

**Parte III**

**Monografia**

**Artificial Intelligence and Machine Learning in Pharmaceutical Sciences**

*Sob orientação da Professora Doutora Cláudia Cavadas*

## Resumo

A inteligência artificial é um ramo em crescente desenvolvimento. Divide-se em várias categorias e estará cada vez mais presente nas diversas áreas existentes, como é o caso da tecnologia, investigação e saúde.

Com o desenvolvimento de medicamentos a atingir um nível de eficácia reduzido, com grandes custos de tempo e recursos, com métodos envolvendo tentativa-erro, a abordagem com inteligência artificial abre um novo leque de possibilidades.

Para o desenvolvimento seguro e consciente da inteligência artificial é necessário que sejam implementadas regulamentações para o bom uso da informação e que sejam analisadas as questões éticas relacionadas, especialmente no que toca às informações sobre a saúde dos cidadãos.

**Palavras-chave:** Inteligência artificial, *Machine learning*, *Deep learning*, Ciências Farmacêuticas, Desenvolvimento de medicamentos, Ética.

## Abstract

Artificial intelligence is a growing sector. It is divided into several categories and will be increasingly present in various areas, such as technology, research and health.

With drug development achieving reduced effectiveness, being time and cost-consuming, using mainly trial-and-error methods, an artificial intelligence approach might open a new range of possibilities.

For the safe and conscious development of artificial intelligence, it is necessary to implement regulations for the good use of information and analyze the ethical issues, especially regarding to the citizen's health.

**Keywords:** Artificial intelligence, Deep learning, Drug development, Ethics, Machine learning, Pharmaceutical Sciences.

## **Abbreviations and Acronyms**

**ADME** – Absorption, Distribution, Metabolism and Excretion

**AI** – Artificial Intelligence

**AMPs** – Antimicrobial Peptides

**AS** – Alternative Splicing

**ANN** – Artificial Neural Network

**BATTLE** – Biomarker-integrated Approaches of Targeted Therapy for Lung cancer Elimination

**CNN** – Convolutional Neural Network

**DL** – Deep Learning

**DNN** – Deep Neural Network

**EU** – European Union

**FDA** – Foods and Drug Administration

**FFNN** – Feed-forward Neural Network

**GDPR** – General Data Protection Regulation

**HD** – Huntington's Disease

**M-CNN** – Multi-layer Convolutional Neural Network

**MDR** – Medical Devices Regulation

**MELLODDY** – Machine Learning Ledger Orchestration for Drug Discovery

**ML** – Machine Learning

**MLP** – Multi-layer Perceptron

**OSAI** – Observatory on Society and Artificial Intelligence

**PPI** – Protein-Protein Interaction

**QSAR** – Quantitative Structure Activity Relationship

**RF** – Random Forest

**RL** – Reinforcement Learning

**RNN** – Recurrent Neural Network

**SMILES** – Simplified Molecular-Input Line-Entry System

**SRIA** – Strategic Research and Innovation Agenda

**SVM** – Support Vector Machine

**TF** – Target Fishing

**TTP** – Technology Transfer Program

**VS** – Virtual Screening

## **I. Introduction**

Artificial intelligence has been around since long ago. However, even after many decades, there is no universal definition. This field is dynamic and has been evolving at a large pace in the last few years<sup>1</sup>. Still, what Hofstadter said, in 1979, can still be applied to today's meaning of artificial intelligence, since it "is whatever hasn't been done yet"<sup>2</sup>. This is the kind of hype that still exists around the area.

Understanding this field when it comes to its activity in research and innovation is important, since with each innovation and new ambition, researchers, companies and other involved in artificial intelligence (like politics), are asked to mind human safety and privacy. Being able to answer these questions is necessary to overcome the potential problems raised in community<sup>1</sup>.

Through the last decade many breakthroughs and policies appeared. Europe has, since 2006, launched Research and Innovation programs (in 2006 was FP7, in 2014 was Horizon 2020). With Europe, the United States and China have done the same when it comes to scientific development. Showing support for innovation makes possible to create new things, as seen by the booming of development that has happened since then<sup>1</sup>. It might not have been in the Health area, but in technology in general. Health is a more sensitive field, where the public has more trust issues to when it comes to autonomous machines. However, each development and discovery in this area makes us a step closer to achieve high quality personalized medicine<sup>3</sup>. Nonetheless, scientific research in drug development has shown advancements in the last five years, with new approaches and discoveries, not only for new drugs, but also in a way that can turn the process a simpler, cost and time-effective one.

All these innovations can become powerful tools to change how certain products and services are made and displayed, and also questioning what will happen to employment, people and data.

In this review, it will be made a general exposition about artificial intelligence, machine learning and other subfields covered by it. Then, we will overview some aspects about how machine learning can change specific subjects in the drug development pipeline. We will also discuss the new data regulation, how it works and what is being done, the ethical challenges and the limitations artificial intelligence faces nowadays.

## **2. Artificial Intelligence**

Although there is no global definition of artificial intelligence (AI), it is possible to say AI simulates human intelligence, through analytical and technological processes, without human intervention<sup>4,5</sup>. With the optimization of computers within recent years, it was possible to evolve AI, especially in the healthcare industry<sup>4</sup>. With the use of algorithms to process data, AI machines can mimic human cognitive functions, allowing them to perform tasks associated with the environments they are inserted, analyzing and identifying patterns. Hence, it can be said that artificial intelligence is the capability a computer or machine have in performing tasks normally done by humans, using data-based programs<sup>6</sup>.

Artificial intelligence is a large group of programs and algorithms with a main field called machine learning (ML), which is divided in many areas and that also has a main subfield, that intertwines with its areas, the deep learning (DL) (Figure 1). Below, it shall be explained each of these categories.

### **2.1. Machine Learning**

When AI uses algorithms and methods that allow the programs to learn and improve without being specifically programmed for it, it enters the subcategory of machine learning. Here, ML is divided in many different areas, such as supervised learning, unsupervised learning, semisupervised learning, active learning, reinforcement learning, transfer learning and multitask learning<sup>7</sup>. In health sciences, the main three used are: the supervised learning, unsupervised learning and reinforcement learning<sup>4</sup>.

#### **2.1.1. Supervised Learning**

Supervised learning uses methods that involve input and output data. Here, with the input data and its responses (output data), it is required, by the algorithm, to predict and develop information from these two data. In order to simplify the understanding, it is possible to see this supervised learning as a formula  $y = f(x)$ , making “y” the output data, “x” the input data and “f(x)” the predictive model<sup>7</sup>. Another way to understand this model is by dividing it in classification and regression methods, where the prediction is based in data from input and output sources<sup>4</sup>. As an example of this predictive model, Gunčar *et al.* used a Random Forest (RF) method to study possible haematological diagnosis and its capability to surpass the performance of haematological and internal medicine specialists. Therefore, the study used two predictive models (both with RF method), changing only the blood test results (and with different numbers of parameters, one with 181 and the other with 61), that were considered input data. To test the predictive models, the output data was 8 233 cases and 20 extra cases,

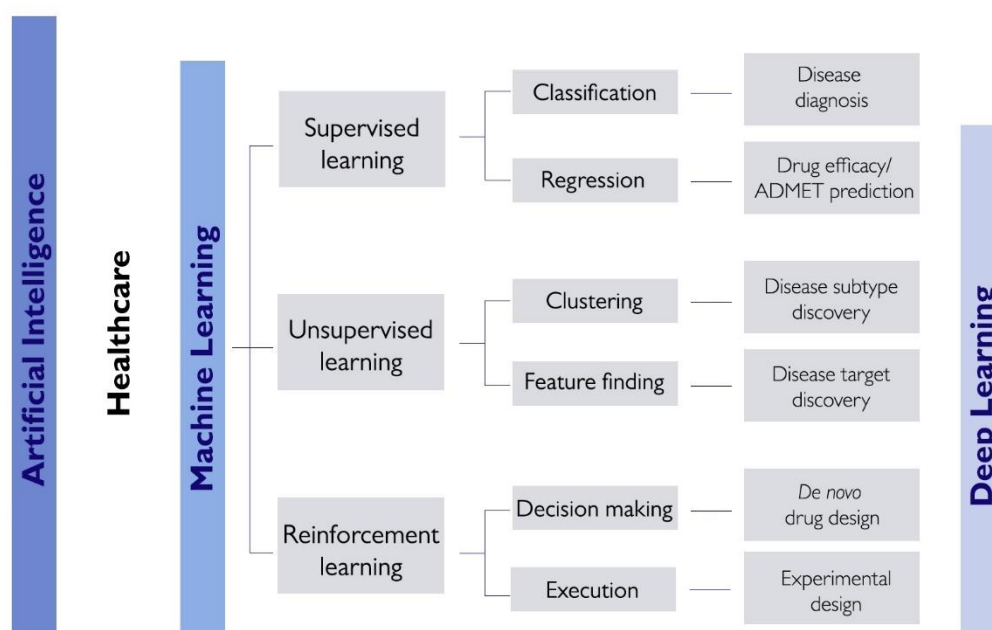
randomly selected. Upon testing them, the results were compared with the analysis of specialists (six haematology specialists and eight non-haematology internal medicine specialists), and they were presented with 20 cases. Comparing the models versus the specialists, the first ones achieved an accuracy of 0.60 (for the model with 181 parameters, the one with 61 obtained 0.55) while the haematology specialists performed a score of 0.62, whereas the non-haematology specialists only obtained an accurate diagnosis of 0.26. However, when the five most likely diseases are the only ones used in the prediction, the accuracy of the random forest methods increases to 0.90 in the RF with 181 parameters and 0.85 to the RF with 61, standing both above the score of the haematology specialists and the non-haematology specialists (the scores were 0.77 for the first ones, and the later ones only predicted correctly one case). This example allowed a way to understand how the supervised learning works and how one of the most common and useful methods is capable of great accuracy, especially when having small amounts of data, in some cases even surpassing the human intelligence (the specialists) <sup>8</sup>.

