msj Medicine & Science Journal

Medicine & Science Journal 2023; 1(1):e202339
https://doi.org/10.53150/msj.v1i1.39

REVIEW ARTICLE

From data to drugs: Harnessing machine learning in drug discovery - A review

Gianluca Gazzaniga[1], Thomas Virdis[2]

[1]Department of Medical Biotechnology and Translational Medicine, Postgraduate School of Clinical Pharmacology and Toxicology, Università degli Studi di Milano, 20122, Milan, Italy.
[2]Department of Oncology and Hemato-Oncology, Università degli Studi di Milano, 20122, Milan, Italy

Correspondence:
Gianluca Gazzaniga, MD
Department of Medical Biotechnology and Translational Medicine, Postgraduate School of Clinical Pharmacology and Toxicology, Università degli Studi di Milano
Via Festa del Perdono 7, 20122 Milan, Italy
email: gianluca.gazzaniga@unimi.it

Financial support: No funding was received in support of this article.

Keywords: Drug Development; Drug Discovery; Drug Repositioning; Machine Learning; Artificial Intelligence.

Abstract

Drug development is a rigorous process essential for improving patient outcomes. However, this complex endeavour requires significant investment and time. The integration of Machine Learning (ML) techniques in drug discovery can revolutionize the field by efficiently processing large amounts of data and accelerating the identification and development of potential drug candidates. This review highlights ML's impact across drug development stages, from design to clinical trials (CTs).

Recently, the availability of high-quality databases and the surge in data digitalization has promoted the development of several ML algorithms, which have proved to be effective in classifying outcomes based on multivariate relationships. Particularly, Deep Learning (DL) architectures such as feedforward networks, Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) neural networks, represent a subset of ML which has been gaining popularity because of its ability to emulate the human brain and handle more complex tasks, thus representing a paradigm shift in data analysis and prediction.

ML plays a vital role in virtual screening, de-novo drug design and drug repurposing. Virtual screening methods can rapidly screen large chemical libraries and identify promising candidates for further investigation. De-novo drug design involves the use of ML-based generative models to produce new chemical structures with desired properties. Drug repurposing aims to identify

Received: July 14, 2023
Revised: October 28, 2023
Accepted: November 07, 2023
Published: November 22, 2023

Article information are listed at the end of this article.

new therapeutic uses for existing drugs. Additionally, ML can improve the efficiency of CTs by addressing challenges related to patient enrolment, study design, and phase transition.

The integration of ML with high-quality datasets can significantly improve drug development process, thereby increasing efficiency and success rates. However, it is important to address issues related to data quality, preprocessing bias, molecular representation, and interpretation of results. Harnessing the power of AI can accelerate drug development, ultimately benefiting patients and the healthcare industry as a whole.

1. Introduction

Drug development is a complex and rigorous process that involves the discovery, design, testing and approval of new drugs for the treatment, prevention or management of diseases and health conditions. It is an important aspect of healthcare that is critical to improve patient outcomes and address unmet medical needs.

Due to its highest standards, drug development process is long and expensive, often lasting several years and requiring significant investment. Up to an amount close to 2.5 billion dollars and five to ten years may be required to pass from bench-side to market [1–4]. Furthermore, despite rigorous testing, not all drug candidates make it through all stages of development due to limited or no therapeutic efficacy in humans or unacceptable toxicity leading to treatment discontinuation. It has been estimated that only 59% of drugs evaluated in phase 3 trials ultimately secure final approval from the Food and Drug Administration (FDA), and astonishingly, when considering the phase 1 trials, a mere 13.8% of drugs entering this phase ultimately attain final regulatory approval [5].

In the realm of drug discovery, the integration of Artificial Intelligence (AI) holds immense potential in terms of improving the efficiency and success rate of drug development process. However, its success heavily relies on the availability of a high-quality databases. In recent years, the pharmaceutical sector, as many others, has witnessed a remarkable surge in data digitalization, revolutionizing the way information is handled. This exponential growth of digitized data presented a formidable challenge of effectively acquiring, scrutinizing, and analysing the vast knowledge available. The development of advanced IT infrastructure has facilitated the organization, and accessibility of these data through user-friendly and widely accessible online databases.

Given this background, AI has emerged as a powerful tool for efficiently managing vast amounts of data through enhanced automation. This capability had a profound impact on the field of drug discovery, resulting in a paradigm shift in the applications of AI techniques. Machine Learning (ML), in particular, has been extensively utilized to analyse clinical data, learn from a large number of examples, and make predictions about the behaviour of unexplored datasets. These advancements have revolutionized the landscape of drug discovery, enabling us to gain valuable insights and make informed decisions based on the power trained models [6].

In this Review, we will try to give a brief overview of how ML may have an impact on different phases of drug development, from drug design to Clinical Trials (CTs). Particularly, we will commence by delineating the requisite technical specifications and operational procedures inherent to ML. Subsequently, our focus will pivot towards a comprehensive exploration of Drug Discovery, starting with its foundation: chemical libraries; this segment will elucidate the storage of potential drug candidates and the pivotal role that AI plays in either facilitating compound screening or engendering novel ones. Following this, we will explore the realm of Drug Repurposing, an alternative approach to conventional new drug development, which can serve as a valuable strategy to address unmet medical needs. Lastly, we will conclude by delving into the latest stages of drug development, where we will discuss the various ways in which AI exerts its influence across different phases, steps, and prospects of CTs.

2. Machine Learning: technical bases

While AI finds extensive application within the biomedical sciences, it predominantly retains its character as a technical discipline grounded in fundamental informatics. In order to furnish readers with a navigational aid within this domain, we provided concise definitions of the most technically nuanced terms employed in this review in **Table 1**.

TABLE 1 - Brief definition of technical AI terms. **1/2**

Terminology	Description
Activation Function	A function used in neural networks that adds non-linearity to the network, enabling it to learn from more complex data.
Artificial Intelligence (AI)	The science of creating intelligent machines capable of performing tasks that typically require human intelligence.
American Standard Code for Information Interchange (ASCII)	Character encoding standard for electronic communication, which is commonly used to represent text in computers and other devices.
Autoencoder Neural Networks (AENs)	Neural networks used for data compression and feature learning, consisting of an encoder and a decoder.
Backpropagation	A method used in artificial neural networks to calculate the error contribution of each neuron after a batch of data is processed, going back from the output layer to the hidden and input layers.
Canonization algorithm	An algorithm used to transform data or structures into a standardized or canonical form, making them more easily comparable or searchable.
Convolutional Neural Network (CNN)	A type of deep learning model primarily used for analyzing visual imagery. It uses convolutional layers to filter inputs for useful information.
Deep Learning (DL)	A subset of ML that uses artificial neural networks with multiple layers (deep structures) to model and understand complex patterns.
Deep Neural Network	Neural networks with multiple hidden layers, allowing them to model complex relationships in data.
Feedforward Networks	A type of neural network where the information flow is unidirectional, moving forward from the input nodes to the output nodes without cycles or loops.
Generative Models (GMs)	ML or DL models used to generate synthetic data, upon being trained on real data and have learned how to optimally approximate them.
Gradient Descent	An optimization algorithm used to find the values of parameters that minimize a given function by iteratively moving in the direction of steepest descent.
Hidden layer	The neural network layers that usually stay in between of the input and output layers.
Kernel Density Estimation (KDE)	Non-parametric method used in statistics to estimate the probability density function of a random variable, through using a non- negative, kernel function, to smooth the data points and generate a continuous and smooth estimate of the underlying distribution.

