistent adoption of standardised data principles," says the company's Mark Davies.

Merck Group expects to see "faster digitalisation and automation in labs."

In comparison with other industries labs are still manufactories and far behind in terms of automatisation. Moving away from manual tasks towards higher automation always increases efficiency and decrease costs per unit. Labs will close-up, simply because they are very important to us as human beings. The main benefit for researchers will be that they can focus on what they decided to do: doing research instead of administrating and documenting their research," says Dr Arne Kusserow of Merck Group.

Benchling's Sanjay Saraf believes that now that the biotech industry has seen the benefits of AI and Machine Learning, organisations will need to begin to incorporate technology into their own operations.

"Currently, access to these AI and Machine Learning tools requires biotechs to partner with other organisations in the technology space to provide these solutions. In the next 10 years, we anticipate a shift in ownership of these technologies, with biotechs investing in developing these technologies themselves to become technology innovators. Others will continue to focus on their biological value proposition, but will need to prioritise the sourcing of AI and machine learning capabilities from a third party," Saraf says.

REFERENCES:

1 https://pubmed.ncbi.nlm.nih.gov/32125404/

2: https://emersioninsights.io/ai-drug-development-market-projected-to-reach-5-1-billion-by-2025/

AI in R&D: From hot topic to practical application

Ashoka Rajendra, Head of Development & Manufacturing Products at Benchling

Artificial Intelligence (AI) is emblematic of the new era of modern biotech — data-driven, collaborative and ultimately, faster than ever. From discovery, to lead optimsation, process development, pre-clinical, and even investigational new drug filings, each stage of R&D stands to benefit from AI.

But as much as we hear about AI and machine learning (ML), not many are taking advantage of it in their scientific work. Labs understand that it's obviously not as simple as sprinkling ML pixie dust into scientific design.

In this guide, we'll talk you through the common hurdles we see, practical advice on how to set-up a strong foundation for AI and also the opportunity of moving AI from theoretical to operational in R&D.

Machine learning: expectations vs. reality

When applying AI, people often think that the majority of their time will be spent writing complex algorithms to find previously unknown insights. The reality is that the majority of the time will be spent acquiring and cleaning data.

In 2018, the CEO of Novartis declared that Novartis would become a data science company. A year later, he discussed the challenges of powering their R&D with data science and the need for good, clean data.

"The first thing we've learned is the importance of having outstanding data to actually base your ML on. In our own shop, we've been working on a few big projects, and we've had to spend most of the time just cleaning the data sets before you can even run the algorithm. It's taken us years just to clean the datasets. I think people underestimate how little clean data there is out there, and how hard it is to clean and link the data," said Vasant Narasimhan, CEO of Novartis.

So what's the solution? If all the work scientists do is captured in software, it ought to be easier to analyse.

The solution is software, but it's not that simple

Many companies have embarked on a digitalisation journey, convinced that their notebook entries, DNA and protein designs, samples used in experiments, and experimental conditions should all be tracked and stored in software. If we go down this path and digitalise everything scientists are doing, we should be well on our way to leveraging all those insights data science can provide.

However, there are challenges with this approach. First, your users are scientists, who are extremely well trained but not software specialists. Second, their workflows are complex, evolving, and in a deep domain area. It's not easy to build high quality software in these conditions.

User-centric is key

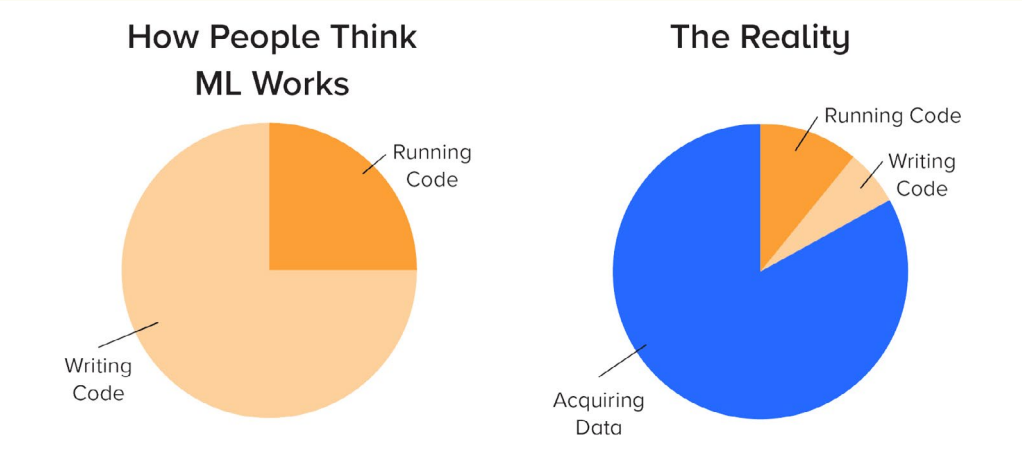
When I was meeting with a leading technologist at a pharmaceutical company recently, he jokingly said: "The best way to get compliance and adoption of tools that you introduce to scientists is to uninstall Excel from all of their computers." It seemed a bit extreme, but I saw where he was coming from. Excel is a very powerful and flexible tool, and it's popular with scientists for a reason. If another system isn't easy for them to use, if it doesn't link data together, and if it doesn't bring them immediate productivity, they'll use Excel instead.