### **2.1.2. Unsupervised Learning**

Another main area of machine learning is the unsupervised learning. These algorithms are used when the only data available is the input one, without corresponding to any output. It can be said that unsupervised learning works with unlabeled data<sup>7,9</sup>. It uses clustering and feature-finding methods, so it can group and interpret the input data, making it possible to find common and different data points that, besides making it easier to organize data, it can be used as a preprocessor of supervised learning and discovering new patterns that might not have been considered before<sup>4,7</sup>. As an example of the unsupervised learning method, the study of Le *et al.* is quite understandable. In the study, the team tried to understand if it was possible to learn face detection using only unlabeled images, randomly selected. Clustering the data of 10 million YouTube videos, without duplicates, in multiple layers and sublayers to filter and pool the information, the unsupervised algorithm was able to achieve an accuracy of 81.7 % detecting faces. This study shows how big of an improvement unsupervised learning is, since it can learn to detect faces even with scale variations and different perspectives, using only unlabeled data<sup>9</sup>. Transporting those algorithms to other cases makes it possible to understand new patterns and possibilities that are lost in unlabeled data, that is a big amount of information that is not treated yet but with this method it can be used for greater meanings and studies.

### 2.1.3. Reinforcement Learning

The last main area of machine learning is the reinforcement learning (RL). It's a method that works under a reward-driven learning. In a simple explanation, is learning what to do to obtain the maximal reward possible. This accomplishment is made combining environment analysis, taking actions to alter that environment and obtaining the outcome of those actions. Although reinforcement learning does not rely on organized data/correct examples, just like unsupervised learning, the two are very different, since reinforcement learning is trying to maximize a signal that exists already rather than finding hidden information or patterns. Another difference from the other two algorithms is that RL has a temporal dimension, which mean that whatever decision is made on the current input, it will affect and determine the next input<sup>7,10</sup>. One application of the reinforcement learning is shown in the study of Olivecrona *et al.*, where RL is used to test molecular *de-novo* design (*de-novo* design is used in drug discovery, to find a molecule that overcomes the many existent criteria, being a fastidious process). The RL algorithm used was inverse Quantitative Structure Activity Relationship (inverse QSAR), that looks for the most promising region of possible activity to the corresponding molecular structures. Upon testing the model, by making it generate sulphur free molecules, against the traditional methods, it was possible to see improvements, making this method a possible way to overcome the problems of *de-novo* design and experimental design<sup>11</sup>.



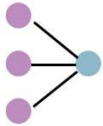
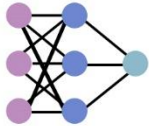
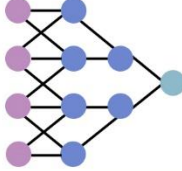
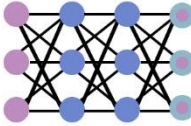
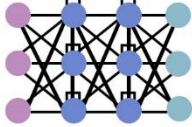





**Figure 1.** Applications of artificial intelligence and its categories. Adapted from Mak *et al.* (2018) <sup>4</sup>.

## 2.2. Deep Learning

Even though machine learning has many areas, there is a subfield that has emerged much alike an evolution of ML – deep learning. As computer power and digital data increase, so does the use of DL<sup>12</sup>.

Deep learning uses representation learning (a set of methods that give the computer raw data that allows it to automatically discover representations needed for detection or classification) methods that convey multiple levels of representation<sup>13</sup>. In this machine learning subgroup, the algorithms use artificial neural networks (ANNs), with several layers of nonlinear processing units to learn data representations. Modern ANNs in DL use as inspiration to its base structure the human brain and can also be called deep neural networks (DNNs), since the latter has more hidden layers than ANNs. This structure has three basic layers: the input layer, the hidden layer and the output layer. Since computers now have big capacity, it can be used a large number of hidden layers, to optimize the algorithm. There are many types of DNNs (Table I), and depending on that, the neurons (or nodes) in the neighboring layers can be completely connected or only partially. In this process, the input data (or variables) is taken by input neurons, these variables are transformed in the hidden neurons and then when they become output values, these values are calculated by the output neurons<sup>12</sup>.

**Table I.** The different types of neural networks. Adapted from Ching *et al.* (2018)<sup>14</sup>.

FFNN Feed-forward neural network	MLP Multi-layer perceptron	CNN Convolutional neural network	Autoencoder	RNN Recurrent neural network
				
Inputs are connected via some function to an output node and the model is trained to produce some output for a set of inputs	A feed-forward neural network in which there is at least one hidden layer between the input and output nodes	A feed-forward neural network in which the inputs are grouped spatially into hidden nodes. In the case of this example, each input node is only connected to hidden nodes alongside their neighbouring input node	A type of MLP in which the neural network is trained to produce an output that matches the input to the network	It is used to allow the neural network to retain memory over time or sequential inputs
				
input node				
hidden node				
output node				
output node (match input)				
edges connecting nodes in different layers or creating cycles within layers, correspond to inputs to mathematical functions				



As an example of the application of deep learning, in particular with deep convolutional neural networks (CNNs), Esteva *et al.* studied its potential to classify types of skin lesions, in order to detect skin cancer, one of the most common human malignancies. For that, this deep learning algorithm was tested against a board of 21 certified dermatologists. Since deep learning allows a large set of data, in this test it was used 129 450 clinical images (1 942 of them were biopsy-labelled as test images), consisting of 2 032 different diseases. By using only one CNN that was trained on general skin lesion classification, the group was able to perform at the same level as 21 dermatologists, across three diagnostic problems: keratinocyte carcinoma, melanoma and melanoma diagnosis under dermoscopy. Since the study showed accuracy on specialist level, the group considered that as an evolution towards the access to low-cost diagnostic care, whereas a simple algorithm could be used on mobile devices to help dermatologists outside clinics and hospitals<sup>15</sup>.

### **3. Drug Development**

In the last few decades, drug discovery and development has showed low rate of success. Even though the probability of success is critical for investors to make decisions about research, the data available (or lack of it) makes it difficult to understand if it is worth taking risks or not, when it comes to certain study areas. This can cause a loss not only for the investors and industries, since they lose a possible compound for new treatments, but also for the patients, that cannot receive a brand-new treatment. After a recent study, that analyzed 406 038 entries of trials, involving 21 143 compounds, from January 1st of 2000 to October 31st of 2015, the group reported that 13.8 % of all the drug programs eventually lead to approval. However, it doesn't mean that it is every year with that percentage of approval, since from 2005 to 2013 the results were lower, but with the evolution and approval of new oncology treatments in recent years (such as Nivolumab), these numbers have increased. Albeit, it might be possible to say that with the new ways to work data in this last decade may have been a way to facilitate some of these processes<sup>16</sup>.

With the evolution of computer hardware, artificial intelligence and machine learning have gained ground to grow, since now there are powerful tools available to process more data, looking for deeper information and, since drug development is time consuming, in a faster way. In recent years, the many stages of the drug discovery pipeline have had successful applications of machine learning techniques<sup>17</sup>.

### **3.1. Applications in drug discovery and development**

As stated before, machine learning uses its capability to adapt and improve the algorithms applied, always trying to augment the performance. Notwithstanding, there are many characteristics necessary to know so it's feasible to apply successfully ML. As appertained in 2.1., there are many techniques within ML, with several methods, for different data treatments. It's important to choose wisely and correctly which algorithm shall be used, to suit the data available and it's crucial to assure the success of the studies.

Since the growth in machine learning and its algorithms, several applications have been accounted throughout the drug development pipeline.