https://doi.org/10.53150/msj.v1i1.39

TABLE 1 - Brief definition of technical AI terms. **2/2**

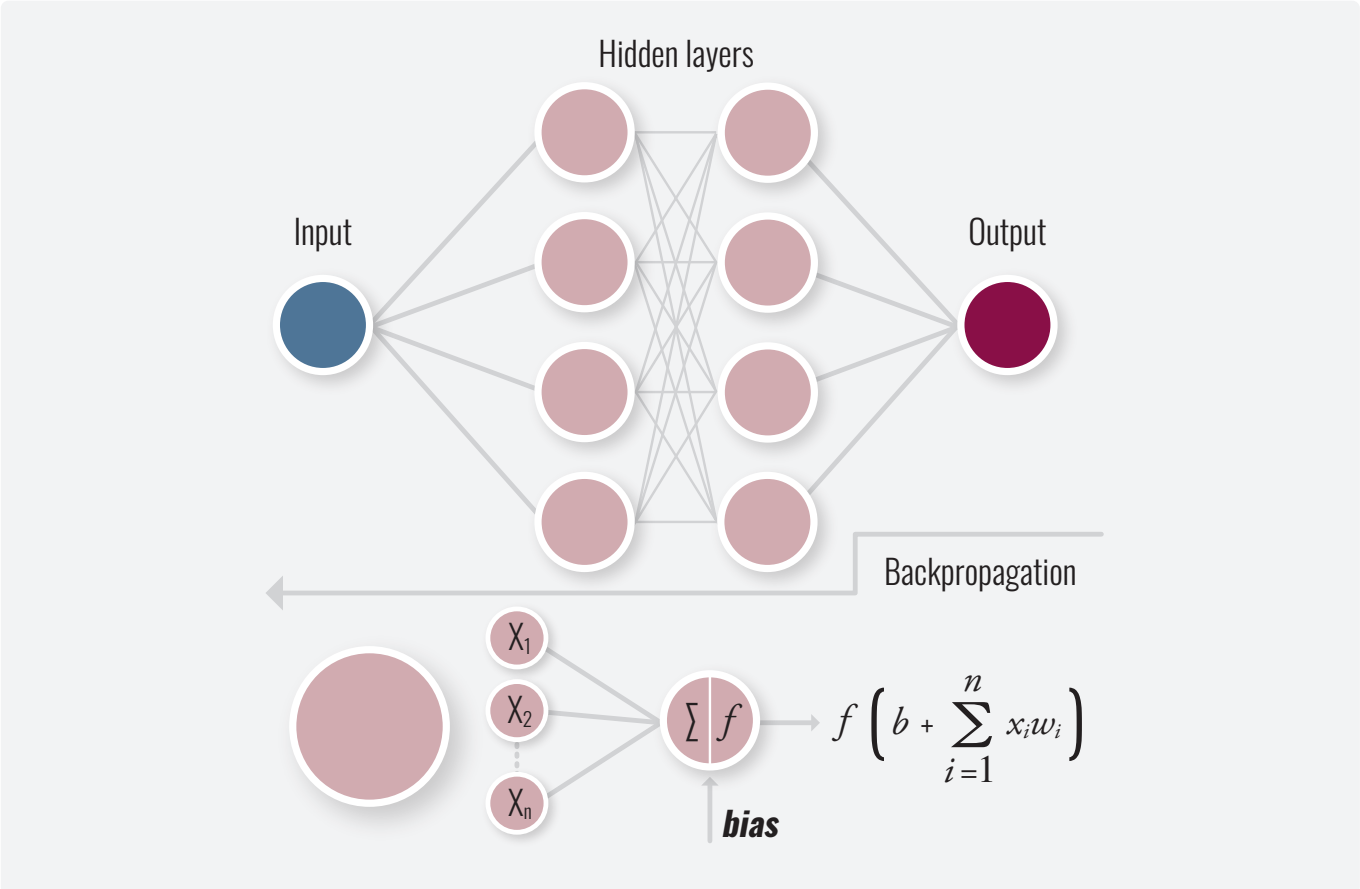
Terminology	Description
Long Short-Term Memory (LSTM) Neural Networks	A special kind of RNN capable of learning long-term dependencies, widely used in tasks involving sequential data and timeseries.
Loss function	A function to measure how well the network is performing with respect to its given training sample and the expected output. It quantifies the disparity between the predicted and actual outcomes, which is what the model seeks to minimize during training.
Machine Learning (ML)	A branch of AI that enables systems to learn and improve from experience without being explicitly programmed.
One-Hot Encoding	A method for representing categorical data as binary vectors, with one element set to 1 and the others set to 0 to indicate the category.
Overfitting	A modeling error in ML which occurs when a function is too closely fit to a limited set of data points and may therefore fail to predict additional data reliably.
Perceptron	A simple type of artificial neuron or node in a neural network, often used as the building block for more complex networks.
Quantitative Structure-Activity Relationship (QSAR) models	Regression or classification models used in the chemical and biological sciences and engineering.
Random Forest (RF), Naive Bayesian (NB), Support Vector Machine (SVM)	ML algorithms used for classification and regression tasks, each with its unique advantages and disadvantages.
Recurrent Neural Networks (RNN)	A type of neural network designed to recognize patterns in sequences of data, such as text, genomes, handwriting, or the spoken word.
Regularization	A technique used in ML to prevent overfitting by adding an additional penalty term to the loss function.
Simplified Molecular Input Line Entry System (SMILE)	A specific line notation for describing the structure of chemical species using short ASCII strings.
Supervised Learning	A type of ML where the model learns from labeled training data and makes predictions based on that learned knowledge.
Underfitting	A situation in ML where a model cannot adequately capture the underlying structure of the data due to its simplicity.
Unsupervised Learning	A type of ML where the model identifies patterns in dataset without any pre-existing labels, often used for clustering and association tasks.

https://doi.org/10.53150/msj.v1i1.39

Among the diverse applications of AI, particular intrigue centers around the utilization of ML algorithms to analyse intricate datasets. Such algorithms are capable to perform pattern recognition in clinical imaging, extraction of fundamental insights from tabular data, and classification of selected outcomes by understanding multivariate relationships. The majority of AI techniques used in drug discovery can be divided into two primary categories: supervised learning and unsupervised learning. Unsupervised learning strategies are frequently employed for exploratory data analysis as they are valuable in identifying hidden patterns in data without pre-labelled information or facilitating data clustering. On the other hand, supervised learning involves training an algorithm with a set of input data to accurately predict specific outputs (such as class labels for classifiers or target values for quantitative outputs) for new, unseen data. In this field, supervised learning may be employed to understand molecular features associated with the bioactivity of compounds; in fact, by training the algorithm with labelled compounds that are either active or inactive, it becomes possible to predict the activity of new pharmacological agents based on their molecular characteristics [7].

ML encompasses several algorithms, most of which have been proven effective in the context of drug discovery. Random Forest (RF), Naive Bayesian (NB) and support vector machine (SVM) are few of the notable examples [8–10], which belong to the class of supervised learning methods. In recent times, Deep Learning (DL) has become popular amongst ML practitioners, due to its intrinsic capability to understand much more complex patterns and relationship. DL is a subset of ML where artificial neural networks with multiple hidden layers (hence the adjective 'deep') are used to model and understand complex patterns in datasets. The main objective of these models, usually referred to as Deep Neural Networks (DNNs) is to mimic the human brain structure, using multiple layers composed of a large number of computational units (perceptrons) **(Figure 1).**

FIGURE 1 - Schematics of a simple Deep Neural Network (DNN); in the bottom window, the mathematic operations of input weighting and output transformation performed by a single neuron (perceptron) are shown.



https://doi.org/10.53150/msj.v1i1.39

Each layer is interconnected upon the previous layer and works towards minimizing the error between the expected and generated outputs, using backpropagation algorithms (e.g., gradient descent) to adjust weights and biases of the model function according to such error measurement. Through many iterations, these hidden parameters are updated and optimized, making the algorithm gradually more accurate. DL models are particularly effective when working with unstructured data such as images, audio, and text, as they become able to automatically learn feature hierarchies and extract most relevant information autonomously, eliminating the need for manual feature extraction which is necessary in traditional ML models. In terms of capability, DL mostly outperform conventional ML algorithms when dealing with complex, heterogeneous data, which is often the case in the domain of healthcare. However, DL models usually relies on large volume of data and their cognitive processes may be difficult to interpret, while other ML models (e.g. decision trees, clustering algorithms) may provide better accuracy when limited data is available (or in particular situations where they are more suitable for solving specific problems).

