Modern Methods

Increased implementation of deep learning is forecast to introduce new opportunities to the drug discovery arena and add value to data held by pharmaceutical companies

By Dr Tom Whitehead at Intellegens

Early-stage drug discovery has been an enthusiastic adopter of computationally aided design in recent years, with computational visualisation and predictions becoming integral to the way chemists work. This forward-looking approach to applying technology in existing workflows makes drug discovery an important field for the ongoing deep learning revolution, where new and innovative tools can make an important difference to real, relevant projects.

A key application of computational measures in drug discovery has been in the development and utilisation of quantitative structure-activity relationship (QSAR) models. These models take features of a compound, known as descriptors, and match them to the compound's activity in an assay of interest. Typically, the compound descriptors capture both whole-molecule properties (such as the molecular weight, topological polar surface area, and McGowan volume [1-2]), as well as sub-structural fragments. The activity in the chosen assay is then expressed as a mathematical function of the descriptors. Many forms for the function have been tried over the years, from simple linear regression fits to machine learning methods, such as support vector machines and random forests (3-4). The predictions have added great value to the drug discovery process, serving to give quantitative confirmation to the intuition and designs of chemists.

Deep Learning

In recent years, one of the most important trends in machine learning has been the development of 'deep learning', where multiple layers of data abstraction are composed together to form very complex and powerful functions of the input data (5). In image recognition and time-series processing tasks, deep artificial neural networks now provide state of the art solutions in the form of convolutional neural networks (CNNs) and recurrent neural networks respectively. More recently, deep artificial neural networks have also been used to construct QSAR models, but this has provided mixed results. At a recent conference in Switzerland, Robert Sheridan from Merck reported that deep-learning QSAR models offered a negligible improvement over traditional approaches across 30 representative QSAR datasets (6). This serves to highlight that deep learning is not a panacea and adds most value when applied to problems where conventional techniques are unable to work effectively at all.

Some of the challenges with applying deep learning to drug discovery are features specific to the pharmaceutical domain. In deep learning, data is king, but, in drug discovery, experimental measurements are often difficult and expensive to obtain, resulting in limited data on the most interesting assays, which complicates the training of accurate deep learning models. While generic image recognition CNNs are frequently trained on hundreds of millions of labelled images, even large pharma companies typically only have a few million compounds in their corporate collections, most measured against a handful of assays. Complicating matters further is that these measured assays are different for different compounds; no compound has been measured in every assay ever devised, and no assay has been run for every compound. The public ChEMBL database only has activity measurements for around 0.07% of the possible compound/assay pairs it contains, and pharma company data is frequently similarly sparse. This sparsity of data makes it difficult for deep learning to capture the relationships between different assays, a problem which is only just beginning to be overcome by modern approaches.

A further challenge with applying deep learning to drug discovery is the variability of the data that is available. Biological data is inherently noisy and uncertain, with three-fold variability between results from the same compound in the same assay not uncommon (7). This makes it impossible for deep learning models to come up with definitive predictions for assay results, which leads to it being vitally important for the uncertainties in predictions to be well captured. However, this, in turn, can lead to complications in analysis and interpretation of results by chemists.

Deep Learning in Practice

Despite these challenges, the rise of deep learning provides a host of opportunities for expanding the toolkit of drug discovery. The first opportunity for deep learning to prove its worth is through the application of multi-target modelling, by constructing a single deep learning model that can simultaneously predict the results of multiple assays. Multi-target

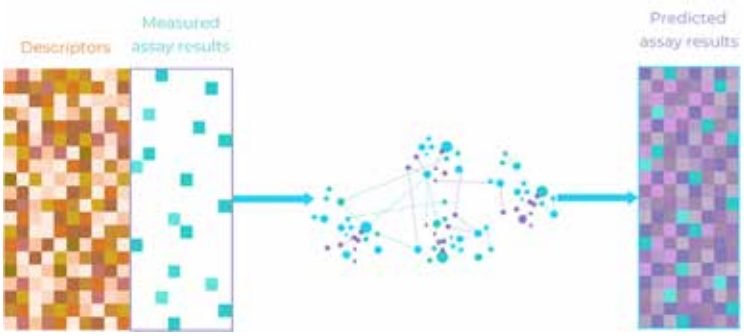

Figure 1: Multi-target modelling can build a single deep learning model for multiple assays simultaneously, learning a much deeper understanding of the relationship between chemical descriptors and assay results and automatically generating selectivity profiles

modelling teaches the algorithm a more profound representation of the chemical properties and their relationship to assay results, enabling the transfer of learnt correlations between assays. These multi-target models extend the concept of a QSAR model and offer the immediate advantage of automatically generating selectivity profiles, rather than just activity levels.