#### **3.1.1. Target identification and validation**

Drug discovery's main objective is to develop drugs (such as small molecules, antibodies, peptides and so on) that can change a disease condition, when it modulates its target. To start developing a new drug, it is needed a target with a possible therapeutic hypothesis – changing the target can change the disease state. To choose a target, researchers use evidence available, following the hypothesis. This type of choosing is called target identification and prioritization. Although the target has been decided, it is necessary to validate it. This validation means to understand how the target will act in modulating the disease. In these two initial steps of the pipeline, machine learning has had successful applications, acting in three important topics: target identification and prioritization based on gene-disease associations, target druggability predictions, and identification of alternative targets (as splice variants) <sup>17</sup>.

#### **Target identification and prioritization based on gene-disease associations**

This topic is relevant to drug discovery, not only since it might prevent development failure due to efficacy reasons (poor association between target and disease), but also because it's a possible way to reduce the initial research time spent. This means it can decrease the space in the first step of the pipeline. As Ferrero *et al.* study shows, using machine learning methods can reduce both costs and time in development. The study tested the hypothesis of ML techniques, using a data platform, being enough to predict therapeutic targets (in this case, the targets were already in the market or being pursued by industries). With that purpose, four different classes of machine learning algorithms were used (random forest, support vector machine, deep neural network and a gradient boosting machine) on partially labelled data, testing their performance. Using a semi-supervised learning (a mix of supervised and unsupervised learning, where less expensive but more abundant unlabeled data is used to train

the algorithm<sup>7</sup>), the deep neural network classifier obtained a precision over 71 %, when predicting therapeutic targets, based on gene-disease association data. An important characteristic of this study is how the predictions are for individual targets, which means it predicts potential targets, not regarding its intended indication. This proves that targets can be predicted merely using disease association data, showing the possibility of “establishing unambiguous causative links between putative targets and diseases of paramount importance to maximize the chances of success of drug discovery programs”<sup>18</sup>.

Another example of successful involves Huntington’s disease (HD). HD is a fatal neurodegenerative disease, where transcriptional regulatory changes occur. These changes can be detected early and they might be related directly to functions of huntingtin protein. With that in mind, and using machine learning, Ament *et al.* reconstructed a model to target transcription factors gene interactions in mouse striatum (where they are more prominent), by integrating data about the binding sites of transcription factors with information from gene co-expression in the striatum. This study showed thirteen transcription factors that could become possible target genes, since they were the most enriched among the differently expressed genes. That means there are new possible experiments that could become validated targets with new functions, that might help understanding more of HD, and attain new approaches to retard this disease’s progression<sup>19</sup>.

### **Target druggability predictions**

The prediction of druggable genes (genes that code proteins that can trigger phenotypic effects on a disease) is a time consuming and fastidious process, making it impossible to have fast information about causal gene-diseases relationships and the druggability of the target. Therefore, with machine learning approaches, it should be possible to predict the target’s druggability and the causal relationships between genes and diseases, in a shorter period, with accuracy. For that, Costa *et al.* trained a ML decision tree-based metaclassifier with datasets learning attributes containing network topological features, tissue expression profile and subcellular localization data, not only for each druggable gene, but also for morbid genes (mutated genes that cause hereditary human diseases), on a genome-wide scale. The metaclassifier was able to correctly recover 78 % of known druggable genes, having a precision of 75 % and for morbid genes, the values were 65 % and 66 %, respectively. After testing it for known genes, it was then used for unknown genes (druggable and morbid) to score druggability and morbidity scores and had a good match between those scores and data from literature. To assess what cellular rules for druggability and morbidity were crucial, another

two decision trees were generated. After analyzing, it was possible for the team to evaluate and conclude which rules were more important for druggability and morbidity, those being plasma membrane localization and number of regulating transcription factors, respectively. Through the years, many different predictive methods regarding druggable genes have been developed, however, only a few used a genome-wide scale (the majority used smaller scales). The methodology with support vector machine by Sugaya and Ikeda was used to select correctly a protein-protein interactions (PPIs) drug target, by assessing the druggability of those interactions. They used data from PPIs' structure (drug and chemical) and functions and used them as parameters for the support vector machine (SVM) – SVM is a supervised learning classifier method<sup>17</sup>. After testing the method, with thirty PPIs known as druggable used as positive samples and 1 295 human PPIs as testing, the SVM model was able to obtain an accuracy of 81 %. Both methods Costa et al. and Sugaya and Ikeda) used wide scale testing, but even though there was a grand amount of data available, it is still necessary to treat and organize it, since it is still quite heterogeneous<sup>20;21</sup>.

### **Identification of alternative targets (splice variants)**

The great majority of genes with multiple exons suffers an alternative splicing. This mechanism plays an important role in gene regulation. A few numbers of genes are estimated to synthesize 250 000 to 1 million proteins, thanks to different splices variants. Besides the diversity of proteins this generates, it also means that different variants generate different proteins and, while they can have positive roles in our system, some can do the opposite effect. Aberrant splice can generate multiple human diseases<sup>22</sup>.

As technology advances, so does the motivation to explore alternative splicing (AS). With machine learning and its methods, computational research about AS has emerged. This method involves splicing codes, which have a role in regulatory mechanisms, also predicting the outcome directly from the genomic sequence. Jha *et al.* studied two questions related to AS: if it was possible to improve previous models for the outcome of alternative splicing prediction and if there was a possibility to integrate more sources of data to improve the predictions for AS factors. For these two challenges, it was used two different machine learning approaches, both being supervised learning techniques, but one a regression analysis method and the other a classifier one (deep neural networks and Bayesian classifier, respectively). With these two techniques, the group was able to develop a new target function for prediction of alternative splicing in exon skipping events, showing an improvement in accuracy. Accessing multiple datasets with information on key splice factors in mouse brain, muscle and heart, they

were able to demonstrate the improvements in prediction models. With these techniques and developments, targeting splice variants might be a way to discover new therapeutic alternatives. While computational power increases, so does the accuracy of these techniques. Many studies compare Bayesian classifier and deep neural networks methods, and even though both have great accuracy, DNNs have surpassed Bayesian, since its characteristics allows the deep neural networks algorithms to exploit more possibilities with less datasets<sup>23;24</sup>.

### **Required data characteristics**

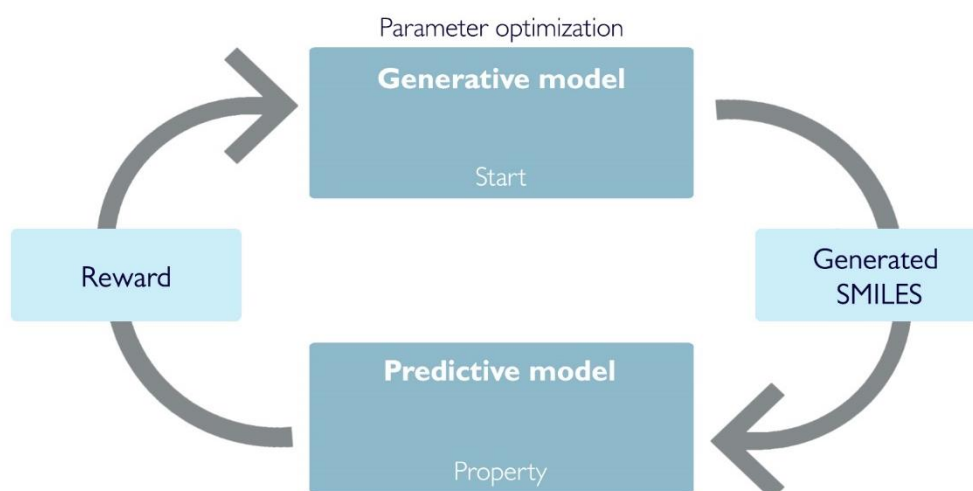
For these artificial intelligence applications to succeed on the first phase of the drug discovery pipeline, it is necessary that a few requirements are met. As it was mentioned along the text, even though it was possible to apply different methods of artificial intelligence, the data currently available is very heterogeneous, and that means it is needed standardized high-dimensional target-disease-drug association information. Besides the need to compensate the heterogeneous data, it is also important to obtain information of comprehensive omics, to have a deeper knowledge about disease and normal states. Although there are databases with information from literature about gene-disease associations, those need to have a high-confidence level so they can be used. Last but not least, to be sure the algorithms can be properly trained, information from positive and negative examples is needed i.e. metadata from clinical trials with positive and negative outcomes<sup>17</sup>.

#### **3.1.2. Compound screening and lead discovery**

After finding targets and validate them, it's still necessary to refine and modify the drug candidates for them to become more specific and selective, always taking into account the pharmacodynamic, pharmacokinetic and toxicological properties associated<sup>17</sup>. High-throughput screening aims to find suitable drug candidates, and it has been evolving, but even with its advancements, there are still problems related to its expensiveness, how it is time-consuming and, with all these in consideration, how it still has a high level of failure rates. Nonetheless, in recent years, a new computational field, called virtual screening (VS) has emerged in order to help in drug discovery, with its capability of estimating unknown biological interactions between compounds and targets. This is already a predictive model, however, machine learning techniques have been applied to it, so its accuracy and predictive power increase<sup>25</sup>.

