The role of artificial intelligence in drug discovery

Personal Pursuit – David Brouwer, s2193698

Over the last few decades, artificial intelligence (A.I.) has grown much in popularity. With the resources becoming more readily available it has been easier than every to develop A.I. models, and continue research in the field. Processors have been becoming faster, and cheaper with the years. This has resulted in a growing interest in A.I. in a variety of different disciplines, of which the health care sector is one. Within healthcare, there are numerous applications for which A.I. can be used, as is displayed in figure 1. An example of a popular application is the diagnosis of diseases, through the interpretation of medical imaging. Yet, another rather interesting domain, and one which I studied during my personal pursuit (P.P.), is the aid of A.I. in drug discovery.

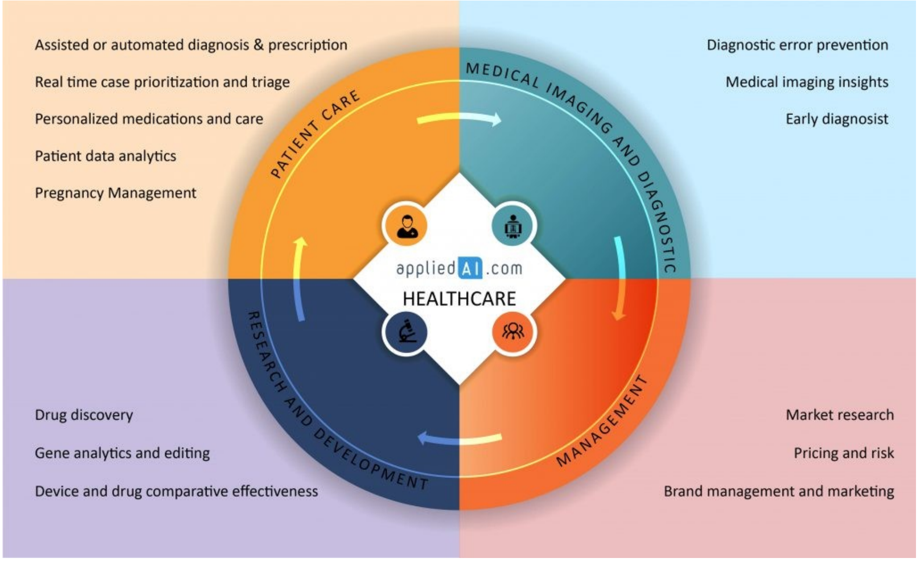
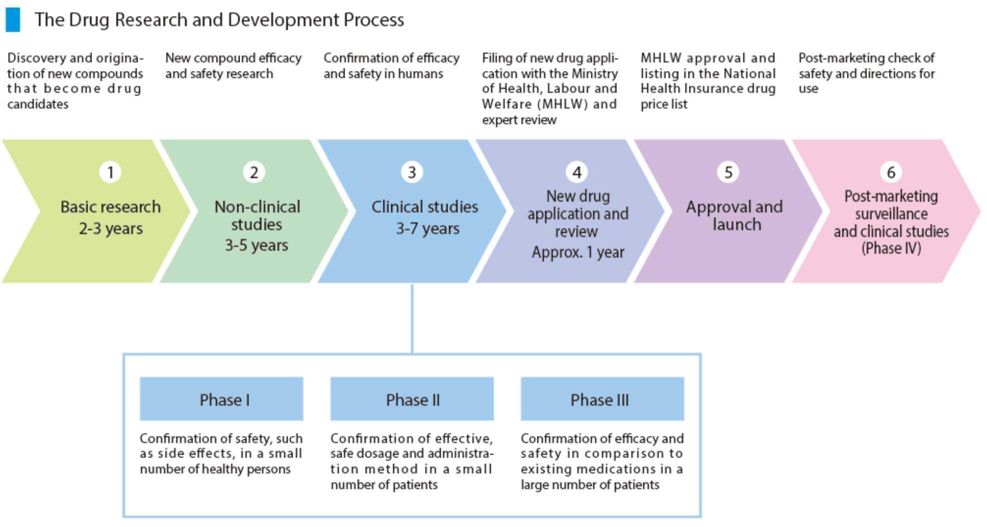

Assisted or automated diagnosis & prescription
Real time case prioritization and triage
Personalized medications and care
Patient data analytics
Pregnancy Management
PATIENT CARE
MEDICAL IMAGING AND DIAGNOSTIC
Diagnostic error prevention
Medical imaging insights
Early diagnosist
appliedAI.com
HEALTHCARE
RESEARCH AND DEVELOPMENT
MANAGEMENT
Drug discovery
Gene analytics and editing
Device and drug comparative effectiveness
Market research
Pricing and risk
Brand management and marketing

Figure 1, source: https://appliedai.com

Developing a new drug is very resource consuming process, it is both expensive and time comsuming. More than a decade goes by from molecule discovery to market approval, as can be seen in figure 2. Within this process, A.I. can be used at multiple stages, such as during clinical studies to aid in the testing process of molecules.


