

In the medical and pharmaceutical areas, discovering different drugs is too important. This process is long, complex, and depends on numerous factors. To avoid that, Machine Learning (ML) techniques are applied. These techniques are the practice of using algorithms to parse data, learn from it and then make a determination or a prediction about the future state of any new data sets. This practice consists of at least 80% data processing and cleaning and 20% algorithm application. Data types can include images, textual information, biometrics and other information from wearables, assay information, and high-dimensional omics data [1]. In relation with the types of techniques that are used to apply ML can separate into two different. The first one is supervised learning, which is used to develop training models to predict future values of data categories or continuous variables. The second one is unsupervised learning, whereas this one is used for exploratory purposes to develop models that enable clustering of the data in a way that is not specified by the user.

ML presents new opportunities for early target identification and validation that will help us to discover drugs sooner. Apart from being used for that, ML can predict biomarkers, which will be a way to better understand the mechanism of action of a drug and for instance to identify the right drug for the right patients. Moreover, it is used to analysis of digital pathology data in clinical trials. It can also predict disease-specific drug effects.

To sum up, nowadays ML approaches are beginning to be commonly used in the various steps of the discovery and development of drugs by pharmaceutical companies, following the steps that appear in the following figure (**Figure 1.**).

| Target identification and validation | Compound screening and lead discovery | Preclinical development | Clinical development |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Successful applications in drug di | iscovery | | |
| Target identification and prioritization based on gene-disease associations Target druggability predictions Identification of alternative targets (splice variants) | Compound design with desirable properties Compound synthesis reaction plans Ligand-based compound screening | Tissue-specific biomarker identification Classification of cancer drug-response signatures Prediction of biomarkers of clinical end points | Determination of drug response by cellular phenotyping in oncology Precise measurements of the tumour microenvironment in immuno-oncology |
| Required data characteristics | | | |
| Current data are highly heterogeneous: need standardized high-dimensional target-disease-drug association data sets Comprehensive omics data from disease and normal states High-confidence associations from the literature Metadata from successful and failed clinical trials | Large amounts of training data needed Models for compound reaction space and nules Gold standard ADME data Numerous protein structures | Biomarkers: reproducibility of models based on gene expression data Dimension reduction of single-cell data for cell type and biomarker identification Proteomic and transcriptomic data of high quality and quantity | Pathology: well-curated expert annotations for broad-use cases (cancer versus normal cells) Gold standard data sets to improve interpretability and transparency of models Sample size: high number of images per clinical trial |

Figure 1. Machine learning applications in the drug discovery and their required data characteristics.

[1]Mamoshina P et al. Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. Front. Genet. 9, 242 (2018). [PubMed: 30050560]