Facultad de **Medicina**

GRADO EN MEDICINA

TRABAJO FIN DE GRADO

Benefits of artificial intelligence in drug discovery

Beneficios de la inteligencia artificial en el descubrimiento de fármacos

Autor: Nicolás Matía González

Director/es: Dra. María Amor Hurlé González

Dra. Raquel García López

Santander, a 3 de Junio de 2024

“In a restless search for new opportunities and new ways of living the mystery and the promise of distant horizons always have called men forward."

Dustin Bates, vocalist of Starset.

ABSTRACT

Artificial intelligence (AI), and mainly one of its fields, machine learning (ML), present themselves as a great tool for drug discovery. Computer-aided drug design (CADD) has revolutionized the traditional drug discovery pipeline with the incorporation of adequate ML techniques at each stage of the process, cheapening and expediting results. Two major strands compose CADD, known as ligand-based drug design and structure-based drug design, each incorporating diverse ML techniques. In this regard, this work will review the main aspects of virtual screening (VS), which is employed to scan for “drug-like” molecules. *In silico* drug design can utilize ligand-based, structure-based or hybrid tools. Additionally, even though AI has centered its applicability at the earlier stages of the drug discovery pipeline, it is true that pre-clinical stages are also becoming promoted. Finally, one field of study where many difficulties for drug development appear is central nervous system (CNS) disorders. Thus, it seems convenient to include a section on recent practical advances in this area. In summary, this review aspires to provide a longitudinal view of the benefits AI can deliver in drug discovery.

Keywords: Artificial intelligence, Machine learning, Drug discovery, Computer-aided drug design, Central nervous system drugs

RESUMEN

La inteligencia artificial (AI), y principalmente uno de sus campos, el aprendizaje automático (ML, de sus siglas en inglés *machine learning*), están siendo una gran herramienta para el descubrimiento de nuevos fármacos. El diseño de fármacos asistido por ordenador (CADD, de sus siglas en inglés *computer-aided drug discovery*) ha revolucionado el proceso tradicional de desarrollo de fármacos con la incorporación de técnicas de ML adecuadas en cada etapa del proceso, abaratando y acelerando los resultados. El CADD se compone de dos vertientes principales, conocidas como diseño de fármacos basado en ligandos y diseño de fármacos basado en estructuras, incorporando cada una diferentes técnicas de ML. En este contexto, este trabajo revisará los principales aspectos del *screening* virtual (VS), que se utiliza para identificar compuestos que presenten una alta probabilidad de unirse a la diana terapéutica. El diseño de fármacos *in silico* puede utilizar herramientas bien basadas en ligandos, en estructuras moleculares o bien un formato híbrido de ambas. Adicionalmente, aunque la AI ha basado su aplicabilidad en las etapas tempranas del proceso de descubrimiento de fármacos, es cierto que las etapas preclínicas de investigación de fármacos también se están favoreciendo por la AI. Finalmente, un campo de estudio donde aparecen muchas dificultades para el desarrollo de fármacos es el de los trastornos del sistema nervioso central (CNS). Por lo tanto, parece conveniente incluir una sección sobre avances prácticos recientes en esta área. En resumen, esta revisión aspira a proveer una visión longitudinal de los beneficios que la AI puede ofrecer en el descubrimiento de fármacos.

Palabras clave: Inteligencia artificial, Aprendizaje automático, Descubrimiento de fármacos, Diseño de fármacos asistido por ordenador, Fármacos del sistema nervioso central

INDEX

1. INTRODUCTION

Artificial intelligence (AI) is steadily and notoriously crawling inside health sciences. More and more medical specialties harness the opportunities it provides, and the doomsayers are slowly subdued by the accomplishments. In the era of big data, AI presents itself as both the necessary and optimal instrument to manage the large amounts of information derived from genomics, proteomics and the never-ending set of novel “multi-omics”. Derivatives of AI like the different machine learning (ML) models have wondrous real-world applications for different contexts in the biomedical scene.

In particular, recent AI advances in the field of drug discovery have fostered the surfacing of the acronym computer-aided drug design (CADD). The once sluggish and “unprofitable” drug discovery pipeline can be better oriented *a priori*, with AI greasing the gears of drug discovery and improving the outcomes at every stage: target identification and validation, high-throughput screening, lead discovery and optimization, and pre-clinical, clinical and post-clinical analyses.

The main improvements CADD has brought in the last five to ten years can be divided into two broad categories: ligand-based drug design and structure-based drug design. Each one in drug design can be approximated to its homologue in virtual screening (VS). Therefore, it is common to find literature on ligand-based VS and structure-based VS. One objective is to examine the main techniques in structure-based VS, concretely molecular docking and molecular dynamics (MD) simulations. Protein function prediction (PFP) plays a major role in target identification, a related structure-based drug design method. Another objective of the present lecture is to review the main techniques in ligand-based VS, namely quantitative structure-activity relationship (QSAR) and pharmacophore modelling. Additionally, ligand-based drug design is also intertwined with *de novo* molecular design, an AI-heavy method for drug conception utilizing diverse deep ML algorithms. As an additional feature, hybrid VS is likewise promising, due to the integration of techniques from both worlds.

Even though the benefits of AI in the earlier stages of drug discovery fairly outshine those in the finale “pre-clinical and clinical analyses” stage it is a goal of this review to look for benefits across the whole drug discovery pipeline. In this regard, the momentum of AI translates into pre-clinical analysis. Nevertheless, other distal areas like approval and post-market analysis can also be enhanced when the adequate AI techniques are employed.

In summary, the aim of this review is to contextualize AI in drug discovery. The wide range of ML techniques utilized are introduced whenever deemed convenient. The leading edge methods for ligand-based and structure-based drug design are introduced and explained. As a final remark, given one of the areas with maximal potential for CADD is CNS drug discovery a brief summary seems convenient. With this topic sufficing as the source of examples due to its magnitude and possible offshoots, a final point is included highlighting novelties in the sector for several CNS disorders including schizophrenia, depression, Alzheimer’s disease and Parkinson’s disease.

2. OBJECTIVES

The objective of the present manuscript is to provide the reader with a comprehensive review of the role of artificial intelligence in drug design. An introductory section about the intricacies of artificial intelligence and a final section including handpicked examples are attached to better illustrate the real-world possibilities of artificial intelligence in the drug discovery pipeline.

3. METHODS

In this review, the author scoped PubMed and Google Scholar for systematic review and other scientific papers available from 2019 up to 2024 on the topic of artificial intelligence and drug discovery. The search was posteriorly widened with recent and relevant books, reviews and specific articles manually chosen by the author. The MeSH terms employed in the search were “artificial intelligence”, “machine learning”, “drug discovery”, “drug design”, “drugs”, and “central nervous system”.

4. LIST of ABBREVIATIONS

AAE	Adversarial autoencoder
AD	Alzheimer's disease
ADMET	Absorption, distribution, metabolism, excretion and toxicity
AI	Artificial intelligence
AMBER	Assisted Model Building with Energy Refinement
ANN	Artificial neural network
BBB	Blood-brain barrier
CADD	Computer-aided drug design
CHARMM	Chemistry at Harvard Macromolecular Mechanics
CNN	Convolutional neural network
CNS	Central nervous system
COCONUT	COlleCtion of Open Natural prodUcTs
DEMON	Deep Dementia Phenotyping
DL	Deep learning
GAN	Generative adversarial network
GO	Gene Ontology
GPT	Generative pre-trained transformer
GPU	Graphics processing unit
GRU	Gated recurrent unit
GWAS	Genome-wide association study
HDAC	Histone deacetylase
LSTM	Long short-term memory
MD	Molecular dynamics
ML	Machine learning
MPO	Multi-property optimization
NLP	Natural language processing
PD	Parkinson's disease

PDB	Protein Data Bank
PFP	Protein function prediction
PPI	Protein-protein interaction
QSAR	Quantitative structure-activity relationship
QSPR	Quantitative structure-property relationship
RNN	Recurrent neural network
SELFIES	SELF-referencing embedded string
SELSER	Sparse EEG Latent SpacE Regression
SMILES	Simplified molecular input line entry system
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVM	Support vector machine
TPU	Tensor processing unit
VAE	Variational autoencoder
VS	Virtual screening

5. STATE of the ART

5. 1. The basics of artificial intelligence

AI is a field of computer science that aspires to create intelligent machines. While the concept was coined in 1956 it has finally bloomed in the past five years. The extent of the influence of AI is notorious in virtually all sciences, including health sciences. With the advent of AI, a huge number of technicalities from computer and data sciences have flooded medical literature. Given the expected preparation of the general medical public regarding AI and other derived concepts, and in order to establish a benchmark to which refer to, an overview of the basics of AI is presented. [1,2]

In the first place, ML is a branch of AI that attempts to create algorithms that can learn from a set of given data (or dataset) and extrapolate those learnings to other problems or situations, gradually becoming better. In the heart of ML there are several paradigms, but the classical and most important ones are three: supervised learning, unsupervised learning, and reinforcement learning. [1-4]

Supervised learning, like every other learning paradigm, seeks to establish an algorithm incorporating features present in the dataset to establish a conclusion. These features are called inputs and the conclusion is called output. One key difference from other ML paradigms is that a supervised learning model utilizes labeled data, that is to say, data tagged with additional information. The dataset is customarily subdivided into three groups: a training dataset that allows the supervised learning model to engender the algorithm, a validation dataset to enhance the algorithm, and a testing dataset to evaluate how well it performs with unknown data. The most common tasks supervised learning algorithms undertake are (i) classification and (ii) regression; while the former elicits a categorical variable, the latter involves a numerical variable (Figure 1). Finally, the denomination “supervised” stems from the fact that the algorithm is informed —usually by a human— of the inputs and the desired output (also called supervisory signal) in the training phase, so that it has certain examples with which to build the algorithm. [1-5]

Unsupervised learning is a paradigm that tries to elucidate patterns upon an unlabeled dataset. Posteriorly, the unsupervised learning model will try to identify those patterns when presented with a new array of data. Some of the most frequent tasks unsupervised learning algorithms conduct are (i) clustering, (ii) association, and (iii) anomaly detection (Figure 1). Clustering creates groups with shared clusters of features and predicts the category of new variables. Association visualizes relationships among variables and predicts the probability of new variables in the dataset presenting certain features. Anomaly detection spots outliers with aberrant features within the dataset and screens for new variables that are inconsistent. In essence, the designation “unsupervised” comes from the absence of human intervention, or in other words, the utilization of unlabeled data. Unsupervised learning is generally considered to grant a more “creative” perspective than supervised learning. Last but not least, another paradigm called semi-supervised learning, halfway through supervised and unsupervised learning, is sometimes used in problems where the dataset is only partially labeled, like a compendium of radiology images where the medical professional labels only a small subset. Other atypical paradigms are under constant scrutiny (e.g. self-supervised learning, weakly-supervised learning, transfer learning). [1-5]

Reinforcement learning is a paradigm that endeavors to obtain a solution to a presented problem where no single answer is right or wrong. Through multiple iterations following

reward-based selection (i.e. selection of preceding behaviors with maximum cumulative reward) the algorithm aspires to solve the problem finding a sweet spot between exploration of new options and exploitation of prior knowledge (Figure 1). However, reinforcement learning is stated to follow a Markov decision process, a mathematical framework where each new iteration is dependent only on the previous one, which condenses all previous learnings the algorithm has made. As far as health sciences are concerned, reinforcement learning remains the least useful of the three main ML paradigms. [1-5]

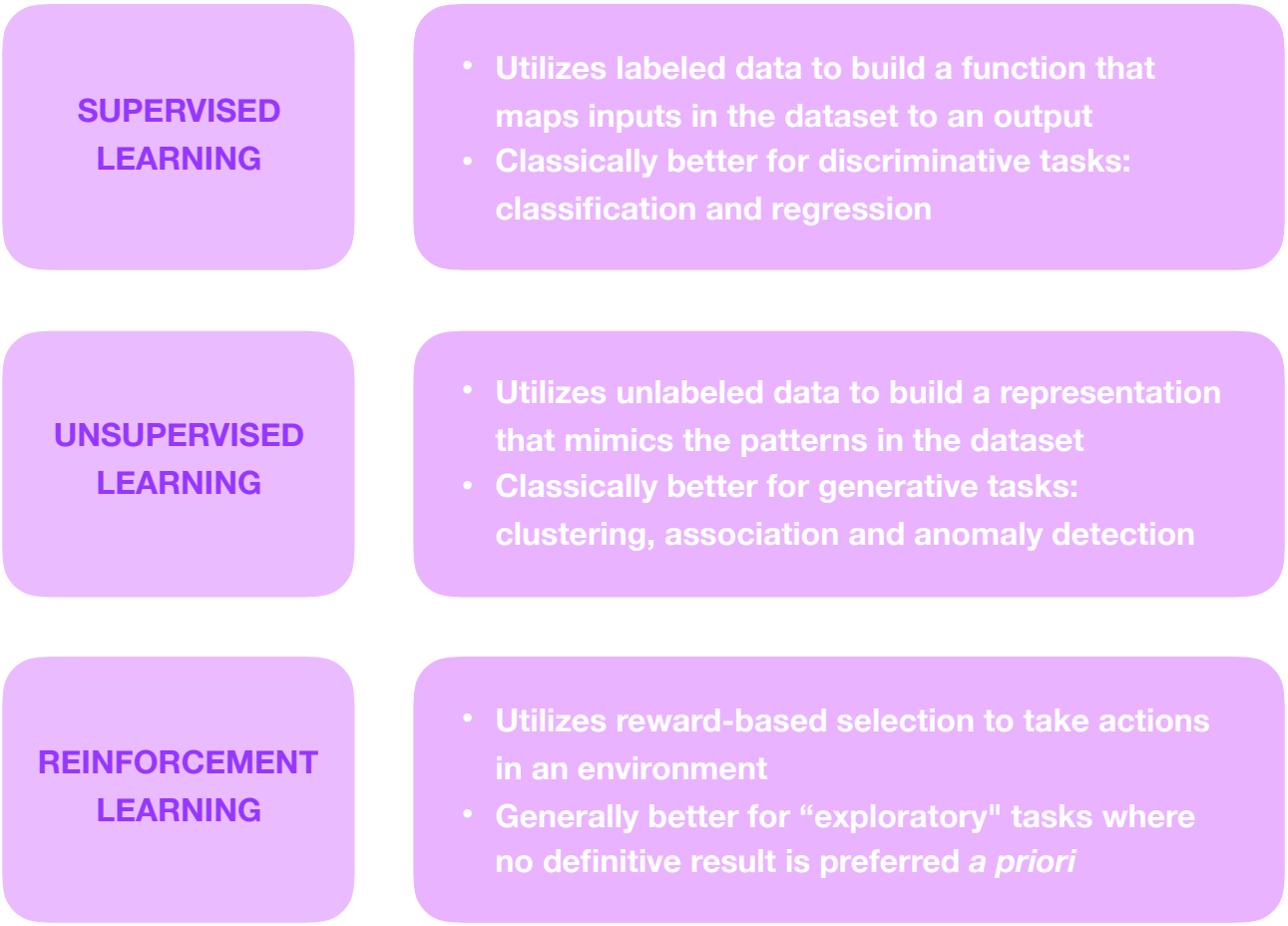


Figure 1. Brief recap about the key differences between the three main ML paradigms.