## Compound design with desirable properties

Compound design is a crucial phase in drug discovery. Sometimes, design hypotheses are often biased (for preferred chemistry or interpretations), which makes computational and automated approaches attractive options, since it is possible to design compounds with desired properties. A way of improving de novo drug design is with machine learning methods. While trying to develop and implement a novel computational strategy for de novo design with desired properties, Popova *et al.* created ReLeaSe (Reinforcement Learning for Structural Evolution), that integrates two deep neural networks, generative and predictive, trained separately but with the aim to generate together novel targeted chemical libraries. For this to be possible, ReLeaSe used a system for simple molecule representation, SMILES (simplified molecular-input line-entry system). To assure a good training of each technique, initially, the predictive and generative DNNs were trained separately with supervised learning algorithm. The generative networks are trained with “a stack-augmented memory network to produce chemically feasible SMILES strings”, while the predictive methods are derived so they can foresee the wanted properties of the de novo generated compounds. On a second approach, the two deep neural networks are trained together using reinforcement learning (Figure 2), so it can purposely bias the generation of new structures, with the desired characteristics. With this new method, ReLeaSe, it is possible to have designing compounds libraries with the desired properties, while using machine learning<sup>26</sup>.



**Figure 2.** General pipeline of reinforcement learning system for novel compound generation. Adapted from Popova *et al.* (2018)<sup>26</sup>.

## Compound synthesis reaction plans

Many machine learning techniques can now be used to plan efficient routes of compound synthesis. To make it possible to plan the synthesis of a target molecule, it's necessary to decompose the compound, using retrosynthesis. Researchers can do this reverse and forward reactions in a laboratory, making it of simple execution to synthesize a target<sup>17</sup>. In a way to understand if it is possible to have a faster way to search and plan retrosynthesis routes, Segler *et al.* tested a method combining Monte Carlo tree search with deep neural networks. Monte Carlo tree search is a “general search technique for sequential decision problems with large branching factors”. Together with three different neural networks the group tested this possibility, extracting transformation rules from 12.4 million single-step reactions and around 301 thousand expansion rules from a chemistry database (Reaxys). After testing this new method, the results showed how it can be used effectively to generate compound synthesis reaction plans, and in a faster way (it only needs a few days, without the necessity of an expert encoding that can be tedious and biased), since it outperformed the established search methods. To test what professionals would prefer, the group did double-blind AB tests, where the professionals considered the quality of retrosynthetic routes generated by ML to be as good as the one's literature-based<sup>27</sup>.

## Ligand-based compound screening

Besides the already spoken virtual screening, machine learning methods have been used in screening and classification of drugs and predicting their toxicity<sup>28</sup>. A ligand-based approach trains on chemical features without modelling target features (as protein structures). An example of how it's possible to find ligand-based compound screening is the research done by Pande *et al.*, where machine learning methods together with Markov state models (models random systems) the group traced the pathway of opiates in binding to the orthosteric site, meaning they were able to find an unknown mechanism involving the binding of the  $\mu$ -opioid receptor, which lead to the revelation of an allosteric site, involved in its activation<sup>17,29</sup>. While machine learning methods were being included in predicting ligand-based targets, target fishing (TF) emerged. This new approach is a prediction method (predicts targets, mechanisms of action and side effects, as examples) that combines machine learning algorithms and cheminformatics, allowing to obtain deeper information about structures of complex compounds, facilitating the design and screening of more complex drug candidates that can face the also complex diseases. In general, these TF methods are based on screening procedures, intertwined with machine learning algorithms, reference ligands, determining if

the targets are appropriated. This use of AI and target fishing might be a way to facilitate one of the fastidious, time-consuming and highly-cost parts of drug development<sup>28</sup>.

### **Computer-guided antibiotic design**

As computer-guided drug design evolves and the world faces a rise on multidrug-resistant organisms, artificial intelligence appears as a new and inventive strategy. Using computational approaches, it may be possible to create new and sturdier antibiotics. Recently, using AI, around 32 new antimicrobial agents have been reported, where 16 of those showed antibacterial activity, with higher activity than some commercial drugs (as fosfomycin, ciprofloxacin), with its bactericidal effect affecting bacterial cell membrane and DNA gyrase. Being machine learning a broad field for research, that can be used with other methods, many see it as a possibility to generate new antimicrobial peptides (AMPs) that can disrupt bacterial membranes. While combining ML algorithms with genetic and *in vitro* evaluation, some studies showed improvements in AMPs efficacy, being able to make them much more active than the wild types<sup>30</sup>. Support vector machine can also be used to design new AMPs, as Lee *et al.* tested while investigating the activity of  $\alpha$ -helical AMPs with activity on the bacteria's membrane<sup>31</sup>.

Deep learning methods are seen as one of the most promising techniques for drug design and development. Using DL to generate new AMPs design, it's possible to predict its activity. However, there are not many studies using deep learning on antibacterial drugs development<sup>30</sup>. Nonetheless, Veltry *et al.* used deep learning to understand if it was possible to recognize antimicrobial activity and with how much accuracy. With deep neural networks, it is possible to obtain a high level of accuracy in AMP recognition, surpassing the state-of-the-art methods currently being used<sup>32</sup>.

With the current need for new antibiotics, these methods bring new possibilities allied with cost and time-effective advantages<sup>30</sup>.

### **Required data characteristics**

As it was mentioned in the previous pipeline topic, to have a successful application of machine learning methods in drug discovery, it's mandatory to fulfill data requirements. Therefore, the requirements for compound screening and lead discovery involve having good models for compound reaction and space rules, and also a numerous amount of protein structures available. Moreover, since dealing with AI involves data, it is crucial to have a large number of training data, so the algorithms are able to perform with major accuracy. Furthermore, the available data should also include a gold standard quality for information on



ADME (absorption, distribution, metabolism and excretion), assuring the developed drug, besides having the desired properties, can act without ADME changing them<sup>17</sup>.

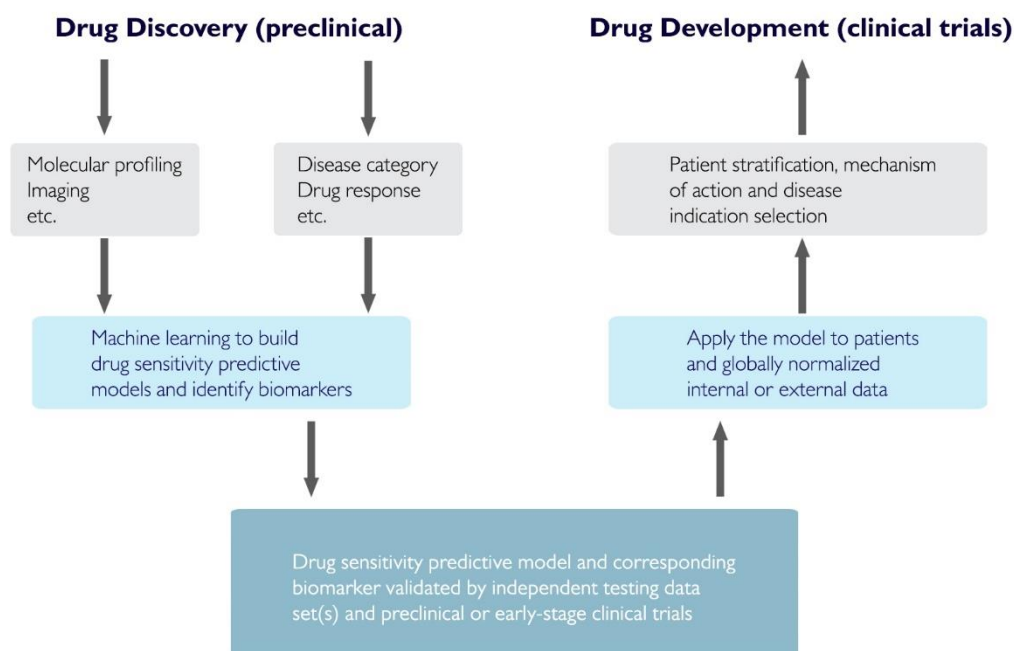
### 3.1.3. Preclinical development

In drug discovery, it's critical to be able to develop responsive drug biomarkers, using the available preclinical data, since that is what dictates the type of patients who will undergo the clinical trials<sup>33</sup>.

Predictive models might be a way to reduce the time and money spent on clinical trials, since using machine learning to discover biomarkers has demonstrated to be effective in helping improve clinical success rates, and also in understanding how the drug's mechanism of action works. It is important to know “the right drug for the right patients”, since a clinical trial might fail for this reason<sup>17</sup>.