Further applications of deep learning will add significant value to the data that pharma companies already hold. Multi-target modelling allows chemists to accurately and confidently 'fill in the gaps' in the sparse databases of compound/assay data, imputing values for what would be measured in each assay for each compound, were the experiment to be carried out. These predictions can be used directly to inform the selection of hits for further analysis, and, as the model predictions are validated in experiment, the data can be fed back into the algorithm to create improved models of the area of chemical space most interesting to the chemist. An entirely automated procedure is also possible where the deep learning algorithm proposes the experiment that will most improve its estimates for a target of interest, iteratively converging to increasingly accurate predictions. One of the most interesting applications of this automated data prediction capability is in the hunt for false negative assay results. High-throughput screening frequently, but incorrectly, identifies active compounds as inactive, and the ability to concretely identify this chemical 'dark matter' would open up new opportunities and understanding (8).

The next step in the cycle of automation is for the deep learning algorithm to be able to propose entirely new compounds for investigation, rather than simply making predictions for existing compounds. In other fields, generative adversarial networks (GANs) have had reasonable success in generating ideas that pass for human-generated, including recently creating artwork that sold for over US $400,000 at Christie's (9). GANs work by setting up two deep learning models, one of which creates suggestions for new ideas, be they artworks or chemical compounds, while the other model then tries to distinguish from real, pre-existing data. GANs are still in their infancy in drug discovery, but they offer the promise of automated design and optimisation of compounds in early-stage projects.

Discovery Developments

Despite these leaps forward in the abilities of deep learning algorithms to generate and test chemical compound proposals, deep learning methods are unlikely to entirely supplant living, breathing chemists. Although machine learning enables very rigorous and detailed analyses of immediate, concrete problems, no machine learning approach developed so far can match the human ability to take a strategic overview of a research project, understanding and directing multiple different strands in pursuit of separate, overlapping objectives simultaneously. This has resulted in the concept of a 'centaur': cooperation between humans and machine learning algorithms, with the human providing high-level direction to advanced deep learning methods. In chess, which was long a leading environment for the development of machine learning methods, a concept has been developed by Garry Kasparov called 'advanced chess', where human chess players are advised by advanced chess programs, with the combination of player and algorithm able to outperform the leading chess software alone (10). A similar development is likely to play out in drug discovery, where chemists

"Further applications of deep learning will add significant value to the data that pharma companies already hold"

supported by powerful deep learning approaches will be more successful than either the algorithms or the chemists alone.

The field of deep learning has much to offer early-stage drug discovery, through the development of more accurate models of chemical activity and selectivity, the triaging and cleaning of existing data, and even the suggestion of new experiments and compounds. Challenges are still present, particularly in the quality and volume of data available for training modern deep learning methods, and, in these areas, drug discovery also has much it will be able to give back to the development of deep learning algorithms, through increased resilience to noise and sparsity in training data. The journey is only just beginning for deep learning in drug discovery, but the future looks set to be productive and engaging for all involved.

References

1. Ertl P *et al,* Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport, *J Med Chem* 43(20): pp3,714-7, 2000
2. Abraham M and McGowan J, The use of characteristic volumes to measure cavity terms in reversed-phase liquid-chromatography, *Chromatographia* 23(4): pp243-6, 1987
3. Doucet J *et al*, Nonlinear SVM approaches to QSPR/QSAR studies and drug design, *Current Computer-Aided Drug Design* 3(4): pp263-89, 2007
4. Gao C *et al,* Selectivity data: assessment, predictions, concordance, and implications, *J Med Chem* 56(17): pp6,991-7,002, 2013
5. LeCun *et al*, Deep learning, *Nature* 521: pp436-44, 2015
6. Sheridan R, What I learned about machine learning – Revisited (again), presented at 'Artificial Intelligence in Chemical Research', 2018
7. Danielson ML *et al,* In Silico and In Vitro Assessment of OATP1B1 Inhibition in Drug Discovery, *Mol Pharm* 15(8): pp3,060-8, 2018
8. Macarron R, How dark is HTS dark matter? *Nature Chemical Biology* 11(12): pp904-5, 2015
9. Visit: www.nytimes.com/2018/10/25/arts/design/ai-art-sold-christies.html
10. Kasparov G, *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*, 2017

Dr Tom Whitehead is a machine learning scientist at Intellegens, a deep learning startup company based in Cambridge, UK. Intellegens focusses on handling sparse, noisy, experimental data, and Tom is leading the application of Intellegens' unique tools to drug discovery. Tom has a PhD in theoretical physics from the University of Cambridge, UK, and now focusses on the development and utilisation of deep learning methods for difficult, high-value data problems.
Email: tom@intellegens.ai