Diving further into ML, there are abundant highly-specific algorithms, models or methods best suited to solve concrete problems. Artificial neural networks, or simply, neural networks, are one of the main and most-developed ones, with tens of applications. An artificial neural network (ANN) is an ML model inspired by a biological neural network composed of nodes called artificial neurons (henceforth neurons) aggregated into layers and communicated among them via connections that emulate synapses. Usually, an ANN has an input layer, an output layer, and an indefinite number of intermediate hidden layers between them (Figure 2). Sometimes, earlier terminology is used; the term perceptron refers to a neuron and the term multilayered perceptron refers to an ANN. The way an ANN works is the ensuing: each neuron in the input layer generates a “signal” and sends it to a number of neurons in the following layer, which balance the sum of inputs by means of an activation function, and in the case of this not being the output layer the procedure is repeated in each layer passing on the modified “signal”. To determine the relative importance of the inputs a neuron receives each connection is given a “weight” by which

the “signal” is multiplied before the activation function combines all those inputs. Similarly to Hebb’s rule, which states that “cells that fire together, wire together”, if a desired output is procured the “weight” of the connection or synapse increases. Once the output layer is reached, the loss or “cost” function is calculated to quantitatively assess the distance between the current output and the expected output, and backpropagation is performed to modify the “weights”. There are two types of neural networks: uni-directional or feed-forward neural networks, and bi-directional or recurrent neural networks. Although the difference is auto-explanatory, —the flow of information can occur in one or both ways— it is important to note that recurrent neural networks also have intra-layer connections. [1-4]

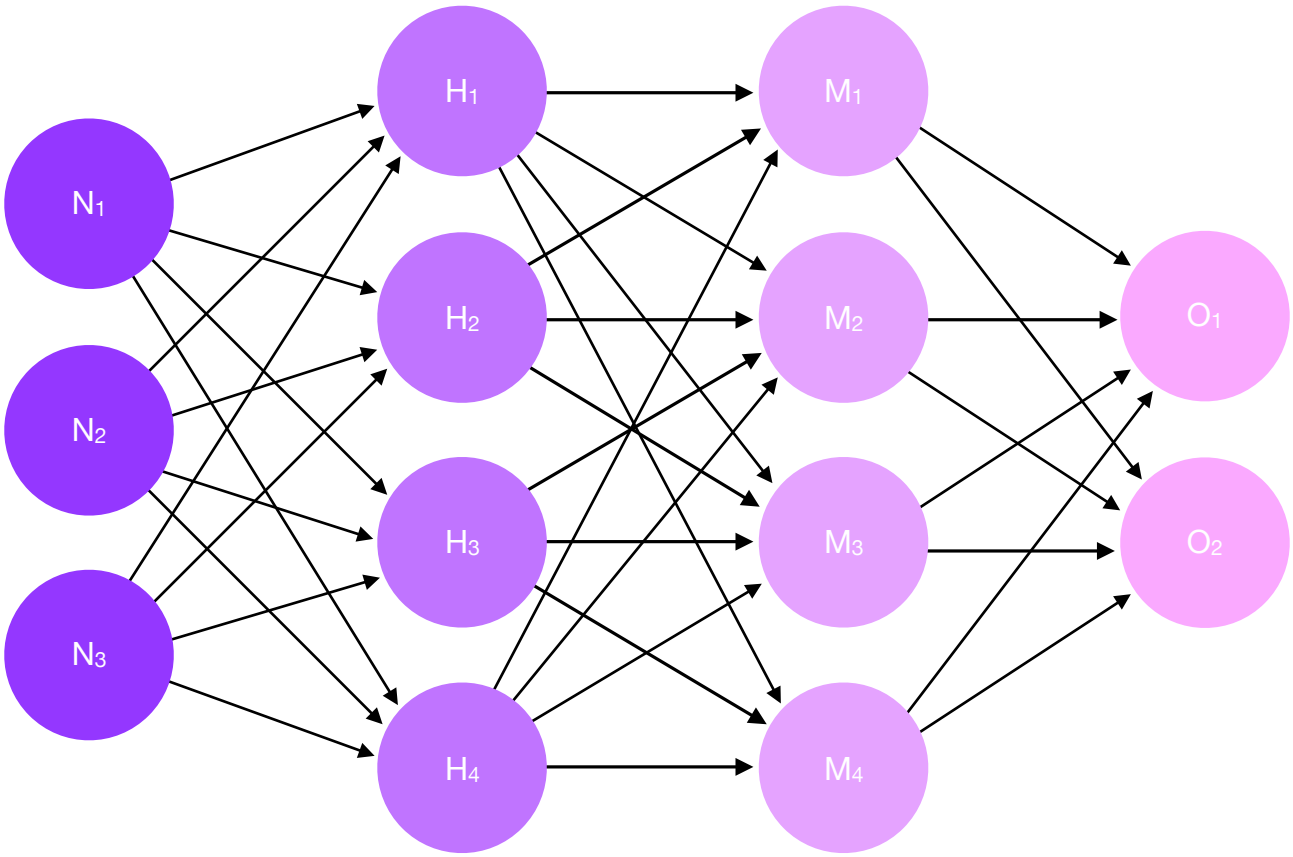


Figure 2. Simplified representation of a basic feed-forward fully-connected ANN. In this example N_X represents the input layer, H_X and M_X represent the hidden layers, and O_X represents the output layer. Each circle or node represents a neuron and each arrow represents a synapse.

Lastly, deep learning (DL) is an ML model that connects deep neural networks (those with two or more hidden layers) and feature or representation learning, an alternative ML paradigm aiming to proficiently discover the representations needed for feature detection. DL algorithms (deep neural network, recurrent neural network, convolutional neural networks, autoencoders) enable much higher, or deeper, capabilities of feature extraction, and consequently deeper levels of abstraction (Figure 3). For example, in the study of diabetic retinopathy, a DL algorithm can be provided with some initial examples, and the algorithm will learn to recognize the signs of the disease (e.g. microaneurysms, intraretinal hemorrhages, macular edema or ischemia, soft exudates, neovascularization…) when no information on what features to look for was gifted. [1,2,6,7]

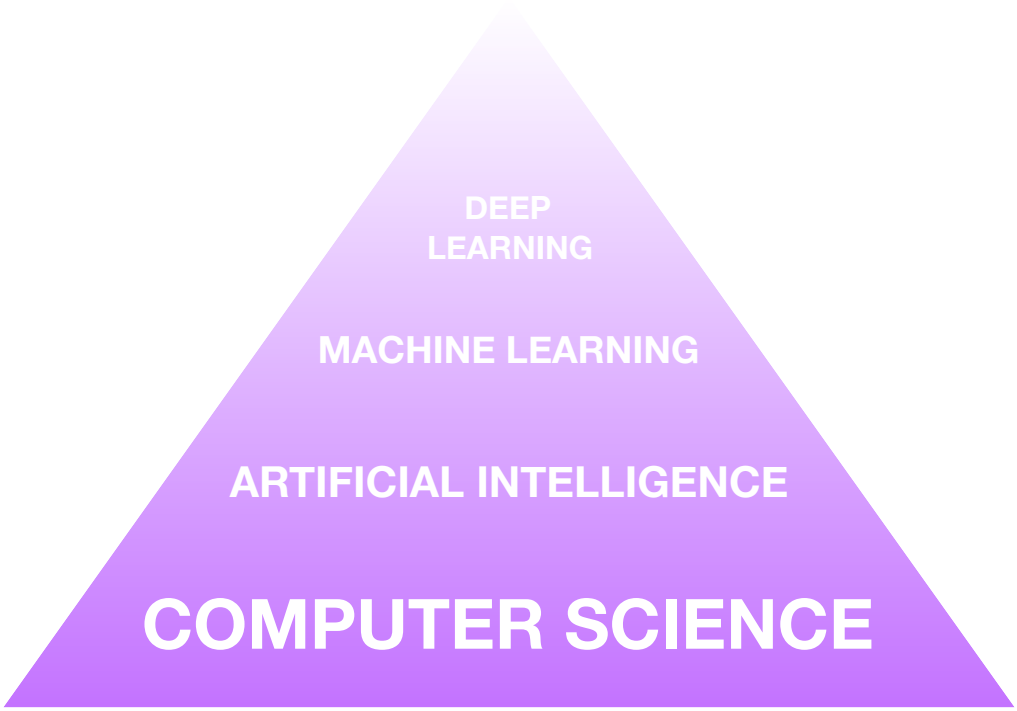


Figure 3. Hierarchical view of emergence of the most common AI-related terms.

5. 2. The drug discovery pipeline

The discovery of a new drug is significantly expensive on all fronts. Most estimations agree somewhere around EUR 2.5 billion and 12 years. As of today, the percentage of novel drugs that reach commercialization is 13%, a dreadful figure. As the brilliancy and refinement of AI permeates other aspects of society the initial unease associated with AI use in health sciences vanishes gradually. Of course, it is self-evident that most drugs in drug discovery are proteins, and for the extent of this review proteins are going to be the main target. Notwithstanding, AI also provides the opportunity to explore other biomolecules (carbohydrates and lipids) as substrates for drug discovery. [8,9]

Figure 4. Stage distribution of the classical drug discovery pipeline.

The drug discovery pipeline contains several relatively well-defined steps for the approval of new drugs on the market (Figure 4). First, the identification and validation of new targets (as a rule, proteins) with therapeutic potential. Second, the examination of multiple molecules with desirable bio-active properties upon the chosen targets, also known as high-throughput screening. Third, the selection of the most promising drug candidates (or hits) and multi-property optimization of the chosen molecules, known respectively as hit-to-lead (lead discovery) and lead optimization in drug discovery. Fourth, the succession of preclinical, clinical, and post-clinical (a.k.a. post-approval, post-marketing) trials. In the case of CADD, where AI is largely applicable, the frontiers between these steps blur considerably, as ML/DL or other models allow multiple steps to be performed in rapid succession or even simultaneously. [2,8,10-13]

Although CADD can be used at any step, it is undoubtedly most common in the earlier stages of drug discovery because one of the main incentives for CADD implementation is the potential for savings if “drug-like” molecules and “druggable” targets are considered from the beginning. Two differentiated strands of CADD stand out: ligand-based drug design and structure-based drug design. Each of them encompasses their “relative” in VS: (i) ligand-based VS and (ii) structure-based VS; and their combination has impulsed (iii) hybrid VS (Figure 5). However, the broader concepts with the surname “drug design” will be used as headings through the narrative to incorporate other aspects like *de novo* molecular design or functional prediction. [2,8,10-13]

5. 3. Structure-based drug design

Structure-based drug design can be seen as an inescapable step, but it is still somehow organic; in an era where big data is the rule rather than the exception, for health sciences to make use of the massive influx of information from different “multi-omics” the parallel expansion of bioinformatics and cheminformatics compulsorily involves AI. It is imperative to understand that while ligand-based VS is the norm, it would take place at a different time in the drug discovery pipeline as a substitute to high-throughput screening (and even lead discovery and lead optimization) and involve its own and distinguishable methods. Molecular docking and MD simulation, the major specific methods for structure-based VS or, circumventing the differences between both concepts, structure-based drug design, will be reviewed in the subsequent subsections. A preface condensing present advances in target identification is included beforehand. [2,10,14]

5. 3. 1. Target identification and protein function prediction

Target identification, defined as the experimental discovery of new “druggable” targets, has also experienced a transition in line with CADD. As of today, only about 3000 out of more than 20000 proteins in the human proteome have known therapeutic potential, and only four protein families condense more than 50% of all current drug targets (G protein-coupled receptors, nuclear receptors, voltage-gated ion channels, ligand-gated ion channels). Hence, the identification of targets to leverage for drug design harbors countless possibilities and, from time to time, provides serendipitous additional information on molecular and pathophysiological mechanisms of diseases (i.e. with the discovery of novel biomarkers). [14-17]

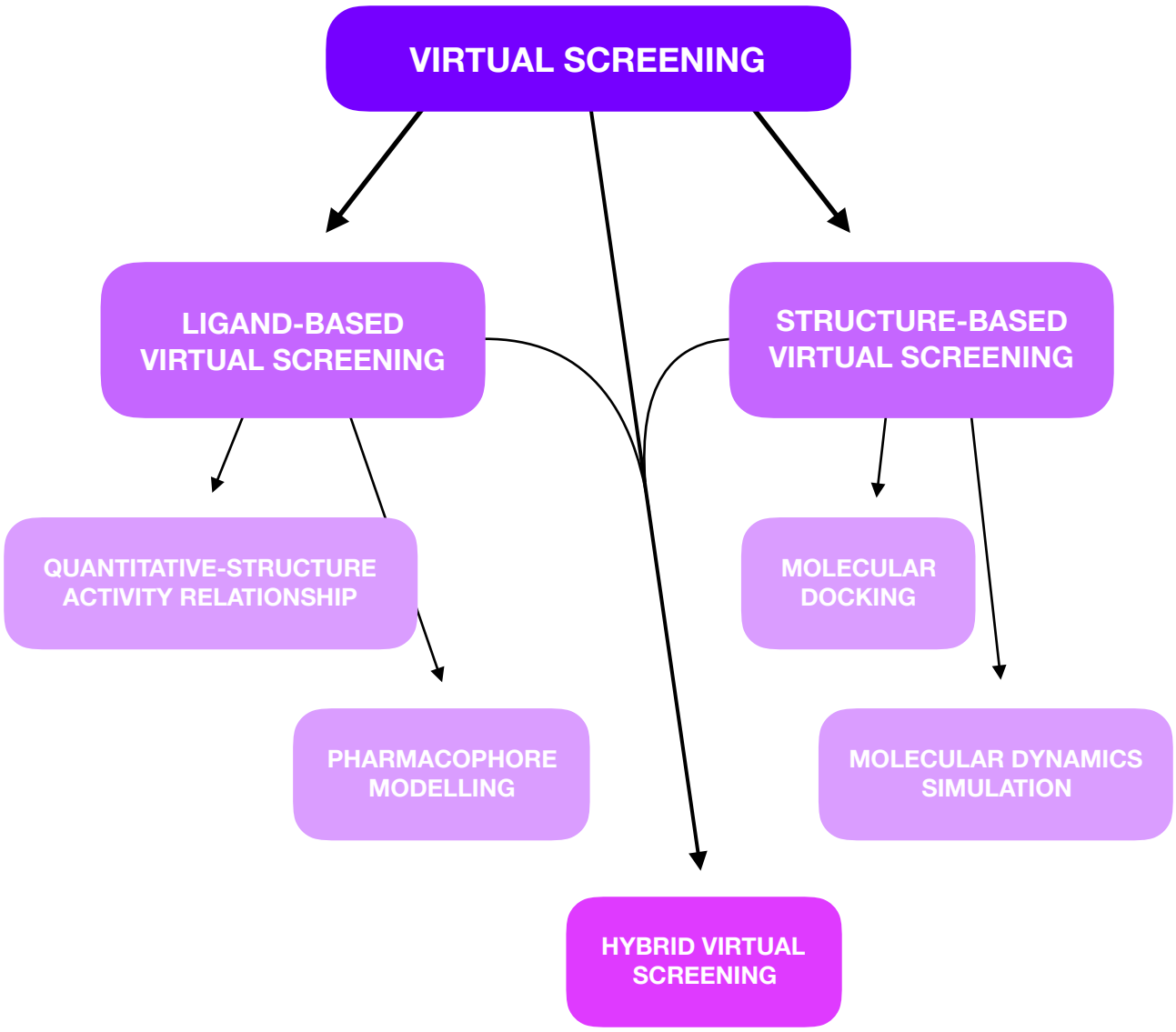


Figure 5. Diagram showing the three varieties of *in silico* methods for compilation and evaluation of compounds from chemical databases.