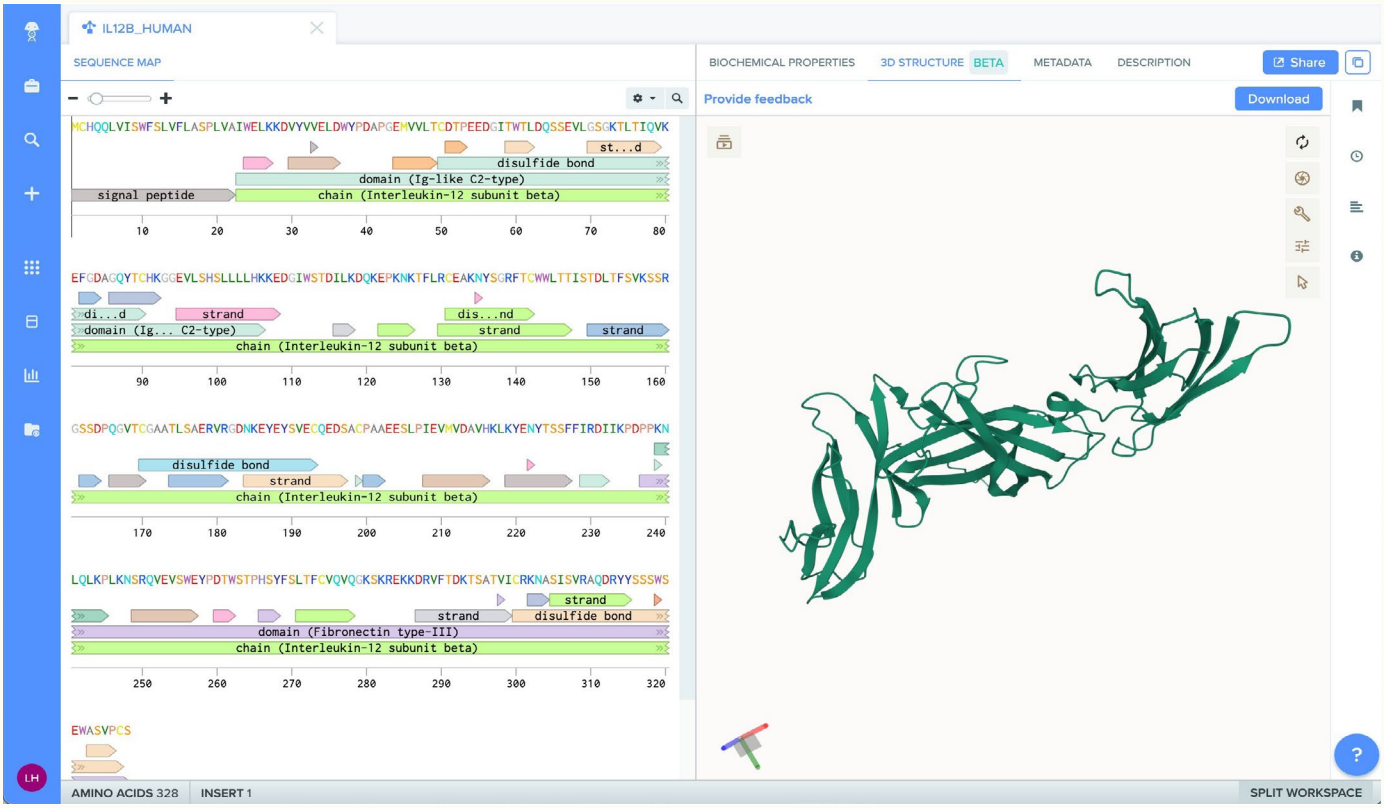
I don't think the solution is to uninstall Excel; I think the solution is to build tools that bring scientists enough value such that they choose to use purpose-built software over Excel.

Software needs to accommodate the complexity of the data that scientists are producing, and it needs to do so with a flexible data structure. If you're over-rotating on the end goals of digitisation, and not putting enough emphasis on the productivity and usability of the end-user scientist, you will not drive adoption.

R&D data is highly complex

Why is this such a challenge in life sciences? Large molecule data is extremely complex. Let's take a relatively simple example: with antibody engineering, you might start with an antibody, its target, and its binding affinity. But you also need the DNA





sequences that encode for the antibody chains, the combination of plasmids used to express the antibody, growth conditions for cell lines to express the antibody, screening data, and so on. All of this needs to be modelled, stored, tracked, and analysed, but most software isn't equipped to deal with the complexity specific to large molecule R&D. Scientists will only use your software if it accommodates the complexity of the data they create and is flexible to change.

Life science R&D data needs to be centralised...

Specialised point solutions for each of your research and development teams can cause significant problems for teams doing data analysis. Each software has a different vendor, models its data differently, and instead of a 'data lake' pooling the data together, you end up with a data swamp. Rather than analysing data, your data science or bioinformatics team will be spending their time linking, reconciling, and cleaning data. Capturing data on a unified platform can dramatically reduce this burden.

Centralising and standardising data across disparate teams is a crucial challenge for life science R&D organisations.

...on a flexible software platform

Another big challenge for software adoption in life science is that while a tool may be initially configured to represent a scientific workflow, that workflow constantly evolves. Scientists may need to test, for example, new versions of a protein purification process in order to improve the quality of the purified sample generated.

Scientists will go where the data leads them, and it's very important to have software that is adaptive to changes in experimental workflows. If the software doesn't keep up with the science, the tools will become out of date and scientists will opt for unstructured notebook entries, Word documents, or pen and paper. Given the pace of the scientific process today, you need software to be able to change in days or weeks, not months.

Companies that want to leverage advanced data science techniques need to digitise their scientists' work. Many companies have embarked on this journey but have had challenges with adoption and fragmentation of tools. At Benchling, we work with companies to consolidate many tools onto a single, unified platform.