DNNs may be structured using different architectures, providing flexibility and adaptability to handle complex scenarios. Notably, feedforward networks have been widely used as they bear the simplest layout, being based on forwarding data from input to output in a streamlined manner. On the other hand, deep convolutional neural networks (CNNs) bear layers that are only locally (rather than globally) connected to the next hidden layer, allowing to perform convolutional transformation to hierarchically compose simple local features into complex models. Another interesting architecture is represented by recurrent neural networks (RNNs), evolving through a series of repeating modules of subnetworks. These loops are suitable to analyse dynamic changes over time where persistent information is needed throughout many iterative cycles. Long short-term memory (LSTM) neural networks are a special kind of RNN, widely applied for their capability of learning long-term dependencies from timeseries data. Data clustering with unsupervised learning is whereas carried out using autoencoder neural networks (AENs), which apply backpropagation with the purpose of dimension reduction, aiming at preserving most relevant variables while removing non-essential information. Some examples of the remarkable performance of DNNs in image recognition and classification tasks are reported in literature. Esteva and co-workers [11] developed a model that could perform skin cancer detection with an accuracy comparable to dermatologists, while Gulshan and his group [12] have used retinal images to train a model capable of detecting diabetic retinopathy in a fast and reliable fashion. Other notable examples are DL models that have been developed to predict the risk of various diseases using electronic health records and patient data, such as those developed by Houssein et al. [13] to predict the onset of cardiovascular events using electronic health records, achieving better accuracy compared to traditional models.

In the field of drug discovery, small drugs are designed by modulating the biological activity according to a specific molecular target. The identification of such target must be supported by a plausible therapeutic hypothesis, often related to a desired modulation of the disease state. Upon identifying the optimal target, the selection has to be validated using physiologically relevant ex vivo and in vivo models (target validation). Nowadays, biological research has produced an astonishing amount of data, including human genomics and proteomic, tabular clinical data and high-content imaging of patients. With the advent of ML, computational models have become more sophisticated and able to discern multivariate correlation patterns within highly dimensional datasets. An interesting example of ML in drug discovery is the use of random forest models to predict drug activity against cancer cells based on minimal genomic information and chemical properties[14]. These models have achieved sensitivities and specificities of around 87%, yielding an area under the receiver operating characteristic curve equal to 0.941. They also develop regression models to predict log (IC 50) values of compounds for cancer cells, achieving a Pearson correlation coefficient of 0.86 for cross-validation and up to 0.65–0.73 against blind test sets. In another study, random forest models were used for drug–target interaction prediction via Kullback–Leibler divergence[15]. This method uses E3FP three-dimensional (3D) molecular fingerprints of drugs as a molecular representation, allowing the computation of 3D similarities between ligands within each target (Q–Q matrix) to identify the uniqueness of pharmacological targets. The 3D similarity matrices are transformed into probability density functions via

https://doi.org/10.53150/msj.v1i1.39

Kernel Density Estimation (KDE) as a nonparametric estimation, successfully predicting Drug-Target Interactions (DTIs) for representative 17 targets (mean accuracy: 0.882, out-of-bag score estimate: 0.876, ROC AUC: 0.990).

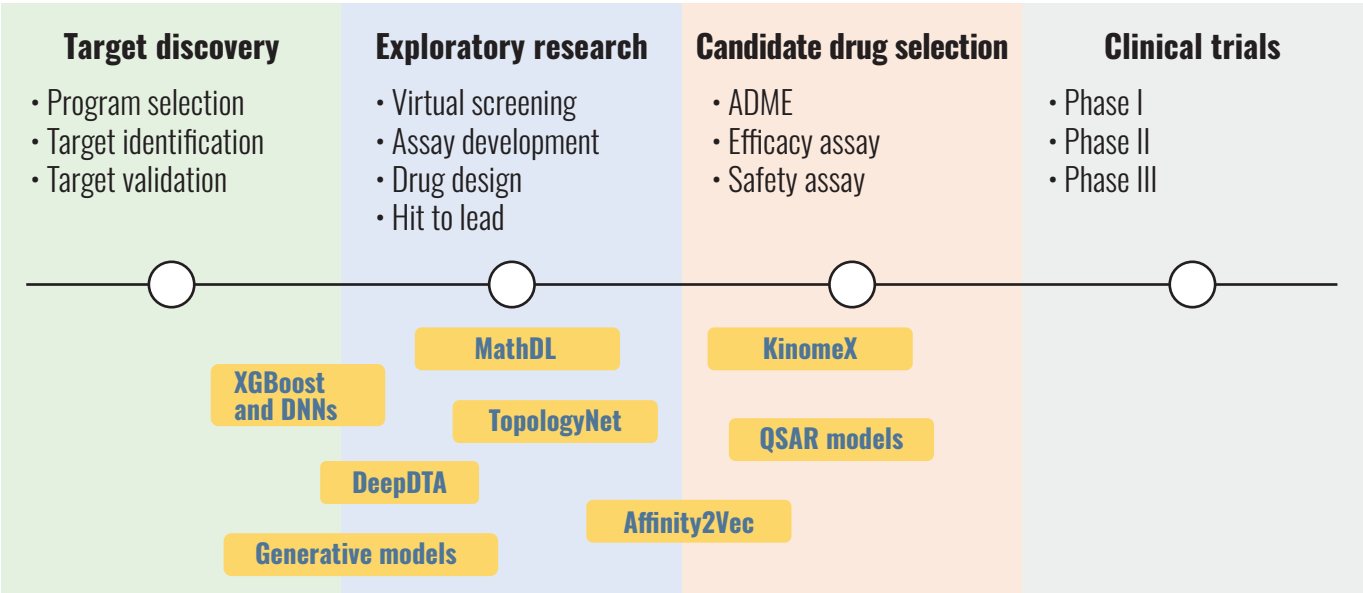
However, either by using DL or other ML methods one must account for considerable drawbacks due to the diversity and uncertainty of the data used as input feed. One clear example is represented by the data scale when considered across multiple variables, leading to a strong necessity to standardize data based on selected criteria. Data pre-processing has a deep influence on the ML outcome and overall performance of the trained model, adding further bias and deviations in the dataset. As the collection of data in the field of drug development can involve millions of compounds, traditional ML tools might not be able to deal with such abundant scale and complexity.

3. Drug Discovery

Drug discovery is a complex and expensive process that involves identifying and developing new drugs to treat medical conditions. The drug development process often begins with extensive research in pharmaceutical sciences. The goal of this step is to identify potential therapeutic targets associated with a particular disease or condition. Once a potential drug candidate is identified, it undergoes preclinical testing in laboratory settings and animal models to evaluate its safety, efficacy, pharmacokinetic and pharmacodynamic profiles. If preclinical studies yield positive results, the drug candidate enters three phases of CTs with an increasing number of participants. In summary, Phase 1 trials often mark the first human testing of a new drug, primarily focusing on establishing safety, tolerability, and appropriate dosage levels for subsequent phases. Phase 2 trials are designed to assess the treatment's effectiveness and safety within a larger group of patients with the specific medical condition of interest. Finally, Phase 3 trials, conducted on a large scale, aim to validate the medication's efficacy, monitor potential side effects, and compare it to established standard treatments or a placebo in a diverse and extensive patient population.