During target validation, multiple-objective criteria can be implemented to better align targets with the objectives of the study. Text data is one source of “multi-omics” information with the development of some medical generative pre-trained transformer (GPT) models. These are generative models intensively trained on huge amounts of data that can create human-like content. Other pure “multi-omics” approaches can procure interconnected molecular information with tools like genome-wide association study (GWAS) analysis. Finally, AI-driven computational approaches for target identification —like pharmacophore screening, reverse docking, and structure similarity assessment— can convert that information through DL algorithms like recurrent neural network (RNN) or generative adversarial network (GAN), or transfer learning techniques. The overall picture is that synchronized application of non-experimental “multi-omics” and computational approaches is enough to help immensely target identification and target validation. [9,14,15,17-19]

PFP is the cornerstone of target identification, and benefits hugely from many ML algorithms, either in isolation or juxtaposition. Most of the time, a conjunction of different ML-processed-sources are entwined, and improved PFP is achieved through means of

consensus. These data sources are usually protein sequences, protein structures or protein-protein interaction (PPI) networks. [14,15,20-22]

Sequence-based PFP is the most common because the amino acid sequence is the easiest data source to obtain information about. Consequently, feature selection (choice of sequence-level and amino acid-level properties to implement, like amino acid frequency) is easier to conduct on amino acid sequences than other data sources to improve on ML approaches to PFP. The main inconvenience with this method is the high false discovery rate. The two ML models primarily used at this stage are convolutional neural network (CNN) and k-nearest neighbors —an alternative supervised learning algorithm that excels at classification—, but as previously indicated multi-algorithm models (generally including one CNN) are both hopeful and fruitful. The primary databases used for this task are UniProt, Gene Ontology (GO), DrugBank, Pfam database, and Therapeutic Target Database. [15,20,23-30]

Structure-based PFP seems less developed due to the recency of 3D graphs and the resultant lack of data. Nevertheless, promising results are appearing with the development of especially graph-adapted databases (i.e. Protein Data Bank [PDB]) and algorithms (i.e. graph convolutional networks and graph attention networks). [15,31-33]

PPI-based PFP offers great insight into the pharmacodynamics of novel drug targets. Foremost, PPIs are highly specific points of biophysicochemical interaction between two or more proteins. The first PFP models were based on the concept of homology; current models retain that core because interacting proteins characteristically have similar functions. Databases used are typically UniProt, GO and Search Tool for the Retrieval of Interacting Genes/Proteins (STRING). ML models employed at this stage are more diverse and dubious. Using an analogy to music bands, it could be argued that each one has a personal music genre that fits it (in this case it is a ML model). Without getting into further details, they can be less-than-ideally classified into three categories up to this point in time: maxout neural networks, network diffusion-based methods, and network embedding-based methods. [15,34-38]

As an addendum to this last paragraph, it is important to note that PPI prediction methods *per se* can also benefit from ML/DL. In the latest systematic review, kernel-based compound-protein interaction (a generalization from PPI) prediction utilizes one SMILES-derived-fingerprints kernel and one three-vectorial-sequence-derived kernel to generate a heterogeneous kernel representation that can be exploited in a classification task. To achieve realistic performance of compound-protein interaction prediction methods, either in the superior kernel-based (e.g. DeepDTA) or in alternative graph-based ML models, four key points were pinpointed: (i) 5-fold non-redundant cross-validation, (ii) validation over true negative examples, (iii) random pairing for negative example generation, and (iv) ”rank of first positive prediction” as the performance evaluation tool. Transient PPIs mediated by short peptides are an ideal drug target, and many ML supervised models around the idea of mass protein-peptide affinity prediction have been tested using sequence-based data. [39-44]

Eventually, multi-information fusion-based PFP is the idealized culmination of function prediction (Figure 6). Sequence homology stands as the cardinal contributor to function prediction, but it has limitations. COFACTOR is a method that makes up for the lack of sequence homology data with the addition of structural homology data that increases functional information. The amalgamation of sequence-based, structure-based and PPI-based functional prediction promises to be the ultimate innovation. Many approaches are

essayed; one of the most prominent methods is named “Quantitative Annotation of Unknown Structure”. This system works *in crescendo*, starting at the sequence level and making structural predictions. Then, structures are used to make functional predictions analogously to COFACTOR, but in this case with the intercurrent contributions of PPI data from the STRING database and functionally discriminative motifs found in the sequence itself. These three pillars prompt a consensus in the form of a quantitative result that is easier to interpret. [15,45-48]

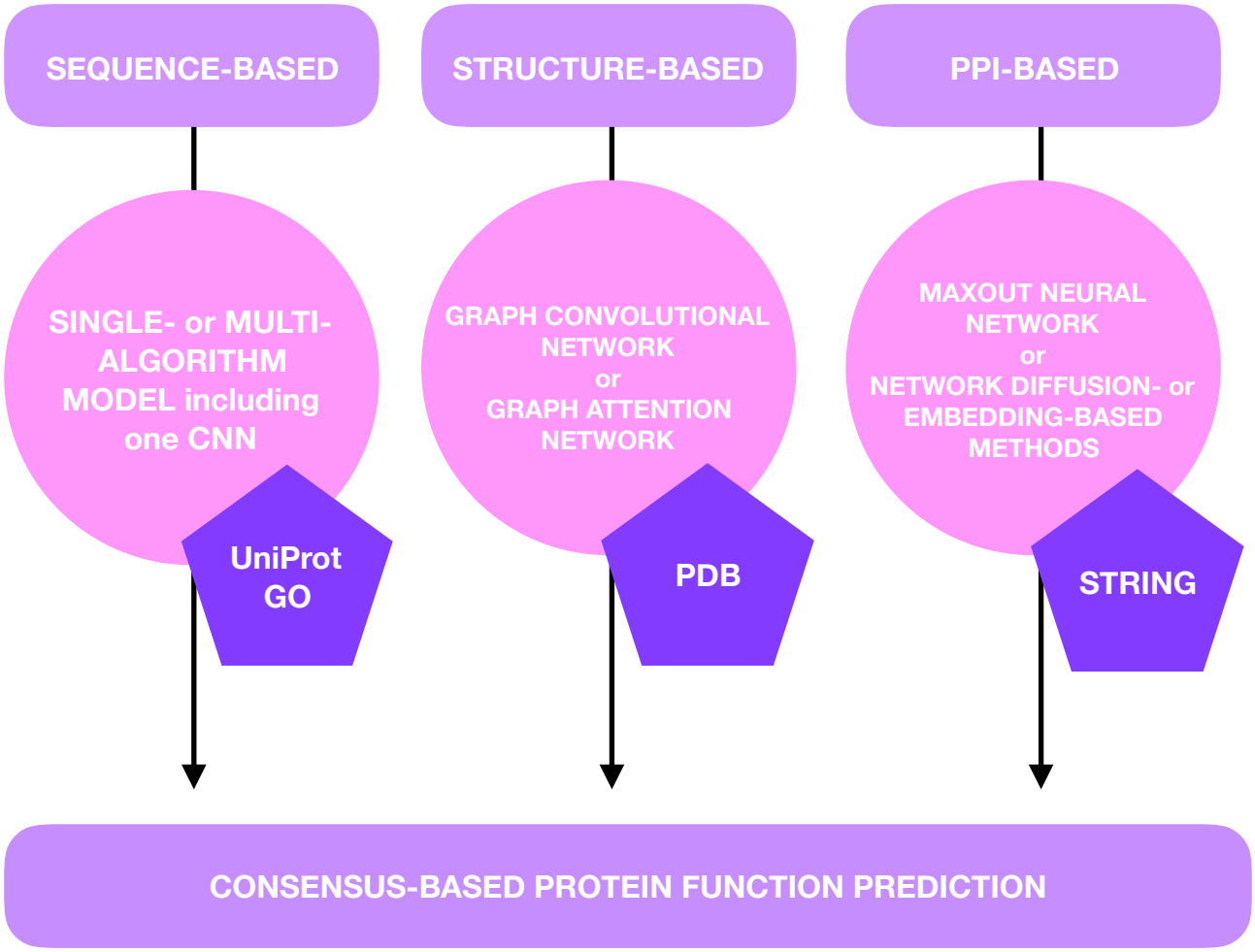


Figure 6. Ideal confluence of sequence-based, structure-based and PPI-based PFP in a fusion-based final prediction. Each individual approximation to PFP (rectangle) appears skewered together with the up-to-date most researched ML models (circle) and the most representative databases utilized (pentagon).

Last but not least, an interesting method named “Iterative Group Function Prediction” distances itself from the classical approaches to PFP. By taking a group of proteins found to work together in a biological context through “multi-omics” techniques like GWAS, turning them into graphs and establishing a network that is repeatedly re-evaluated, it is indeed possible to put forward the global function of the group with accuracy based on those inconclusive “multi-omics” cues. [15,49]

5. 3. 2. Molecular docking

Molecular docking, or simply, docking, is a molecular modelling method to envisage the preferred orientation of a ligand (almost always a small molecule) with regard to a target and determine the binding affinity between the pair. The orientation of the ligand relative to the target altogether with the conformation of each while forming the complex is called binding mode, and each particular binding mode is called a pose. Two concepts are of utmost importance in molecular docking: the docking algorithm and the scoring function. [2,14,50,51]

On the one hand, the docking algorithm is a computational approach for scouting the nearly limitless binding mode search space. There are two options for docking algorithms. Some are systematic and involve intensive exploration of poses; the rest are stochastic and involve random modifications of the binding mode in order to generate new possible poses. [2,10,14,50,51]

On the other hand, the scoring function is a measurable estimation of the binding affinity of the ligand-target complex, or, to put it bluntly, the confidence in the binding free energy of the pose being the lowest possible, since a low binding free energy means a high stability of the complex. There are three categories for scoring functions. They can be force field scoring functions (these work by decomposition of the binding free energy into force field parameters), empirical scoring functions (these work by summation of the energy terms involved) or knowledge-base scoring functions (these work by experimental determination via gathering of structural information). [2,10,14,50,51]

Most ML models focus on the scoring function because the docking algorithm requires considerably more computing power, becomes much slower to train and/or modify, and also because currently the obtainment of original poses is in large part dependent on external software, and these ML algorithms require datasets of up to millions of examples. But, with the scoring function in mind, there have been novel ML algorithms essayed (random forest, support vector machine [SVM], CNN), and novel approaches. To name but a few: creation of new scoring functions, modification of known scoring functions, or adaptation of existing scoring functions to particular instances. [2,8,10,14,52]

Protein flexibility refers to the dynamic nature of the binding mode. Although molecular docking is considered a static approach to structure-based VS, some methods have been developed to account for protein flexibility. The core ones are (i) soft docking, (ii) side-chain flexibility, (iii) molecular relaxation and (iv) protein ensemble docking. All operate by increasing permissiveness at some point, being it Van der Waals forces, side chains, or other. [2,52]

5. 3. 3. Molecular dynamics simulation

MD simulation is a computer simulation method to observe *in silico* the interactions between molecules throughout time. In comparison, it can be said that docking allows for the envisioning of static interactions, although protein flexibility is considered. MD simulations offer a good grasp of the bio-activity of a drug-target complex. The main setback with MD simulations is that they are quite computationally expensive, and therefore, require copious resources for a simulation generally spanning no more than nanoseconds. DL algorithms can learn the implicit patterns present in an MD simulation and give a better perspective of the binding mode and the binding affinity of the poses. In a

similar trend, the parallel development of hardware (i.e. GPUs, TPUs) due to the rise of AI has enabled longer and more expensive fine-grained simulations that minimize the room for error; for example, involving quantum mechanics of the electron cloud. [2,14,53]

MD simulations calculate the variations in protein conformation across the passage of time using Newtonian physics and force fields. The resonant concept of force field can be abbreviated as a computational model to assemble all inter-molecular or intra-molecular forces at play. As such, force fields are helpful to determine the free energy of the complex (incorporating all potential energies involved), and consequently estimate its binding affinity. Within this framework, the most notable programs are Assisted Model Building with Energy Refinement (AMBER), Chemistry at Harvard Macromolecular Mechanics (CHARMM), optimized potentials for liquid simulations, Groningen Molecular Simulation and coarse-grained force fields. [2,14,54-56]

ML-based models can work as a lever to gain insight into drug response. Also, ML can accelerate MD simulations by predicting the free energy of a complex and channelling them towards it, a strategy sometimes referred to as “ML force field”. [14,57]

5. 3. 4. Computational geometry of the binding site

The geometry of the binding site heavily affects the “druggability” of the target and the effect the ligand has upon it. That is to say, the binding site is often a cavity or pocket concave in shape because it has a greater potential contact surface area —and higher binding affinity— than a protrusion convex in shape would have. Henceforth, in a typical structure-based drug design scheme, a target will be selected from one database (e.g. PDB) and a group of candidate ligands will be selected from another database (e.g. DrugBank) to conduct molecular docking followed by MD simulation. Data about the binding mode, stability and binding free energy can be collected, and DL methods can be implemented to learn these features and extract information about the geometry of the binding site (Figure 7). For example, dose-response curves are used in drug discovery to obtain and represent information about the pharmacokinetics and the pharmacodynamics of a drug-target complex. Parameters of functional potency like EC50 or IC50 are established by carrying out costly and unreliable *in vivo* experiments. Instead, DL methods can be used to draw predictions from the overall binding site geometry and its interactions with the ligand molecule using pharmacophore-based schemes. [2,10,14,58-60]

5. 4. Ligand-based drug design

Although the preceding section has necessarily a liaison with this one, this section will refer only to ligand-based VS, going into detail about QSAR and pharmacophore modelling, and, to a much lesser extent, some information on hybrid VS will be provided. Moreover, a brand-new possibility that AI enables is *de novo* drug design or *de novo* molecular design, defined as the development of new molecules already meeting the requirements of the screening process. Different DL applications including RNN, GAN, variational autoencoder (VAE), and CNN appear worthwhile for the moment. Both central points of this section continuously move the spotlight in the drug discovery pipeline from the screening phase towards lead discovery and optimization. [2,10,61,62]

With the ever-growing popularity of AI-driven models in medicine, there is a consequent increase in resources that can mimic the previously long-lasting and inefficient process of sorting through thousands of chemical compounds known as drug screening. The classic procedure at this stage is *in vitro* high-throughput screening, and while the results are more realistic it is costly and only allows for the exploration of a little fraction of the vast chemical search space, estimated to contain up to 10^{60} pharmacologically active molecules. More and more studies are inclined to look into the benefits of (*in silico*) VS since ML can be useful for this task. Nevertheless, both central concepts in this section (*de novo* molecular design and ligand-based VS) are unavoidably interrelated and go together in their evolution. [2,10,13,61,62]

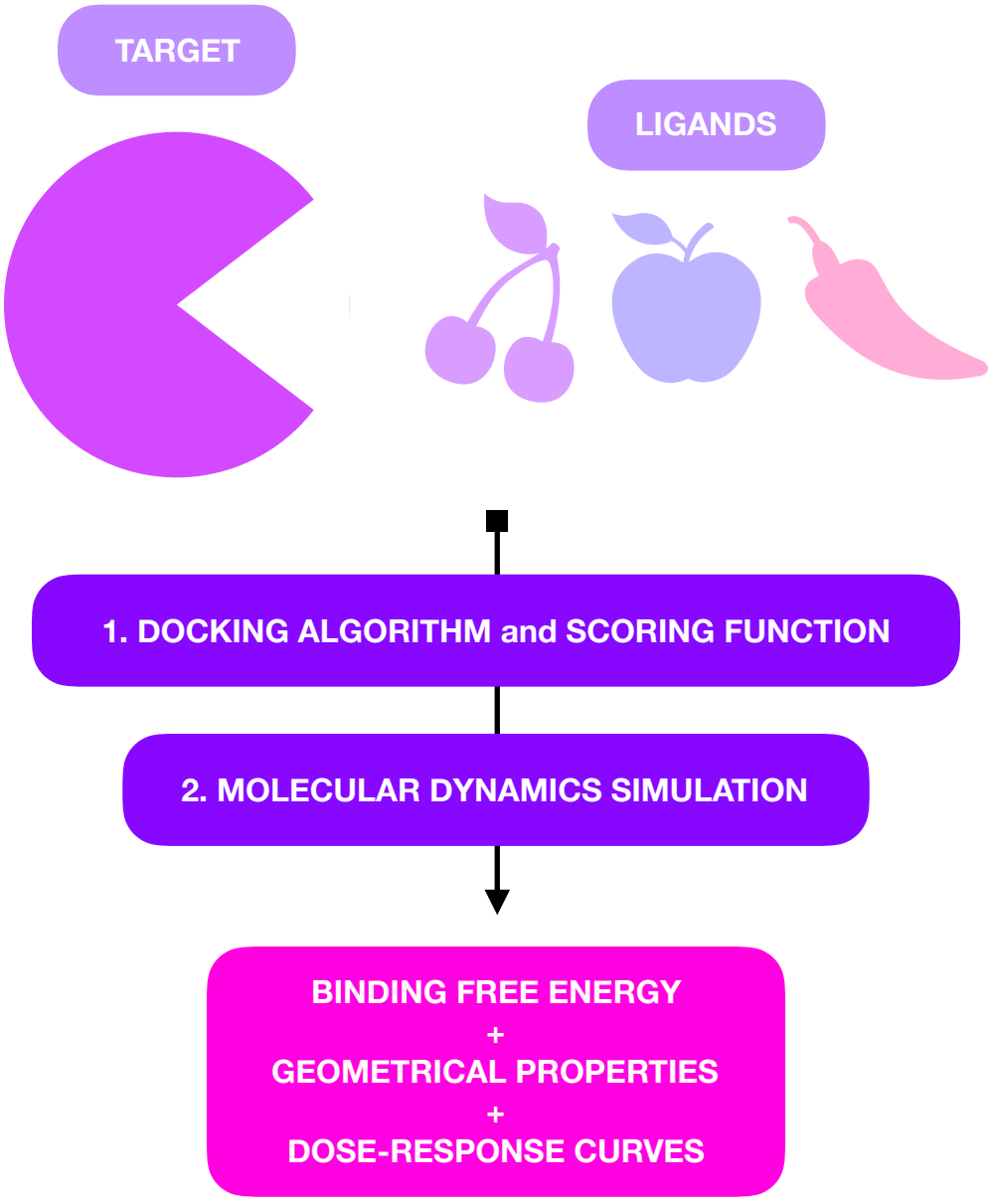


Figure 7. Diagram of the usual workflow with both major structure-based VS techniques molecular docking and MD simulation. Once the set of ligands is decided upon, the less demanding molecular docking is performed first and the MD simulation is conducted afterwards. The results are then processed with DL algorithms to obtain estimations of the binding free energy, computational geometrical properties and dose-response curves.