Unlocking new capabilities

Laying this foundation with a suite of applications on a scientifically aware platform allows companies to more easily layer on advanced analysis techniques that are internally or externally developed.

Benchling's rich application functionality sits on top of a platform that allows data to be accessed via our API and data warehouse. Your data scientists and engineers can write code that pulls unified data out of Benchling, integrates it with internal systems, and feeds data through analysis pipelines. Scientists can get recommended experimental conditions, managers can look across programs and determine where more resources are needed, and executives can flag programs that are promising or risky.

AlphaFold is an exciting example of how externally developed breakthroughs can be easily applied to your scientific data. AlphaFold is an AI system built by Google's DeepMind that predicts 3D structure from an amino acid sequence with relatively high accuracy. It has been described as a significant achievement in computational biology and a game-changing tool that has the potential to speed up protein structure characterisation, work that generally takes months or years at the bench.

Benchling's scientifically aware data model allows us to easily add this functionality for scientists, without the need for investing in major computational know how or power. Having put in place this foundation and having achieved this level of data analysis, our clients are already beginning to apply AI to their R&D efforts using data from Benchling.

Precision medicine in the era of artificial intelligence

Dr. Lotfi Chouchane and **Dr. Javaid Sheikh**, Weill Cornell Medicine-Qatar, outline the challenges and opportunities ahead for drug discovery.

Precision medicine is an emerging model for the next generation of clinical care that will capitalise on the dynamic interaction between individual biology, lifestyle, behaviour, and environment. It holds huge promise for healthcare and the drug discovery sector in particular.

An essential objective of precision medicine is quantifying an individual's risk for any disease and tailoring personalised prevention and therapeutic strategies. This includes improving diagnosis, designing therapeutic interventions and determining prognosis through the use of large complex datasets that incorporate individual gene, function and environmental variations[1]. A targeted and more effective drug design based on the integration of multiple sources of data for each individual, including longitudinal multi-omic datasets, can lead to more personalised treatments.

There is much progress still to be made before the potential of precision medicine is fully realised. The recent Economist Intelligence Unit report on precision medicine[2], commissioned by Qatar Foundation, highlighted several challenges for precision medicine implementation, including the challenge of harnessing the vast data pools that already exist in order to produce actionable insights for clinicians within health systems.

Nevertheless, artificial intelligence (AI), machine learning (ML), deep learning (DL) and big data analytics are evolving to be a great aid to precision medicine. Furthermore, information and communication technologies in general and wearable sensors are helping to promote a greater level of precision in healthcare[3].

Precision medicine and biobanks

Precision medicine uses diverse technologies to collect and interpret personalised data for the sole purpose of an individual's treatment. The ability to use intelligent algorithms to mine vast stores of unstructured and structured data for better insights has empowered providers with the tools to design personalised interventions for individual patients. Recent advancements in the fast collection of data has led to an incredible increase in the volume of biological and medical data collected from human populations, with UK Biobank[4], the "All of Us" research programme[5] and the China Kadoorie Biobank[6] generating extremely thorough and deep phenotypic reports of health trajectories for millions of individuals.

Similarly, personalised healthcare initiatives in Qatar

About the authors:

Dr Lotfi Chouchane is a professor in the departments of genetic medicine, and microbiology and immunology at Weill Cornell Medicine-Qatar. He received his Ph.D in immunology from the Pasteur Institute of Paris and the University of Paris VII. He also holds a D.Sc in human genetics and immunology.

Dr Javaid Sheikh is the Dean of Weill Cornell Medicine-Qatar. He joined WCM-Q as Vice Dean for Research and Professor of Psychiatry in 2007 from Stanford University School of Medicine, where he was an Associate Dean and Professor of Psychiatry and Behavioral Sciences.

are part of a coordinated and comprehensive precision medicine strategy to deliver world-class future healthcare. The Qatar Biobank has been conducting a large population-based cohort study, which was initiated in 2012 by Qatar Foundation. The biobank's broad data sets already cover exposomes and whole genome sequencing from 20,000 individuals[7], and Qatar Genome Programme plans to sequence complete genomes of around 300,000 native Qataris. This will provide multitudes of rich source data to fulfil the aims of applying and advancing precision medicine powered with AI in Qatar[8].

Recently, the Qatar National Research Fund (QNRF), the main research funding agency in Qatar, and Qatar Genome Programme also launched the 'Path Towards Precision Medicine' research programme. This initiative aims to support genomics research to promote drug discovery and use patient specific genomic variants for tailored, personalised therapies for the Qatari population.

AI and machine learning in drug discovery and development

Historically, drug discovery is a long and very expensive process and highly prone to failure due to unexpected/unpredictable toxicity, poor pharmacokinetics or insufficient activity of potential therapeutic molecules. The launch of a new drug on the market costs between several billion to tens of billions of dollars, typically taking between three to 20 years. A research survey among 106 new drugs developed by 10 pharmaceutical companies found that they cost on average $2.7 billion (£2 billion) to develop[9].

In the current omics era of big data, the implementation of AI/ML based algorithms has leveraged the paradigm of 'one gene, one target, one drug' into a framework of unselective targets, even for one drug[10]. In this context, AI/ML/DL algorithms can learn from heterogeneous datasets and discover new drug targets, repurpose the current existing ones or eventually guide the decision-making protocol. Recently, this was demonstrated

> **Historically, drug discovery is a long and very expensive process.**

with international clinical data sharing programmes for Covid-19, which has opened up an enlightened vision in which AI/ML can guide clinicians to a rapid classification of the severity of the infection and thus the most effective treatment[11].