Once these steps are successfully completed, drug developers collect all preclinical and CT data and submit them to regulatory authorities, which ensure that the drug's benefits outweigh its risks and meet strict safety and efficacy standards. Eventually, upon approving the medicine, it can be marketed and made available to healthcare professionals and patients. DL methodologies are able to assist drug design by predicting optimal molecules from previously learned relationship amongst large datasets that may include chemical structures, biological activities, pharmacokinetics, and toxicological profiles. Deep models can optimize every step involved in the long process of drug discovery, from the identification of target to the analysis of CTs data. **(Figure 2).**

FIGURE 2 - Drug discovery phases and Deep Learning models which may be used.



https://doi.org/10.53150/msj.v1i1.39

While DL methods have shown outstanding potential in the domain of drug discovery, traditional ML models may still hold an advantage in certain research scenarios. For instance, decision trees are frequently employed in drug discovery due to their interpretability [16,17]. They can be utilized to identify crucial features that contribute to a drug's effectiveness. The branches of the tree can provide insights into the decision-making process, such as "if a drug has feature X and Y, it is likely to be effective" [16,17]. Moreover, ML models require significantly less training data compared to neural networks and are often less computationally demanding than DL methods [16]. Support Vector Machines (SVMs), for example, have been used to predict drug toxicity with limited data[18]. Logistic regression, a relatively simple and computationally efficient ML model, can be used for binary classification problems in drug discovery, such as predicting whether a compound will be active or inactive against a specific biological target [19].

Usually, datasets available for drug development in pharmaceutical companies include millions of compounds. Quantitative structure-activity relationship (QSAR)-based computational model can easily predict large numbers of compounds or simple physicochemical parameters, but their accuracy may vary when predicting complex biological properties, such as the efficacy and adverse effects of drugs. In addition, QSAR-based models also suffer from small training sets, experimental data error in the latter and lack of experimental validations. In 2012, Merck supported a QSAR ML challenge to endorse the deployment of DL methodologies in the drug discovery process, showing that those are significantly better at predicting absorption, distribution, metabolism, excretion, and toxicity (ADMET) of drug candidates, when compared to traditional ML methods [20].

As it has been mentioned earlier in this review, DL methodologies have shown great promise in the field of drug discovery. However, their implementation is not without challenges. One of the primary hurdles is the requirement for large and diverse datasets of high-quality chemical and biological data. In drug discovery, obtaining such data can be challenging due to issues such as experimental noise, missing values, data imbalance, data heterogeneity, and data privacy [21]. Another challenge lies in the interpretability of DL models. Often considered as black boxes, these models do not provide much insight into how they make predictions or what features they use [21,22]. This can limit their usefulness in drug discovery, where understanding the molecular mechanisms and the reasoning behind predictions is crucial for generating new hypotheses and designing new experiments. DL models can sometimes be prone to overfitting, a phenomenon where the model learns the training data too well, to the point that it performs poorly when presented with new or unseen data. This can lead to false positives or false negatives in drug discovery, where the chemical space is extremely vast and complex, and many molecules share similar physicochemical properties. Overfitting is directly opposed to underfitting, where the model is not powerful enough to minimize the error between true labels and predicted labels and therefor extract any useful information from the given data. Lastly, DL models need to be rigorously validated and evaluated using appropriate metrics and methods to ensure their predictive power and applicability domain. However, there is no consensus on how to best validate and evaluate DL models in drug discovery, especially when dealing with imbalanced or sparse data, multiple targets or tasks, or novel compounds [23].

3.1 Chemical libraries

Chemical libraries are repositories of molecules which are largely used in chemical industries and academic centres. Molecules included in the database are atomized into several descriptors, such as structural and physicochemical ones, providing information on chemical structure, molecular weight, atoms and bonds type, as well as solubility and acidity/basicity. Depending on the library, the wealth of information about each molecule may include pharmacokinetics (how the compound is absorbed, distributed, metabolized, and excreted by the body), pharmacodynamics (its biochemical and physiological effects), and toxicology. Other potential features could be the synthetic accessibility of the molecule (how easy it is to synthesize), commercial availability, or even its predicted activity against specific biological targets. Moreover, with the advancement of cheminformatics, new methods for molecular description have been developed. These include various 2D and 3D molecular descriptors, such as path-based fingerprints, extended-connectivity fingerprints, 2D pharmacophore fingerprints and extended 3D fingerprints [24].

Data for pharmacological features are manually extracted

https://doi.org/10.53150/msj.v1i1.39

from published literature and are routinely updated. Moreover, regulatory agency documents are periodically checked for schedule of administration, indications and warnings of drugs. At present, some compound databases are available online and extensively used in DL (i.e., PubChem, ChEMBL), containing over 105 million compound candidates [25,26]. Such databases have integrated advanced information regarding drugs biological activity, most in the form of QSAR descriptors [27]. **(Table 2)**

TABLE 2 - Molecular descriptors available in chemical libraries based on their dimensionality [27].

Descriptor Dimensions	Properties
NON-DIMENSIONAL	• Molecular weight • Atom number • Atom-type count
1D DESCRIPTORS	• Functional groups • Substituent atoms
2D DESCRIPTORS	• Molecular topology • Connectivity bonds
3D DESCRIPTORS	• Steric properties • Molecular geometry • Surface area and volume • Binding site properties

In particular, the ChEMBL database maintained by the European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL), contains 1.6 million distinct compounds, 14 million bioactivities, 11 thousand biological targets, and other related data. Furthermore, it is equipped with toolkits for data mining [28], including tailored resources for specific tasks, as for example Kinase SARfari (chemogenomics workbench focused on kinases) or ADME SARfari (tool for predicting and comparing cross-species ADME targets). Another notable database is the QM9, which is a widely used in the field of computational chemistry and includes quantum mechanical calculations for a diverse set of small organic molecules [29]. The QM9 dataset provides important molecular properties such as atomization energies, equilibrium geometries, dipole moments, and other complex molecular parameters. It has been used for the development and validation of ML models for molecular property prediction and drug discovery.

While these databases may comprise millions of different molecules, they are far from covering the entire chemical space. As a matter of fact, both empirical and simulated molecules may exist in available databases. In this context, empirical molecules refer to those that have been experimentally observed and characterized in the laboratory, which data is derived from actual experimental results, and their properties and behaviours are known based on empirical evidence. On the other hand, simulated molecules are those that have been predicted or generated using computer simulations, such as molecular dynamics simulations using mathematical models and algorithms to predict the properties and effects of molecules according to their

https://doi.org/10.53150/msj.v1i1.39

atomic composition and structure. Predicted molecules are often collected in designated databases known as virtual libraries. Such libraries may be divided in static and dynamic [30]. Static libraries aim at enumerating all possible virtual molecules that may exist in a specific field. As an example, GDB-17 report about 116 billion virtual organic molecules [31]. Other libraries, such as ZINC, has far less compounds (around 22 million) but are free, focused on ready-made molecules and provide three- dimensional conformations, therefore are widely used in ligand- and structure-based virtual screening studies [32]. Aiming at completeness and performing a screening afterward as in static libraries may be good strategies to systematically screen chemical space; however, it must be considered that libraries by definition may not be completed, and the largest the library, the most difficult it becomes to perform a screening. For this reason, dynamic libraries have been developed, which may be seen as an algorithm capable of investigating only the chemical space around molecules of user interest, thus enabling fastest and more efficient screening. A common example is PINGUI, a free tool which identifies reactive site of a molecule and recombines resulting fragments as building blocks to complement the core fragment and generating a tailored chemical space defined by the scientist [33]. Both types of libraries have pros and cons and should be used combinedly. Future challenges will involve how to store and screen such a large amount of data and how to make them available to a larger audience for research purpose [30].

3.2 Virtual Screening of molecules

In the last decades, virtual screening (VS) has emerged as a more efficient way to physical screening; it consists **of** a computational screening of large libraries of molecules which compared biochemical properties, such as structure similarity, and gave a rank of most promising drugs [34,35]. Overall, VS strategies may be divided in two groups:

1. ligand-based methods: starting from an already existing input molecule, these methods try to find similar molecules in an extensive chemical library basing on atoms types and reciprocal connection and three-dimensional configuration.
2. structure-based methods: in this case, the search in chemical libraries aims at finding pharmacological agents that may fit a known binding site. Since these latter methods do not require an input compound, they are more useful than ligand-based ones in case of biological target with no binders available yet. For the same reason, as a drawback, structure-based virtual screening may be less accurate [35].

