Presenting a New Paradigm for Drug Discovery: Combining Computational Biophysics and AI through MatchMaker

Jan. 15, 2019   by Naheed Kurji

In this Special Perspective, our fourth in an ongoing series, we will be presenting MatchMaker™, a novel deep proteome screening technology that we have developed and validated over the past 2 years to identify DTIs. MatchMaker builds on Cyclica's passions of combining protein, chemistry, and genomic data, and augmenting it with high performance computing and algorithm development supported on the cloud.

The Pharmaceutical Industry is Changing.

Fast, accurate, and generalizable drug-target interaction (DTI) predictions have the potential to transform pharmaceutical R&D. In this Special Perspective, our fourth in an ongoing series, we will be presenting MatchMaker™, a novel deep proteome screening technology that we have developed and validated over the past 2 years to identify DTIs. MatchMaker builds on Cyclica's passions of combining protein, chemistry, and genomic data, and augmenting it with high performance computing and algorithm development supported on the cloud.

MatchMaker combines molecular biophysics and deep learning to predict binding of new drug molecules to all proteins, i.e. the "cell", with high accuracy and high throughput, moving beyond the reliance on molecular docking. MatchMaker will be the engine that powers our Ligand Express proteome screening platform, and our yet to be released Differential Drug Design (DDD) technology for lead optimization as well as single and multi-target target drug design. MatchMaker will increase the speed, predictive power, and generalizability of these technologies, enabling an integrated in silico first workflow and turning our attention to designing drugs for patients, not just one protein. Prior to releasing this Special Perspective and the accompanying validation notes, we shared the story with a number of partners and investors during the annual JP Morgan Healthcare Conference between January, 7-9, 2019. The response was energizing. We are therefore thrilled to share with you MatchMaker, but before diving into details, let's take a quick step back.

Cyclica combines knowledge about proteins, drug design, '-omics' big data, and artificial intelligence to bring a revolutionary platform to the market.

A Review of the First Three Special Perspectives

Over the past 1.5 years, we have released three Special Perspectives with the goal of exciting the world about what we are doing at Cyclica:

In the first Perspective, entitled **A Look at Yesterday, Today, and Tomorrow**, we provided a historical overview of computational drug discovery, including a review of Computer Assisted Drug Discovery (CADD), High Throughput Screening (HTS), Virtual Screening, and Artificial Intelligence (AI) applications. We also introduced Ligand Express®, Cyclica's first-in-class *in silico* proteome screening technology.

In the second note, entitled **In Silico Polypharmacology**, we explained why designing drugs for only one protein target, often driven by HTS, virtual screening or AI-enhanced virtual screening, is limited. We pointed out that those methods do not consider the upwards of 300 off-target protein interactions of a small molecule, many of which are not known up front. Even the most carefully designed molecules, with promising inhibition constants for a given target, often fail due to unanticipated off-target activity. We

presented Torcetrapib as a costly example of a promising drug that failed in late stage clinical trials. We then introduced the notion of *in silico* polypharmacology - that using computational methods to understand how a drug will interact with *all* proteins, scientists will be able to form more robust hypotheses, make more informed decisions, and thereby reduce the risk associated with drug discovery and development.

In the third note, entitled **Artificial Intelligence for Drug Discovery**, we discussed the growing use of AI within the pharmaceutical industry, with emphasis on the area of drug discovery. We also presented important limitations of AI, and provided our perspective on how best to drive value in drug discovery by leveraging AI. There are two primary challenges (opportunities) that we are focused on regarding the use of AI in pharmaceutical R&D: i) interpretability, and ii) generalization. As professor Tommi Jaakkola from MIT stated in the MIT Artificial Intelligence: Implications for Business Strategy course: "deep learning architectures…remain quite opaque about how they actually solve the problem. So, there is a question, then, how we can learn to interpret and trust the predictions of these methods". He then goes on to conclude: "the key problem here [with AI/machine learning/ deep learning], fundamentally, is a problem of generalization - how do I learn from a training set so that I can generalize and do well on the task that I have not yet seen".