Moreover, using AI/ML can enable the drug discovery community to benefit from large sets of expression data from target tissues or organs. This data can help to identify cell membrane receptors with a regulatory role in disease-related gene expression. This allows medicinal chemists to elucidate the mechanism of action of a disease, trace back the target(s), data mine existing databases for drugs with an inhibitor, and computationally predict the effectiveness, potency and selectivity of different drugs.

For example, in an effort to improve the accuracy of deep DTnet – a deep learning algorithm for the identification of new drug targets – Zeng and co-workers experimentally predicted and validated topotecan, an approved drug for ovarian carcinoma, as a promising treatment for multiple sclerosis[12]. In addition, a recent study into the inhibitors of triple-negative breast cancer, which made use of deep neural networks (DNN), demonstrated which compounds were most efficient[13]. This implies that AI/ML has a chance to be applied in drug selection, repurposing and thus accelerating the process of drug discovery without repeatedly reverting to de novo design.

We have also been seeing growing interest of big pharma companies in applying AI/MI in precision medicine for drug discovery and treatment. Aiming to apply genome sequencing powered by AI/ML to a large population, the pharma company

Roche acquired private company Bina in 2014. Additionally, GNS healthcare announced the launch of a collaboration with Genetech to boost the development of novel cancer therapies using the GNS REFS (Reverse Engineering and Forward Simulation) causal machine learning and simulation platform. Furthermore, the progressive partnership between Biogen, EMBL-European Bioinformatics Institute, GlaxoSmithKline and the Wellcome Trust Sanger Institute established the Open Targets validation platform. This platform is a public-private partnership that uses genetics and genomics data for systematic drug target identification and prioritisation, and this large database trained on four different ML classifiers. Medicinal chemists can now use the platform for drug discovery[14].

AI and machine learning in precision oncology

Precision oncology treatments rely heavily on the patient's genomic data to make treatment decisions. Whole genome sequencing has already improved our understanding of tumours; the unprecedented molecular detail has enabled highly efficient targeted therapy, coupled with a new generation of drug development.

At the same time, cancer drug development is evolving rapidly thanks to precision medicine, with efforts centred on matching drugs or treatments to predictive marker(s) for selected patients[15]. The use of AI/ML has already proven to be successful in choosing the drug combination based on a patient's own biopsy, and to make recommendations for N-of-1 medications[16]. In cancer treatment, it is crucial to identify reliable drug targets and the driver genes for personalised medicine. AI/ML has begun to play a role in generating novel drug candidates and repurposing existing drugs. As for cancer drug development, a critical demand for the agents to target low incidence mutation is inevitable.

> **In high income countries, AI powered healthcare practices have already been put into place.**

Although AI/ML can improve the design of preclinical experiments and clinical trials, help to match correct patients with clinical trials and even optimise clinical trials, current regulatory requirements indicate that having enough patients for low incidence events is a big obstacle. Fortunately, AI also provides a solution, which is in silico patients. With AI patients, we can run in silico clinical trials to identify responders and optimise therapy combinations, the line of therapy and treatment sequences.

Challenges and future directions

One of the key challenges in the process of drug development is ensuring drug safety. Translating knowledge on the known effects of drugs to anticipate their side effects is a difficult process. Scientists and engineers from academic institutions and pharmaceutical industries such as Roche and Pfizer have sought to use AI/ML to extract useful knowledge from data gathered in clinical trials. Currently, an active area of research is the analysis of this data in the context of drug safety[17].

Yet, despite all the promises of AI/ML technology in precision oncology, obstacles and pitfalls are formidable in the real world of cancer patient care. A recent example is IBM's AI algorithm, Watson for Oncology[18]. This algorithm was based on a small number of synthetic cases with very limited real data. Many of the recommendations were shown to be erroneous, such as suggesting the use of bevacizumab in a patient with severe bleeding – which represents an explicit contraindication for the drug.

However, the efficiency and usefulness of AI/ML algorithms depends heavily on the accuracy and consistency of the data they are trained on. This doesn't mean the failure of AI in clinical care. In fact, various algorithms essentially share similar principles and no revolutionary algorithm method has emerged so far.

More comprehensive and accurate training data is the key to the success of AI in precision medicine. Therefore, greater effort should be put into generating data and collecting data from diverse populations. Exactly how this revolutionary role of AI will improve the real clinical world remains to be demonstrated and will be dependent on the availability of comprehensive, trustworthy and diverse patient data[19].

To this end, crowd-source challenges have been designed to tackle cancer genomics, using experimental data to objectively and transparently evaluate the accuracy. Collaboration between many research entities and pharmaceutical companies has also enhanced access to a range of candidate drugs and a large number of tumour-specific consortia[20,21,22]. While such collaborations improve the chances of success, they increase the complexity of precision medicine for heterogenous diseases.

Genomic medicine and ML have been successful in identifying rare diseases with a different frequency of occurrence too (eg. rheumatologic versus rare inborn errors of metabolism). Data in the literature of the last 10 years alone contains

at least 74 different cases of rare diseases[23]. To this end, a possible route could be establishing a reward for the pharmaceutical industries for investing in medicines that target a small number of patients with rare diseases. Nevertheless, there are encouraging examples of pharmaceutical companies supporting rare diseases. Novartis collaborated with the private company Pharming to work on a molecule for the treatment of Activated PI3K delta syndrome (APDS), an ultra-rare autoimmune disease. Similarly, GlaxoSmithKline expressed interest in the treatment of the same pathology and is testing the use of a drug administered by inhalation.