Ligand-based methods have revolutionized the field of drug discovery by delivering new drug candidates more quickly and at a lower cost. These methods allow for rapid VS of molecules, utilizing pharmacophoric techniques and alignment methods based on ligand shape and electrostatic similarity. Such techniques have proven their efficiency in identifying novel potentially active compounds [36,37]. On the other hand, structure-based techniques have unveiled a plethora of potential drug targets. These techniques facilitate enhanced meticulous target identification and validation, thereby reducing efficacy-related drug attrition. Furthermore, once the crystal structure of a target is obtained through structural biology techniques such as X-ray crystallography, Nuclear Magnetic Resonance (NMR), neutron crystallography (NC), cryo-Electron Microscopy (cryo-EM), and mass spectrometry (MS) among others, researchers can swiftly establish the structure-activity relationship of the compound. This accelerates the process and reduces the number and time of compound synthesis [38,39].

Overall, VS allows for a fast screening of large chemical libraries going beyond empirical screening capabilities and pragmatically providing lists of likely candidates for further studies. However, some challenges must still be faced. First, spatial conformation of molecules is not fixed: molecular flexibility in aqueous environment and dynamic structure of receptors make it far more complex to establish ligand-receptor absolute binding energies and energetically accessible conformations of receptors; this process may require relevant resources in terms of time and CPU power [34,35]. Second, although efficient, VS may still make mistakes by incorrectly missing a true efficacy of an agent or, more importantly, by anticipating activity of an inactive drug. In drug development, even a minimal false positive rate of VS may lead to a large number of compounds tested, with both an increase of expenses for testing and an obscuration of the signal of truly active molecules [35,40].

DL methods play a vital role in evaluating physicochemical properties, target affinity, and pharmacokinetic (PK) profiles of potential drug candidates, becoming a valuable tool to speed up the drug discovery process. In particular, DTI prediction has been of great help for drug repositioning

https://doi.org/10.53150/msj.v1i1.39

and virtual drug screening. Binding affinity prediction has been explored using DL methods. One of the biggest challenges when manipulating the complex molecular data consist in the identification of the most optimal encoding method. Molecules are often encoded according to the Simplified Molecular Input Line Entry System (SMILES), which represent and efficient way of storing information from the molecular graph using string characters [41]. SMILES allows for the linearization of the molecular graph by enumerating the nodes and edges on the bases of a certain path. However, it is affected by the randomness attributed to the selection of the starting atom in the 2D graph, meaning that multiple SMILES may exist for one molecule. For example, the canonical SMILE notations for water and ethanol are O and CCO, respectively. [42,43] SMILES are usually taken in consideration as they may be standardized through canonization algorithm. This latter is a procedure that generates a unique and unambiguous representation of a molecule, usually by assigning a priority to each atom based on its connectivity, atomic number, chirality, and other properties, and then generating a SMILES string that follows the priority order. While there are different canonization algorithms, they all aim to ensure that the same molecule always has the same canonical SMILES string. Another method is represented by labelling and "one hot encoding", a process by which categorical variables are converted into a new categorical feature with binary values assigned (1 if present or 0 if absent), allowing to represent each integer value as a binary vector [44].

Several DL methods have shown to outperform conventional ML approaches, due to their generally better capability of handling high-dimensional data that is particularly useful in domains with large datasets including hundreds of features. DL methods are also largely superior at pattern recognition, which proves to be extremely helpful when discerning patterns and discriminative features in molecules with complex structure and topology. In some cases, DL models are enhanced by using algorithms that have been widely validated in conventional ML, as is the case of Cheng Chen and co-workers, who have applied XGBoost algorithm together with multi-layered DNNs to build a drug-target interactions predictor, achieving an accuracy above 98% and outclassing other state-of- the-art prediction systems [45]. A real-life example is high-performance DL models is brought by the Affinity2Vec, a drug-target binding affinity prediction model based on representation learning [46].

Another network-based approach is DeepDTA [47], which uses a heterogeneous graph attention (HGAT) model coupled with bidirectional ConvLSTM to learn topological information of compound molecules and modelling spatial-sequential information based on the molecular SMILES sequences. This is yet a further examples of superior deep networks performances, due to the ability to learn hierarchical representations. This means that they can learn multiple levels of abstraction from data, where lower layers hold information about general molecular shape and topology, while deeper layers may gain insight on more complex physicochemical properties. Beside the model layout, other approaches may be undertaken to improve the capability of DL; such is the case of two models, namely MathDL and TopologyNet, which make use of algebraic topology to identify interactions between protein and ligand. In particular, MathDL [48] exploited advanced mathematical techniques such as graph theory to encode the physicochemical interactions into lower-dimensional rotational and translational invariant representations. TopologyNet [49] is created as an ensemble of multi-channel topological CNNs to represent the protein-ligand complex geometry through 1D topological invariants for affinity prediction and protein mutation. Finally, some DL applications have been focused on specific segments of the drugs chemical space, such is the case of KinomeX. The latter is an online platform to predict polypharmacology effects of kinases solely based on their chemical structures. A multi-task DNN model trained with over 140 000 bioactivity data points for 391 kinases carries out predictions for the users, enabling them to create a comprehensive kinases interaction network for designing novel chemical modulators [50].

3.3 De Novo Drug Design

It has been estimated that the total number of organic compound that may potentially be synthesized ranges from 10^{30} to 10^{60} [35]. Searching for a potentially useful drug amidst this vast compound space is akin to looking for a needle in a haystack for pharmaceutical industries, as it accounts for a large amount of money and time. For this reason, rational de novo design approach has always been used. De novo drug design essentially involves the creation of new chemical structures with specific desired properties, such as a particular biological response.

Traditionally, computational methods proceeded one

https://doi.org/10.53150/msj.v1i1.39

molecular fragment at a time to highlight worse or better biochemical properties of the new drug. This process may obviously benefit from automated methods for constructing these novel structures. This is the reason why ML may be extremely helpful in this field as well and Generative Models (GMs) have been proposed. A generative model is a ML algorithm which may retrieve data of existing chemical compounds to identify patterns and chemistry rules which may be employed to generate new molecules [51]. Technically, GMs are made of three parts linked by a neural network: first, an encoding module converts a set of molecules into a continuous vectorized representation; then a decoding module reconstructs the continuous vector representation back into a molecule; lastly, a predictive module computes one or multiple properties for vectors derived from the continuous representation [35]. Gomez-Bombarelli and colleagues were among the first to propose this new method based on encoding of compounds and showing good prediction power [52].

The choice of molecular representation to be presented to the encoder significantly influences the learning process of the model, determining how molecular information is acquired. There are three primary types of representations: 3D (e.g., coordinate-based), two-dimensional (e.g., molecular graphs) and one-dimensional. The commonly used 1D representation is SMILES notation, which is particularly suitable for neural network architectures designed for language processing [7]. For example, a generative neural network trained on SMILES was recently developed to de novo design a ligand for nuclear receptor related 1 (Nurr1), a transcription factor involved in neurodegenerative disease pathways; these strategies led to the synthesis of two candidates with desired activity, one of which with a notable potency, even if this network was trained with a relatively limited number of molecules due to the lack of active drugs in this setting [53].

However, ML models often struggle to fully comprehend the intricacies of SMILES grammar, resulting in the generation of invalid SMILES that cannot be translated into meaningful molecular structures; therefore, other strategies have been proposed [54]. Similarly, considering molecular 3D structure may represent an issue, as different forcefields and interactions are involved in determining spatial conformation. To address this challenge, recent efforts have focused on training 3D generative models using extensive sets of conformers. For example, the geometric ensemble of molecules (GEOM) comprises over 37 million molecular conformations for about half million molecules [55].