### Biomarker identification and prediction of biomarkers

Using machine learning, it's possible to predict biomarkers on preclinical phases, making it possible to obtain easier patients' stratification, to understand and suggest the mechanisms of action of the drug, and to identify potential drug indications (Figure 3)<sup>17</sup>.



**Figure 3.** The use of predictive biomarkers to support drug discovery and development. Adapted from Vamathevan *et al.* (2019)<sup>17</sup>.

Biomarker identification and prediction can lead to better outcomes for patients. With that in mind and thinking of personalized medicine, Kim et al. developed the novel II Biomarker-integrated Approaches of Targeted Therapy for Lung cancer Elimination (BATTLE) program, since this type of cancer is one of the cancers that has major mortality rates, with a majority of clinical trials failing in improving the clinical outcomes of patients. The study was developed with an initial equal randomization time and then, with machine learning algorithms (Bayesian classifier) used together with BATTLE, the method adaptively randomized the patients for the different treatments (erlotinib, vandetanib, erlotinib with bexarotene, or sorafenib), with the data gathered from molecular biomarkers of the non-small-cell lung cancer pathogenesis, collected from biopsy. This adaptative method allowed more patients to be assigned for a more effective treatment, since of the 255 patients assign for the respective treatments, 244 were eligible for the disease control rate (the primary end point), after 8 weeks. Besides this major improvement in patients stratification, the group showed how important personalized trials can be, since in this case of the non-small-cell lung cancer, the previous established predictive biomarker is epidermal growth factor receptor mutations, but only 10 % to 15 % of the lung cancer population has this type of mutation, making the majority of population not being treated with a correct biomarker<sup>34</sup>.

### **Classification of cancer drug-response signatures**

When dealing with cancer biomarkers, these have specific DNA/RNA/protein characteristics that can be associated with prognostic (if there is a risk of cancer progression), or with understanding if the patient is responding to therapy (prediction). With the advancement in technology, the identification of cancer biomarkers has been evolving, changing disease classification and therapies. Recently a major advancement has been made, when the U.S. Food and Drug Administration (FDA) approved the use of pembrolizumab for cancer treatment based not on histology but on a biomarker. As Gulley *et al.* stated, pembrolizumab was the first anticancer drug to receive tissue-agnostic approval (from FDA). A major reason why that happened has to do with the way the trials were conducted – during that time, tissue collection was a priority, acquiring a high amount of data, allowing the investigators to retrospectively test the hypothesis. This states how biomarker- based disease classification is important for the future, since it can create different treatment options, and how precision medicine can help to get a better treatment, being more specialized and individualized<sup>35</sup>.

## **Required data characteristics**

As stated before, in 3.1.1. and 3.1.2., for every phase of drug development using artificial intelligence, there are data requirements. When dealing with biomarkers, it means gene expression data is required. But only having gene expression data would not suffice since it is necessary that this information can be used to create models that are possible to reproduce, or it would not be an important set for preclinical trials and patient care. However, not only gene expression is an important dataset. Omics data are also needed, with a demand for a high number of proteomic and transcriptomic data, accompanied with their high quality. Nevertheless, another major characteristic is necessary to consider: to be able to identify a biomarker for a specific cell-type, the data available shall be reduced to single-cell data, which means, it cannot be a pool of mixed information on different cells, it has to be dimensionally reduced. Having the required data characteristics on this phase allows a major assurance on the possibility to diminish the failure rates of clinical trials, guaranteeing the development of critical data associated with machine learning techniques and stepping towards a future where patients have personalized treatments<sup>17</sup>.

### **3.1.4. Clinical development**

When in the last phase of the drug development pipeline, it is mandatory to look at pathologies, since from a pathology it is possible to obtain a large amount of information. Even though a pathologist's analysis is important and provides data, machine learning techniques allow high-throughput generation of features involving a vast number of cells and its special relationships, something that is too hard and time-consuming for pathologists. Now, many tasks of computational pathology are done by machine learning methods, especially in image-recognition<sup>17</sup>, where deep learning appears in segmentation (as seen in tubule nuclei quantification, where it's correlated with risk categories that appear on pathologic images associated with breast cancer<sup>36</sup>), detection (per example, detection of mitotic activity in breast cancers, using a deep learning classifier<sup>37</sup>) and classification (as example, the use of convolutional neural networks for gastric carcinoma<sup>38</sup>).

## **Phenotyping of cellular images**

Godinez *et al.* studied the possibility of a multi-scale convolutional neural network (M-CNN) be used for phenotyping cellular images. When using a conventional image analysis, there are several limitations such as the several steps involved, with the need to have customization of almost every parameter, requiring a priori knowledge. However, using deep learning, it is possible to obtain a better performance with fewer hindrances. For this study,

the group used M-CNN in order to classify raw images into phenotypes, turning it into one unbiased and automatic step. With this method, it is possible to identify different patterns of phenotypes, at different scales and spatial levels and, as a major difference from conventional techniques, without requiring a priori knowledge about the expected imaging phenotypes<sup>39</sup>.

### **Required data characteristics**

As it is recurrent in all the phases of drug development using artificial intelligence applications, the data and samples available shall be of high number, to allow a better performance. Nonetheless, so that the methods used are not biased, it is crucial that the standard data is improved to guarantee the interpretability and transparency of the models. As said initially, understanding of pathologies is important so it is possible to have as much information as needed. With expert annotations, well screened and curated, they can be used for many different cases, broadening the image-recognition spectrum using deep learning<sup>17</sup>.

### **3.2. DREAM Challenges**

DREAM Challenges is an open platform, where science is shared openly, as a non-profit community. Many researchers and pharmaceutical companies contribute to DREAM Challenges, and since its beginning, it has only increased. Since one of the main problems involving artificial intelligence is how to filter the ever-increasing data available, and how complex it can be, DREAM tries to create a community that can solve these questions, improving computational models<sup>40</sup>.

DREAM Challenges has partnered up with Sage Bionetworks, which is also a non-profit research organization, created from a subsidiary of a big pharma company, Merck & Co, Inc<sup>41</sup>. Using Sage's Synapse platform, it is possible to support large scale research, analyzing it openly, and having access to reproducible data, since the answers for the challenges can be viewed by any user, so are the algorithms and methods<sup>40</sup>.

As an example of how this platform works, from July to October of 2017, a challenge was organized, having the name of "Parkinson's Disease Digital Biomarker DREAM Challenge". Since it is an open community, for this challenge, DREAM was able to have a top journal partner, Nature Biotechnology, meaning the results from it would be published on the journal. With the advancements in mobile health and in computational power, it might now be possible to measure and control diseases through the life of the patients, in an easier and quicker way than clinical exams. The problem that arises involves the conversion of the mobile sensor data to digital biomarkers, and this is where the challenge appears. Imposing a set of rules and giving to all participants the same initial data, the challengers asked for benchmark methods for

processing sensor data, using standard machine learning algorithms. The Parkinson's Disease Digital Biomarker Challenge was divided in two parts (sub-challenge 1 and sub-challenge 2). In the first one, data participants could use was from accelerometers, gyroscopes and magnetometers, while in the second part of the challenge involved gathering different data from the disease's symptoms. Each sub-challenge has a scoreboard and a winner, and at the end of the challenge, many of the participants (including the winners) will work together on a next phase, that will be published later (after testing algorithms and methods in a more practical way)<sup>42;43</sup>.

### **3.3. TensorFlow**

With the increase in popularity and use of machine learning in the world, a platform where beginners and professionals have an easier way to develop ML models emerged<sup>44</sup>. Through the use of several tutorials (even introducing neural networks models) and different algorithms (depending on the experience level of the user), anyone can use it<sup>45</sup>. TensorFlow also created a lighter version, for mobile devices, where it's still possible to apply machine learning<sup>44</sup>.

With this kind of platform, it might be easier for beginners to experiment theories and try new strategies using machine learning and deep learning, without the need of a vast amount of time and costs. Besides, it shows how artificial intelligence may become part of our daily lives, being of easy access to everyone.

### **3.4. MELLODDY**

On the 1<sup>st</sup> of June, the Machine Learning Ledger Orchestration for Drug Discovery (MELLODDY) was launched. Ten companies from the European Federation of Pharmaceutical Industries and Associations (EFPIA), like Bayer, Janssen, Merck and Novartis, participate in this project, that aims to create a machine learning platform that would allow learning from different datasets (of different companies), but without the fear of losing confidential information. This platform project aims to demonstrate the practicability of using a large amount of competitive preclinical data (over a billion drug-development relevant data and hundreds of terabytes of image data that has information about biological effects of more than ten million compounds) and, since this ML is not disease-specific, it should be able to be used in any pharmacological area. During these three years, MELLODDY hopes to create a solution that will make drug development a faster process, with less costs (the cost of the project is approximately 19 million euros, representing less than the price of developing two new drugs,

which is around 24 million euros), where it is possible to identify which small molecules are the most promising ones<sup>46,47</sup>.