Rare diseases contribute to a significant proportion of morbidity and mortality in populations with high rates of consanguinity, including in Arab populations. They are estimated to be the second leading cause of infant mortality in Qatar. Many genetic diseases that plague Arab populations are not greatly shared with other populations; therefore, genetic testing developed for other populations are of limited value for Arab communities. Qatar is becoming a precision medicine hub for rare diseases, and we expect to see more biotech and pharmaceuticals companies setting-up joint business ventures here in the coming years.

What's next

The ways in which AI/ML can support precision medicine are truly innovative and could certainly lead to major scientific achievements. Precision medicine is already bringing significant improvement in rare diseases, with previously undiagnosed diseases being identified, and patients with potentially lethal cancers are now seeing higher life expectancies and, in some cases, cures.

AI is evolving by itself but with all these trials generating big data, it will also foster the evolution of AI further. Human populations have been combating chronic diseases such as diabetes, obesity, cancer and rare diseases for the last several decades. Will AI prove to be the potent weapon that will turn the tide in this long-ranging battle? Many countries and stakeholders are betting on it.

In high income countries, AI powered healthcare practices have already been put into place. For example, in the UK and Singapore, national AI-based initiatives have been introduced to effectively deal with the burden of many diseases. Qatar is joining in this battle by launching national initiatives to implement precision medicine powered with AI. The country is on its way to becoming a hub of experimentation and hopes to lead neighbouring countries in the Middle East and North Africa in this most important endeavour[24].

Acknowledgments

In addition to Dr. Chouchane and Dr. Sheikh, the following people contributed to this article: Murugan Subramanian, AtilioReyes Romero and Jingxuan Shan from the Genetic Intelligence Laboratory, Weill Cornell Medicine-Qatar.

The authors would like to thank Qatar National Research Fund and the World Innovation Summit for Health (WISH) for their support.

REFERENCES:

1 Melamud, E., et al., The promise and reality of therapeutic discovery from large cohorts.J Clin Invest, 2020. 130(2): p. 575-581.

2 Doing well? Fullfilling the promise of precision medicine. The Economist Intelligence Unit 2020. (https://www.eiu.com/n/fulfilling-the-promise-of-precision-medicine/)

3 Ho, D., et al., Enabling Technologies for Personalized and Precision Medicine.Trends Biotechnol, 2020. 38(5): p. 497-518.

4 Sudlow, C., et al., UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.PLoS Med, 2015. 12(3): p. e1001779.

5 All of Us Research Program, I., et al., The "All of Us" Research Program.N Engl J Med, 2019. 381(7): p. 668-676.

6 Wang, M., et al., Associations between stressful life events and diabetes: Findings from the China Kadoorie Biobank study of 500,000 adults.J Diabetes Investig, 2019. 10(5): p. 1215-1222.

7 Al Thani, A., et al., Qatar Biobank Cohort Study: Study Design and First Results.Am J Epidemiol, 2019. 188(8): p. 1420-1433.

8 Fakhro, K.A., et al., The Qatar genome: a population-specific tool for precision medicine in the Middle East. Human Genome Variation, 2016. 3(1): p. 16016.

9 DiMasi, J.A., H.G. Grabowski, and R.W. Hansen, Innovation in the pharmaceutical industry: New estimates of R&D costs.J Health Econ, 2016. 47: p. 20-33.

10 Vamathevan, J., et al., Applications of machine learning in drug discovery and development.Nat Rev Drug Discov, 2019. 18(6): p. 463-477.

11 Ning, W., et al., Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning.Nature Biomedical Engineering, 2020.

12 Zeng, X., et al., Target identification among known drugs by deep learning from heterogeneous networks. Chemical Science, 2020. 11(7): p. 1775-1797.

13 Tsou, L.K., et al., Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery.Sci Rep, 2020. 10(1): p. 16771.

14 Koscielny, G., et al., Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res, 2017. 45(D1): p. D985-D994.

15 Azuaje, F., Artificial intelligence for precision oncology: beyond patient stratification.npj Precision Oncology, 2019. 3(1): p. 6.

16 Ho, D., Artificial intelligence in cancer therapy.Science, 2020. 367(6481): p. 982-983.

17 Ekins, S., et al., Exploiting machine learning for end-to-end drug discovery and development.Nat Mater, 2019. 18(5): p. 435-441.

18 Miller, A., The future of health care could be elementary with Watson. CMAJ, 2013. 185(9): p. E367-8.

19 Wilkinson, J., et al., Time to reality check the promises of machine learning-powered precision medicine. The Lancet Digital Health, 2020. 2(12): p. e677-e680.

20 Zardavas, D., et al., The AURORA initiative for metastatic breast cancer. Br J Cancer, 2014. 111(10): p. 1881-7.

21 Khan, S.S., A.P. Chen, and N. Takebe, Impact of NCI-MATCH: a Nationwide Oncology Precision Medicine Trial. Expert Review of Precision Medicine and Drug Development, 2019. 4(4): p. 251-258.

22 Shah, R. and B. Khan, The MATRIX trial.Lancet, 2019. 393(10183): p. 1803.

23 Schaefer, J., et al., The use of machine learning in rare diseases: a scoping review.Orphanet J Rare Dis, 2020. 15(1): p. 145.

24 Qoronfleh, M.W., et al., THE FUTURE OF MEDICINE, healthcare innovation through precision medicine: policy case study of Qatar.Life Sciences, Society and Policy, 2020. 16(1): p. 12.