3.4 Chemical synthesis of drugs

Not only must new drugs be designed, but they have also to be effectively synthesized; only a relatively small number of chemical reactions have been shown to be used in recent years and few have been introduced compared to 40 years ago. On the one hand, these commonly used reactions are reliable and have led to a wide range of commercially available building blocks that can be effectively exploited. On the other, despite the growth in synthesized drugs, the limited variety of reactions used has led to certain regions of the chemical space being densely populated with structurally similar molecules, while other regions remain unexplored [56].

Recent developments include the use of AI to process large databases of organic reactions and propose new synthetic pathways. This process consists of two activities: research and reaction prediction. The research involves identifying a series of chemical reactions to form a retrosynthetic pathway between a target compound and starting materials. Afterwards, reaction prediction determines the feasibility of those reactions basing on the context [51,57].

Two approaches may be used to investigate steps of organic synthesis: template-based and template-free methods. Template-based methods use hand-coded reaction templates to describe determined steps of bond formation among atoms: they have been widely used for decades, but can be computationally expensive and limited by the quality of the templates [58]. More recently, template-free methods have been introduced: they use neural networks to learn the relationship between reactants and products. In particular, neural networks consider chemical blocks as a sequence of characters, such as in the SMILES strings. After a proper training on a large database of chemical compounds, the neural network tries to generate reactants given the proposed products, thus suggesting new chemical reactions [59]. Additionally, neural networks may be trained by associating a large number of molecules to known originating reactions in order to predict the probability of a further reaction to produce the desired product [60]. However, assessing the feasibility of unprecedented reactions can be difficult, as the same reaction in different conditions (i.e. time, temperature and catalysts) may lead to different yields; therefore, it may still takes resources to be validated [51].

https://doi.org/10.53150/msj.v1i1.39

4. Drug repurposing

Drug repurposing is the process of identifying new therapeutic applications for existing drugs. This strategy is achievable as drugs may have several targets, which may induce different effects according to contexts and diseases. Drug repositioning has various advantages over traditional drug development, including shorter development times and lower costs, given that the safety and pharmacokinetic profile of the drugs have already been established [61].

However, finding novel uses for currently available medications is a difficult endeavour as it necessitates a thorough comprehension of the drug's mechanism of action as well as the underlying molecular pathways of the condition of interest. AI has emerged as a viable technique for drug repurposing, as it can scan vast volumes of data on pharmacological characteristics and disease pathways in order to find new therapeutic targets and indications. As an example, drug repositioning strategies were proposed in the first phases of Coronavirus disease (COVID-19) pandemics. In this setting, challenges in drug repurposing included limitations of preclinical assays, suboptimal CT designs, lack of appropriate clinical endpoints, and the absence of reproducible preclinical animal models. ML was thought to provide in the shortest time pharmacological agents to treat the disease while bypassing traditional development steps required for new drugs [62]. Similarly, an integrative deep network that combined multiple relationships and leveraged a vast amount of information embedded as vectors from the PubMed and DrugBank databases was proposed [63]. 41 drugs, including dexamethasone and indomethacin, were predicted to have repurposing potential as therapeutic agents against for treating SARS-CoV-2.

Furthermore, several AI-based drug repurposing methods have been developed, including ML algorithms, network-based approaches, and natural language processing methods. These strategies typically involve the integration of multiple data sources and the use of advanced statistical and computational techniques to identify potential drug-disease associations. One of the main benefits of AI-based drug repurposing is the ability to integrate multiple data sources with Real-World data, such as electronic health records and CT data, to identify new drug-disease associations that might not be immediately obvious using traditional approaches and statistical methods which are limited in handling a substantial amount of data. Liu and colleagues proposed a framework for screening of on-market drug candidates by retrospectively analysing Real-World data [64]. In particular, they emulated randomised CTs to systematically evaluate drug efficacy on a panel of diseases; moreover, DL techniques were used to control for confounders in real-world data through a propensity score estimation model. Additionally, they provide an example which demonstrates the effectiveness of the proposed computational drug repurposing framework in identifying potential drug candidates with beneficial effects on disease outcomes for coronary artery disease patients, even outreaching the performances of other existing pre-clinical drug repurposing methods. This new approach may be exploited whenever real RCTs are not available despite a large volume of observational data.

However, further progresses are required. Data and model harmonization are essential for the development of broadly applicable and interoperable computational tools for drug repositioning. Similarly, data security and privacy concerns must be addressed through careful consideration of the data lifecycle and the implementation of regulations and transparency [62].

5. Clinical Trials

CTs are the gold standard to prove efficacy of a drug and are required by regulatory agencies for releasing marketing authorisation. Although CTs require a large amount of time and resources to be performed, a positive result is not guaranteed even in case of a true effect of the drug; in fact, many pitfalls are described in terms of patients' enrolment, study design, trial conduction and phase transitions, which may lead to trial unsuccess. Recently, ML has been discussed as a way to overcome these problems and offering several opportunities for improving the efficiency of different trial process steps [65,66].

First, patients' enrolment is one of the major challenges in CT conduction. This is due to both complex protocol designs, which makes it difficult to include subjects and fully adhere to the conduction, and lack of interest of the patients. It has been shown that only 15% of CTs manage to entirely avoid patients dropout, while average dropout rates are about 30% [66]. Notably, a poor recruitment with a high number of dropouts may lead to trial discontinuation due to the inability to demonstrate the outcome. In this setting, ML algorithms can identify potential CT participants by analysing electronic health records, large datasets, and patient files. By more effectively identifying the right

https://doi.org/10.53150/msj.v1i1.39

participants, they can help streamline the recruitment process and speed up trial enrolment, thus leading to faster and more reliable results [67]. Additionally, this may reduce early stopping of CTs due to insufficient patient enrolment or high rates of dropout. It has been suggested that AI may predict likelihood of dropouts from CT: this information may be used to focus effort on these patients to provide additional education to encourage longer participation [68,69].

Second, CT design may be optimized as well under different points of view. It has been shown that AI may simulate different scenarios according to trial design, sample size, randomization strategy, and statistical analysis plan and evaluate their likely outcomes [66]. Moreover, AI may even permit the substitution of control group with an artificial one in future: given a large amount and variety of data from each enrolled patient, AI might predict individual natural history and disease progression; in particular, this study design emulates the impact of placebo on virtual patients and compares it to the intervention group of RCT; by employing synthetic control arms, this approach ensures that all enrolled participants receive the experimental intervention, thereby addressing concerns related to treatment assignment and the potential for unblinding [69,70]. However, these strategies need to be fully validated in order to ensure an efficacy comparable with traditional trials.

Third, trial conduction may be improved as well. ML may provide real-time monitoring of patient data to ensure safety and efficacy parameters are met. Continuous data analysis might identify patterns, outliers, adverse events and treatment responses in a timely manner, thus enabling prompt intervention and improved patient safety. Non-stop monitoring may be further permitted by wearable devices which monitor patient's parameters 24 hours a day, thus reducing number of missing data points [71]. Even traditional tests performed during CTs may be affected: ML techniques can automate the evaluation of trial endpoints, such as analysis of radiographic images or pathology slides, thus reducing manual workload, and enabling more efficient data analysis. For instance, Erdaw et al. developed a model to accurately make diagnosis of COVID-19 using digital chest X-ray image. Remarkably, the model achieves an accuracy of over 97% [72].

Therefore, by integrating patients' characteristics, historical data, biomarkers and treatment outcomes, ML may help optimize treatment options by identifying patient subgroups that may respond differently to interventions; this may be relevant considering the increasing interest raised by personalised medicine in order to tailor pharmacological treatment to each patients' characteristics [73].