#### **4. Regulatory Agencies and Data Protection**

Artificial intelligence relies on data to create all of these new approaches. However, some of the information could be called “sensitive personal data”, as it might contain personal information about the ethnic origin, the genes or health of the patients (data from clinical trials, per example). This brings to the surface questions about data protection<sup>48</sup>. To answer these questions, the European Union (EU) updated its General Data Protection Regulation (GDPR) in 2016, making it officially implemented by all Member States on May 25<sup>th</sup> of 2018 (article 51 of GDPR). The new regulation has stricter obligations for data protection and more requirements, with higher fines for infringements, having sanctions with fines up to 20 million euros or 4 % of the full annual turnover (whichever is higher), for the most serious cases (article 83 of GDPR)<sup>49</sup>.

As Torne and Binns pointed, there are some important changes within the updated GDPR, as: improved consent obligations; greater territorial scope; enhanced data subject notification obligations; right to be forgotten; privacy by design and default; contractual relationships between data processors and data collectors; registers of data processing activities; and the requirement for an appointed data protection officer if the core activities of the data controller involve regular and systematic monitoring of the data subject on a “large scale”<sup>48</sup>. Some particular changes are stated below.

##### Improved consent obligations

As stated in sections 32 and 42, consent for data use can only be called so when there is a clear positive act, showing free will, being specific, informed and unambiguous, by the owner of the data. This act can be done with physical or verbal proofs. The consent shall comprehend all the activities that the data will be used for, specifying each one of them, not being a general consent<sup>49</sup>.

##### Greater territorial scope

To ensure all personal data, if the responsible for the data treatment is within the EU, it shall abide by the regulation, even if the treatment doesn't happen in the European Union. The regulation also covers the possibility of the responsible not being from the EU, but if the data involves any citizen from any member-state, the regulation will apply to it (article 3 of GDPR)<sup>49</sup>. This makes it possible for companies headquartered inside and outside the EU to be accounted responsible for the use of data, allowing the tracking of the use of data.

### Right to be forgotten

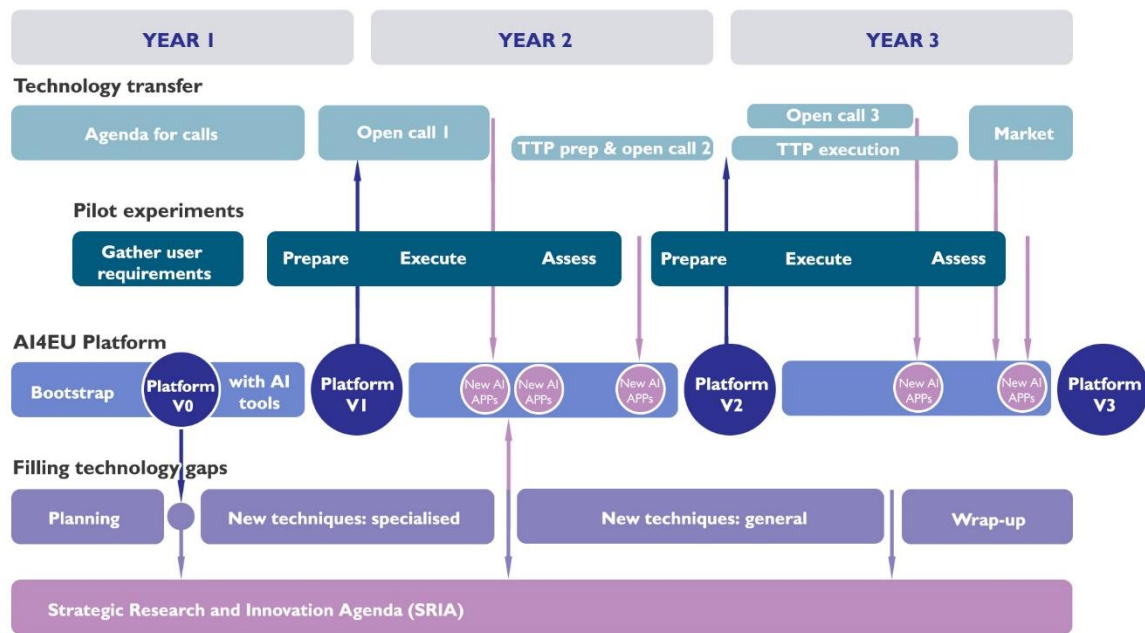
Under the regulation, the data subject has the right to ask to have their data deleted (as in, for the subject to be forgotten), if certain circumstances apply, as per example, if the data subject withdraws their consent to data treatment and there are no lawfully interests to justify that processing (article 17). However, when it comes to public interest, scientific and historical investigation and statistical means, there are assurances and derogations in the GDPR. If the data controllers can guarantee that no data subject can be identified because of relevant data, then it is possible to proceed with its processing<sup>49</sup>, since the GDPR cannot go against scientific innovation (article 89 of the GDPR).

### Privacy by design and default

To assure data protection, the data controllers have to guarantee it through the whole process (and before and after it), demonstrating they have proper technical and organizational measures, like pseudonymization and minimization (to ensure only the data needed will be used for the processing). By default, data should be secured, and personal data cannot be available without human intervention (article 25 of GDPR)<sup>49</sup>.

## **4.1. AI4EU**

Although the EU updated GDPR in 2016 (and to be fully applied in 2018), on January 2019, the EU launched its Artificial Intelligence project, AI4EU. With over 80 partners, from 21 countries, comes to facilitate AI research in Europe. This project has two main goals to achieve during the next three years: to mobilize the European AI community to make artificial intelligence plans come out of the paper; and create a leading community platform for AI in the UE, to promote economic and scientific growth. Since AI4EU is still in the middle of its first year, not much can be said about its evolution, though the EU launched a timeline for the ambitions of the project (Figure 4)<sup>50</sup>.



**Figure 4.** Timeline for the ambitions of the AI4EU project. Adapted from AI4EU (2019)<sup>50</sup>. TTP, as Technology Transfer Program.

## 4.2. Medical devices regulations

With the advance of AI use in digital solutions and devices in health, there was also needed to create a new regulation for medical devices (MDR). Community pharmacists could aid in monitoring a patient's health through this kind of devices. However, even though this is an important development, it is necessary to have regulations, so it is possible to understand what can be considered a medical device (to help a patient) or not<sup>48</sup>. The new MDR will be fully effective on 26<sup>th</sup> May of 2020, assuring to regulate the digital therapeutic solutions and devices, to ensure that they have identification, the general safety and performance requirements, conformity assessment procedures and many other regulatory and clinical aspects<sup>51</sup>.

## 5. Ethical and Social Issues

In terms of ethics, artificial intelligence is a very controversial topic. Even though this is controversial, Tran *et al.* study showed that only 204 papers (0.7 %) on their dataset were related to ethics, demonstrating how little debated this theme is<sup>52</sup>.

There is an ethical paradigm around the Turing Test about “thinking machines”. If during a conversation between computer, humans and examiner, the latter isn't able to understand without doubt if is talking with a computer or a human, the machine passes the test. On a positive remark, it means the machine is sophisticated; on the other side, some



argue it could mean the AI has a consciousness. For that to happen, it would mean what is considered to be human (the thinking and consciousness, ethical and values), might not be only created by the brain. However, for that to happen, it would be needed an AI that was able to act in multiple areas, which doesn't happen right now, since each algorithm focus on a particular area/question<sup>53</sup>.

Artificial intelligence consciousness may still be far, but human consciousness can be reflected in AI algorithms. Even though machine learning methods can reduce human bias and errors, it can also be a way to enhance those, since it depends on the data it's fed on. This means that, unconsciously, a team of researchers could use biased data for certain populations (per example, training the computer about a type of cancer with data mostly from Europe, and then reflecting as a general information for the world – a disease in a part of a world cannot represent a global information about it). While programming an algorithm, the AI developer can embed their beliefs and prejudices in it, consciously or not. In a way to decrease this possibility, some ask that development teams of AI algorithms should be more diverse<sup>54</sup>.

Even though AI can promise better health, it does not mean it will be well-distributed. Countries and populations in which data is less abundant may have difficulties to collect health information<sup>54</sup>. Another example could be if an algorithm is trained with data mostly from older Caucasian women – it would have poor predictions on young African men<sup>55</sup>.

Other fears and disbeliefs could be referred about the use of artificial intelligence and machine learning, such as jobs losses, malicious use of AI, undetectable fatal errors, possible loss of human contact, loss of privacy, and so on<sup>56</sup>. With proper population education, studying and improving ethical and regulatory issues, some of these problems and concerns might be solved<sup>54</sup>.