Generalization has been a focus of recent applications of deep learning for conventional virtual screening. For example, convolutional neural networks (CNN) trained on atomically detailed 3D structures of protein/ligand complexes have been proposed by Wallach et al. in 2015 in a non-peer reviewed paper, and Ragoza et al. in 2017 in the ACS Journal of Chemical Information and Modeling. Ragoza et al. describe an open source implementation, but also raise substantial concerns about the approach. Ragoza et al. found that such CNN-enhanced virtual screening trained on publicly available data does not generalize well and therefore struggles to deliver value outside of the training set because the available data is too limited. Opening up a proverbial black box is crucial to convincing scientists, who are sceptical by nature, about the integrity and fidelity of the approach. It also offers some reassurance of true predictive power, as deep learning approaches are particularly prone to various forms of overfitting and over-estimation of predictive power due to limitations in testing protocols and data nuances like compound series bias.

The world of AI for pharmaceutical R&D has recently become very exciting, and is growing fast. According to Forbes, in articles from November and December 2018, there are now over 125 companies applying "AI" for Drug Discovery, with Cyclica listed amongst the top 20. While this is encouraging, we have set our sights on achieving something much bigger than that.

Polypharmacology Through In Silico Proteome Screening

Real AI transformation to the drug-discovery process will not come from simple point solutions that merely substitute for an existing technology (like ultra-HTS), but from higher order advances that modify and redefine the approach to drug discovery. At Cyclica, we believe this will arise from a holistic understanding of drug effects on the human body, i.e. all proteins in the cell. The efficacy, polypharmacology, toxicity, and side effects of drugs are mediated through interactions with tens to hundreds of proteins found within the human proteome. The desire to profile the physiological effect of a drug on cellular systems has inspired numerous *in silico* and experimental approaches, each with their own strengths and weakness. Experimental approaches to investigate polypharmacology, such as Thermal Proteome Profiling (TPP), CETSA or other Mass Spectrometry based approaches, are restricted to proteins that happen to be expressed in the cell type used, and thus are not able to scale up to the entire proteome or uncover previously unknown targets.

Taking a computational approach as an alternative to reveal a drug's polypharmacology has been widely recognized as important, but no commercially viable solution has emerged. In the early days at Cyclica we developed, validated, and commercialized a novel *in silico* proteome screening technology that formed the core of our Ligand Express platform. With proteome screening, which is based on the use of pocket profiling, surface matching, and classical docking, Ligand Express screens small molecules against repositories of structurally-characterized proteins to identify their significant targets. Ligand Express then leverages AI to determine the biological relevance of those targets, and applies systems biology to link them to particular biological processes or disease states. By leveraging large scale, proteome-wide structure data, Ligand Express provides a panoramic view of all activities of a small molecule, providing scientists with valuable insights into the polypharmacology of a given drug. The platform is used to investigate the mechanism of drug action, including the elucidation of both desired and undesired drug effects. Ligand Expressis is also used to identify potentially deleterious side effects earlier in the drug development process. Our technology has been widely tested and accepted, and is actively being used by many pharmaceutical companies. Over the past several months we have announced deals with Bayer Pharma AG, Merck KGaA, Eurofarma, WuXi Apptec, and the National Research Council (NRC) of Canada, and many others.

While Ligand Express has driven substantial value for our pharmaceutical partners, we realized that it had two important limitations that we saw as opportunities for future innovation: i) throughput was limited; only one molecule could be submitted for proteome screening at a time, only a handful of molecules could be processed simultaneously, and it took several days for a run to be finished; and ii) proteome screening was reliant on classical molecular docking to score and rank protein hits that emerged from surface matching, which we recognized are limited by a number of shortcomings. Despite those limitations, the combination of surface matching and docking performed well, as can be seen in our early validation notes - the data we presented in those early notes was encouraging, but not without room for improvement. Stated differently, our initial approach was *good* but not *great*. Over the past two years, we have set out to address these limitations, and we believe we have made a substantial leap in computational chemistry and drug discovery. This has been achieved with MatchMaker.

Ligand Express evaluates *all* proteins within a cell, paving the way for a wholistic drug design. Now that Ligand Express is powered with MatchMaker, it is possible to start designing for patients, not for a single protein at a time.