Lastly, ML can be employed to assess the likelihood of success during CT phase transitions. This aspect holds great importance considering that only one out of every five drugs that enter phase 1 CTs ultimately receives marketing authorization [66]. Specific algorithms have shown an average accuracy of approximately 80% in predicting the outcome of these transitions. This capability can offer advantages to both trialists, who can utilize the information to refine protocol designs, and pharmaceutical industries, which can save resources and allocate budgets more effectively [74].

6. Conclusions

ML methodologies can have a deep impact on drug development; DL, especially through the use of generative models, has demonstrated significant advancements in drug design and investigating protein-ligand binding processes, leading to a renewed enthusiasm in the field.

Despite progress made in recent years, at the moment GMs are unlikely to automatically produce drug candidates with optimal properties due to the complexity of molecular interactions. In fact, these interactions hinge on the three-dimensional conformation of receptors, which, in turn, relies both on the structural composition of aminoacidic chain, and environmental factors, including pH, temperature and oxidation. These factors influence molecular flexibility, receptor affinity, and other features that algorithms may not consistently predict automatically in every instance.

For these reasons, the input of human expertise continues to be crucial. Therefore, further advancements in next years are required in terms of:

1. data curation and quality,
2. appropriate model validation and ways to handling uncertainty in predictions;
3. improving data accessibility, ensuring clearer interpretation, better readability, and the potential for reuse, with the aim of reducing resources waste, accelerating innovation and providing new drugs for unmet clinical needs.

Automation and combining multiple approaches - generative, predictive, synthesis planning - will surely be promising research directions. Another challenge is represented by performance evaluation: comparing

https://doi.org/10.53150/msj.v1i1.39

experimentally determined and predicted values for physical properties or biological activity may represent a benchmark for ML models evaluation; additionally, comparison among techniques should be performed as well, in order to minimise costs and maximise results.

The final phases of drug development have also benefited from these new techniques. Incorporating AI in CTs holds great promise, as it is predicted to advance medical research and make it more sustainable. Continuous efforts have been made to investigate the potential of AI in order to conduct more efficient and profitable trials. To fully validate these methods, however, much more research is needed. Additionally, as there is still a lack of specific ethical and regulatory guidance focused on AI's use in CTs, adoption of digital transformation will likely be cautious and slow. Sponsors, investigators, and regulators may successfully tackle all of these issues by working together closely and using a patient-centered, ethical strategy.

In conclusion, a considerable amount of effort is still needed to fully integrate AI algorithms into standard drug discovery processes. However, this emerging technology holds the potential to substantially enhance drug development, even in areas where unmet medical needs persist. In this context, there is a reasonable expectation that ML may become a reliable tool to effectively address the challenges that both basic and clinical researchers currently have encountered for years.

Conflict of Interest

The authors declare no competing interests.

References

1. *Brown DG, Wobst HJ, Kapoor A, Kenna LA, Southall N. Clinical development times for innovative drugs. Nature Reviews Drug Discovery. 2021 Nov 10;21(11):793–4.*
2. *DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. Journal of Health Economics. 2016 May 1;47:20–33.*
3. *Wouters OJ, McKee M, Luyten J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. JAMA. 2020 Mar 3;323(9):844–53.*
4. *Schlander M, Hernandez-Villafuerte K, Cheng CY, Mestre-Ferrandiz J, Baumann M. How Much Does It Cost to Research and Develop a New Drug? A Systematic Review and Assessment. Pharmacoeconomics. 2021 Nov;39(11):1243–69.*
5. *Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics. 2019 Apr 1;20(2):273–86.*
6. *Chen W, Liu X, Zhang S, Chen S. Artificial intelligence for drug discovery: Resources, methods, and applications. Molecular Therapy - Nucleic Acids. 2023 Mar 14;31:691–702.*
7. *Cerchia C, Lavecchia A. New avenues in artificial-intelligence-assisted drug discovery. Drug Discov Today. 2023 Apr;28(4):103516.*
8. *Breiman L. Random Forests. Machine Learning. 2001 Oct 1;45(1):5–32.*
9. *Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995 Sep 1;20(3):273–97.*
10. *Sammut C, Webb GI, editors. Encyclopedia of Machine Learning and Data Mining [Internet]. Boston, MA: Springer US; 2017 [cited 2023 Jun 30]. Available from: https://doi.org/10.1007/978-1-4899-7687- 1_100507*
11. *Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb 2;542(7639):115–8.*
12. *Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016 Dec 13;316(22):2402–10.*
13. *Houssein EH, Mohamed RE, Ali AA. Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques. Sci Rep. 2023 May 3;13(1):7173.*
14. *Lind AP, Anderson PC. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. PLOS ONE. 2019 Jul 11;14(7):e0219774.*
15. *Ahn S, Lee SE, Kim M hyun. Random-forest model for drug–target interaction prediction via Kullback– Leibler divergence. Journal of Cheminformatics. 2022 Oct 3;14(1):67.*
16. *Xuan P, Sun C, Zhang T, Ye Y, Shen T, Dong Y. Gradient Boosting Decision Tree-Based Method for Predicting Interactions Between Target Genes and Drugs. Frontiers in Genetics [Internet]. 2019 [cited 2023 Oct 20];10. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2019.00459*
17. *Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine Learning in Drug Discovery: A Review. Artif Intell Rev. 2022 Mar 1;55(3):1947–99.*
18. *Rodríguez-Pérez R, Bajorath J. Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. J Comput Aided Mol Des. 2022 May 1;36(5):355–62.*