The next step in big data and AI equals personalised, predictive, preventative medicine

Big data and AI are all the rage. But, beyond these buzzwords and the quantity of data, what's next for these powerful tools? **Troy Groetken** of intellectual property and technology law firm McAndrews, Held & Malloy tells **Lu Rahman** how pharma companies can take this data from where it currently stands (drug design and development) to a more clinical level.

Troy Groetken says that data has the ability to create personalised, predictive and preventative medicine tailored to individual patients and identified patient groups/populations. Groetken has 25 years' legal experience in the intellectual property (IP) field and more than 25 years' technical experience in the pharmaceutical, biotechnological, and chemical fields.

As a registered US patent attorney, Groetken is known globally as a 'go-to' intellectual property attorney for Fortune 500 clients on complex and cutting-edge IP matters and on strategic global patent portfolio development, implementation, and enforcement.

"I love innovation and working with incredible minds in the pharmaceutical and life science fields who create amazing new technologies and methods of treatment. I am never bored being a patent attorney," he says.

His work provides huge insight into AI and big data in relation to pharma and drug discovery – now and in the future. "AI and big data are now commonly involved with drug discovery and development. The next phase, however, is the utilisation of AI and big data as to precision therapy. In doing

so, AI and big data can be used to assess treatment modalities, especially in the clinical trial dynamic currently, to better tailor those treatments to various patient de-mographics (particular genders, ethnicities, age, comorbidities, and so many additional varia-bles)," he says.

According to Groetken the overall goal is to better understand the incredible amount of data that is generated from clinical activities and to use that data "to better design treatments for individual patients and patient groups by improving results, reducing side effects, among other outcomes".

> **The next phase, however, is the utilisation of AI and big data as to precision therapy.**

Key opportunties

Drug development companies are always keen to uncover new market opportunities and Groetken is clear about the ways they can maximise on the commercial opportunities created by AI and data. His advice? Up front collection and controlling your data!.

He adds: "The better data set that can be developed short- and long-term allows for maximum AI utilisation of that data. Better data helps to prevent garbage-in-and-garbage-out when AI is applied. As a result, better technologies and treatments are developed. Further, improved intellectual property can be obtained as well. From an ROI perspective, AI and big data is definitely the new frontier for many pharmaceutical and life science companies."

So what advice would he suggest for pharma and life sciences companies looking to take current data (ie drug design and development) to a more clinical level? And what kind of investment / skill level would be required for this to take place?

"For a pharmaceutical company to take advantage of AI and big data, a well-constructed team of professionals is needed. The standard team of today will now also include software engineers, AI specialists and big data specialists who bring a wealth of new skills to the

table to help design clinical programs that collect patient outcomes and then model new and improved therapies. The kind of investment will be significant and ongoing for those pharmaceutical and life science companies who wish to develop within this arena. However, the rewards of such investment could be very substantial as well," he says.

Groetken sees many opportunities to utilise big data and AI in the pharma sector. For example, in therapeutics he acknowledges that AI and big data already allow for high throughput outcomes to identify pharmaceutical and life science candidates with potential value. Where clinical trials are concerned, he sees the opportunity for improved data collection and patient/patient group modeling to enhance therapy development at an accelerated pace.

"AI and big data allows for enhanced modeling analysis in a real time environment to determine ways to improve the therapeutic outcomes, reduce negative outcomes, and assess if a particular approach should be continued or discontinued," he says.

Of course there are always barriers to the uptake of any technology and Groetken is open about this: "The main barriers to using AI and big data in a more precision-therapy-manner are ownership, use constraints, regulatory constraints, and privacy considerations. As data are generated, who owns it? The patient? The clinical facility? The company paying for the clinical trial?"

He believes we need to take all these owners into account and the data needs to be treated accordingly bearing each owner in mind. "Additionally, as AI uses that initial data and may interact with the internet to cross consider further data, who owns those resultant outcomes?" he asks.

Groetken believes we need to consider cross-ownership concerns of comingled data by the AI mechanism and raises some key questions. "Does this lead to patent infringement concerns? Does this also lead to other legal ownership and misappropriation concerns? Alternatively, a number of concerns are raised if only one source 'owns' the data or AI mechanism. If so, what about governmental interests? What about access by others to move the data and AI resultant outcomes forward in their own creative ways? Isn't the whole point of data and AI to accelerate innovation? Yet, ownership and ROI are always factors that must be considered as well," he says.

> **Groetken sees many opportunities to utilise big data and AI in the pharma sector.**

He adds that a new balance has to be considered between the investment and return on investment that must be capitalised for big data and AI and the public needs to use and access that same data, AI mechanism and resultant outcomes once AI is applied to that big data.

"In many ways, attorneys are addressing these various issues currently and trying to draft improved legal documentation to address these ownership and related issues," he says, adding that the 'how' has to be answered as well.

"If data is generated, how will it be utilised with the AI mechanism? Is the patient, patient population, clinician, clinical facility, company, governmental entity in agreement as to how such data will be implemented with AI?" he asks.

This of course, then interacts with privacy concerns, especially in the clinical setting. "How will the privacy of the patient be maintained? How will resultant information be kept in a manner that allows for innovation and precision therapy development, while ensuring patient/patient populations privacy," Groetken states and says that this is not so easily achieved when the overall goal of precision therapy is to tailor therapies to a particular patient or patient population, which will inevitably already have various identifiers, characteristics and the like.

"Here again are a number of considerations for attorneys as they work within these areas of the pharmaceutical and life science fields," he says.