5. 4. 1. Strategies for *de novo* molecular design

An RNN is a type of neural network that works bi-directionally. In a simplified sense, a neuronal network receives inputs (small data pieces in a database) that are then multiplied by randomly generated “weights”, and the outputs (inferences made at each layer) are used to further modify these “weights” applied to the inputs until the outputs are closer to the desired result. In the case of an RNN, it possesses a memory called internal state. When an input is provided for the first time, the internal state records this as the first iteration and it receives the greatest “weight”. With each sequential iteration, the RNN incorporates a new input —each of them carrying a lesser “weight” than the previous one — and uses the preceding input to refine the final output. The one problem with RNNs is that due to their bi-directional nature the final output is also affecting all previous inputs, and the “weight” it carries fades with each iteration while the “cost” grows. This backpropagation concept is called the vanishing gradient problem (the weight diminishes progressively) or the exploding gradient problem (the cost increases progressively). Two types of RNN are used in drug discovery to mitigate these issues: long short-term memory (LSTM-RNN) and gated recurrent unit (GRU-RNN). The difference between them is that a GRU-RNN is more compact and utilizes fewer parameters than an LSTM-RNN. One immediate benefit of this approach is that the RNN is not required to *in silico* navigate the endless chemical search space —which is called a brute force search—, and instead it perfects the molecule it is “looking for” with each new output based on the original input it received, such as pharmacokinetic properties. [61-68]

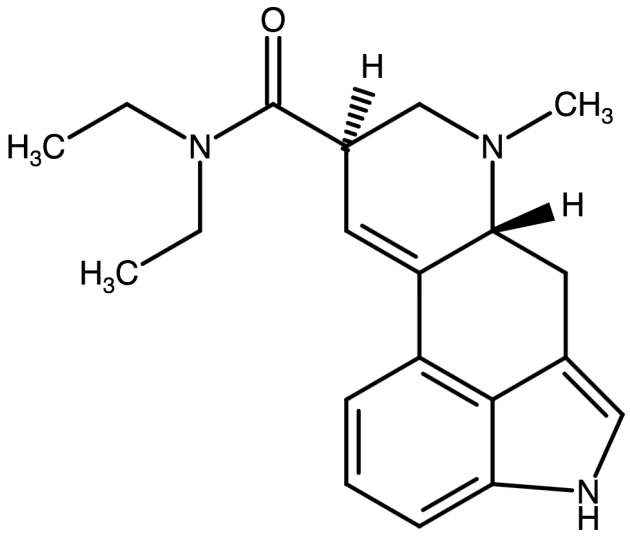
A GAN is a type of ML algorithm that was initially devised for unsupervised learning. In broad terms, two opposing neural networks compete against each other until one of them wins. Essentially, it involves a zero-sum game where the first neural network (which is called generative neural network) tries to generate an output resembling one from an array of data, while the second neural network (which is called discriminative neural network) tries to discern if the input received belongs to the original array of data or is a fabrication of the generative neural network. [61,63,69-71]

An autoencoder is also a type of ML algorithm that uses unsupervised learning. In short, an autoencoder is a type of neural network utilized to create efficient representations of data. Two components are principal: the encoder and the decoder. An autoencoder generates a simplified version of the data or “latent representation” through the encoder, from which it is capable of deterministically regenerating an output matching the original input through the decoder. To succeed in this task, the “latent representation” must achieve a finer level of detail called granularity, and is said to be fine-grained. A VAE is a variant that goes one step further, because it is capable of generating a “latent representation” that is not static. Thus, probabilistically regenerating an output similar to the original input, but with slight modifications each time due to it sampling the “latent representation” and not decoding it at face value. An adversarial autoencoder (AAE) is another probabilistic variant that combines a GAN after the decoding process. A discriminative neural network tries to discern if the slightly different outputs come from the original array or a different set of data. For example, in one pioneering study employing an AAE called DruGAN and comparing it against a VAE for the generation of molecular fingerprints depicting novel anticancer molecules the VAE was outclassed by the AAE in generation power and efficiency. The general utility of these autoencoder-based methods resides in the possibility of guiding the modifications made to optimize the molecule, and typically a third component —the predictor— evaluates the novel molecules (in the form of vectors) emerging from the “latent representation”. [8,61,62,72-74]

A CNN is a type of feed-forward neural network. The basics are the same as in any neural network, the peculiarity of the CNN being that it performs convolution in at least one of the layers (which is called convolutional layer). Using a considerably non-mathematical approximation it can be said that convolution in this scenario means that a neuron from a convolutional layer is affected only by inputs with similar features named “receptive field” in the previous layer, instead of having this neuron be affected by all inputs in the previous layer like in a typical fully-connected neural network. In some cases, another type of layer (which is called pooling layer) can be interspersed after a convolutional layer to merge all outputs from a group of neurons into a single one in the next one. The importance of CNN models just at this stage is enormous, because they enable the learning of molecules simplifying immensely the impracticalities of neural networks using fully-connected layers. [1,62,63,75-77]

5. 4. 2. Data representation and findings

Molecules can be represented or encoded in multiple ways, but the most employed ones are string-based representations and graphs, especially 3D graphs. The “nomenclature” most widely used for string-based representations is one-dimensional, the Simplified Molecular Input Line Entry System (SMILES). One SMILES string provides information about the atoms, bonds, branches, and cyclic structures of the original molecule (Figure 8). Further changes have been made to these string-based representations, making it



[H][C@@]1(CN(C)[C@]2([H])CC3=CNC4=CC=CC(=C34)C2=C1)C(=O)N(CC)CC

[H][C@@][Branch2][Ring1][P][C][N][Branch1][C][C][C@][Branch1][C][H][C][C][=C][N][C][=C][C][=C][C][=Branch1][=Branch1][=C][Ring1][=Branch2][Ring1][=Branch1][C][Ring1][=N][=C][Ring2][Ring1][C][C][=Branch1][C][=O][N][Branch1][Ring1][C][C][C][C]

Figure 8. Skeletal formula made with the Chemical Sketch Tool of the Protein Data Bank (PDB) database and correspondent SMILES and SELFIES strings of the molecule lysergic acid diethylamide.

plausible to reflect other characteristics of the original molecules. In particular, inside the field of *de novo* molecular design, it is interesting that while not all randomly generated SMILES strings are valid, its recent successor SELF-referencing Embedded String (SELFIES) is completely robust in the sense that every randomly generated SELFIES will depict an existing molecule. On another note, 2D and 3D molecular graphs employ matrices to show the relative positions of atoms and their relationships, and can be used to represent their original molecules in ML operations. [14,61-63,78-81]

The results of the current approximations to *de novo* drug design are briefly summarized in the present and next paragraphs. Primarily, AI-derived models can offer great results, but can also meet derailment if the datasets employed are not tailored to the problem and thence impede generalizability. There are copious examples of independent researchers who have tried utilizing RNN models for *de novo* drug design (e.g. ReLeaSE, DrugEx), and many examples of studies where *in vivo* validation of the new molecule was proven. The current consensus seems to be that LSTM-RNN models are superior to GRU-RNN models, even though both of them have been designed and employed successfully to create valid SMILES strings. In regard to VAE and AAE models, they are the second most-used strategy incorporating AI in this step of drug discovery. GAN is the least used method of these initially discussed, but depending on the concrete example there are also some that reach almost 100% of valid SMILES strings. Other models have been tried to a lesser extent for *de novo* drug design, the most common ones being CNN models and evolutionary algorithms. [10,61,62,64-77,82]

From the array of molecular representations available SMILES strings are the most frequent ones. There is evidence that suggests AI trained with enumerated SMILES —where all possible SMILES forms of the molecule are laid over— yield better results than those trained with canonicalized SMILES —where the SMILES form is absolute—. With regard to SELFIES, its implementation is not major due to its recency, but it seems promising due to its robust nature. Graphs (especially three-dimensional graphs) are a rising trend because they provide AI with additional information about the conformation of the molecule and additional opportunities to learn the implicit rules of molecular structure, especially when in conjunction with a CNN. Finally, other string-based representations like molecular fingerprints (bit sequences carrying information about molecular features) are seldom used. These molecular fingerprints, which can be any-dimensional —even zero-dimensional—; for example, the 2D Molecular ACCess System or extended connectivity fingerprints, seem to render suboptimal results and should be discouraged with the current knowledge. [8,61-63,79,81]

On a similar note, it is interesting to add that AI tools utilized in ligand-based drug design nowadays include the following: AlphaFold2, DeepChem, DeepBind, DeepBar, Chemputer, Chemical VAE, PPB2, InnerOuterRNN, DeltaVina… It is substantive to note that the large majority of them are free to use and accessible online, as well as most databases employed in drug discovery. Thus, the irreproducibility and the unverifiability of most AI research seems rather awkward. [2,14,83-86]

5. 4. 3. Lead discovery and lead optimization

Once there are means by which to represent a molecule and to evaluate if it is a feasible candidate for a new drug the previous points are inescapably connected to VS. In the traditional drug discovery pipeline it is not clear where the transition from high-throughput

screening to lead discovery to lead optimization happens when AI enters the picture. Through diverse traditional or ML/DL applications (mainly RNN, GAN, VAE, CNN) new molecules with the potential of becoming a drug are created. These selected molecules have optimized properties including bio-activity (such as anticancer, antiviral or antibacterial activity), but must also meet basic requirements like validity (the adherence to the laws of physics and chemistry), novelty (the presence of molecules distinct from those in the dataset or the drug market) and synthetic feasibility (the possibility of being synthesized *ex silico*). Other requirements can be bypassed depending on the setting (uniqueness, diversity, similarity). This overlying concept is referred to as multi-property optimization (MPO). [10,61-63,87,88]

Due to the goal-oriented nature of *de novo* molecular design, MPO is translated to proxy scoring functions in order to carry out comparisons among competing molecules. The basics of MPO are the classical pharmacological concepts of absorption, distribution, metabolism, excretion and toxicity (ADMET), although ADMET properties are similarly involved in the alternative *in vitro* high-throughput screening phase (Figure 9). Further options for MPO involve imposing restrictions on the molecules created in order to turn them even more “drug-like”. For instance, it is true that physicochemical descriptors remain an important part of MPO. The most common conditions imposed tend to relate to Lipinski’s rule of five, minted in 1997, which states that an active ingredient candidate for an orally administered drug should not have: more than 5 hydrogen bond donors, more than 10 hydrogen bond acceptors, more than 500 daltons of molecular mass, or a calculated octanol-water partition coefficient (clogP) greater than 5. The most important prerequisite is undoubtedly a small size, and virtually every studied ligand in drug discovery is less than 500 daltons. [8,10,61-63,87,88]

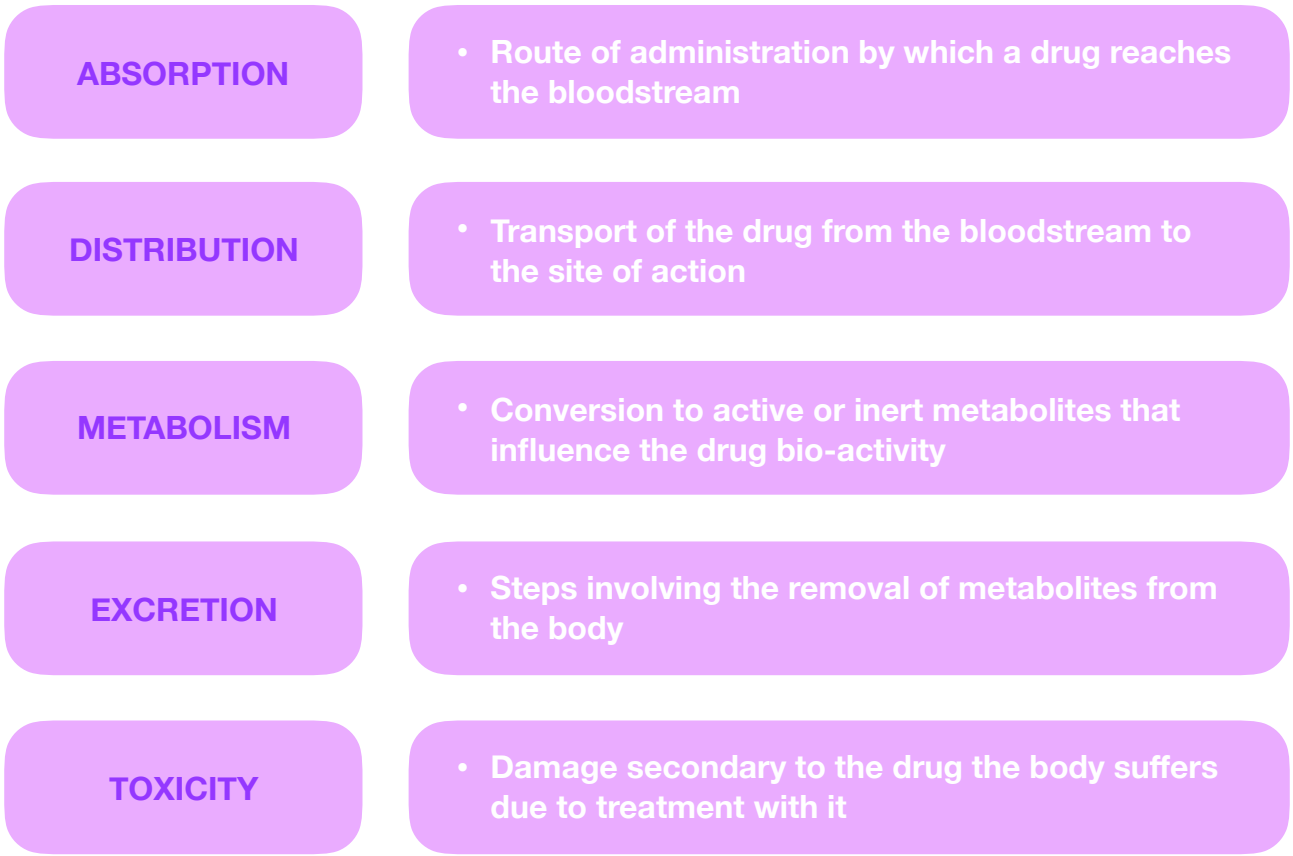


Figure 9. Reminder of the concepts integrating ADMET, the pivotal aspect of MPO.