19. Gala DV, Gandhi VB, Gandhi VA, Sawant V. Drug Classification using Machine Learning and Interpretability. 2021 Smart Technologies, Communication and Robotics (STCR). 2021 Oct 9;1–8.
20. Ciallella HL, Zhu H. Advancing Computational Toxicology in the Big Data Era by Artificial Intelligence: Data-Driven and Mechanism-Driven Modeling for Chemical Toxicity. Chem Res Toxicol. 2019 Apr 15;32(4):536–47.
21. Whang SE, Roh Y, Song H, Lee JG. Data Collection and Quality Challenges in Deep Learning: A Data- Centric AI Perspective [Internet]. arXiv; 2022 [cited 2023 Oct 20]. Available from: http://arxiv.org/abs/2112.06409
22. Nag S, Baidya ATK, Mandal A, Mathew AT, Das B, Devi B, et al. Deep learning tools for advancing drug discovery and development. 3 Biotech. 2022 Apr 9;12(5):110.
23. Koutroumpa NM, Papavasileiou KD, Papadiamantis AG, Melagraki G, Afantitis A. A Systematic Review of Deep Learning Methodologies Used in the Drug Discovery Process with Emphasis on In Vivo Validation. International Journal of Molecular Sciences. 2023 Jan;24(7):6573.
24. Bragina ME, Daina A, Perez MAS, Michielin O, Zoete V. The SwissSimilarity 2021 Web Tool: Novel Chemical Libraries and Additional Methods for an Enhanced Ligand-Based Virtual Screening Experience. International Journal of Molecular Sciences. 2022 Jan;23(2):811.
25. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012 Jan;40(Database issue):D1100-1107.
26. Li Q, Cheng T, Wang Y, Bryant SH. PubChem as a public resource for drug discovery. Drug Discov Today. 2010 Dec;15(23–24):1052–7.
27. Todeschini R, Consonni V. Handbook of Molecular Descriptors [Internet]. John Wiley & Sons, Ltd; 2008 [cited 2023 Jun 30]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527613106.fmatter
28. Bellis LJ, Akhtar R, Al-Lazikani B, Atkinson F, Bento AP, Chambers J, et al. Collation and data-mining of literature bioactivity data for drug discovery. Biochem Soc Trans. 2011 Oct;39(5):1365–70.
29. Glavatskikh M, Leguy J, Hunault G, Cauchy T, Da Mota B. Dataset's chemical diversity limits the generalizability of machine learning predictions. J Cheminform. 2019 Nov 12;11(1):69.
30. van Hilten N, Chevillard F, Kolb P. Virtual Compound Libraries in Computer-Assisted Drug Discovery. J Chem Inf Model. 2019 Feb 25;59(2):644–51.
31. Ruddigkeit L, van Deursen R, Blum LC, Reymond JL. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model. 2012 Nov 26;52(11):2864–75.
32. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: a free tool to discover chemistry for biology. J Chem Inf Model. 2012 Jul 23;52(7):1757–68.
33. Chevillard F, Rimmer H, Betti C, Pardon E, Ballet S, van Hilten N, et al. Binding-Site Compatible Fragment Growing Applied to the Design of ß2-Adrenergic Receptor Ligands. J Med Chem. 2018 Feb 8;61(3):1118–29.
34. Shoichet BK. Virtual screening of chemical libraries. Nature. 2004 Dec 16;432(7019):862–5.
35. Walters WP. Virtual Chemical Libraries. J Med Chem. 2019 Feb 14;62(3):1116–24.
36. Vázquez J, López M, Gibert E, Herrero E, Luque FJ. Merging Ligand-Based and Structure-Based Methods in Drug Discovery: An Overview of Combined Virtual Screening Approaches. Molecules. 2020 Jan;25(20):4723.
37. Umar AB, Uzairu A, Shallangwa GA, Uba S. Ligand-based drug design and molecular docking simulation studies of some novel anticancer compounds on MALME-3M melanoma cell line. Egyptian Journal of Medical Human Genetics. 2021 Jan 18;22(1):6.
38. Batool M, Ahmad B, Choi S. A Structure-Based Drug Discovery Paradigm. International Journal of Molecular Sciences. 2019 Jan;20(11):2783.
39. Herrera-Acevedo C, Perdomo-Madrigal C, de Sousa Luis JA, Scotti L, Scotti MT. Drug Discovery Paradigms: Target-Based Drug Discovery. In: Scotti MT, Bellera CL, editors. Drug Target Selection and Validation [Internet]. Cham: Springer International Publishing; 2022 [cited 2023 Oct 20]. p. 1–24. (Computer-Aided Drug Discovery and Design). Available from: https://doi.org/10.1007/978-3-030-95895-4_1
40. Jain AN. Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. J Comput Aided Mol Des. 1996 Oct;10(5):427–40.
41. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988 Feb 1;28(1):31–6.
42. PubChem. Ethanol [Internet]. [cited 2023 Oct 25]. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/702
43. PubChem. Water [Internet]. [cited 2023 Oct 25]. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/962
44. Goodfellow I, Bengio Y, Courville A. Deep Learning [Internet]. MIT Press; 2016. Available from: http://www.deeplearningbook.org
45. Chen C, Shi H, Jiang Z, Salhi A, Chen R, Cui X, et al. DNN-DTIs: Improved drug-target interactions prediction using XGBoost feature selection and deep neural network. Comput Biol Med. 2021 Sep;136:104676.
46. Thafar MA, Alshahrani M, Albaradei S, Gojobori T, Essack M, Gao X. Affinity2Vec: drug-target binding affinity prediction through representation learning, graph mining, and machine learning. Sci Rep. 2022 Mar 19;12(1):4751.
47. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. Bioinformatics. 2018 Sep 1;34(17):i821–9.

48. Nguyen DD, Gao K, Wang M, Wei GW. MathDL: mathematical deep learning for D3R Grand Challenge 4. J Comput Aided Mol Des. 2020 Feb;34(2):131–47.
49. Cang Z, Mu L, Wei GW. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. PLoS Comput Biol. 2018 Jan;14(1):e1005929.
50. Li Z, Li X, Liu X, Fu Z, Xiong Z, Wu X, et al. KinomeX: a web application for predicting kinome-wide polypharmacology effect of small molecules. Bioinformatics. 2019 Dec 15;35(24):5354–6.
51. Walters WP, Barzilay R. Critical assessment of AI in drug discovery. Expert Opin Drug Discov. 2021 Sep;16(9):937–47.
52. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Cent Sci. 2018 Feb 28;4(2):268–76.
53. Ballarotto M, Willems S, Stiller T, Nawa F, Marschner JA, Grisoni F, et al. De Novo Design of Nurr1 Agonists via Fragment-Augmented Generative Deep Learning in Low-Data Regime. J Med Chem. 2023 Jun 22;66(12):8170–7.
54. Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. Mach Learn: Sci Technol. 2020 Oct;1(4):045024.
55. Axelrod S, Gómez-Bombarelli R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. Sci Data. 2022 Apr 21;9(1):185.
56. Brown DG, Boström J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? J Med Chem. 2016 May 26;59(10):4443–58.
57. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature. 2018 Mar 28;555(7698):604–10.
58. Coley CW, Green WH, Jensen KF. Machine Learning in Computer-Aided Synthesis Planning. Acc Chem Res. 2018 May 15;51(5):1281–9.
59. Lin K, Xu Y, Pei J, Lai L. Automatic retrosynthetic route planning using template-free models. Chem Sci. 2020 Mar 3;11(12):3355–64.
60. Segler MHS, Waller MP. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. Chemistry. 2017 May 2;23(25):5966–71.
61. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004 Aug;3(8):673–83.
62. Zhou Y, Wang F, Tang J, Nussinov R, Cheng F. Artificial intelligence in COVID-19 drug repurposing. Lancet Digit Health. 2020 Dec;2(12):e667–76.
63. Zeng X, Song X, Ma T, Pan X, Zhou Y, Hou Y, et al. Repurpose Open Data to Discover Therapeutics for COVID-19 Using Deep Learning. J Proteome Res. 2020 Nov 6;19(11):4624–36.
64. Liu R, Wei L, Zhang P. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. Nature machine intelligence. 2021 Jan;3(1):68.
65. Askin S, Burkhalter D, Calado G, El Dakrouni S. Artificial Intelligence Applied to clinical trials: opportunities and challenges. Health Technol (Berl). 2023;13(2):203–13.
66. Harrer S, Shah P, Antony B, Hu J. Artificial Intelligence for Clinical Trial Design. Trends in Pharmacological Sciences. 2019 Aug 1;40(8):577–91.
67. Vazquez J, Abdelrahman S, Byrne LM, Russell M, Harris P, Facelli JC. Using supervised machine learning classifiers to estimate likelihood of participating in clinical trials of a de-identified version of ResearchMatch. Journal of Clinical and Translational Science. 2021 Jan;5(1):e42.
68. Krittanawong C, Johnson KW, Tang WW. How artificial intelligence could redefine clinical trials in cardiovascular medicine: lessons learned from oncology. Personalized Medicine. 2019 Mar;16(2):87– 92.
69. Lee CS, Lee AY. How Artificial Intelligence Can Transform Randomized Controlled Trials. Transl Vis Sci Technol. 2020 Feb 12;9(2):9.
70. Thorlund K, Dron L, Park JJH, Mills EJ. Synthetic and External Controls in Clinical Trials - A Primer for Researchers. Clin Epidemiol. 2020;12:457–67.
71. Weissler EH, Naumann T, Andersson T, Ranganath R, Elemento O, Luo Y, et al. The role of machine learning in clinical research: transforming the future of evidence generation. Trials. 2021 Aug 16;22(1):537.
72. Erdaw Y, Tachbele E. Machine Learning Model Applied on Chest X-Ray Images Enables Automatic Detection of COVID-19 Cases with High Accuracy. Int J Gen Med. 2021 Aug 28;14:4923–31.
73. Goldstein BA, Rigdon J. Using Machine Learning to Identify Heterogeneous Effects in Randomized Clinical Trials—Moving Beyond the Forest Plot and Into the Forest. JAMA Network Open. 2019 Mar 8;2(3):e190004.
74. Feijoo F, Palopoli M, Bernstein J, Siddiqui S, Albright TE. Key indicators of phase transition for clinical trials through machine learning. Drug Discov Today. 2020 Feb;25(2):414–21.

https://doi.org/10.53150/msj.v1i1.39