### **5.1. Observatory on Society and Artificial Intelligence**

As part of the AI4EU project (see 4.1.), the Observatory on Society and Artificial Intelligence (OSAI) was created to “act as a clearinghouse for information and research on the Ethical, Legal, Social-Economic and Cultural issues” involving the development and research of AI in Europe. OSAI's mission is to educate the EU citizens on AI and its benefits and effects on society and also reflect and discuss the values and questions associated with AI development and research in Europe. The observatory is the part of AI4EU that will deal with Ethics and Legal questions concerning the European society<sup>57</sup>.

## 6. The Limits of Artificial Intelligence

Artificial intelligence has many areas of interest and application. Within each area, there's a more appropriate algorithm to use and to fully benefit the study. However, no algorithm is free of challenges and limitations.

The three main categories in machine learning are supervised learning, unsupervised learning and reinforcement learning (deep learning has also one of these in it). In supervised learning, some limitations involve the necessity of labeled data for training, and how this method tends to overfit (failing to generalize accordingly in cases the algorithm wasn't trained<sup>10</sup>). When it comes to unsupervised learning, this technique may not be able to generally specify output space. Last, but not least, the reinforcement learning. RL uses reward function, which is sometimes difficult to create (a good reward). Besides that, there can be sample inefficiency, interfering with the process and results<sup>7</sup>.

Since artificial intelligence is a data-mining process, the quality and quantity of data available affects the outcome performance. Even if there is a vast amount of data, sometimes it's not qualified for use (some data in public databases don't meet the necessary requirements to be used). If the data is not organized in the through the same methods/characteristics, they cannot be compared through algorithms, besides the fact that, when using public databases, there isn't a filter to detect what is good data or bad data. For last, there is deep learning. An immediate question arises when thinking about DL: is it possible to know what happens during the processing? The answer is no. A big problem (limitation) is that even the developer of the DL algorithm may not understand what is being inspected and processed during the intermediate phases of deep learning and how and why does the computed gets a specific conclusion<sup>6</sup>.

Many of these limitations can and will probably be solved during the next years, along with the evolution of artificial intelligence.

## 7. Concluding remarks

Artificial intelligence is around since a few decades ago, rapidly growing in the last years. AI can be a key to change the health world, be it in drug development, diagnosis, treatments and follow-ups.

For this to happen, it is necessary good quality data, in a good amount. But data alone is not enough. It is important to have good research teams, qualified in the traditional drug discovery but allied with professionals that are able to use machine learning algorithms. It is also important to have diversity in these teams, so that the algorithms and data chosen are not biased. Data can also be collected from different professionals. In community pharmacies and health centers, it could be possible to gather information about certain disease's in the local community using machine learning techniques. That way, it might be possible to overcome the problems associated with not having enough data about certain minorities in society.

Artificial intelligence can play an important role in pharmaceutical sciences and therefore, in drug development. Before the introduction of AI in healthcare, drug discovery pharmacists had no other option besides the trial-and-error methods to develop new drugs. However, the many new ways to apply machine learning in the drug discovery pipeline could allow a more fluid process.

The innovation on medical devices could allow new ways of monitoring a patient's health. Community pharmacists can have a major role in helping patients with these new devices.

Although inside the scientific community AI is a strong growing world of possibilities, for the general public it is more as an unknown world, that books and movies depict as evil and that would overtake humans. With this in mind, educating the population and letting them understand the benefits, the pros and cons of artificial intelligence is a way to advance society. Discussion of ethics and values is vital.

The European Union is ensuring the digital safety of its citizens by updating their GDPR, but still assuring that scientific progress prevails. Despite this kind of measures, EU created a project with 21 countries that would allow data to flow inside the community and, with data, find new plans for AI development, not forgetting about ethical and social issues.

To think in the future of AI is important. Currently, even though artificial intelligence is booming, it's still far from the whole range of possibilities. With the evolution of machine learning, it's mandatory to create policies about the issue. To create AIs more intelligent than humans is a delicate process, since it must assure no harm is caused to humans, morally

relevant being, and that this advanced technology will be used for good and not with ill intentions<sup>52</sup>.

## References

1. ELSEVIER - **Artificial Intelligence: How knowledge is created, transferred, and used**. Amsterdam, The Netherlands : [s.n.]
2. HOFSTADTER, D. R. - **An Eternal Golden Braid**. New York : [s.n.]. ISBN 0465026567.
3. FRÖHLICH, H., BALLING, R., BEERENWINKEL, N., KOHLBACHER, O., KUMAR, S., LENGAUER, T. - **From hype to reality : data science enabling personalized medicine**. BMC Medicine. (2018) 1–15.
4. MAK, K., PICHKA, M. R. - **Artificial intelligence in drug development : present status and future prospects**. Drug Discovery Today. 00:00 (2018) 1–8.
5. DIEBOLT, V., AZANCOT, I., ADENOT, I., BALAGUE, C., BARTHÉLÉMY, P., BOUBENNA, N., COULONJOU, H., FERNANDEZ, X., LONGIN, J., METZINGER, A., MERLIÈRE, Y., PHAM, E., PHILIP, P., MARCHAL, T. - “ **Artificial intelligence ”: Which services, which applications, which results and which development today in clinical research? Which impact on the quality of care? Which recommendations?**. Elsevier. (2019).
6. CHAN, H. C. S., SHAN, H., DAHOUN, T., VOGEL, H., YUAN, S. - **Advancing Drug Discovery via Artificial Intelligence**. CellPress Reviews. (2019) 1–13.
7. YANG, X., WANG, Y., BYRNE, R., SCHNEIDER, G., YANG, S. - **Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery**. Chemical Reviews. (2018).
8. GUNČAR, G., KUKAR, M., NOTAR, M., BRVAR, M., ČERNELČ, P., NOTAR, M., NOTAR, M. - **An application of machine learning to haematological diagnosis**. Scientific Reports. 8:1 (2018) 1–12.
9. LE, Q. V., MONGA, R., DEVIN, M., CORRADO, G. S., CHEN, K., RANZATO, M. A. A., DEAN, J., NG, A. Y. - **Building high-level features using large scale unsupervised learning Dataset constructions**. arXiv preprint arXiv:1112.6209. (2011) 1–10.
10. BACH, F. - **Reinforcement Learning: An Introduction**. Second Edition. Massachusetts : MIT Press, 2008. ISBN 9780262039246.
11. OLIVECRONA, M., BLASCHKE, T., ENGVIST, O., CHEN, H. - **Molecular de-novo design through deep reinforcement learning**. Journal of Cheminformatics. 9:1 (2017) 1–14.

12. CHEN, H., ENGVIST, O., WANG, Y., OLIVECRONA, M., BLASCHKE, T. - **The rise of deep learning in drug discovery**. Drug Discovery Today. 23:6 (2018) 1241–1250.
13. LECUN, Y., BENGIO, Y., HINTON, G. - **Deep learning**. Nature. 521:7553 (2015) 436–444.
14. CHING, T., HIMMELSTEIN, D. S., BEAULIEU-JONES, B. K., KALININ, A. A., DO, B. T., WAY, G. P., FERRERO, E., AGAPOW, P. M., ZIETZ, M., HOFFMAN, M. M., XIE, W., ROSEN, G. L., LINGERICH, B. J., ... GREENE, C. S. - **Opportunities and obstacles for deep learning in biology and medicine**. ISBN 0000000305396.
15. ESTEVA, A., KUPREL, B., NOVOA, R. A., KO, J., SWETTER, S. M., BLAU, H. M., THRUN, S. - **Dermatologist-level classification of skin cancer with deep neural networks**. Nature. 542:7639 (2017) 115–118.
16. WONG, C. H., SIAH, K. W., LO, Andrew W. - **Estimation of clinical trial success rates and related parameters**. Biostatistics. 20:2 (2019) 273–286.
17. VAMATHEVAN, J., CLARK, D., CZODROWSKI, P., DUNHAM, I., FERRAN, E., LEE, G., LI, B., MADABHUSHI, A., SHAH, P., SPITZER, M., ZHAO, S. - **Applications of machine learning in drug discovery and development**. Nature Reviews Drug Discovery. 18:6 (2019) 463–477.
18. FERRERO, E., DUNHAM, I., SANSEAU, P. - **In silico prediction of novel therapeutic targets using gene-disease association data**. Journal of Translational Medicine. 15:1 (2017) 1–16.
19. AMENT, S. A., PEARL, J. R., CANTLE, J. P.; BRAGG, R. M., SKENE, P. J., COFFEY, S. R., BERGEY, D. E., WHEELER, V. C., MACDONALD, M. E., BALIGA, N. S., ROSINSKI, J., HOOD, L. E., CARROLL, J. B., PRICE, N. D. - **Transcriptional regulatory networks underlying gene expression changes in Huntington's disease**. Molecular Systems Biology. 14:3 (2018) 1–16.
20. COSTA, P. R., ACENCIO, M. L., LEMKE, N. - **A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data**. BMC Genomics. 11:SUPPL. 5 (2010) S9.