As mentioned above, ADMET profiles constitute the base of MPO. In principle, toxicity remains the most important ADMET property to bear in mind, because roughly 33% of drug candidates are rejected in pre-clinical and clinical development due to toxicity issues. Efforts have accomplished multi-task deep neural networks that have proved to foretell hallmarks of drug toxicity better than their predecessors and with a reduction in the rate of false positives, like DeepTox (which paired with the specific dataset Tox21 works wonders). Additionally, a subfield of computer science utilizing statistical and AI-based approaches known as natural language processing (NLP) can be utilized to “understand” literature and mine for specific information, in this case about the subject of drug toxicity. [8,10,63,87,89,90]

When using a proxy scoring function in MPO for *de novo* molecular design —for example, using a QSAR (see later) for the prediction of EC50 or IC50 for a novel set of molecules created by a generative model— it must be considered that it is heavily database-dependent. As in a paradox, the closer the QSAR (or any other proxy scoring function) of each novel molecule resembles those in the database, the higher the validity and the lower the novelty, and vice versa. Performance evaluation in MPO is complicated, and barely-grounded ML results should be interpreted extra cautiously since they can misrepresent reality and hinder interpretability. [10,91,92]

To date, the most widely used —generally small-molecule but target-inclusive as well— databases in the realm of CADD are PubChem, ChEMBL, DrugBank, UniProt database, PDB, BindingDB, BindingMOAD, ChemSpider, COlleCtion of Open Natural ProdUcTs (COCONUT) and ZINC. These and many other databases provide information on the pharmacokinetics and pharmacodynamics of the included molecules, that can be incorporated into VS by AI through the aforementioned and other ML applications. A common strategy in supervised ML is cross-validation, to employ one of these datasets for the primary training of the algorithm and refine it with a smaller and more specific dataset on a second occasion. However, a valid argument is to be made that VS incorporates bias, as even the biggest databases —containing hundreds of millions of legitimate “drug-like” compounds— fall quite short against the size of the search space at hand. Another source of bias seems to be that, as new efforts to discover useful molecules concentrate around previous hits, the known search space grows in related spurts, while the majority of it remains obscure. [2,10,14,58,59,61]

Due to the patent inability of the prevailing databases to encompass enough labeled data for its use in certain situations, ML techniques around the idea of adaptability have been studied. These are interrelated with the semi-supervised learning paradigm, yet employ different methods to essentially continually incorporate knowledge into a growing pool. First, transfer learning implicates utilizing information from a previously already-solved task to solve the one ahead by fine-tuning the conclusions. Second, multi-task learning implicates utilizing more than one dataset at a time to solve many tasks at once, and it is especially useful for recycling rather small datasets that are not useful in themselves. Third, self-supervised learning —which can be best classified as an alternative ML paradigm— implies that the very algorithm is the one assigning sloppy easy-to-generate labels to the dataset and training itself on that same dataset afterwards. The best alternative is unclear since evidence is conflicting and ever-changing in most of the field, but ML techniques with a certain degree of flexibility seem to pose an enormous advantage in data scarcity settings. [10,93-95]

Drug repurposing or drug repositioning is another concept tangentially related to VS, but with great leverage at the stages of lead discovery and optimization. The idea behind drug

repurposing is multi-faceted; by promoting repurposing of known drugs with copious amounts of data available, a “speedrun” of the drug discovery pipeline can be performed, shortening both the time and costs inverted before commercialization. A typical molecule can have multiple targets, and exploitation of compiled data from known drugs via ML/DL can guarantee a higher rate of “hits” for new purposes and a faster hit-to-lead conversion. [8,96-99]

5. 4. 4. Strategies for ligand-based virtual screening

In the absence of structural information of the target —as it is the case in structure-based VS—, the so-called ligand-based VS strategies can be engaged. Two common strategies take the spotlight, the former is QSAR and the latter is pharmacophore modelling. A recurring concept at the stage of hit-to-lead or MPO is called QSPR (Quantitative Structure-Property Relationship), an empirical linear regression model that allows the expression of a physical or chemical property (typically biological activity, giving rise to the acronym QSAR) as a function of the structure of a molecule, in quantitative terms. Hence, the crucial components of a QSPR/QSAR are the origination of molecular descriptors and the formulation of the mathematical paradigm. The steps in a QSAR are the following: (i) preparation of the ligand battery, (ii) selection of molecular descriptors in the training dataset (habitual molecular descriptors are electronic effects, hydrophobic parameters, steric effects, and substructural effects), (iii) calculation of the correlation to biological activity for the chosen descriptors, and (iv) evaluation of internal and external validity (Figure 10). [2,10,62,100]

The outcome of a QSAR is a column of cells with the predicted values of the biological activity of a range of molecules for a single target. In this manner, ML can be applied at this stage to fit even tampered nonlinear regression models and promote ADMET-appropriate drug candidate profiles using QSARs or other QSPRs. The deployment of ML in QSAR analysis has been essayed successfully through various DL approaches: random forest, SVM, multi-task deep neural network… SVM, one of the principal supervised learning algorithms, can create a division in an array of data either linear (when there are two perfectly distinct classes) or nonlinear (when there is some overlapping between the classes). [2,8,10,62,100,101]

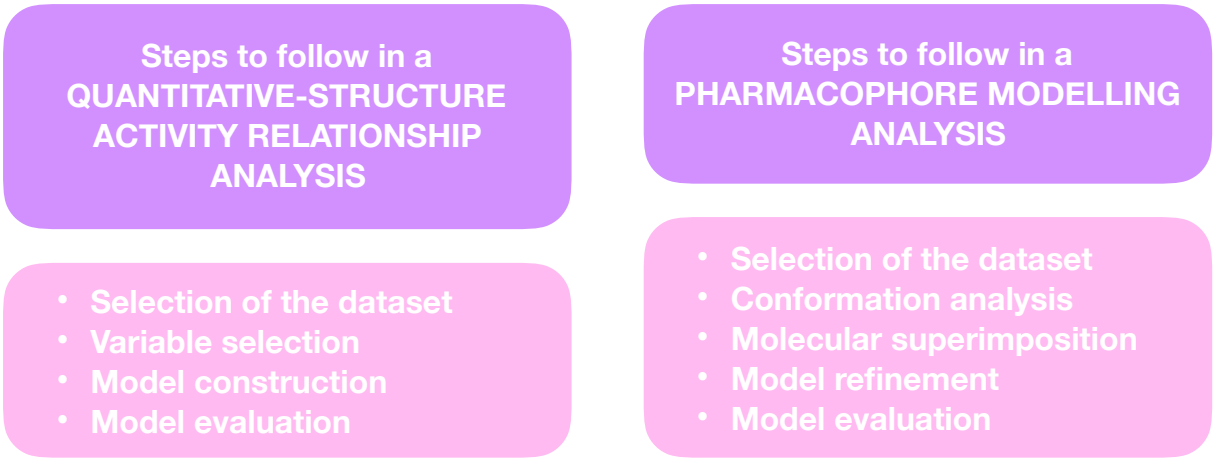


Figure 10. Comparison of the workflow of a QSAR *versus* a pharmacophore modelling analysis.

The pharmacophore can be defined as the ensemble of molecular features —steric and electronic— that grant biological activity to specific ligands over a target, that is to say, the requirements for a key to be recognized by the lock. Pharmacophore modelling can be used to confirm the presence of biological activity and guide generative ligand-based VS strategies. In a nutshell, there are five steps to pharmacophore modelling, similar to those in a QSAR: (i) preparation of the ligand battery, (ii) creation of a low-energy conformation list, (iii) superimposition of low-energy conformations, (iv) abstraction of pharmacophore elements, and (v) evaluation of internal & external validity (Figure 10). [2,62,100,102,103]

5. 4. 5. Hybrid virtual screening

Although diverting slightly from the main topic of this segment, hybrid VS should be bestowed some representation. This in-between concept refers to the conjunction of ligand-based VS (primarily QSAR-oriented) and certain coarse target-based techniques. The main hybrid VS techniques derive from proteochemometrics. This type of analysis with a bit of a bizarre name is almost interchangeable with multitarget-QSAR, and differs from a traditional QSAR in that it makes use of multiple targets instead of one and incorporates target descriptors. Hence, the outcome of a proteochemometric model is, in contrast with a QSAR, a matrix where the additional dimension is composed of a number of parallel QSARs, but its usefulness increases exponentially thanks to the exploitation of the emerging synergies. In essence, since hybrid VS can be seen as an expansion of ligand-based VS, both methods share pros and cons. Some enlightening examples of achievements via hybrid VS are: the learning of protein representations from amino acid sequences, the unification of uncertainty for data from heterogeneous sources, and the hybridization of the workflow in very ambitious structure-based VS projects (structure-based techniques are performed on a small subset which is used as a training dataset in an unsupervised model). [10,104-106]

5. 5. Pre-clinical and clinical development

Many AI-related models can be helpful at the final stages of the drug discovery pipeline, but no single one is “*El Dorado*”. NLP and one derivative sub-task called named entity recognition (based on localization of named entities in a text and their categorization in different semantic fields) can be helpful altogether with GPT models at summarization and generation of documents. Prediction of cell responses to drugs is one key step in the drug discovery pipeline. Two methods stand out with this intention, similarity-based (based on the premise that similar drugs act on similar targets) and feature-based (based around the idea of drug-target feature vectors informing convolutional to attentional algorithms). AI can be also tried at other tasks related to drug marketing like drug manufacturing, quality control, clinical trial blueprinting, and post-clinical analysis and development. [8,10]

5. 5. 1. Approval and post-market analysis

The tentacle-like possibilities of AI can reach beyond development of novel drugs, helping approval and market forecasting and positioning. Data mining approaches like NLP can help understand the selling space and concrete needs of the population by leveraging big data gained through business-to-business internal surveying. As a matter of fact, ML can

even help at making use of scattered data indirectly related to financial expenditure and profits during drug development, establishing more competitive and fair prices for the benefit of the company and society. [8,107,108]

Another post-market niche where AI has found its way is drug antagonism/synergism prediction, since experimental study of drug interactions is inefficient. Supervised ML advances like a Bayesian ANN and a random forest algorithm have been essayed for scouting interactions, albeit DL-boosted models like DeepSynergy have overpowered them, as it is more often than not the case due to the might of DL. Anyhow, ML/DL has the faculties to reform the antagonism/synergism interaction scouting process much more effective and applicable in clinical practice. In particular, anticancer drug combo optimization is the area where the stimulus of drug antagonism/synergism prediction can be greater, for the reason that many anticancer schemes require up to six or seven drugs. [8,63,109-111]

5. 6. AI in drug discovery for central nervous system disorders

One of the fields that could capitalize on AI the most includes the group of agents specifically targeting the CNS. Disorders of the CNS mostly include neuropsychiatric conditions such as schizophrenia or dementia, but CADD in this setting spreads out over areas like anesthesia, neuropathic pain, or substance abuse. CNS drugs require even longer than the usual 10-15 years on average to gain approval —but could see this period remarkably shortened thanks to AI-enabled drug discovery—, and they are subjected to especially tight control. By the way, CNS drugs also have a lower success rate due to other factors encompassing insufficient trial length, substandard knowledge of the underlying pathophysiology, “dirty” target engagement, and the presence of the blood-brain barrier (BBB). Target identification in CNS drugs appears to follow the maxim that few protein families (characteristically G protein-coupled receptors) condense more than 50% of drug target identification efforts. [5,112-115]

The BBB is a selective semi-permeable membrane comprised of capillary endothelial cells that are lined by a basement membrane made from structural proteins, pericytes, astrocytic end-feet, and microglial cells. Given that the physiologic action of this barrier is to not let exogenous substances come into contact with the brain, the success rate of CNS drugs is extremely dependent on good BBB permeability. BBB permeability prediction is the epicentre of CNS drug discovery, and many ML algorithms propelled by the abovementioned Lipinski’s rule of five or other physicochemical descriptors have been employed to account for passive diffusion. Accomplishments notwithstanding, particular molecules follow specialized drug-transporter interactions that cannot be translated into simple physicochemical descriptors. Transporters like the ATP-binding cassette transporter and efflux pump P-glycoprotein appear to be the primary determinants of achieving therapeutic concentrations in the blood and regulation of pharmacoresistance in the CNS and should not be disregarded. Efforts in this direction have not provided an omnipotent algorithm for the prediction of BBB permeability yet, but definitively DL has set a course on the right track. [5,114,116-120]

5. 6. 1. Examples of advancements in central nervous system disorders

Schizophrenia is currently the biggest enigma in psychiatry. While AI is indeed not the panacea, it offers opportunities to tackle the many challenges this disorder proposes, although data scarcity hinders the resolution of the heterogeneity of the disorder. AI has been essayed satisfactorily in target identification, ligand-based VS, structure-based VS, prediction of therapeutic adherence and drug repurposing. [5,113-115,121] In detail, schizophrenia target genes were identified employing an SVM to initially rank genes from public microarray datasets —built with the help of GWAS analyses—, paired with another algorithm so-called recursive feature elimination to posteriorly discard genes devoid of significance. [122] For VS, SVMs were also applied for the prediction of presynaptic dopamine overactivity and the formation of gamma-aminobutyric acid uptake inhibitor QSAR models. Another VS method entailed the creation of a pharmacophore model of α7 nicotinic acetylcholine receptor agonists coupled with a recursive partitioning model to trim undesirable ligands. [123] The SVM methodology was successful as well for drug repositioning in schizophrenia. [124]

Depression treatment is already benefiting from AI. [5,112,115,121] Gradient boosting models have been attempted to predict if a patient will achieve symptomatic remission using both citalopram and escitalopram. [125,126] This gradient boosting model was repeated in an ulterior study coupling it with a hierarchical clustering algorithm to determine groups of symptoms that could benefit from treatment with a particular antidepressant. [127] Since these studies do not explain the degree of response to each particular antidepressant, another group engendered the “Antidepressant Response Prediction Network” model, which achieved positive antidepressant response prediction and assessed if the patient would reach symptomatic remission and if the patient could benefit from treatment with additional antidepressants. [128] Electroencephalography and functional magnetic resonance imaging can be used as biomarkers in psychiatric research and have also been tried to predict depression treatment response, for example, in the former via an SVM —for escitalopram— and in the latter via a “homebrew” algorithm called Sparse EEG Latent SpacE Regression (SELSER) —for sertraline—. [129,130]

Alzheimer’s disease (AD) is the leading neurodegenerative disorder but astonishingly it has over a 99% failure rate in drug development, allegedly chiefly due to its complex and misunderstood pathogenesis. [5,113-116,121] ML designs ranging from ANN to SVM to random forest algorithms have been surveyed for VS of AD-implicated proteins histone deacetylase (HDAC), acetylcholinesterase and S100 calcium-binding protein A9, respectively. In this regard, it is intriguing that there has been an instance where ML was used to ascertain that random forest algorithms would be the best from an assortment of miscellaneous ML models for the prediction of AD drugs and targets. [131-133] In another vein, hyper-predictive ML designs have originated in the form of ANN and random forest trained on PPIs and MD simulations data of binding modes including caspase-8 —another AD-implicated protein— as the target. [134] Posteriorly, following the guidelines laid by these predecessors, a novel graph CNN model has been essayed for VS of the AD-implicated protein beta-secretase 1. [135] Despite these prospects in single-target inhibitors, the complexity of the AD pathogenesis enforces the pursuit of multitarget drugs, which could benefit hugely from hybrid VS through the use of proteochemometrics archetypes. The endless endeavors in that direction, cyclically based on the foregoing, have so far yielded various HDAC inhibitors with the support of diverse ML algorithms (ANN, Bayesian algorithm, recursive partitioning…), but non-specificity and non-selectivity remain nasty issues due to the numerous isoforms in the HDAC family. [136] Another remarkable breakthrough enabled by ML has been the reprofiling of AD drugs for vascular

dementia, given the frequent overlap of both entities. [137,138] In the end, collaborative networks between biotechnology companies, academia, regulators and health care professionals like the Deep Dementia Phenotyping (DEMON) Network appear to be the greatest route to crack the code of better dementia patient treatment. [113]

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder. A few AI-enabled drugs have been essayed to mitigate its symptoms. [5,114,115] Side effects of current medications (mainly levodopa-induced dyskinesia) have also been faced with drug repurposing techniques based upon literature mining. [137,138] Both ligand-based and target-based VS appear efficacious; for instance, the first through the inference of QSAR models of putative inhibitors of leucine-rich repeat kinase 2 protein (a key risk factor in familial and sporadic PD) with an array of ML techniques, and the second through identification of compounds at a time binding to the two receptors —adenosine A2A receptor and dopamine D2 receptor— implicated in the pathophysiology of PD with an SVM model. [139] Additionally, docking and MD studies revealed that an additional ring in piperine-like compounds augments the inhibitory potency against monoamine oxidases A and B. [140] More modern propositions paradoxically involve the use of *in vivo* (i.e. human midbrain organoid model) or *in vitro* (i.e. zebrafish model) experimental designs with ML-based backbone or background reinforcements. [141,142]

6. DISCUSSION & CONCLUSIONS

AI represents a powerhouse to be reckoned with in modern drug discovery. With the floodgates open AI promises to shake the traditional drug discovery pipeline, although current progress is not there yet. AI can push CADD to break past the 13% estimated rate of commercialization, and bring down both the cost (EUR 2.5 billion) and time (12 years) it takes on average to bring a new drug to the market. All at once, ML strategies can reach unexplored areas in drug discovery. For example, the assessment of carbohydrates or lipids as drug candidates or the assessment of proteins outside of the four flagship protein families (G protein-coupled receptors, nuclear receptors, voltage-gated ion channels, and ligand-gated ion channels) in target identification.

As previously stated, the traditional drug discovery pipeline is composed of four distinctive steps: target identification and target validation, high-throughput screening, lead discovery and lead optimization, and pre-clinical, clinical and post-clinical analyses. With the advent of CADD these stages are bound to change. Target identification is currently dependent on protein function prediction strategies, which seem to flourish when sequence-based, structure-based and PPI-based are used in tandem to integrate "multi-omics" information. *In vitro* high-throughput screening has already been dethroned by VS, and is the stage where the greatest progress has been made, because the potential for savings is higher if molecules with considerable "drug-like" properties and targets with a high probability of being "druggable" are tested from the beginning. *In silico* techniques for ligand-based VS (QSAR and pharmacophore modelling), structure-based VS (molecular docking and MD simulation) and hybrid VS are already employed regularly with success and permeate the next stage in the drug discovery pipeline. Lead discovery and lead optimization have bi-faceted AI ramifications, inasmuch as AI can be advantageous for MPO through the promotion of adequate ADMET profiles (with a special interest in toxicity, a leading cause of drug rejection in pre-clinical and clinical studies) and fulfillment of other requirements like Lipinski's rule of five, and AI can also be favorable for drug repurposing. Pre-clinical and clinical analysis is the area where AI has prompted a lesser push. Finally, post-clinical

analysis is likewise understudied but NLP and DL-derived literature mining techniques seem hopeful and drug antagonism/synergism prediction techniques are one area of recent growth (Figure 11).