Regulatory issues

Finally, current regulatory frameworks are a barrier to the AI and big data process and progress. According to Groetken such frameworks are built upon methodologies that do not involve AI. "As a result," he says, "many regulatory schemas are being reviewed currently and analyses are being done as to how AI can be evaluated and validated. Again, if the AI dynamics change (as they can during the drug development and precision therapy development process, etc), what are the impacts of such changes to the safety and efficacy analysis from a regulatory perspective for the therapy being developed?"

He says that the analysis and review may not always involve clinicians but a much larger professional team involving software engineers and the like, which is a new landscape for regulatory bodies.

"As a result, regulatory bodies such as the FDA are trying to adapt their processes for drug and medical device reviews when AI and big data is involved in the development of those products and treatments," says Groetken.

The global view

Groetken believes that those countries that currently have a well-developed pharma/life science systems in combination with AI systems have a significant advantage and head start. "Thus, the United States, Europe, Japan, among others would appear to have a significant leg up on the competition since each system already has mechanisms in place that they can bring to bear upon precision therapy development with AI and big data. Already, we are seeing significant investment in these countries and regions regarding AI and big data for drug and therapy development."

About the author:

Troy Groetken has around 25 years' legal experience in the intellectual property (IP) field and more than 25 years' technical experience in the pharmaceutical, biotechnological, and chemical fields. He is recognised in the IAM Patent 1000: The World's Leading Patent Professionals, and has been listed as one of the Best Lawyers in America since 2012.

As a registered US patent attorney, Groetken is known globally as a 'go-to' intellectual property attorney for Fortune 500 clients and others on complex and cutting-edge IP matters, and on strategic global patent portfolio development, implementation, and enforcement.

He also advises upon and institutes multi-level, front-end and back-end diligence and licensing programs coordinated with a client-focused business modeling approach. In addition, he assists clients with advanced acquisition, divestiture, and platform-positioning transactions designed for institutional growth and enhanced business valuations.

He strategically addresses multi-jurisdictional intellectual property integration issues ranging from core and non-core patent portfolio competitive white-space analyses to global portfolio coordination.

Making life science data work for the digital age

The life sciences data landscape is growing in complexity and scale, with organisations generating increasingly high volumes of data. Yet, these data are still stored and searched using outdated methods. As a result, there are vast amounts of unstructured data stored in various siloed locations. **Vladimir Makarov**, Consultant at Pistoia Alliance, discusses this challenge and ways to move forward.

One example of this challenge lies in current approaches to bioassay protocols, which provide information about the methods for biological research. This information is hard to find, compare, analyse, or use in data mining. Considerable time investments are required along with a specialised expertise. In fact, the Pistoia Alliance conducted a number of interviews with scientists on this subject, finding that they spend up to 12 weeks per assay selecting and planning new experiments.

The challenge of siloed data is symptomatic of a larger failure to collaborate in the life sciences, and it also represents the reluctance of life science organisations to fully take advantage of digitalisation. In many ways, digitalisation has transformed the life sciences, yet it seems data management systems are the last remaining mark of legacy approaches.

Making data FAIR

It's essential that any new method of storing and searching life science information ensures

that data are made FAIR: findable, accessible, interoperable and reusable. To facilitate a more collaborative approach and ensure that organisations and individuals all work to the same standards, the FAIR principles should be applied across the industry. If data are made FAIR, they become more easily retrievable and sharable, preventing unnecessary duplication of research, and perhaps more importantly, repetition of experiments that failed in the past.

Taking again the earlier example of bioassay protocols illustrates the broader issue with current data systems. At this time bioassay data are not recorded in the FAIR format. The assay protocols are widely accessible, as they are stored in public data banks, either in form of research papers or a metadata attached to scientific results, however, both exist in plain-text formats. This means assay protocols are not machine-readable, and therefore require manual review. Many of the current assay protocol annotations also lack the depth or quality needed to drive the research forward. As a result, scientists spend huge amounts of time manually sifting through vast libraries of old records, rather than conducting new research or being able to apply AI and machine learning to datasets.

A case study

The FAIR principles enable data to be machine-readable and so mitigate this challenge. With this change, it becomes far easier to implement AI and machine-learning, which will transform the data-searching process by saving time and reducing room for error. One example of this in practice is the DataFAIRyproject, which demonstrates the advantages offered by an approach that combines FAIR with AI and ML. In the DataFAIRy project the assay unstructured metadata is processed by an automated Natural Language Processing engine, and the output is then vetted by human experts (a 'human in the loop' approach) to ensure the quality of annotations. To develop the DataFAIRy method, the Pistoia Alliance first conducted extensive analysis of the needs of a typical scientist in the pharmaceutical industry. Then, the project team developed an ontology-based model that would allow the typical data mining questions to be answered.

> **Considerable time investments are required along with a specialised expertise.**

Perhaps the greatest advantage offered by adopting a DataFAIRy type approach to data management is the vast potential for time saving. As the cost of developing new drugs continues to rise, it is vital for scientists to work more productively, spending more time on analysis and as little as possible on preliminary research.

Why change now?

As datasets generated by organisations grow in volume and complexity, there is a need for new search and storage methods that assist scientists working in R&D, rather than slowing them down. With the acceleration of digitalisation, we should look to new technologies and standards to solve the problems presented by manual search methods.

