Confluence of Technologies Can Bring “Virtual Pharmacology” to the Next Level

Feb. 20, 2019 by Andrii Buvailo

In 1970-80s, the idea of virtual screening was regarded as a conceptual way to substitute costly and time-consuming experimental “screen-everything-you-have” approaches with a much faster and cheaper predictive modelling to cherry-pick only the best molecules for subsequent synthesis and validation in a lab. A great number of computational tools and approaches emerged, aiming at “pre-screening” new promising molecules, so called “hits”, or augmenting experimental screening programs to optimize efforts.

One of the most popular and powerful methods of this kind is molecular docking, which has been widely used ever since the early 1980s. The idea of molecular docking is essentially to identify a perfect “key” among many diverse options, fitting to a “lock”, a biological target of choice -- using computational 3D representations of the interacting moieties.

With all the early promise, molecular docking stumbled upon several fundamental issues, peculiar for the early decades of “digital revolution”: a lack of sufficient computing power, imperfect predictive algorithms, a limited choice of available molecules for virtual screening, and a small number of structurally characterized biological targets of sufficient pharmaceutical significance.

Last but not least, the virtual screening approach contained substantial “synthesizability risk” of not being able to physically get to all the hits, picked by a computer, in a cost-efficient and timely manner. As Asher Mullard, a freelance journalist at Nature, described it: “Many computational approaches also have an annoying habit of suggesting candidates that are nightmares to cook up in a lab.”

A lucky confluence of technologies opens doors in universe of synthesizable “hits”

Over the past several decades enormous progress has been achieved in many technologies, essential for a truly successful virtual screening effort. This finally allowed to open doors in large-scale "virtual pharmacology" of the 21st century -- with the emergence of the largest docking-friendly database of synthetically accessible compounds, outlined in a recent influential article in Nature: “Ultra-large library

docking for discovering new chemotypes".

The new virtual pharmacology platform, developed by scientists from the University of California San Francisco (UCSF) in collaboration with their colleagues from (UNC), currently contains hundreds of millions of 3D representations of novel synthetically accessible drug-like molecules, immediately suitable for large-scale docking studies, promising to change the landscape of modern hit discovery.

To illustrate the immediate practical benefit of the new "virtual pharmacology" platform, the scientists docked 99 million compounds against the AmpC, a bacterial enzyme, beta-lactamase, which is involved in antibiotic resistance, and another 138 million compounds -- against another unrelated target, the D4 dopamine receptor of brains cells, known to be involved in psychosis and addictive behavior.

The results were pretty amazing, with at least one-third of all predicted hits appearing active in subsequent laboratory validations. Scientists found a number of highly potent hits to both targets, including one of the most potent non-covalent inhibitors ever reported for AmpC, and highly active and selective partial agonists and antagonists for D4. Further extrapolation suggested presence of around half a million other potentially active hits against those targets.

As Daniel Erlanson of Practical Fragments noted about this achievement: "...if the method proves generalizable, the question a decade hence may not be how to find hits, but rather how to choose between hundreds of thousands of them".

The emergence of this kind of platform, unimaginable in early days of computational drug discovery, has become possible due to fascinating progress in the following areas:

Advances in computing power

For the past several decades there has been a steady growth in computing power, with the costs of computing dropping exponentially after 2000s.

The team at UCSF managed to create 3D representations of hundreds of millions molecules and conduct docking studies with them using commodity Intel/AMD CPU hardware -- a relatively small machine with 1500 cores, compared to the current leadership standards. It is easy to imagine a rate of progress in the nearest future, should the team upgrade to using modern GPUs or even grid infrastructures for massively parallel computing… we might see orders of magnitude larger resource within a couple of years.

Advances in docking models and software

A lot of progress has been made since 1970s in how scientists model drug-target interactions. The team at UCSF used sophisticated DOCK software, a powerful and modern tool for virtual screening to study millions of compounds with pretty amazing predictive accuracy.