Moreover, the drug discovery pipeline can now be shortened by performing several steps at a time. *De novo* drug design enables almost simultaneous target identification, VS and MPO with the intervention of DL algorithms. Among those algorithms discussed, LSTM-RNN, AAE and CNN appear the most promising, and the candidate molecules were even tested *in vivo* in some cases. When speaking of molecule representation the string-based enumerated SMILES lead but SELFIES is a very promising 2D representation due to its robustness. Nonetheless, the future is 3D graph-based for sure, since graphs convey a lot more information which in turn makes DL models better. There is at the moment no ground to suggest a favoritism for any other algorithm or molecular representation.

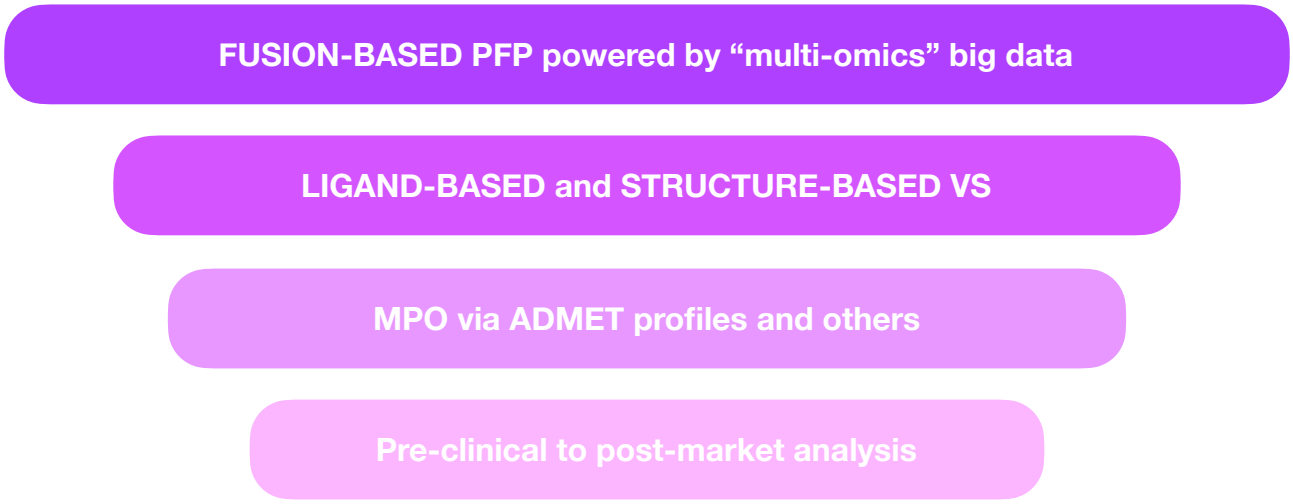


Figure 11. Feasible stage distribution of the modern drug discovery pipeline. See Figure 4 for a reference of the traditional drug discovery pipeline.

One point to highlight from the present review is the overwhelming dominance of DL when compared to other ML models for the resolution of all kinds of problems. In the instances where DL can be applied in the form of CNN, GAN, VAE, AAE, RNN or any other DL model it is nearly doubtless that it is the best option. The only setback is that due to the nature of DL it can not be applied to every problem, and even if it can be applied it is sometimes not optimal from a computing power perspective because the task is minor. In consequence, shallower ML models like SVM, random forest, k-nearest neighbors, decision tree, Bayesian models or many other ML algorithms prevail in the majority of current research because they are easier to train, apply and modify. In fact, in the provided examples of drug discovery in CNS disorders it is apparent that in schizophrenia, depression, AD and PD research an SVM model (modified or not) was attempted at least once and conversely DL methodologies are scarce.

Relatedly, another interesting idea is that there seems to be a perfect match between task and AI tool towards which efforts should strive to reach. All basic ML paradigms initially discussed have certain tasks at which they excel, and the crystal clear evidence is that the bulk of experiments in drug discovery employ models under the supervised learning paradigm, which is classically better for the customary classification and regression tasks in this field. But, in spite of the differences, each research group tends to focus on the

development of a somewhat personal algorithm or model, distancing itself from the “typical” ML/DL nomenclature, like in the case of SELSER. It is perhaps early to fully understand how to classify AI innovations, but the more specific an algorithm is for a task the better results it aspires to achieve.

Although AI in drug discovery presents many advantages from afar, some nuances remain. One such con is the enormous size (approx. 10^{60} pharmacologically active molecules) of the small-molecule search space, together with the current exploratory initiatives orbiting closely around previous hits. Like a large fishing vessel fishing in a pond, no results can emanate if this situation continues. The ocean awaits and should be explored. Secondly, given the size of the chemical search space, data scarcity is a problem and will be for a while even with the magnitude of current databases. To circumvent this issue, flexible ML techniques like transfer learning, multi-task learning or self-supervised learning have been launched that improve outcomes in data-deprived situations. Thirdly, since ML algorithms devised by research groups act like a black box nothing is known from the exterior and results are often irreproducible and performance evaluation impossible. Plus, as previously stated a paradox is established when novelty increases due to a parallel decrease in validity, and vice versa. For now and the coming years, results must always be interpreted conservatively in the setting of AI.

In summary, AI is a wonderful revolution and provides the means to unlock a new era in drug discovery, which has already started. Due to the novelty, nobody should jump to conclusions yet and all possibilities should be explored and re-explored until solid axioms to AI in CADD are revealed by scientific efforts. Be that as it may, fabulous and vibrant results are already at the door, either in CNS drug discovery or other domains, and fear of the unknown should not stop scientific progress. In the end, the mystery and the promise of distant horizons always have called men forward.

7. ACKNOWLEDGEMENTS

Thanks to A. Vicario and H. Briongos for their active participation in the revision of certain areas of their expertise. Thanks as well to my family and girlfriend, and to my friends at the CM Torres Quevedo, who supported me during the making of this thick boy. Final thanks, perchance, to me, myself, and I, for enduring the hardships of manufacturing this beautiful albeit obnoxious concoction.

8. BIBLIOGRAPHY

1. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. Transl Vis Sci Technol. 2020 Feb 27;9(2):14. doi: 10.1167/tvst.9.2.14. PMID: 32704420; PMCID: PMC7347027.

2. Vemula D, Jayasurya P, Sushmitha V, Kumar YN, Bhandari V. CADD, AI and ML in drug discovery: A comprehensive review. Eur J Pharm Sci. 2023 Feb 1;181:106324. doi: 10.1016/j.ejps.2022.106324. Epub 2022 Nov 5. PMID: 36347444.

3. Klambauer G, Hochreiter S, Rarey M. Machine Learning in Drug Discovery. J Chem Inf Model. 2019 Mar 25;59(3):945-946. doi: 10.1021/acs.jcim.9b00136. PMID: 30905159.

4. Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, Cedrón F, Novoa FJ, Carballal A, Maojo V, Pazos A, Fernandez-Lozano C. A review on machine learning approaches and trends in drug discovery. Comput Struct Biotechnol J. 2021 Aug 12;19:4538-4558. doi: 10.1016/j.csbj.2021.08.011. PMID: 34471498; PMCID: PMC8387781.

5. Vatansever S, Schlessinger A, Wacker D, Kaniskan HÜ, Jin J, Zhou MM, Zhang B. Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: State-of-the-arts and future directions. Med Res Rev. 2021 May; 41(3):1427-1473. doi: 10.1002/med.21764. Epub 2020 Dec 9. PMID: 33295676; PMCID: PMC8043990.

6. Coyner AS, Swan R, Campbell JP, Ostmo S, Brown JM, Kalpathy-Cramer J, Kim SJ, Jonas KE, Chan RVP, Chiang MF; Imaging and Informatics in Retinopathy of Prematurity Research Consortium. Automated Fundus Image Quality Assessment in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. Ophthalmol Retina. 2019 May;3(5):444-450. doi: 10.1016/j.oret.2019.01.015. Epub 2019 Jan 31. PMID: 31044738; PMCID: PMC6501831.

7. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, Askham H, Glorot X, O'Donoghue B, Visentin D, van den Driessche G, Lakshminarayanan B, Meyer C, Mackinder F, Bouton S, Ayoub K, Chopra R, King D, Karthikesalingam A, Hughes CO, Raine R, Hughes J, Sim DA, Egan C, Tufail A, Montgomery H, Hassabis D, Rees G, Back T, Khaw PT, Suleyman M, Cornebise J, Keane PA, Ronneberger O. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018 Sep;24(9):1342-1350. doi: 10.1038/s41591-018-0107-6. Epub 2018 Aug 13. PMID: 30104768.

8. Sarkar C, Das B, Rawat VS, Wahlang JB, Nongpiur A, Tiewsoh I, Lyngdoh NM, Das D, Bidarolli M, Sony HT. Artificial Intelligence and Machine Learning Technology Driven Modern Drug Discovery and Development. Int J Mol Sci. 2023 Jan 19;24(3):2026. doi: 10.3390/ijms24032026. PMID: 36768346; PMCID: PMC9916967.

9. Pun FW, Ozerov IV, Zhavoronkov A. AI-powered therapeutic target discovery. Trends Pharmacol Sci. 2023 Sep;44(9):561-572. doi: 10.1016/j.tips.2023.06.010. Epub 2023 Jul 19. PMID: 37479540.

10. Thomas M, Boardman A, Garcia-Ortegon M, Yang H, de Graaf C, Bender A. Applications of Artificial Intelligence in Drug Design: Opportunities and Challenges. Methods Mol Biol. 2022;2390:1-59. doi: 10.1007/978-1-0716-1787-8_1. PMID: 34731463.

11. Tekade RK. The future of pharmaceutical product development and research. London: Academic Press, An Imprint Of Elsevier; 2020.

12. Hinkson IV, Madej B, Stahlberg EA. Accelerating Therapeutics for Opportunities in Medicine: A Paradigm Shift in Drug Discovery. Front Pharmacol. 2020 Jun 30;11:770. doi: 10.3389/fphar.2020.00770. PMID: 32694991; PMCID: PMC7339658.

13. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012 Jan;40(Database

issue):D1100-7. doi: 10.1093/nar/gkr777. Epub 2011 Sep 23. PMID: 21948594; PMCID: PMC3245175.

14. Qureshi R, Irfan M, Gondal TM, Khan S, Wu J, Hadi MU, Heymach J, Le X, Yan H, Alam T. AI in drug discovery and its clinical relevance. Heliyon. 2023 Jul;9(7):e17575. doi: 10.1016/j.heliyon.2023.e17575. Epub 2023 Jun 26. PMID: 37396052; PMCID: PMC10302550.

15. Yan TC, Yue ZX, Xu HQ, Liu YH, Hong YF, Chen GX, Tao L, Xie T. A systematic review of state-of-the-art strategies for machine learning-based protein function prediction. Comput Biol Med. 2023 Mar;154:106446. doi: 10.1016/j.compbiomed.2022.106446. Epub 2022 Dec 21. PMID: 36680931.

16. Bakheet TM, Doig AJ. Properties and identification of human protein drug targets. Bioinformatics. 2009 Feb 15;25(4):451-7. doi: 10.1093/bioinformatics/btp002. Epub 2009 Jan 21. PMID: 19164304.

17. Househ M, Borycki E, Kushniruk A, editors. Multiple Perspectives on Artificial Intelligence in Healthcare: Opportunities and Challenges. 1st ed. Cham, Switzerland: Springer Nature; 2022.

18. Yin X, Bose D, Kwon A, Hanks SC, Jackson AU, Stringham HM, Welch R, Oravilahti A, Fernandes Silva L; FinnGen; Locke AE, Fuchsberger C, Service SK, Erdos MR, Bonnycastle LL, Kuusisto J, Stitziel NO, Hall IM, Morrison J, Ripatti S, Palotie A, Freimer NB, Collins FS, Mohlke KL, Scott LJ, Fauman EB, Burant C, Boehnke M, Laakso M, Wen X. Integrating transcriptomics, metabolomics, and GWAS helps reveal molecular mechanisms for metabolite levels and disease risk. Am J Hum Genet. 2022 Oct 6;109(10):1727-1741. doi: 10.1016/j.ajhg.2022.08.007. Epub 2022 Sep 1. PMID: 36055244; PMCID: PMC9606383.

19. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, Liu TY. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022 Nov 19;23(6):bbac409. doi: 10.1093/bib/bbac409. PMID: 36156661.

20. Hu Y, Zhao T, Zhang N, Zhang Y, Cheng L. A Review of Recent Advances and Research on Drug Target Identification Methods. Curr Drug Metab. 2019;20(3): 209-216. doi: 10.2174/1389200219666180925091851. PMID: 30251599.

21. Bonetta R, Valentino G. Machine learning techniques for protein function prediction. Proteins. 2020 Mar;88(3):397-413. doi: 10.1002/prot.25832. Epub 2019 Nov 14. PMID: 31603244.

22. Lv Z, Ao C, Zou Q. Protein Function Prediction: From Traditional Classifier to Deep Learning. Proteomics. 2019 Jul;19(14):e1900119. doi: 10.1002/pmic.201900119. Epub 2019 Jul 11. PMID: 31187588.

23. Kaleel M, Zheng Y, Chen J, Feng X, Simpson JC, Pollastri G, Mooney C. SCLpred-EMS: subcellular localization prediction of endomembrane system and secretory pathway proteins by Deep N-to-1 Convolutional Neural Networks. Bioinformatics. 2020 Jun 1;36(11):3343-3349. doi: 10.1093/bioinformatics/btaa156. PMID: 32142105.

24. Kulmanov M, Zhapa-Camacho F, Hoehndorf R. DeepGOWeb: fast and accurate protein function prediction on the (Semantic) Web. Nucleic Acids Res. 2021 Jul

2;49(W1):W140-W146. doi: 10.1093/nar/gkab373. PMID: 34019664; PMCID: PMC8262746.

25. Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics. 2020 Jan 15;36(2):422-429. doi: 10.1093/bioinformatics/btz595. Erratum in: Bioinformatics. 2021 May 23;37(8):1187. PMID: 31350877; PMCID: PMC9883727.

26. Törönen P, Holm L. PANNZER-A practical tool for protein function prediction. Protein Sci. 2022 Jan;31(1):118-128. doi: 10.1002/pro.4193. Epub 2021 Oct 14. PMID: 34562305; PMCID: PMC8740830.

27. Liu YW, Hsu TW, Chang CY, Liao WH, Chang JM. GODoc: high-throughput protein function prediction using novel k-nearest-neighbor and voting algorithms. BMC Bioinformatics. 2020 Nov 18;21(Suppl 6):276. doi: 10.1186/s12859-020-03556-9. PMID: 33203348; PMCID: PMC7672824.

28. Xia W, Zheng L, Fang J, Li F, Zhou Y, Zeng Z, Zhang B, Li Z, Li H, Zhu F. PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. Comput Biol Med. 2022 Jun;145:105465. doi: 10.1016/j.compbiomed.2022.105465. Epub 2022 Mar 28. PMID: 35366467.

29. Zhao C, Liu T, Wang Z. PANDA2: protein function prediction using graph neural networks. NAR Genom Bioinform. 2022 Feb 2;4(1):lqac004. doi: 10.1093/nargab/lqac004. PMID: 35118378; PMCID: PMC8808544.

30. Hakala K, Kaewphan S, Bjorne J, Mehryary F, Moen H, Tolvanen M, Salakoski T, Ginter F. Neural Network and Random Forest Models in Protein Function Prediction. IEEE/ACM Trans Comput Biol Bioinform. 2022 May-Jun;19(3):1772-1781. doi: 10.1109/TCBB.2020.3044230. Epub 2022 Jun 3. PMID: 33306472.

31. Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S. Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. Methods Mol Biol. 2017;1607:627-641. doi: 10.1007/978-1-4939-7000-1_26. PMID: 28573592; PMCID: PMC5823500.

32. Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, Chandler C, Taylor BC, Fisk IM, Vlamakis H, Xavier RJ, Knight R, Cho K, Bonneau R. Structure-based protein function prediction using graph convolutional networks. Nat Commun. 2021 May 26;12(1):3168. doi: 10.1038/s41467-021-23303-9. PMID: 34039967; PMCID: PMC8155034.

33. Lai B, Xu J. Accurate protein function prediction via graph attention networks with predicted structure information. Brief Bioinform. 2022 Jan 17;23(1):bbab502. doi: 10.1093/bib/bbab502. PMID: 34882195; PMCID: PMC8898000.

