Man versus machines

Who will spot the next blockbuster?

evolvalor

ACCELERATE BY INSIGHT

This issue of Evolvalor research dives into effects of artificial intelligence (AI), Machine Learning (ML) and Deep Learning (DL) in drug discovery. It includes a review of the literature, discussing effects of ML and DL; as well as, an overview of current big pharma collaborations around AI in drug discovery.

AI, ML and DL in drug discovery

Research and application of Artificial Intelligence (AI), machine learning (ML) and deep learning (DL) in the field of new drug discovery is continuously developing (1). In short, the drug discovery phases can be grouped into three segments: 1) target identification & validation; 2) hit identification & confirmation; 3) lead generation & optimization, as shown in Figure 1.

Figure 1. Overview of drug discovery process.

The process

Once an unmet medical need has been found, the first phase of drug discovery starts: identification (TI) and validation (TV) (2) with the aim to find and select a potential druggable target that will be validated for its function in a disease. After a druggable target is found, the next phase becomes apparent: hit identification (HI) and confirmation (HC). In this stage the biological activity of the therapeutic target is measured versus 'drug-like' compounds to asses its response mechanisms compared to the desired outcome (for example inhibition, enhancement or no effect on the target) (3). Lastly, once the desired effect is confirmed, the subject is further analyzed on their structure activity relationship (SAR) on various factors like activity, absorption, distribution, metabolism, and excretion (ADME) values, and toxicity (2, 4) before being identified as a pre-clinical drug candidate. This is called the lead generation (LG) and optimization (LO) phase.

Machine Learning vs Deep Learning

The several A.I. methods can be classified by underlying methodology resulting in 'machine learning' and 'deep learning', with the latter being the more complex method over the first. While both ML and DL are mainly based on statistical methods (support vector machine (SVM), random forest (RF), gradient boosting machine (GBM), and others with the ability to learn with or without being explicitly programmed (1, 5), DL can be seen as the next stage of AI technology as a subfield of AI. DL uses mainly artificial neural networks that adapt and learn from the vast amount of experimental data (1).

Machine- and deep learning in drug development stages

When comparing the AI methods opposed to the drug development stages of target identification and validation, hit identification and confirmation, and lead generation and optimization, different results can be observed; shown in Table 1.

***Table 1**. AI method compared to research results per drug development phase*

Phase	Main results	Pro's and con's
TI & TV	AI can be used to predict novel targets, DTIs & CPIs	+ AI may lead to a more cost-effective TI & TV process
	DL statistically outperforms ML	+ DL can process massive amounts of data and generates improved predictions + With increase in data availability and GPU computing, DL shows potential for usage in drug repositioning
	DL has several limitations in TI & TV	- Training DL models is less efficient than ML models - DL models are prone to overfitting - Designing DL models remains challenging
HI & HC	DL & ML outperform human ratings in HTS	+ DL shows possibility to automate HTS - HTS generally suffers from high levels of noise, lowering performance of DL
	Both DL & ML can reduce time of VS	- DL & ML show limited performance in ligand-based VS + In structure-based VS, DL outperforms the often used AutoDock Vina - DL lacks performance on intra-target ranking for pose prediction

LG & LO	AI can predict and analyze SARs with high accuracy	+ Reduces time and costs in LG & LO process
	DNNs outperform RF models (currently most used) for analyzing SARs	+ Accuracy of DNNs increase with an increase in training data + With complex biological data, DNNs show their advantageous performance over ML
	DNNs show some limitations in LG & LO	- DNNs are more difficult to use and construct than ML models - SAR datasets are large, diverse and may contain noisy or erroneous data, affecting DNN performance

DL: deep learning, DNN: deep neural network, GPU: graphical processing unit, HC: hit confirmation, HI: hit identification, HTS: high throughput screening, LG: lead generation, LO: lead optimization, ML: machine learning, SAR: structure activity relationship, TI: target identification, TV: target validation, VS: virtual screening

Target identification and validation

Even though machine learning methods are generally outperformed by deep learning methods in most studies, machine learning still shows its predictive performance over traditional target prediction methods (6). Six different studies focusing on this phase show that AI methods, in combination with transcriptome data of drugs and genes, can classify drugs for specific therapeutic targets (7) and can predict novel targets effectively (8). Also, AI methods can predict potential drug-target interactions (DTIs) (9, 10), which can be used for the prediction whether a new drug will interact with an existing target or known drug with a new therapeutic target (9). This feature can be utilized to design a more cost-effective drug repositioning process, since compounds and properties (e.g. activity, toxicity, ADME, etc.) of these repurposed drugs or targets are already known (9, 10). Next to DTIs, AI can be used to learn features of compound protein pairs to predict potential compound-protein interactions (CPIs) (6, 11). Predicting DTIs or CPIs both have the potential to predict and identify new drug targets and their interaction (6).