Advances in creating ultra-large spaces of feasible molecules

Importantly, the virtual pharmacology platform by UCSF, is largely based on another innovative chemical technology -- REAL database (REAL DB), developed by Enamine chemists over a decade of R&D work in synthesis of small molecules. Any selection of compounds from REAL database can be quickly synthesized with high success rate (above 80%, according to Enamine), and available via flat-rate pricing model. This eliminates major synthesizability risks, when using virtual pharmacology platform for large-scale hit discovery programs.

(Read also: "Navigating In REAL Chemical Space To Find Novel Medicines (Now 3.8 Billion Molecules)")

What about available biological targets?

Clearly, ultra-large library docking could be a game-changer for targets that have been structurally characterized. But here is another bottleneck…

As was nicely summarized by Christopher VanLang in his Quora post, there are 19,613 proteins identified to be encoded by the human genome, 74% of which have no known linkage with a disease. Among the rest 5,068 proteins, having a known link to a disease, 3131 have been deemed undruggable (either they have no clear pocket for small molecules, or they are not available outside cells). Thus, it yields 1,937 druggable targets, roughly 10% of the entire protein landscape, of which 672 targets already have approved drugs -- leaving 1,265 potentially interesting targets to focus on.

On top of all struggles associated with revealing promising targets, obtaining a well-characterized molecular structure suitable for virtual screening is another challenge in its own right.

Advances in cryogenic electron microscopy (cryo-EM)

One way to push the limits of structural characterization of molecular targets arises from a technology, called cryogenic electron microscopy, or cryo-EM. While it was invented in 1968, only recently, with the advent of other technologies, it regained a wider traction, accurately modelling 3D protein structures to a

higher resolution than most x-ray crystallography methods can achieve. It allows substantially accelerating structural characterization of promising targets, thereby expanding the scope of possibilities of virtual screening. It give hope that the horizon of virtual pharmacology will be substantially expanded within the nearest several years.

Artificial intelligence is the next step

Applying artificial intelligence (AI) is a likely next step in building on top of the UCSF "virtual pharmacology" platform, powered by REAL technology, because it appears to be the largest publicly available dataset of unprecedented quality in terms of synthesizability, a fruitful opportunity for AI-developers in this space.

Artificial intelligence (AI), or to be more precise, one of its sub-disciplines -- deep learning (DL), has re-gained enormous attention lately in virtually every industry dealing with lots of data: finance, e-commerce, security, communication, surveillance, etc. This attention is fueled by an undeniable fact that AI-driven tools already delivered striking practical results in real life, with lots of commercially available applications emerging every month.

(Read also: "The “Why”, “How” and “When” of AI in Pharmaceutical Innovation")

While the first self-learning program emerged in distant 1950s, it took more than half a century, with many ups and downs, for the idea to become a transformative global technological movement. Its skyrocketing progress after 2000s was conditioned by many factors, including the emergence of high performing GPUs in late 90s; many important breakthroughs in machine learning algorithms and model architectures (e.g. back propagation, reinforcement learning, LSTM, deep neural networks, etc); the rise of cloud-based infrastructures and distributed computing frameworks; and the increasing availability of big data of high quality, on which machine learning models would be trained efficiently enough (e.g. ImageNet database catalyzed breakthroughs in image processing algorithms).

Pharmaceutical industry is a relative “late-adopter” in this context, due to its natural conservatism, and also a much more complex nature of scientific datasets, requiring intricate domain-specific feature engineering. However, certain progress is happening in this area as well, and we might expect pretty transformative results to emerge in the coming years.

(Read also: "How Big Pharma Adopts AI To Boost Drug Discovery")

AI-driven algorithms could efficiently solve such challenges of “virtual pharmacology” as increasing accuracy of docking experiments, discriminating binders and non-binders from docking results, or expanding the synthetically available chemical space via “REAL-trained” de-novo design. Multiple other challenges might be solved via AI-guided automation.

On the other hand, the emergence of a database like the one UCSF managed to aggregate, might have catalyzing effect for the progress in chemical and drug discovery AI tools, the “ImageNet moment for AI in medicinal chemistry”. Another powerful confluence of great technologies.