34. Wan C, Cozzetto D, Fa R, Jones DT. Using deep maxout neural networks to improve the accuracy of function prediction from protein interaction networks. PLoS One. 2019 Jul 23;14(7):e0209958. doi: 10.1371/journal.pone.0209958. PMID: 31335894; PMCID: PMC6650051.

35. Jiang B, Kloster K, Gleich DF, Gribskov M. AptRank: an adaptive PageRank model for protein function prediction on bi-relational graphs. Bioinformatics. 2017 Jun 15;33(12):1829-1836. doi: 10.1093/bioinformatics/btx029. PMID: 28200073.

36. Devkota K, Schmidt H, Werenski M, Murphy JM, Erden M, Arsenescu V, Cowen LJ. GLIDER: function prediction from GLIDE-based neighborhoods. Bioinformatics. 2022 Jun 27;38(13):3395-3406. doi: 10.1093/bioinformatics/btac322. PMID: 35575379; PMCID: PMC9237677.

37. Pan X, Chen L, Liu M, Niu Z, Huang T, Cai YD. Identifying Protein Subcellular Locations With Embeddings-Based node2loc. IEEE/ACM Trans Comput Biol Bioinform. 2022 Mar-Apr;19(2):666-675. doi: 10.1109/TCBB.2021.3080386. Epub 2022 Apr 1. PMID: 33989156.

38. Gligorijevic V, Barot M, Bonneau R. deepNF: deep network fusion for protein function prediction. Bioinformatics. 2018 Nov 15;34(22):3873-3881. doi: 10.1093/bioinformatics/bty440. PMID: 29868758; PMCID: PMC6223364.

39. Yaseen A, Amin I, Akhter N, Ben-Hur A, Minhas F. Insights into performance evaluation of compound-protein interaction prediction methods. Bioinformatics. 2022 Sep 16;38(Suppl_2):ii75-ii81. doi: 10.1093/bioinformatics/btac496. PMID: 36124806.

40. Li Z, Miao Q, Yan F, Meng Y, Zhou P. Machine Learning in Quantitative Protein-peptide Affinity Prediction: Implications for Therapeutic Peptide Design. Curr Drug Metab. 2019;20(3):170-176. doi: 10.2174/1389200219666181012151944. PMID: 30317994.

41. Hashemifar S, Neyshabur B, Khan AA, Xu J. Predicting protein-protein interactions through sequence-based deep learning. Bioinformatics. 2018 Sep 1;34(17):i802-i810. doi: 10.1093/bioinformatics/bty573. PMID: 30423091; PMCID: PMC6129267.

42. Corbi-Verge C, Kim PM. Motif mediated protein-protein interactions as drug targets. Cell Commun Signal. 2016 Mar 2;14:8. doi: 10.1186/s12964-016-0131-4. PMID: 26936767; PMCID: PMC4776425.

43. Tsubaki M, Tomii K, Sese J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics. 2019 Jan 15;35(2):309-318. doi: 10.1093/bioinformatics/bty535. PMID: 29982330.

44. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: predicting drug-target binding affinity with graph neural networks. Bioinformatics. 2021 May 23;37(8):1140-1147. doi: 10.1093/bioinformatics/btaa921. PMID: 33119053.

45. Roy A, Yang J, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W471-7. doi: 10.1093/nar/gks372. Epub 2012 May 8. PMID: 22570420; PMCID: PMC3394312.

46. Smaili FZ, Tian S, Roy A, Alazmi M, Arold ST, Mukherjee S, Hefty PS, Chen W, Gao X. QAUST: Protein Function Prediction Using Structure Similarity, Protein Interaction, and Functional Motifs. Genomics Proteomics Bioinformatics. 2021 Dec;19(6):998-1011. doi: 10.1016/j.gpb.2021.02.001. Epub 2021 Feb 23. PMID: 33631427; PMCID: PMC9403031.

47. MacCarthy EA, Zhang C, Zhang Y, Kc DB. GPU-I-TASSER: a GPU accelerated I-TASSER protein structure prediction tool. Bioinformatics. 2022 Mar 4;38(6):1754-1755. doi: 10.1093/bioinformatics/btab871. PMID: 34978562; PMCID: PMC8896630.

48. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021 Jan 8;49(D1):D605-D612. doi: 10.1093/nar/gkaa1074. Erratum in: Nucleic Acids Res. 2021 Oct 11;49(18):10800. PMID: 33237311; PMCID: PMC7779004.

49. Khan IK, Jain A, Rawi R, Bensmail H, Kihara D. Prediction of protein group function by iterative classification on functional relevance network. Bioinformatics. 2019 Apr 15;35(8):1388-1394. doi: 10.1093/bioinformatics/bty787. PMID: 30192921; PMCID: PMC6477972.

50. Li Q, Shah S. Structure-Based Virtual Screening. Methods Mol Biol. 2017;1558:111-124. doi: 10.1007/978-1-4939-6783-4_5. PMID: 28150235.

51. Maia EHB, Assis LC, de Oliveira TA, da Silva AM, Taranto AG. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. Front Chem. 2020 Apr 28;8:343. doi: 10.3389/fchem.2020.00343. PMID: 32411671; PMCID: PMC7200080.

52. Chen YC. Beware of docking! Trends Pharmacol Sci. 2015 Feb;36(2):78-95. doi: 10.1016/j.tips.2014.12.001. Epub 2014 Dec 24. Erratum in: Trends Pharmacol Sci. 2015 Sep;36(9):617. PMID: 25543280.

53. Liu W, Schmidt B, Voss G, Müller-Wittig W. Accelerating molecular dynamics simulations using Graphics Processing Units with CUDA. Comput Phys Commun. 2008;179(9):634–41. doi: 10.1016/j.cpc.2008.05.008

54. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J Comput Chem. 2004 Jul 15;25(9):1157-74. doi: 10.1002/jcc.20035. Erratum in: J Comput Chem. 2005 Jan 15;26(1):114. PMID: 15116359.

55. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, Mackerell AD Jr. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J Comput Chem. 2010 Mar;31(4):671-90. doi: 10.1002/jcc.21367. PMID: 19575467; PMCID: PMC2888302.

56. Meier K, Schmid N, van Gunsteren WF. Interfacing the GROMOS (bio)molecular simulation software to quantum-chemical program packages. J Comput Chem. 2012 Oct 5;33(26):2108-17. doi: 10.1002/jcc.23047. Epub 2012 Jun 27. PMID: 22736402.

57. Unke OT, Chmiela S, Sauceda HE, Gastegger M, Poltavsky I, Schütt KT, Tkatchenko A, Müller KR. Machine Learning Force Fields. Chem Rev. 2021 Aug 25;121(16):10142-10186. doi: 10.1021/acs.chemrev.0c01111. Epub 2021 Mar 11. PMID: 33705118; PMCID: PMC8391964.

58. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids

Res. 2008 Jan;36(Database issue):D901-6. doi: 10.1093/nar/gkm958. Epub 2007 Nov 29. PMID: 18048412; PMCID: PMC2238889.

59. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235-42. doi: 10.1093/nar/28.1.235. PMID: 10592235; PMCID: PMC102472.

60. Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations. Nat Mach Intell. 2021;3(12):1023–32. doi: 10.1038/s42256-021-00418-8

61. Martinelli DD. Generative machine learning for de novo drug discovery: A systematic review. Comput Biol Med. 2022 Jun;145:105403. doi: 10.1016/j.compbiomed. 2022.105403. Epub 2022 Mar 13. PMID: 35339849.

62. 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. Int J Mol Sci. 2023 Mar 31;24(7):6573. doi: 10.3390/ijms24076573. PMID: 37047543; PMCID: PMC10095548.

63. Chen W, Liu X, Zhang S, Chen S. Artificial intelligence for drug discovery: Resources, methods, and applications. Mol Ther Nucleic Acids. 2023 Feb 18;31:691-702. doi: 10.1016/j.omtn.2023.02.019. PMID: 36923950; PMCID: PMC10009646.

64. Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de-novo design through deep reinforcement learning. J Cheminform. 2017 Sep 4;9(1):48. doi: 10.1186/ s13321-017-0235-x. PMID: 29086083; PMCID: PMC5583141.

65. Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. Sci Adv. 2018 Jul 25;4(7):eaap7885. doi: 10.1126/sciadv.aap7885. PMID: 30050984; PMCID: PMC6059760.

66. Gupta A, Müller AT, Huisman BJH, Fuchs JA, Schneider P, Schneider G. Generative Recurrent Networks for De Novo Drug Design. Mol Inform. 2018 Jan;37(1-2):1700111. doi: 10.1002/minf.201700111. Epub 2017 Nov 2. Erratum in: Mol Inform. 2018 Jan; 37(1-2): PMID: 29095571; PMCID: PMC5836943.

67. Merk D, Friedrich L, Grisoni F, Schneider G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. Mol Inform. 2018 Jan;37(1-2):1700153. doi: 10.1002/minf.201700153. Epub 2018 Jan 10. PMID: 29319225; PMCID: PMC5838524.

68. Xia W, Zheng L, Fang J, Li F, Zhou Y, Zeng Z, Zhang B, Li Z, Li H, Zhu F. PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. Comput Biol Med. 2022 Jun;145:105465. doi: 10.1016/j.compbiomed.2022.105465. Epub 2022 Mar 28. PMID: 35366467.

69. Sanchez-Lengeling B, Outeiral C, Guimaraes GL, Aspuru-Guzik A. Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). ChemRxiv. 2017; doi:10.26434/ chemrxiv.5309668.v3

70. Putin E, Asadulaev A, Vanhaelen Q, Ivanenkov Y, Aladinskaya AV, Aliper A, Zhavoronkov A. Adversarial Threshold Neural Computer for Molecular de Novo Design.

Mol Pharm. 2018 Oct 1;15(10):4386-4397. doi: 10.1021/acs.molpharmaceut.7b01137. Epub 2018 Mar 30. PMID: 29569445.

71. Putin E, Asadulaev A, Ivanenkov Y, Aladinskiy V, Sanchez-Lengeling B, Aspuru-Guzik A, Zhavoronkov A. Reinforced Adversarial Neural Computer for de Novo Molecular Design. J Chem Inf Model. 2018 Jun 25;58(6):1194-1204. doi: 10.1021/acs.jcim.7b00690. Epub 2018 Jun 12. PMID: 29762023.

72. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Cent Sci. 2018 Feb 28;4(2):268-276. doi: 10.1021/acscentsci.7b00572. Epub 2018 Jan 12. PMID: 29532027; PMCID: PMC5833007.

73. Lim J, Ryu S, Kim JW, Kim WY. Molecular generative model based on conditional variational autoencoder for de novo molecular design. J Cheminform. 2018 Jul 11;10(1):31. doi: 10.1186/s13321-018-0286-7. PMID: 29995272; PMCID: PMC6041224.

74. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Polykovskiy DA, Kuznetsov MD, Asadulaev A, Volkov Y, Zholus A, Shayakhmetov RR, Zhebrak A, Minaeva LI, Zagribelnyy BA, Lee LH, Soll R, Madge D, Xing L, Guo T, Aspuru-Guzik A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019 Sep;37(9):1038-1040. doi: 10.1038/s41587-019-0224-x. Epub 2019 Sep 2. PMID: 31477924.

75. Hosseini MP, Lu S, Kamaraj K, Slowikowski A, Venkatesh H. Deep Learning Architectures. Studies in computational intelligence. 2019;1–24.

76. Berrhail F, Belhadef H, Haddad M. Deep Convolutional Neural Network to improve the performances of screening process in LBVS. Expert Syst Appl. 2022;203(117287):117287. doi: 10.1016/j.eswa.2022.117287

77. Huang K, Fu T, Glass LM, Zitnik M, Xiao C, Sun J. DeepPurpose: a deep learning library for drug-target interaction prediction. Bioinformatics. 2021 Apr 1;36(22-23):5545-5547. doi: 10.1093/bioinformatics/btaa1005. PMID: 33275143; PMCID: PMC8016467.

78. 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;1(4):045024. doi: 10.1088/2632-2153/aba947

79. Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods. 2015 Jan;71:58-63. doi: 10.1016/j.ymeth.2014.08.005. Epub 2014 Aug 15. PMID: 25132639.

80. David L, Thakkar A, Mercado R, Engkvist O. Molecular representations in AI-driven drug discovery: a review and practical guide. J Cheminform. 2020 Sep 17;12(1):56. doi: 10.1186/s13321-020-00460-5. PMID: 33431035; PMCID: PMC7495975.

81. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28(1):31–6. doi: 10.1021/ci00057a005

82. Liu X, Ye K, van Vlijmen HWT, Emmerich MTM, IJzerman AP, van Westen GJP. DrugEx v2: de novo design of drug molecules by Pareto-based multi-objective reinforcement learning in polypharmacology. J Cheminform. 2021 Nov 12;13(1):85. doi: 10.1186/s13321-021-00561-9. PMID: 34772471; PMCID: PMC8588612.

83. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583-589. doi: 10.1038/s41586-021-03819-2. Epub 2021 Jul 15. PMID: 34265844; PMCID: PMC8371605.

84. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015 Aug;33(8): 831-8. doi: 10.1038/nbt.3300. Epub 2015 Jul 27. PMID: 26213851.

85. Karimi M, Wu D, Wang Z, Shen Y. DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. Bioinformatics. 2019 Sep 15;35(18):3329-3338. doi: 10.1093/bioinformatics/btz111. PMID: 30768156; PMCID: PMC6748780.

86. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021 Jan;26(1):80-93. doi: 10.1016/j.drudis.2020.10.010. Epub 2020 Oct 21. PMID: 33099022; PMCID: PMC7577280.

87. Bueno J. ADMETox: Bringing nanotechnology closer to Lipinski's rule of five. In: Nanotechnology in the Life Sciences. Cham: Springer International Publishing; 2020. p. 61–74.

88. Morgan P, Brown DG, Lennard S, Anderton MJ, Barrett JC, Eriksson U, Fidock M, Hamrén B, Johnson A, March RE, Matcham J, Mettetal J, Nicholls DJ, Platz S, Rees S, Snowden MA, Pangalos MN. Impact of a five-dimensional framework on R&D productivity at AstraZeneca. Nat Rev Drug Discov. 2018 Mar;17(3):167-181. doi: 10.1038/nrd.2017.244. Epub 2018 Jan 19. PMID: 29348681.

89. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity Prediction using Deep Learning. Front Environ Sci. 2016;3. doi: 10.3389/fenvs.2015.00080

90. Awale M, Reymond JL. Polypharmacology Browser PPB2: Target Prediction Combining Nearest Neighbors with Machine Learning. J Chem Inf Model. 2019 Jan 28;59(1):10-17. doi: 10.1021/acs.jcim.8b00524. Epub 2018 Dec 31. PMID: 30558418.

91. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Polykovskiy DA, Kuznetsov MD, Asadulaev A, Volkov Y, Zholus A, Shayakhmetov RR, Zhebrak A, Minaeva LI, Zagribelnyy BA, Lee LH, Soll R, Madge D, Xing L, Guo T, Aspuru-Guzik A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019 Sep;37(9):1038-1040. doi: 10.1038/s41587-019-0224-x. Epub 2019 Sep 2. PMID: 31477924.

92. Walters WP, Murcko M. Assessing the impact of generative AI on medicinal chemistry. Nat Biotechnol. 2020 Feb;38(2):143-145. doi: 10.1038/s41587-020-0418-2. PMID: 32001834.

93. Cai C, Wang S, Xu Y, Zhang W, Tang K, Ouyang Q, Lai L, Pei J. Transfer Learning for Drug Discovery. J Med Chem. 2020 Aug 27;63(16):8683-8694. doi: 10.1021/acs.jmedchem.9b02147. Epub 2020 Jul 24. PMID: 32672961.

94. Sosnin S, Vashurina M, Withnall M, Karpov P, Fedorov M, Tetko IV. A Survey of Multi-task Learning Methods in Chemoinformatics. Mol Inform. 2019 Apr;38(4):e1800108. doi: 10.1002/minf.201800108. Epub 2018 Nov 28. PMID: 30499195; PMCID: PMC6587441.

95. Li X, Fourches D. Inductive transfer learning for molecular activity prediction: Next-Gen QSAR Models with MolPMoFiT. J Cheminform. 2020 Apr 22;12(1):27. doi: 10.1186/s13321-020-00430-x. PMID: 33430978; PMCID: PMC7178569.