Introducing MatchMaker: A Leap Forward in Pharma R&D

To progress beyond the limitations of molecular docking, we have developed, validated, and patented MatchMaker, a novel deep learning approach that leverages the millions of known Drug/Target Interactions (DTI) found in public databases while retaining the generalizability advantages of structure-based approaches. MatchMaker has been incorporated in Ligand Express to achieve proteome screening on a larger scale than previously possible, and in our yet-to-be released DDD technology to accelerate the *de novo* design of smart NCEs. DDD will be formally announced in early 2019. By synthetically augmenting DTI datasets with structure-based information using an appropriate abstraction of biophysical interactions, we can integrate high-volume and high-dimension data with neural networks. This creates generalizable models that learn drug-target compatibility. MatchMaker achieves a high degree of accuracy, outperforming experimental screening approaches, demonstrated in the face of stringent testing criteria mimicking real-world applications in pharmaceutical R&D.

Other computational DTI predictions focus on single target virtual screening, and include ligand-based approaches and structure-based approaches. Unlike target-centric virtual screening technologies that screen millions of molecules against one protein target, MatchMaker is the first technology that permits the rapid screening of millions of molecules against the entire human proteome in an ***all-by-all*** analysis. Consistent with our philosophy that AI is best when it complements computational biophysics, MatchMaker is based on structure-based and knowledge-based technology, combining the best of both.

When we first observed the high accuracy our deep learning models achieved, we were skeptical and spent a long time searching for flaws in our benchmarking. Once we convinced ourselves, we knew that our claims would understandably be met with skepticism externally, so we undertook an extensive and rigorous set of benchmarks, and worked with some of our partners to perform blind-evaluations. The results have been very exciting, far exceeding our own expectations. Our validation note, which can be found here, shows that MatchMaker clearly outperformed Secure DTI, a comparable approach recently published by Hie et al. (Science 2018), despite the use of more stringent testing conditions. In short, we were able to do more, even when stressing our system intensely. Our application note, which can be found here shows that MatchMaker offers accuracy comparable to, and possibly better than, costly experimental methods such as thermal proteome profiling (TPP).

We began sharing matchmaker data with a number of pharmaceutical collaborators in the form of re-ranked protein interaction lists. Their feedback was that the results were consistently superior to previous docking based rankings and, where applicable, comparable and confirmative of experimental results. Over the next few months we will continue to collaborate with our partners to perform more scientific validation to be reported along with more details on the methodology in an upcoming peer-reviewed publication.

Looking Forward: Driving Drug Discovery by Doing More with AI

Now, more than ever before, we are seeing demand from pharmaceutical companies for novel technological insights. They are showing a willingness to adopt *validated* technology platforms to improve the drug discovery process, to enhance the value chain, and introduce new opportunities for scientific exploration and product development. Specifically, we are seeing a demand for technologies that can identify and validate novel targets, design higher quality drugs more efficiently, and personalize the drug discovery and development process. We are also seeing the emergence of many companies who *claim* that their AI will swoop in and solve these very complex problems. We believe that AI (including machine learning and deep learning) will undeniably play a role in the future of drug discovery, but that it is not a "silver bullet". We also believe that the pharmaceutical industry is looking for more than just another technology that does one thing. Based on our market analysis, and conversations with existing and prospective partners, we believe that the pharmaceutical industry is looking for a holistic, integrated, and end-to-end enabling set of technologies that leverage AI to drive value at various stages of drug discovery.

Cyclica is on a trajectory that will take us to our moonshot and beyond.

Looking forward: we don't want to solve just one particular problem really well, using AI or any other particular technology. That doesn't really excite us. With MatchMaker integrated within Ligand Express to enhance proteome screening, and within DDD for the generation of NCEs, we believe that Cyclica will catalyze a fundamental paradigm shift in the industry by offering a holistic, end-to-end enabling suite of technologies focused on whole biological systems. In the near term, we will offer the pharmaceutical industry an integrated *in silico* workflow that combines computational biophysics and AI, built on an appreciation of the importance of polypharmacology. This cloud-based, and AI-augmented set of technologies will transform how scientists design, screen, and personalize medicines, by significantly enhancing productivity by reducing costly, time consuming, high risk trial-and-error based strategies. Our near term vision is to reduce the drug discovery lead times by 7 years to 2 years, while also enhancing the quality of the lead molecules. By providing a highly validated, more integrated workflow, we will be poised to realize our ultimate vision of designing drugs for **patients**, and not for **proteins**. We appreciate that this is a lofty goal to achieve using computational methods, but this is what drives us every day.

- Cyclica