AI in Drug Discovery

Artificial intelligence (AI) is driving transitions in healthcare. A major area where it is driving this transition is in Precision Medicine, where the goal is to enhance efficacies by improving the accuracy of predicting treatment and prevention strategies - based on the characteristics of particular maladies, drug physiochemical properties, and the genetic, environmental and lifestyle factors of individuals or groups of people (MedlinePlus, 2020). An important component to precision medicine is the facility to generate drug profiles that are adapted to the variability in disease and patient profiles. AI-driven approaches are finding and fueling success in this area.

Bioactivity prediction

Computational methods have been used in drug development for decades (Gregory Sliwoski, 2014). The emergence of high-throughput screening (HTS), in which automated equipment is used to conduct large assays of scientific experiments on molecular compounds in parallel, has resulted in generation of enormous amounts of data that require processing. Quantitative structure activity relationship (QSAR) models for predicting the biological activity responses to physiochemical properties of predictor chemicals, regularly use machine learning models like support vector machines (SVM) and random decision forests (RF) for this processing (Hongming Chen, 2018) (Delora Baptista, 2020).

While deep learning (DL) approaches have an advantage over single-layer machine learning methods, when predicting biological activity responses to properties of predictor chemicals, they have only recently been used for this (Hongming Chen, 2018). The need to interpret how predictions are made through computationally-oriented drug discovery, is seen - in part - as a factor to why DL approaches have not been adopted as quickly in this area (Erik Gawehn, 2016). However, because DL models can learn complex non-linear data patterns, using their multiple hidden layers to capture patterns in data, they are better suited for processing complex life sciences data than other machine learning approaches (Erik Gawehn, 2016).

For example, DL models were found to perform better than standard RF models (Junshui Ma, 2015) in predicting the biological activities of molecular compounds, using datasets from the Merck Molecular Activity Challenge on Kaggle (Kaggle, n.d.). Deep neural networks were also used in models that won NIH's Toxi21 Challenge (National Institute of Health, 2014) on using chemical structure data only to predict compounds of concern to human health (Andreas Mayr, 2016).

Their applications have included profiling tumors at molecular level, and predicting drug response based on pharmacological and biological molecular structures, functions and dynamics. This is attributed to their ability to handle high dimensionality in data features, making them appealing for use in predicting drug response (Delora Baptista, 2020)

De novo molecular design

DL is also finding new uses in developing novel chemical structures. Methods that employ variational autoencoders (VAE) have been used to generate new chemical structures by 1) encoding input string molecule structures, 2) reparametrizing the underlying latent variables and then 3) searching for viable solutions in the latent space, by using methods such as Bayesian optimizations. The final step involves decoding the results back into simplified molecular-input line-entry system (SMILES) notation, for

recovery of molecular descriptors. A variation to this involves using generative adversarial networks (GAN), as a subnetwork in the architecture, to generate the new chemical structures (Hongming Chen, 2018).

Other methods for developing new chemical structures include use of recurrent neural networks (RNN) to generate new valid SMILES strings, after training the RNNs on large quantities of known SMILES datasets. The RNNs use probability distributions learned from training sets, to generate new strings that correspond to new molecular structures (Marwin H. S. Segler, 2018). A variation to this approach incorporates reinforcement learning to reward models for new chemical structures, while punishing them for undesirable results (N Jaques, 2017)

The promise of precision medicine, and gains demonstrated through the infusion of AI approaches in drug discovery, will likely continue to fuel growth in this area of transition.

References

Andreas Mayr, G. K. (2016). Deeptox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science*.

Delora Baptista, P. G. (2020). Deep learning for drug response prediction in cancer. *Briefings in Bioinformatics*, 22, 2021, 360–379.

Erik Gawehn, J. A. (2016). Deep Learning in Drug Discovery. *Molecular Informatics*, 3 - 14.

Gregory Sliwoski, S. K. (2014). Computational Methods in Drug Discovery. *Pharmacol Rev*, 334 - 395.

Hongming Chen, O. E. (2018). The rise of deep learning in drug discovery. *Elsevier*.

Jacobs, V. S. (2019). Deep learning and radiomics in precision medicine, Expert Review of Precision Medicine and Drug Development. In *Expert Review of Precision Medicine and Drug Development: Personalized medicine in drug development and clinical practice* (pp. 59 - 72). Informa UK Limited, trading as Taylor & Francis Group.

Junshui Ma, R. P. (2015). Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *Journal of Chemical Information and Modeling*, 263-274.

Kaggle. (n.d.). *Merck Molecular Activity Challenge*. Retrieved from Kaggle.com: https://www.kaggle.com/c/MerckActivity

Marwin H. S. Segler, T. K. (2018). Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *America Chemical Society*.

MedlinePlus. (2020, September 22). *What is precision medicine?* Retrieved from https://medlineplus.gov/: https://medlineplus.gov/genetics/understanding/precisionmedicine/definition/

N Jaques, S. G. (2017). Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control. *Proceedings of the 34th International Conference on Machine Learning, PMLR* (pp. 1645-1654). MLResearchPress.

National Institute of Health. (2014, November 14). *Tox21 Data Challenge 2014*. Retrieved from tripod.nih.gov: https://tripod.nih.gov/tox21/challenge/

Vishwa S. Parekh, M. A. (2018). MPRAD: A Multiparametric Radiomics Framework. *ResearchGate*.