96. Mak KK, Pichika MR. Artificial intelligence in drug development: present status and future prospects. Drug Discov Today. 2019 Mar;24(3):773-780. doi: 10.1016/j.drudis.2018.11.014. Epub 2018 Nov 22. PMID: 30472429.

97. Koromina M, Pandi MT, Patrinos GP. Rethinking Drug Repositioning and Development with Artificial Intelligence, Machine Learning, and Omics. OMICS. 2019 Nov;23(11):539-548. doi: 10.1089/omi.2019.0151. Epub 2019 Oct 25. PMID: 31651216.

98. Park K. A review of computational drug repurposing. Transl Clin Pharmacol. 2019 Jun;27(2):59-63. doi: 10.12793/tcp.2019.27.2.59. Epub 2019 Jun 28. PMID: 32055582; PMCID: PMC6989243.

99. Lotfi Shahreza M, Ghadiri N, Mousavi SR, Varshosaz J, Green JR. A review of network-based approaches to drug repositioning. Brief Bioinform. 2018 Sep 28;19(5):878-892. doi: 10.1093/bib/bbx017. PMID: 28334136.

100. Mouchlis VD, Afantitis A, Serra A, Fratello M, Papadiamantis AG, Aidinis V, Lynch I, Greco D, Melagraki G. Advances in de Novo Drug Design: From Conventional to Machine Learning Methods. Int J Mol Sci. 2021 Feb 7;22(4):1676. doi: 10.3390/ijms22041676. PMID: 33562347; PMCID: PMC7915729.

101. Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, Oprea TI, Baskin II, Varnek A, Roitberg A, Isayev O, Curtarolo S, Fourches D, Cohen Y, Aspuru-Guzik A, Winkler DA, Agrafiotis D, Cherkasov A, Tropsha A. QSAR without borders. Chem Soc Rev. 2020 Jun 7;49(11):3525-3564. doi: 10.1039/d0cs00098a. Epub 2020 May 1. Erratum in: Chem Soc Rev. 2020 May 22;:. Curtalolo, Stefano [corrected to Curtarolo, Stefano]. PMID: 32356548; PMCID: PMC8008490.

102. Gao Q, Yang L, Zhu Y. Pharmacophore based drug design approach as a practical process in drug discovery. Curr Comput Aided Drug Des. 2010 Mar;6(1):37-49. doi: 10.2174/157340910790980151. PMID: 20370694.

103. Zhao S, Li X, Peng W, Wang L, Ye W, Zhao Y, Yin W, Chen WD, Li W, Wang YD. Ligand-based pharmacophore modeling, virtual screening and biological evaluation to identify novel TGR5 agonists. RSC Adv. 2021 Mar 2;11(16):9403-9409. doi: 10.1039/d0ra10168k. PMID: 35423434; PMCID: PMC8695346.

104. Bongers BJ, IJzerman AP, Van Westen GJP. Proteochemometrics - recent developments in bioactivity and selectivity modeling. Drug Discov Today Technol. 2019 Dec;32-33:89-98. doi: 10.1016/j.ddtec.2020.08.003. Epub 2020 Sep 20. PMID: 33386099.

105. van Westen GJP, Wegner JK, IJzerman AP, van Vlijmen HWT, Bender A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. Medchemcomm. 2011;2(1):16–30. doi: 10.1039/c0md00165a

106. Cortés-Ciriano I, Ain QU, Subramanian V, Lenselink EB, Méndez-Lucio O, IJzerman AP, et al. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. Medchemcomm. 2015;6(1):24–50. doi: 10.1039/c4md00216d

107. Davenport T, Guha A, Grewal D, Bressgott T. How artificial intelligence will change the future of marketing. J Acad Mark Sci. 2020;48(1):24–42. doi: 10.1007/s11747-019-00696-0

108. Syam N, Sharma A. Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. Ind Mark Manag. 2018;69:135–46. doi: 10.1016/j.indmarman.2017.12.019

109. Li P, Huang C, Fu Y, Wang J, Wu Z, Ru J, Zheng C, Guo Z, Chen X, Zhou W, Zhang W, Li Y, Chen J, Lu A, Wang Y. Large-scale exploration and analysis of drug combinations. Bioinformatics. 2015 Jun 15;31(12):2007-16. doi: 10.1093/bioinformatics/btv080. Epub 2015 Feb 8. PMID: 25667546.

110. Wildenhain J, Spitzer M, Dolma S, Jarvik N, White R, Roy M, Griffiths E, Bellows DS, Wright GD, Tyers M. Prediction of Synergism from Chemical-Genetic Interactions by Machine Learning. Cell Syst. 2015 Dec 23;1(6):383-95. doi: 10.1016/j.cels.2015.12.003. Epub 2015 Dec 23. PMID: 27136353; PMCID: PMC5998823.

111. Preuer K, Lewis RPI, Hochreiter S, Bender A, Bulusu KC, Klambauer G. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. Bioinformatics. 2018 May 1;34(9):1538-1546. doi: 10.1093/bioinformatics/btx806. PMID: 29253077; PMCID: PMC5925774.

112. Brady LS, Lisanby SH, Gordon JA. New directions in psychiatric drug development: promising therapeutics in the pipeline. Expert Opin Drug Discov. 2023 Jul-Dec;18(8):835-850. doi: 10.1080/17460441.2023.2224555. Epub 2023 Jun 23. PMID: 37352473.

113. Doherty T, Yao Z, Khleifat AAL, Tantiangco H, Tamburin S, Albertyn C, Thakur L, Llewellyn DJ, Oxtoby NP, Lourida I; Deep Dementia Phenotyping (DEMON) Network; Ranson JM, Duce JA. Artificial intelligence for dementia drug discovery and trials optimization. Alzheimers Dement. 2023 Dec;19(12):5922-5933. doi: 10.1002/alz.13428. Epub 2023 Aug 16. PMID: 37587767.

114. Dorahy G, Chen JZ, Balle T. Computer-Aided Drug Design towards New Psychotropic and Neurological Drugs. Molecules. 2023 Jan 30;28(3):1324. doi: 10.3390/molecules28031324. PMID: 36770990; PMCID: PMC9921936.

115. Gautam V, Gaurav A, Masand N, Lee VS, Patil VM. Artificial intelligence and machine-learning approaches in structure and ligand-based discovery of drugs affecting central nervous system. Mol Divers. 2023 Apr;27(2):959-985. doi: 10.1007/s11030-022-10489-3. Epub 2022 Jul 11. PMID: 35819579.

116. Qiu Y, Cheng F. Artificial intelligence for drug discovery and development in Alzheimer's disease. Curr Opin Struct Biol. 2024 Apr;85:102776. doi: 10.1016/j.sbi.2024.102776. Epub 2024 Feb 8. PMID: 38335558.

117. Doniger S, Hofmann T, Yeh J. Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms. J Comput Biol. 2002;9(6):849-64. doi: 10.1089/10665270260518317. PMID: 12614551.

118. Hindle SJ, Munji RN, Dolghih E, Gaskins G, Orng S, Ishimoto H, Soung A, DeSalvo M, Kitamoto T, Keiser MJ, Jacobson MP, Daneman R, Bainton RJ. Evolutionarily Conserved Roles for Blood-Brain Barrier Xenobiotic Transporters in Endogenous Steroid Partitioning and Behavior. Cell Rep. 2017 Oct 31;21(5):1304-1316. doi: 10.1016/j.celrep.2017.10.026. PMID: 29091768; PMCID: PMC5774027.

119. Miller DS. Regulation of ABC transporters blood-brain barrier: the good, the bad, and the ugly. Adv Cancer Res. 2015;125:43-70. doi: 10.1016/bs.acr.2014.10.002. Epub 2015 Jan 8. PMID: 25640266.

120. Miao R, Xia LY, Chen HH, Huang HH, Liang Y. Improved Classification of Blood-Brain-Barrier Drugs Using Deep Learning. Sci Rep. 2019 Jun 19;9(1):8802. doi: 10.1038/s41598-019-44773-4. PMID: 31217424; PMCID: PMC6584536.

121. Rema J, Novais F, Telles-Correia D. Precision Psychiatry: Machine Learning as a Tool to Find New Pharmacological Targets. Curr Top Med Chem. 2022;22(15):1261-1269. doi: 10.2174/1568026621666211004095917. PMID: 34607546.

122. Yang QX, Wang YX, Li FC, Zhang S, Luo YC, Li Y, Tang J, Li B, Chen YZ, Xue WW, Zhu F. Identification of the gene signature reflecting schizophrenia's etiology by constructing artificial intelligence-based method of enhanced reproducibility. CNS Neurosci Ther. 2019 Sep;25(9):1054-1063. doi: 10.1111/cns.13196. Epub 2019 Jul 27. PMID: 31350824; PMCID: PMC6698965.

123. Hsu KC, Wang FS. Model-based optimization approaches for precision medicine: A case study in presynaptic dopamine overactivity. PLoS One. 2017 Jun 14;12(6):e0179575. doi: 10.1371/journal.pone.0179575. PMID: 28614410; PMCID: PMC5470743.

124. Zhao K, So HC. Drug Repositioning for Schizophrenia and Depression/Anxiety Disorders: A Machine Learning Approach Leveraging Expression Data. IEEE J Biomed Health Inform. 2019 May;23(3):1304-1315. doi: 10.1109/JBHI.2018.2856535. Epub 2018 Jul 16. PMID: 30010603.

125. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, Cannon TD, Krystal JH, Corlett PR. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016 Mar;3(3):243-50. doi: 10.1016/S2215-0366(15)00471-X. Epub 2016 Jan 21. PMID: 26803397.

126. Tian S, Sun Y, Shao J, Zhang S, Mo Z, Liu X, Wang Q, Wang L, Zhao P, Chattun MR, Yao Z, Si T, Lu Q. Predicting escitalopram monotherapy response in depression: The role of anterior cingulate cortex. Hum Brain Mapp. 2020 Apr 1;41(5):1249-1260. doi: 10.1002/hbm.24872. Epub 2019 Nov 22. PMID: 31758634; PMCID: PMC7268019.

127. Chekroud AM, Gueorguieva R, Krumholz HM, Trivedi MH, Krystal JH, McCarthy G. Reevaluating the Efficacy and Predictability of Antidepressant Treatments: A Symptom Clustering Approach. JAMA Psychiatry. 2017 Apr 1;74(4):370-378. doi: 10.1001/jamapsychiatry.2017.0025. PMID: 28241180; PMCID: PMC5863470.

128. Chang B, Choi Y, Jeon M, Lee J, Han KM, Kim A, Ham BJ, Kang J. ARPNet: Antidepressant Response Prediction Network for Major Depressive Disorder. Genes (Basel). 2019 Nov 7;10(11):907. doi: 10.3390/genes10110907. PMID: 31703457; PMCID: PMC6895829.

129. Zhdanov A, Atluri S, Wong W, Vaghei Y, Daskalakis ZJ, Blumberger DM, Frey BN, Giacobbe P, Lam RW, Milev R, Mueller DJ, Turecki G, Parikh SV, Rotzinger S, Soares CN, Brenner CA, Vila-Rodriguez F, McAndrews MP, Kleffner K, Alonso-Prieto E, Arnott SR, Foster JA, Strother SC, Uher R, Kennedy SH, Farzan F. Use of Machine Learning for Predicting Escitalopram Treatment Outcome From Electroencephalography Recordings in Adult Patients With Depression. JAMA Netw Open. 2020 Jan 3;3(1):e1918377. doi: 10.1001/jamanetworkopen.2019.18377. PMID: 31899530; PMCID: PMC6991244.

130. Wu W, Zhang Y, Jiang J, Lucas MV, Fonzo GA, Rolle CE, Cooper C, Chin-Fatt C, Krepel N, Cornelssen CA, Wright R, Toll RT, Trivedi HM, Monuszko K, Caudle TL, Sarhadi K, Jha MK, Trombello JM, Deckersbach T, Adams P, McGrath PJ, Weissman MM, Fava M, Pizzagalli DA, Arns M, Trivedi MH, Etkin A. An electroencephalographic signature predicts antidepressant response in major depression. Nat Biotechnol. 2020 Apr;38(4):439-447. doi: 10.1038/s41587-019-0397-3. Epub 2020 Feb 10. PMID: 32042166; PMCID: PMC7145761.

131. Hung TC, Lee WY, Chen KB, Chan YC, Lee CC, Chen CY. In silico investigation of traditional Chinese medicine compounds to inhibit human histone deacetylase 2 for patients with Alzheimer's disease. Biomed Res Int. 2014;2014:769867. doi: 10.1155/2014/769867. Epub 2014 Jun 23. PMID: 25045700; PMCID: PMC4090436.

132. Cavas L, Topcam G, Gundogdu-Hizliates C, Ergun Y. Neural Network Modeling of AChE Inhibition by New Carbazole-Bearing Oxazolones. Interdiscip Sci. 2019 Mar; 11(1):95-107. doi: 10.1007/s12539-017-0245-4. Epub 2017 Dec 13. PMID: 29236214.

133. Lee J, Kumar S, Lee SY, Park SJ, Kim MH. Development of Predictive Models for Identifying Potential S100A9 Inhibitors Based on Machine Learning Methods. Front Chem. 2019 Nov 25;7:779. doi: 10.3389/fchem.2019.00779. PMID: 31824919; PMCID: PMC6886474.

134. Jamal S, Grover A, Grover S. Machine Learning From Molecular Dynamics Trajectories to Predict Caspase-8 Inhibitors Against Alzheimer's Disease. Front Pharmacol. 2019 Jul 12;10:780. doi: 10.3389/fphar.2019.00780. PMID: 31354494; PMCID: PMC6639425.

135. Miyazaki Y, Ono N, Huang M, Altaf-Ul-Amin M, Kanaya S. Comprehensive Exploration of Target-specific Ligands Using a Graph Convolution Neural Network. Mol

Inform. 2020 Jan;39(1-2):e1900095. doi: 10.1002/minf.201900095. Epub 2019 Dec 9. PMID: 31815371; PMCID: PMC7050504.

136. Kleandrova VV, Speck-Planche A. PTML Modeling for Alzheimer's Disease: Design and Prediction of Virtual Multi-Target Inhibitors of GSK3B, HDAC1, and HDAC6. Curr Top Med Chem. 2020;20(19):1661-1676. doi: 10.2174/1568026620666200607190951. PMID: 32515311.

137. Nam Y, Kim M, Chang HS, Shin H. Drug repurposing with network reinforcement. BMC Bioinformatics. 2019 Jul 24;20(Suppl 13):383. doi: 10.1186/s12859-019-2858-6. PMID: 31337333; PMCID: PMC6651901.

138. Johnston TH, Lacoste AMB, Visanji NP, Lang AE, Fox SH, Brotchie JM. Repurposing drugs to treat l-DOPA-induced dyskinesia in Parkinson's disease. Neuropharmacology. 2019 Mar 15;147:11-27. doi: 10.1016/j.neuropharm.2018.05.035. Epub 2018 Jun 1. PMID: 29907424.

139. Sebastián-Pérez V, Martínez MJ, Gil C, Campillo NE, Martínez A, Ponzoni I. QSAR Modelling to Identify LRRK2 Inhibitors for Parkinson's Disease. J Integr Bioinform. 2019 Feb 14;16(1):20180063. doi: 10.1515/jib-2018-0063. PMID: 30763264; PMCID: PMC6798859.

140. Shao YM, Ma X, Paira P, Tan A, Herr DR, Lim KL, Ng CH, Venkatesan G, Klotz KN, Federico S, Spalluto G, Cheong SL, Chen YZ, Pastorin G. Discovery of indolylpiperazinylpyrimidines with dual-target profiles at adenosine A2A and dopamine D2 receptors for Parkinson's disease treatment. PLoS One. 2018 Jan 5;13(1):e0188212. doi: 10.1371/journal.pone.0188212. PMID: 29304113; PMCID: PMC5755735.

141. Monzel AS, Hemmer K, Kaoma T, Smits LM, Bolognin S, Lucarelli P, Rosety I, Zagare A, Antony P, Nickels SL, Krueger R, Azuaje F, Schwamborn JC. Machine learning-assisted neurotoxicity prediction in human midbrain organoids. Parkinsonism Relat Disord. 2020 Jun;75:105-109. doi: 10.1016/j.parkreldis.2020.05.011. Epub 2020 May 8. PMID: 32534431.

142. Hughes GL, Lones MA, Bedder M, Currie PD, Smith SL, Pownall ME. Machine learning discriminates a movement disorder in a zebrafish model of Parkinson's disease. Dis Model Mech. 2020 Oct 16;13(10):dmm045815. doi: 10.1242/dmm.045815. PMID: 32859696; PMCID: PMC7578351.