Hit identification and confirmation

Research indicates that AI is shown to have an impact on improving both computational approaches of high throughput screening (HTS) and virtual screening (VS). For HTS, AI is mostly used to classify images in different categories of hits (enhancing, inhibiting or no hit), and both deep- and machine learning outperform human ratings (12, 13). In addition, two separate studies (12, 12) show that deep learning methods (CNN and DNN) significantly outperform machine learning approaches (kNN, LR, NBC and SVM).

In terms of HTS, AI models tend to create high levels of noise. Koutsoukas et al. (2017) shows how their AI models dealt with the high-level noisy data. Their study demonstrates that the performance of the DNN model drops significantly, even below that of other machine learning methods when noise was increased in the dataset (12). This indicates that deep learning might be more susceptible to noise than other AI methods, making 'traditional' machine learning techniques like LR and NBC preferable in real-life HTS (12).

Yet, multiple studies show the promises AI methods might have in improving VS, reducing the time required for this process (13 – 18). Especially when handling large volumes of CPIs within assay detection (14) and utilizing large datasets (15), deeper models appear to yield benefits for these tasks. However, on datasets to assess VS performance, both AI methods performed poor, and machine learning outperformed deep learning slightly (13, 15, 17).

Lead generation and optimization

AI has the potential to predict and analyze SARs with a high accuracy, and therefore reduce time and costs in this drug discovery phase (19 – 23). RF appears to be the most common used approach for analyzing SARs of lead compounds by applying quantitative structure-activity relationship (QSAR) studies, not many machine learning algorithms are suitable for this (21). However, all 8 studies concerning lead generation and optimization show that deep learning outperforms various machine learning methods; of which most of the studies compared DNNs with RF models and showed significant improvement (13, 16, 19 – 22). Furthermore, accuracy of DNNs is rising steadily with an increase in training data, while accuracy of machine learning methods diminishes with an increase of data (16, 20). Additionally, AI also shows strong performance on its predicting ability when fingerprints, which act as molecular descriptors, where added together with SAR datasets (12, 16, 20, 21).

Caution

Even though these studies show promising results, conclusions should be interpreted with caution (20, 21). DNNs are generally difficult to use due to its multiple hidden layers, large number of adjustable parameters, and it requires computationally intensive training with large amounts of training data (12, 19, 21). Besides difficulties in using deep learning models, constructing DNNs for drug discovery purposes remains challenging (12, 16). Noisy data and problems with overfitting are recurring issues with deep learning in drug discovery, including the LG and LO phase. The large and diverse SAR datasets may contain noise or other erroneous measurements, which can affect deep learning's performance (12, 20). In terms of overfitting, mentioned by Korotcov et al. (2017), their deep learning model showed significant differences in performance metrics between training and test values. Nevertheless, with deep learning's predictive capability in large SAR datasets, it can become comparable if not more advantageous over other machine learning methods in practice (16, 19, 20, 23). Especially when dealing with complex biological data (12, 13, 22).

Industry implementations

Even though AI-powered drugs have yet to be introduced to the market, the investments of pharma companies in IT firms show their expectations regarding the potential impact AI may have on drug discovery. Currently, nearly all (23) big pharma companies are strongly invested and applying AI in drug discovery, collaborating with various AI specialized companies. Moreover, there are nearly 200 startup companies in the space of AI for drug discovery. An overview of the big pharma companies and their collaborations around AI in drug discovery, is depicted below.

