The past 10 years have seen an amazing change in the miniaturisation, cost reduction, high-fidelity and data acquisition of modern instrumentation; the surge of robotic-controlled processes enabling DNA synthesis, genome editing, screening, plating and cell culture have led to a data explosion. In the nineties and early 2000s, the introduction of automation and high-throughput screening transformed the way in which drug discovery research was performed, leading to a rise in the number of compounds tested against a target of interest and a significant amount of investment in the quest to produce the ultimate screening factory. Massive repetition led to consistency – lower error (greater precision with higher number of experiments), better fidelity and the ability to quickly generate enough data to run \textit{in silico} or ‘virtual’ experiments. This boon has generated yet another problem, that of ‘Big Data’ \footnote{1} where sophisticated algorithms must infer patterns from large warehouses of data to distil wisdom from gathered information.

Now, as many researchers struggle with the ever-increasing complexity of drug development and the rise of personalised medicine approaches, the increasing use of Artificial Intelligence (AI)/Deep Learning (DL) presents one of the most promising and transformative opportunities for the life sciences and medical industries \footnote{2}. The emergence of AI/DL in drug discovery provides many advances over traditional techniques in genomics, image analysis and medical diagnostics \footnote{3} and is one of the reasons that pharmaceutical companies such as Merck, Sanofi, AZ and Takeda are placing big bets on the ability of AI to deliver improvements in quality, clinical success rates and reduced costs \footnote{4}.

This short perspective will show the reader how recent revolutions chemical and biological automation produce enough data and learning to build a...
deep learning model pipeline, making science faster, more efficient and more accurate.

Chemistry automation – computation, synthesis and prediction

Prognosis: available and maturing

As a discipline, chemistry has existed for thousands of years and its practice on industrial scales in recent history dates back at least a few centuries. As such, AI/ML and machine learning (ML) can more readily be applied here: multiple large data sets exist in the public domain, albeit of varying quality and reproducibility, and all large companies have proprietary molecular databases containing molecular descriptors, predicted properties, safety thresholds and assay data against multiple targets of interest.

Most applications in artificial intelligence and the subset field of machine learning in chemistry are concerned with predicting three things: which molecules to make, how to make them and predicting properties or safety panel data for the molecules thus synthesised. To be clear, applying algorithmic intelligence to synthetic pathways is not new: Nobelist E.J. Corey explored symbolic logic and computational route-finding with his LHASA and OCSS already in the late 1960s5. Chemistry AI generally, and machine learning from text-mining specifically, has been used to great effect. One of us (MAT) assisted a Novartis team with a text-mining and predictive reaction assignment of 30 years of patent literature, seeking to describe en masse the types of molecules and physical properties produced by pharmaceutical and biotech firms6. Retrosynthetic analysis now commonly runs on multiple commercial platforms, such as SciFindern (formerly ChemPlanter)7 or MilliporeSigma’s Synthia (formerly Chematica). Both systems utilise machine learning to predict routes based on combinations of expert rule-sets and literature-derived reaction capture8.

Multiple companies have begun to use fully AI-based approaches to target selection, drug design, property prediction or synthetic execution. Cyclofluidic utilised AI-based algorithms to analyse product output to inform future synthesis rounds, an approach adapted and refined by Exscientia. Nimbus Therapeutics claims to utilise bespoke computational tools to shorten drug discovery development times by 75%. Revolution Medicines, which spun out from synthetic automation efforts at the University of Illinois – Urbana-Champaign – uses a proprietary computational pipeline against oncology targets, while Recursion Pharmaceuticals mines phenotypic data to construct models for ideal drugs9. Though not itself AI, data collection and automation efforts on an academic scale; fully digital synthetic ‘engines’ have sprung up in labs such as Lee Cronin’s at The University of Glasgow10.
One risk for chemistry is that in many small firms and academic labs, data is still recorded in paper lab notebooks, which slows the amount of AI adoption to early discovery and more risky reactions. Automation provides consistent data templates and inputs for algorithms, which means a gap to overcome would be to decide which processes are automatable. Perhaps due to the large data warehouses and mandated electronic storage of experimental records and instrument data, practical innovations in this space from industry have outpaced academia.

Another cautionary tale has come from the use of sophisticated AI systems such as IBM’s Watson, famous previously for accomplishments in beating human competitors in Jeopardy and assisting doctors with medical literature gathering. However, IBM announced in early 2019 that it had halted drug discovery applications of its system. This may also be an indication of the maturity of AI in this field; IBM’s exit was not seen as an algorithmic issue, but rather one of insufficient high-quality data on which to base drug predictions\(^1\). To be certain, as observed in Ian Davies’ seminal Nature perspective\(^2\), we face many challenges to wider adoption of AI in chemistry automation – consistent data formats, standard automation sets, cultural shifts in reporting and practice.

**Discovery biology – from DNA to screening**

**Prognosis: gaining traction**

Various methods of AI have been successfully applied in drug discovery outside the small molecule world. AI, and more specifically deep learning, are useful on several levels for life science and biologists in particular: AI improves the quality of the obtained insight and allows scientists to handle the large volume and speed of data coming from automation that would otherwise require ‘superhuman’ effort from a single bench scientist.

Biological deep learning allows scaling-up of effort and expedites traditionally manual steps. These successes depend heavily on several factors: the amount of data, the type of data to analyse and the methods used for the analysis. Unlike most bench chemistry, diversity in biology data is the main hurdle. Another obstacle, machine learning, requires large amounts of data across the range of expected measurements of a given experiment, yet also a large amount of data necessary for the training, validation, testing and finally execution of the machine learning method.

Though a general ‘biology agent’ does not yet exist, deep learning methods have been applied to more narrow questions with success.