Projects like DataFAIRy encourage a collaborative approach between scientists and organisations, so that data can be accurately shared between teams and organisations, thus reducing the time wasted by errors in data, or by repeating experiments that have already been completed. With the Pistoia Alliance aiming to scale up DataFAIRy's annotation process to thousands of assay protocols at a time in the next phase of the project, this method – and others like it – has the potential to transform the way in which bioassay protocols, as well as other types of important data, are recorded and searched in life sciences.

About the author:

Vladimir Makarov is a consultant at Pistoia Alliance, a global, not-for-profit members' organisation collaborating to lower barriers to R&D in life sciences. He has experience working in informatics and biotechnology, and has a PhD in Computational Biology.

Leveraging AI and Machine Learning for smarter clinical trial design

By **Rafael Rosengarten**, CEO of Genialis and Executive Member of the Board of Directors for the Alliance for Artificial Intelligence in Healthcare (AAIH).

How does a late-stage clinical trial fail when the preceding trials showed such promise? The likely reason is that the patients in the earlier trials were not representative of the larger population.

Clinical trial participation is shaped by stringent enrolment criteria, and access to trial centres, as well as socioeconomic and demographic factors, are key determinants. To improve clinical trial representativeness, the medical community must broaden community access, focus on diversity and inclusion in enrolment, and factor in adaptive trial designs and precision patient-trial matching.

Working its way to the forefront to aid in improving clinical trial design is artificial intelligence (AI) and machine learning (ML), a subset of AI that involves using mathematical algorithms to infer patterns from some input data and perform tasks on new data based on the previously identified patterns. For example, an ML model might be constructed to predict which patients are most likely to respond to a specific class of drug based on molecular data from blood or tissue biopsies and/or prior treatment history derived from a body of clinical records.

The development of ML-based biomarkers may be used for the diagnosis of disease, prognostic risk assessment and predictive stratification of patients for treatment. These biomarker models often learn patterns from molecular data such as genomic variants or gene expression signatures, or they may perform digital pathology analysis, or use health history, demographics, and other patient data. Collectively, the application of ML to develop clinically actionable biomarkers aims to make precision medicine the "default" approach to medicine.

Another application in ML-enabled clinical trial innovation is the digital twin. Each patient's entire health record is modelled to create a statistical twin of the live human and then used to predict the control or placebo effect in a trial. One benefit of the digital twin is significantly fewer enrolees, as each person can enter a treatment arm with a matched control. This technology may also be desirable in therapeutic areas where a placebo, or even standard of care, is unethical due to the severity of the disease.

We also can't overlook an AI solution modelling of real-world evidence (RWE) to bridge between clinical trial findings and actual clinical practice.

RWE, which derives from the aggregation and analysis of real-world data (RWD), aims to learn all we can from a broad population of patients. While RWE presents the opportunity to assemble large datasets from a more generalised subset of people, this application area must still pay attention to the actual patient diversity captured in the records, not to mention tackle challenges associated with harmonising disparate, often incomplete, and unstructured data.

The same phenomenon that can derail a promising clinical trial programme—that is, failure to represent the intent to treat population—is of central concern in the application of AI/ML in healthcare settings. AI fails to be applicable when the data used to construct, train and validate the model does not represent the larger population of data. Virtually any modeling exercise rests on the core assumption that new data to be evaluated derives from the same distribution as the training data. Sadly, violating this assumption is the norm.

Given that much of the data used to model patients comes from the same under-representative trials, ML models risk learning exactly the same biases, rather than uncovering some hidden biological signal. Likewise, RWD collected from major academic clinical centres risk excluding patients who do not have the means to access those facilities but rather seek treatment at their local community practice. Practitioners of ML must take care to confront, identify and address sources of bias in the datasets they model. One should deliberately seek out

> **Practitioners of ML must take care to confront, identify and address sources of bias in the datasets they model**

> **Another application in ML-enabled clinical trial innovation is the digital twin.**

independent datasets on which to train models. Independent data are critical for model validation as well. For too many ML models, the reported performance is limited to cross-fold validation within the training set. That is appropriate for initial model development, but additional validations should assess how well the model does on all new data derived from different sources.

The responsible adoption of AI in this arena is a team effort, with contributions coming from all sectors. Providers, payors, patient and family groups, as well as many biotechs, are beginning to put AI to work. The FDA has taken a lead role in devising the necessary frameworks to ensure the safe and effective development and deployment of AI within medical software and devices. These stakeholders are committed to the shared promise of AI to improve patient outcomes by enabling smarter clinical trial design, facilitating enrolment, identifying reliable biomarkers, and bridging the gap between the realities of trial participation and the broader patient population.

About the author:

Rafael Rosengarten, CEO of Genialis, leads the company's effort to integrate and mine vast and diverse sources of biomedical knowledge to realise the promise of precision medicine and therapeutic discovery. He spent nearly 20 years in biomedical research prior to Genialis, publishing on the evolution of innate immune systems, bioengineering of microbes, and genetics of development. He has also nurtured a specialty in developing software for high-throughput molecular design and analyses, co-inventing the j5 DNA assembly design automation tool (which has since been commercialised by TeselaGen Biotechnology). Rafael attended Dartmouth College and then Yale University, where he was an NSF Graduate Research Fellow. He went on to postdoctoral training in Jay Keasling's synthetic biology group at Lawrence Berkeley National Laboratory, Joint BioEnergy Institute (JBEI), followed by a National Library of Medicine fellowship in Biomedical Informatics at Baylor College of Medicine.

Modern software to accelerate modern science

Trusted by the world's largest biotechs and the next generation of scientists to accelerate the path from research to breakthrough.

Benchling

Your Single Source of Truth for Biotech R&D