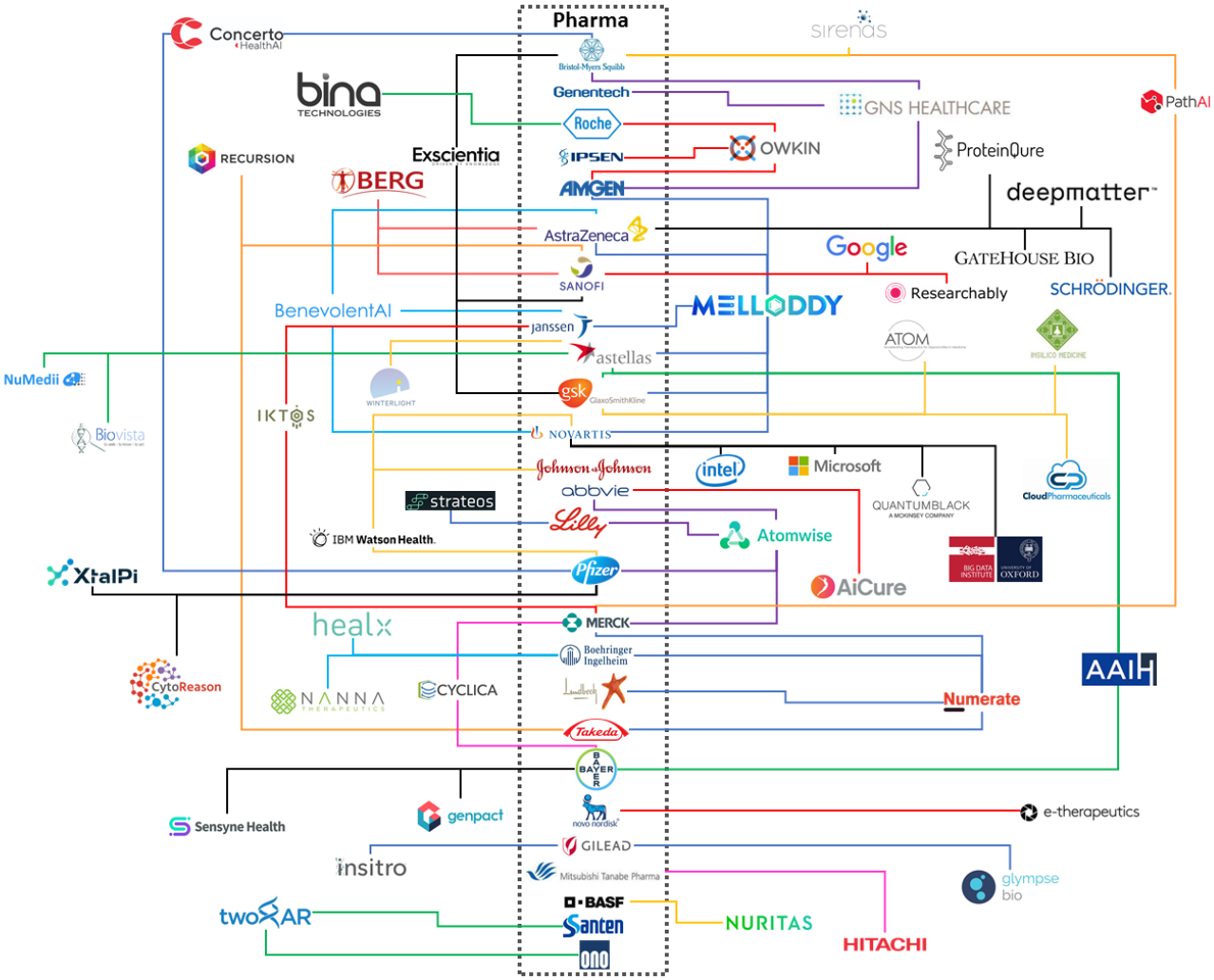

Figure 2. *Overview of big pharma companies and their collaborations around AI in drug discovery*

Conclusion

AI, in the forms of deep- and machine learning, has the potential to reduce time of the drug discovery process and to generate a more cost-effective R&D approach for pharmaceutical companies. Even though AI-powered drugs have yet to be introduced to the market, the investments of pharma companies in IT firms show their expectations regarding the potential impact AI may have on drug discovery. When comparing the different approaches in drug discovery, deep learning statistically outperforms machine learning and other 'traditional' methods. Although deep learning shows promising results, it remains unclear how it will perform in practice. Deep learning models are generally difficult to use and manufacture, as they require computationally intensive training with large training datasets. Therefore, for deep learning to make a meaningful impact in drug discovery, large training sets need to become publicly available and more research is required to assure its consistent performance. Furthermore, certain practical implications in terms of industrial partnerships and changes in employee competencies must follow. Nonetheless, with the potential shown by deep- and machine learning, it is only a matter of time for an AI drug to reach the market.

Research accountability

This research focussed on analysing the use of Artificial Intelligence for Drug Discovery. Specifically, the use of AI in three different phases of the drug discovery process was analysed. These phases consisted of Target Identification & Target Validation, Hit Identification & Hit Validation, and Lead Generation & Lead Optimization. A thorough and systematic literature database analysis was performed, focusing on publications within the last 5 years. The databases used for this analysis included PubMed and Scopus.

Although our analysis aimed to be thorough, systematic and bias-free, we can in no form ensure that the data was 100% complete.