One application? Phenotypic screening. Just like image recognition of cats, we can leverage similar algorithms for microscopy image analysis. Image-based screening is particularly suitable for high-throughput cell biology where microscopy images provide scientists with means to determine different types of cells in a sample, counting the cell or if proteins are expressed or to identify cell anatomy.

One deep learning model, the convolutional neural network (CNN), was developed for image process-
ing and has been applied to microscopy image analysis\textsuperscript{13,14}. Other methods of deep learning have been used in different contexts for image analysis: histopathology images can be analysed to determine phenotype associated with gene expression\textsuperscript{15}. Such work would have previously required manual inspection of the slides. Finally, morphological classifiers of cell phenotypes upon small molecule exposure have recently been analysed with CNNs\textsuperscript{16}.

Another successful area of application for deep learning is genomics. Large amounts of data in the field allow leverage of machine learning models to identify biomarkers associated with phenotype\textsuperscript{17} or level of gene expression\textsuperscript{18}. The amount of data generated from whole genome sequencing is massive and collaborations between academic and tech firms to capitalise on their analysis have increased rapidly, for example in that between Google and the Broad Institute to implement the Broad’s open-source genomics toolkit on Google’s cloud servers\textsuperscript{19}.

The same methods can be used to search for therapeutics with the potential to silence genes or inhibit specific gene activity. For example, artificial neural networks (ANN) have been used to predict siRNA activity on predefined target sequences\textsuperscript{20}.

Another interesting deep learning application covers the design of new peptides with high bioactivity. Combinatorial peptide libraries challenge creators through the sheer numbers of generated compounds. It is then crucial to select the peptides with the highest activity for synthesis. Machine learning provided a method to predict that activity and sort the selected peptides\textsuperscript{21}. Other machine learning methods provided a way to predict protein binding site for small molecules\textsuperscript{22} or prediction of protein expression and solubility of a large dataset to optimise high-throughput experiments\textsuperscript{23}.

These projects would not have been possible without the use of both automation and machine learning. Specific models and methods from machine learning are keys to answer specific scientific questions. The diversity of methods is key to the diversity of biological questions. It is why the industry has been turning to technology companies for models such as GAN (generative adversarial networks)\textsuperscript{24}, for customised hardware\textsuperscript{25} or for
collaboration with experts in machine learning algorithms coupled with large-scale cloud capabilities\textsuperscript{26}. Entire companies’ business models are based on machine learning methods for drug discovery: the unique combination of cloud technology, automation and machine learning provides opportunities in life science at a scale we could not manage before (vide supra).

\textbf{Instrumentation – image analysis, IoT, virtual help}

Prognosis: imaging and voice interaction prototypes. Wider adoption pending

The concept of a ‘Robotic Scientist’ is not new and was first conceptualised in 2004\textsuperscript{27} as a combination of computational methods, automated instruments linked to complex laboratory robotic systems and the need for AI and machine leaning to test and iterate on a hypothesis in real time.

Cellular imaging provides an ideal opportunity to showcase the power of AI. High content screening (HCS) or cellular imaging and analysis is today widely used across many parts of the drug discovery process, driven by the need to gain greater understanding of the phenotypic nature of interactions between potential new drug candidates and the cells found in the human body. The ability to predict and visualise potential unwanted cellular interactions increases the potential of success when a candidate enters the clinical phase of testing, however the vast quantity of information provided by even the most basic of HCS platforms opens itself to the need for more intelligent forms of data analysis. AI techniques provide the ability to provide enhanced segmentation and even to identify specific organelle structures in the absence of segmentation, precluding the need for cell labelling\textsuperscript{3}. In addition, so called image-based cell profiling strategies\textsuperscript{28} provide a path to high-throughput quantitation of phenotypic differences in cell populations through the collection and analysis of hundreds of morphological changes caused by prior treatment with a chemical or biological agent. This analysis approach provides a path to identification of novel targets or mechanisms of actions for prospective new drugs.

Given the number of sensors and facile communication among various lab instrumentation, a growing movement seeks to capitalise in two distinct but complementary ways: building ‘internet of things’ networks to alert scientists\textsuperscript{29}, or two-way communication using robotic lab assistants powered by natural language processing and virtual models\textsuperscript{30}.

In summary, automation itself has provided the high-quality, high volume data required to train neural networks and recognise patterns. As new tasks are automated for both chemistry and biological processes, the instruments also gain sensor data: location, position, temperature, torque. Soon, it will not be even that a gene is properly functioning or a reaction works, but more subtle ways of discerning how the robot or instrumentation runs the assay that will be the rate-limiting step to scientific innovation.

\textbf{Authors’ note}

This piece was commissioned in part to celebrate the inaugural ‘AI in Process Automation Symposium’ which will be held in Boston, MA, USA in October 2019. The Society for Laboratory Automation and Screening (SLAS), of which all authors are volunteer leaders or staff, appreciates \textit{Drug Discovery World}'s kind invitation and facilitation to produce this perspective piece.  

\textit{Drug Discovery World}’s kind invitation and facilitation to produce this perspective piece.  

---

Dr Michael Tarselli trained as a synthetic organic chemist and worked in CROs, start-ups and multiple pharma companies as a chemist, before leading a team in chemical information systems development at Novartis. Since 2018 Mike has served as the Scientific Director for SLAS.

Dr Yohann Potier trained as a computational chemist and spent time in business analysis and informatics before landing at Novartis. He leads a team creating new solutions to support bench scientists across the biological and bioinformatics community. Yohann is also involved in local consortia such as Pistoia Alliance.

Dr Alan Fletcher trained as a pharmacologist, then spent a decade running HTS and robotics for MSD in the United Kingdom. On the business side, he has served as a Director and VP GM for Life Sciences at two technology firms. He is the 2019 President of the SLAS Board of Directors.
References