The Drug Research and Development Process
Discovery and origination of new compounds that become drug candidates
New compound efficacy and safety research
Confirmation of efficacy and safety in humans
Filing of new drug application with the Ministry of Health, Labour and Welfare (MHLW) and expert review
MHLW approval and listing in the National Health Insurance drug price list
Post-marketing check of safety and directions for use
1
Basic research
2-3 years
2
Non-clinical studies
3-5 years
3
Clinical studies
3-7 years
4
New drug application and review
Approx. 1 year
5
Approval and launch
6
Post-marketing surveillance and clinical studies (Phase IV)
Phase I
Confirmation of safety, such as side effects, in a small number of healthy persons
Phase II
Confirmation of effective, safe dosage and administration method in a small number of patients
Phase III
Confirmation of efficacy and safety in comparison to existing medications in a large number of patients

Reference: "JPMA Guide," Japan Pharmaceutical Manufacturers Association

Figure 2: Drug pipeline

However, my focus was at the very beginning of this process, on drug molecule discovery. With the billions of possible compounds, effective molecules can be challenging to find, and because it is such a numbers game, a well trained algorithm could prove very effective in the field, cutting years off of the drug pipeline. There are various ways to frame this challenge with regards to machine learning, such as a classification or discriminative problem, where an algorithm is tasked to classify a certain molecules to predefined groups, for example based on drug targets. Yet another way to frame it would be as a generative problem, where an algorithm is trained to generate new potential drug molecules based on training conditions. For this purpose, a general adversarial network (GAN) may be well suited. The architecture of a GAN consists of two artificial neural networks (ANN). One of these networks serves as the generator, and the other as the discriminator. The generator strives to generate an output mimicking real data as closely as possible. For example a GAN trained on a data set of paintings would ideally generate a piece of art distinguishable from a human created one. The ANN strives too classify a given data point as either real of fake. In the ideal scenario, the discriminator is unable to accurately classify the output of the generator, giving a classification with a probability of .5, meaning, a wild guess.

As drug discovery with A.I. is a hot topic, various research papers have been published on the matter. The paper that has inspired me most is called *"Generating Focussed Molecule Libraries for Drug Discovery with Recurrent Neural Networks" (Segler, Kogej, Tyrchan, & Waller, 2017).* They have used a natural language processor, in the form of a recurrent neural network, to process SMILES formatted molecules. Their model finally reproduced 14% and 28% of the per-designed test molecules for two distinct drug targets.

A common challenge in drug discovery is the validation of the output molecules. As the goal is to find new molecules that are effective against certain drug targets, the only way to truly validate the generated molecules would be to test them in the lab. This is however a very expensive processes, and will only be done for the few most promising compositions. This challenge makes it hard to validate an A.I. model aimed at creating new drugs, yet various researchers deal with this problem in different ways. For example Segler et. al. (2017), compare their generated samples against a set of drug molecules that have been designed by chemists against a specific drug target. Benhenda (2017) quantifies the internal chemical diversity as a metric for such purposes. By creating a better way to validate drug molecules, the output of A.I. generated molecules would be substantially higher.

GANs are renowned for their great performance with images, such as creating art work and colorizing gray scale images. A common approach in drug discovery is to train a GAN based on molecules represented in the SMILES format, which is a common ascii format to represent molecules as such:

C[C@@H]1CN(C(=O)c2cc(Br)cn2C)CC[C@H]1[NH3+]

Figure 3: a molecule in SMILES format

Because GANs work so well with images, I decided to convert 25.000 SMILES molecules to images, and use that as training data for a GAN. The molecule in figure 3 then results in the image in figure 4. To achieve this, the API from http://hulab.rxnfinder.org/ has been utilized. With a small script *smiles_to_img.py*, 25.000 SMILES molecules have been converted, of which 23.382 have been converted successfully to images in about 40 minutes. To decrease the number of channels, and thus decrease the amount of data that needs to be processed, the images are converted to grayscale.

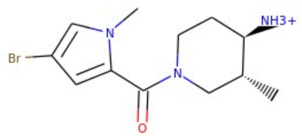


Figure 4: visual representation of the molecule in figure 3

To then create a model to train this data on, I consulted https://github.com/rasbt/deeplearning-models and used a convolutional GAN as template. This template has been trained on an image set of digits, of 28x28 pixels. As my images are 300x300px, the model has to be tweaked to allow the processing of such images. This however turned out to be quite challenging to get exactly right, which resulted in me getting stuck.

Discussion

Over the course of my P.P. I have gained knowledge in on various subjects, as fully explained in my separate P.P. reflection. Looking back on what I have learned about A.I. and my subject, I think that drug discovery is a very interesting topic to study. There is a lot of potential, and still a lot needs to be improved. The simple model I attempted to create does not do justice to the domain and the possibilities. In the best case scenario, if my model had worked as expected, and was able to create valid molecule structures, using it as input for novel drug discovery would be rather difficult. I do think GANs can be of great value in drug discovery, yet much needs to be improved on the validation side. As the generator is only as strong as the discriminator, without proper metrics to judge on, the discriminator will not be reliable. In the ideal scenario I would have created a discriminator that would know the laws of chemistry, and how molecules and atoms interact with each other, and would thus be able to simulate how the molecules would interact with the drug target. Such a GAN would be of great value to the drug discovery pipeline, where eventually most drug testing will be done digitally through simulation, cutting massively on the development costs.

During my time studying A.I. for my P.P. I have learned that it is relatively easy to reason about A.I. on an abstract level, thinking about architectures and the ways A.I. can be applied. I noticed that while watching interviews with the top researchers in the industry such as Ian Goodfellow, I will follow most of the discussions. However I have also discovered that implementing an actual complex algorithm is more challenging than expected. Setting up a simple linear regression model is far different from building and configuring a complete and functional GAN. But I think most importantly, my love for A.I. has grown even stronger over the course of my P.P., as I started to discover more and more of what is possible, and the current challenges. I always found myself excited as I learned new things about the field. Therefore, this P.P. has proven to be only the start of my journey into the rabbit hole of A.I, and I will most likely continue with this project, either during my next P.P., or in my spare time.

Benhenda, M. (2017). ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity?

Segler, M. H. S., Kogej, T., Tyrchan, C., & Waller, M. P. (2017). Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Central Science*, *4*(1), 120–131. doi: 10.1021/acscentsci.7b00512