Authors:
L.H.C. Geelen, MSc.
T. Oussoren, MSc.
M. Griffioen, MSc.
Dr. K.D.S. Fernald

References

1. Mak, K. K., & Pichika, M. R. (2018). Artificial intelligence in drug development: present status and future prospects. Drug discovery today.
2. Vohora, D., & Singh, G. (2017). Pharmaceutical Medicine and Translational Clinical Research (Academic Press).
3. Chen, X. P., & Du, G. H. (2007). Target validation: A door to drug discovery. Drug Discov Ther, 1(1), 23-29.
4. Myint, K. Z., & Xie, X. Q. (2015). Ligand biological activity predictions using fingerprint-based artificial neural networks (FANN-QSAR). In Artificial Neural Networks (pp. 149-164). Springer, New York, NY.
5. Oussoren, T. (2019). Artificial Intelligence in Drug Discovery: A Systematic Literature Review.
6. Tian, K., Shao, M., Wang, Y., Guan, J., & Zhou, S. (2016). Boosting compound-protein interaction prediction by deep learning. Methods, 110, 64-72.
7. Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., & Zhavoronkov, A. (2016). Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. Molecular pharmaceutics, 13(7), 2524-2530.
8. Ferrero, E., Dunham, I., & Sanseau, P. (2017). In silico prediction of novel therapeutic targets using gene‐disease association data. Journal of translational medicine, 15(1), 182.
9. Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., & Lu, H. (2017). Deep-learning-based drug‐target interaction prediction. Journal of proteome research, 16(4), 1401-1409.
10. Xie, L., He, S., Song, X., Bo, X., & Zhang, Z. (2018). Deep learning-based transcriptome data classification for drug-target interaction prediction. BMC genomics, 19(7), 667.
11. Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., ... & Hochreiter, S. (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chemical Science.
12. Koutsoukas, A., Monaghan, K. J., Li, X., & Huan, J. (2017). Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. Journal of cheminformatics, 9(1), 42.
13. Urban, G., Bache, K. M., Phan, D., Sobrino, A., Shmakov, A. K., Hachey, S. J., ... & Baldi, P. (2018). Deep Learning for Drug Discovery and Cancer Research: Automated Analysis of Vascularization Images. IEEE/ACM Transactions on Computational Biology and Bioinformatics.
14. Hamanaka, M., Taneishi, K., Iwata, H., Ye, J., Pei, J., Hou, J., & Okuno, Y. (2017). CGBVS-DNN: Prediction of Compound-protein Interactions Based on Deep Learning. Molecular informatics, 36(1-2), 1600045.
15. Imrie, F., Bradley, A. R., van der Schaar, M., & Deane, C. M. (2018). Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data. Journal of chemical information and modeling, 58(11), 2319-2330.
16. Korotcov, A., Tkachenko, V., Russo, D. P., & Ekins, S. (2017). Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. Molecular pharmaceutics, 14(12), 4462-4475.
17. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., & Koes, D. R. (2017). Protein‐Ligand scoring with Convolutional neural networks. Journal of chemical information and modeling, 57(4), 942-957.
18. Wallach, I., Dzamba, M., & Heifets, A. (2015). AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv:1510.02855.
19. Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., Mai, H., Li, Y., ... & Pande, V. S. (2018). PotentialNet for Molecular Property Prediction. ACS central science, 4(11), 1520-1530.
20. Hop, P., Allgood, B., & Yu, J. (2018). Geometric Deep Learning Autonomously Learns Chemical Features That Outperform Those Engineered by Domain Experts. Molecular pharmaceutics.
21. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). Deep neural nets as a method for quantitative structure‐activity relationships. Journal of chemical information and modeling, 55(2), 263-274.
22. Turk, S., Merget, B., Rippmann, F., & Fulle, S. (2017). Coupling Matched Molecular Pairs with Machine Learning for Virtual Compound Optimization. Journal of chemical information and modeling, 57(12), 3079-3085.
23. Li, H., Yu, L., Tian, S., Li, L., Wang, M., & Lu, X. (2017). Deep learning in pharmacy: The prediction of aqueous solubility based on deep belief network. Automatic Control and Computer Sciences, 51(2), 97-107.

About evolvalor

It all starts with believing – in yourself and in the value you can offer. We believe in solving many of today's problems through innovation, entrepreneurship, and partnerships.

We are a team of energetic no-nonsense business consultants with significant business experience in Healthcare, IT and Life Sciences. And we are committed to your success.

Throughout the journey of our clients we develop customized roadmaps of opportunities in terms of business and funding. We support our clients in accelerating their innovation projects and plans, securing long term growth and profitability. We offer our dedicated services, drawing from hands-on business experience as well as academic research and insights.

evolvalor
Entrada 100 (4th floor)
1114 AA Amsterdam-Duivendrecht
The Netherlands