21. SUGAYA, N., IKEDA, K. - **Assessing the druggability of protein-protein interactions by a supervised machine-learning method.** BMC Bioinformatics. 10:(2009) 263.
22. PANWAR, B., MENON, R., EKSI, R., LI, H. D., OMENN, G. S., GUAN, Y. - **Genome-wide functional annotation of human protein-coding splice variants using multiple instance learning.** Journal of Proteome Research. 15:6 (2016) 1747–1753.
23. JHA, A., GAZZARA, M. R., BARASH, Y. - **Integrative deep models for alternative splicing.** Bioinformatics. 33:14 (2017) i274–i282.
24. LEUNG, M. K. K., XIONG, H. Y., LEE, L. J., FREY, B. J. - **Deep learning of the tissue-regulated splicing code.** Bioinformatics. 30:12 (2014) 121–129.
25. RIFAIIOGLU, A. S., ATAS, H. MARTIN, M. J., CETIN-ATALAY, R., ATALAY, V., DOĞAN, T. - **Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases.** Briefings in Bioinformatics. May (2018) 1–36.
26. POPOVA, M., ISAYEV, O., TROPSHA, A. - **Deep reinforcement learning for de novo drug design.** Science Advances. 4:7 (2018) 1–15.
27. SEGLER, M. H. S., PREUSS, M., WALLER, M. P. - **Planning chemical syntheses with deep neural networks and symbolic AI.** Nature. 555:7698 (2018) 604–610.
28. HASSANZADEH, P., ATYABI, F., DINARVAND, R. - **The significance of artificial intelligence in drug delivery system design.** Advanced Drug Delivery Reviews. (2019).
29. BARATI FARIMANI, A., FEINBERG, E., PANDE, V. - **Binding Pathway of Opiates to  $\mu$ -Opioid Receptors Revealed by Machine Learning.** Biophysical Journal. 114:3 (2018) 62a-63a.
30. TORRES, M. D. T., LA FUENTE-NUNEZ, C. - **Toward computer-made artificial antibiotics.** Current Opinion in Microbiology. 51:(2019) 30–38.
31. LEE, E. Y., FULAN, B. M., WONG, G. C. L., FERGUSON, A. L. - **Mapping membrane activity in undiscovered peptide sequence space using machine learning.** Proceedings of the National Academy of Sciences of the United States of America. 113:48 (2016) 13588–13593.

32. VELTRI, D., KAMATH, U., SHEHU, A. - **Deep learning improves antimicrobial peptide recognition**. *Bioinformatics*. 34:16 (2018) 2740–2747.
33. LI, B., SHIN, H., GULBEKYAN, G., PUSTOVALOVA, O., NIKOLSKY, Y., HOPE, A., BESSARABOVA, M., SCHU, M., KOLPAKOVA-HART, E., MERBERG, D., DORNER, A., TREPICCHIO, W. L. - **Development of a drug-response modeling framework to identify cell line derived translational biomarkers that can predict treatment outcome to Erlotinib or Sorafenib**. *PLoS ONE*. 10:6 (2015) 1–20.
34. MA, Q.-L., TETER, B., UBEDA, O. J., MORIHARA, T., DHOOT, D., NYBY, M. D., TUCK, M. L., FRAUTSCHY, S. A., COLE, G. M., KIM, E. S., HERBST, R. S., WISTUBA, I. I., LEE, J., ... HONG, W. K. - **The BATTLE trial: Personalizing Therapy for Lung Cancer**. *Cancer Discovery*. 27:52 (2011) 14299–14307.
35. BOYIADZIS, M. M., KIRKWOOD, J. M., MARSHALL, J. L., PRITCHARD, C. C., AZAD, N. S., GULLEY, J. L. - **Significance and implications of FDA approval of pembrolizumab for biomarker-defined disease**. *Journal for ImmunoTherapy of Cancer*. 6:1 (2018) 1–7.
36. ROMO-BUCHELI, D., JANOWCZYK, A., GILMORE, H., ROMERO, E., MADABHUSHI, A. - **Automated Tubule Nuclei Quantification and Correlation with Oncotype DX risk categories in ER+ Breast Cancer Whole Slide Images**. *Scientific Reports*. 6:April (2016) 1–9.
37. ROMO-BUCHELI, D., JANOWCZYK, A., GILMORE, H., ROMERO, E., MADABHUSHI, A. - **A deep learning based strategy for identifying and associating mitotic activity with gene expression derived risk categories in estrogen receptor positive breast cancers**. *Cytometry Part A*. 91:6 (2017) 566–573.
38. SHARMA, H., ZERBE, N., KLEMPERT, I., HELLWICH, O., HUFNAGL, P. - **Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology**. *Computerized Medical Imaging and Graphics*. 61:(2017) 2–13.
39. GODINEZ, W. J., HOSSAIN, I., LAZIC, S. E., DAVIES, J. W., ZHANG, X. - **A multi-scale convolutional neural network for phenotyping high-content cellular images**. *Bioinformatics*. 33:13 (2017) 2010–2019.
40. DREAM CHALLENGES - **About DREAM** [Consult. 28 ago. 2019]. Disponível na



Internet: <http://dreamchallenges.org/about-dream/>

41. DREAM CHALLENGES - **Sage/Synapse** [Consult. 28 ago. 2019]. Disponível na Internet: <http://dreamchallenges.org/sagesynapse/>

42. DREAM CHALLENGES - **Parkinson's Disease Digital Biomarker DREAM Challenge** [Consult. 28 ago. 2019]. Disponível na Internet: <https://www.synapse.org/#!/Synapse:syn8717496/wiki/422884>

43. WIRE, B. - **Sage Bionetworks in Collaboration with The Michael J. Fox Foundation Announce Winners in the DREAM Parkinson's Disease Digital Biomarker Challenge**, atual. 2018. [Consult. 28 ago. 2019]. Disponível na Internet: <https://www.businesswire.com/news/home/20180117006187/en>

44. TENSORFLOW - **Introduction to TensorFlow** [Consult. 29 ago. 2019]. Disponível na Internet: <https://www.tensorflow.org/learn>

45. TENSORFLOW - **TensorFlow Core** [Consult. 29 ago. 2019]. Disponível na Internet: <https://www.tensorflow.org/overview/>

46. BURKI, T. - **Pharma blockchains AI for drug development**. Lancet (London, England). 393:10189 (2019) 2382.

47. INNOVATIVE MEDICINES INITIATIVE - **MELLODDY**, atual. 2019. [Consult. 28 ago. 2019]. Disponível na Internet: <https://www.imi.europa.eu/projects-results/project-factsheets/melloddy>

48. TORNE, L., BINNS, R. - **Drug development and therapeutic solutions in the digital age**. Drug Discovery Today. 23:12 (2018) 1922–1924.

49. EUROPEAN PARLIAMENT - **EU Directive 2016/679 - General Data Protection Regulation (GDPR)**. Official Journal of the European Union. May 2016 (2016) 6,8,32,33,43,44,65,82-85.

50. AI4EU - **About AI4EU**, atual. 2019. [Consult. 29 ago. 2019]. Disponível na Internet: <https://www.ai4eu.eu/#about>

51. EUROPEAN PARLIAMENT - **Regulation (EU) 2017/746 – of the European Parliament and of the Council on in vitro diagnostic medical devices and repealing**

**Directive 98/79/EC and Commission Decision 2010/227/EU.** Official Journal of the European Union. L177:(2017) 1,2.

52. TRAN, B., VU, G., HA, G., VUONG, Q.-H., HO, M.-T., VUONG, T.-T., LA, V.-P., HO, M.-T., NGHIEM, K.-C.; NGUYEN, H., LATKIN, C., TAM, W., CHEUNG, N.-M., ... HO, R. - **Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study.** Journal of Clinical Medicine. 8:3 (2019) 360.

53. KESKINBORA, K. H. - **Medical ethics considerations on artificial intelligence.** Journal of Clinical Neuroscience. 64: (2019) 277–282.

54. NUFFIELD COUNCIL ON BIOETHICS - **Artificial intelligence ( AI ) in healthcare and research.** Bioethics Briefing Note. (2018) 1–8.

55. VAYENA, E., BLASIMME, A., COHEN, I. G. - **Machine learning in medicine: Addressing ethical challenges.** PLoS Medicine. 15:11 (2018) 4–7.

56. BECKER, A. - **Artificial intelligence in medicine: What is it doing for us today?.** Health Policy and Technology. 8:2 (2019) 198–205.

57. EUROPEAN CENTRE FOR LIVING TECHNOLOGY - **OSAI**, atual. 2019. [Consult. 29 ago. 2019]. Disponível na Internet: <https://www.unive.it/pag/36811/>