CAPÍTULO 1

BIG DATA AND ARTIFICIAL INTELLIGENCE (AI) AS COMPUTATIONAL TOOLS IN THE DISCOVERY OF COMPOUNDS WITH MEDICINAL POTENTIAL

https://doi.org/10.22533/at.ed.394122608011

Vagner Marques de Moura
Faculdade Eficaz – Departamento de Ciência de Dados e TI
Maringá – PR
https://orcid.org/0000-0002-8463-9594

Angélica de Almeida Moura
Faculdade Eficaz – Departamento de Biomedicina
Maringá – PR
http://lattes.cnpq.br/3914076782213841

Patrícia Silva Furlan
Universidade UNICESUMAR – Departamento de História
Maringá – PR
http://lattes.cnpq.br/1892497616995514

ABSTRACT: The use of big data and artificial intelligence (AI) has been revolutionary across all fields, particularly in the search for new pharmacological prototypes for the pharmaceutical industry. The storage of large datasets, especially in the field of cheminformatics, has contributed significantly to the advancement of research focused on the design and synthesis of new drugs through molecular modeling and chemometrics. As a result, there has been a need to develop new algorithms and architectures to access these databases and meet the specific demands of the medical and chemical-pharmaceutical sectors, especially in terms of virtual and *in silico* screening.The emergence and development of learning neural networks and their variants, combined with extensive chemical and biological knowledge, as well as their associated datasets, have led to a paradigm shift in the way information is captured and stored in these fields. This review aims to briefly report on the role and advancements of big data, deep generative models (DGMs), and AI techniques in the molecular design of compounds with medicinal potential, exploring various algorithms and contributing to the development of drugs with greater therapeutic efficacy for disease treatment.

KEYWORDS: Big data, Artificial intelligence, Chemical compounds, Medicinal potential.

INTRODUCTION

The processes of synthesis and discovery of new drugs by the pharmaceutical industry for disease treatment are considered highly complex, costly, and challenging. It is estimated that each drug requires an investment of approximately US$ 2.5 to 2.8 billion and around 10 years to reach the market. The entire theoretical and experimental research process involves stages ranging from the identification and characterization of new molecules to biological, preclinical, and clinical trials, all properly registered and approved by regulatory agencies. Therefore, it is crucial to develop and promote efficient strategies that address and minimize these challenges faced by the industry (Gandwal & Lavecchia, 2024).

In the modern era, technological development combined with the reduction of instrumentation costs has led to a significant increase in the generation of data in both quantity and diversity, enabling the acquisition of numerous datasets (Dash et al., 2019). Thus, the collection of data commonly large in volume and complexity can be encompassed and analyzed through big data. The substantial growth of data has resulted in its availability across various platforms, spanning public and private sectors as well as commercial and industrial domains. In this context, the resulting data-centered environment has required the acquisition, integration, and analysis of big data to elucidate complex and challenging issues across multiple fields, particularly in the scientific, pharmaceutical, and medical communities, among others (De Mauro et al., 2016; Sivarajah et al., 2017).

The emergence of big data has revolutionized the processes and strategies involved in drug discovery, development, and the identification of new bioactive molecules. It is evident how the translation of discoveries from basic research to clinical practice has become faster and more efficient, and how data-driven approaches in drug discovery have been successfully achieved (Qian et al., 2019). The availability of large volumes of data has enabled the exploration of artificial intelligence (AI), which mimics human intelligence, to strategically address multidimensional challenges and problems in the process of discovering new pharmacological agents. This process encompasses all stages from the design and identification of new chemical structures, molecular modeling, and biological and pharmacological testing, to clinical trials and ultimately the use of these compounds as medicines for disease treatment (Schneider et al., 2020).

Thus, AI applications related to big data analysis in the chemical and pharmaceutical fields have shown effective and increasingly promising results. However, some gaps still need to be addressed, which remain challenges despite the many advances made, thereby allowing for further enhancement of data-driven and AI led innovations (Zhao et al., 2020).

In this context, cheminformatics is included as a field of information technology that uses computational resources for the collection, storage, analysis, and manipulation of large volumes of chemical data, such as formulas, structures, properties, spectroscopic and spectrometric data, as well as information on the biological and pharmacological activities of compounds. Furthermore, it has been characterized as an interdisciplinary science that employs tools from computer science, data science, and information technology, with applications and contributions across all areas of chemistry. The use of cheminformatics has demonstrated several benefits in advanced research, mainly by facilitating the use of computational models to estimate molecular activity, reducing costs during the drug discovery process, decreasing the number of animals used in experiments, and contributing to green chemistry (Wishart, 2016; Alves et al., 2018).

The rapid progress of big data and AI has reorganized and enhanced strategies for drug design and development, particularly in terms of time efficiency and cost reduction in synthesis stages. Computational algorithms and models used in these processes known as virtual screening (VS), utilize data from chemical compound libraries to perform more reliable analyses, from identifying potential drug candidates to determining the final synthesis route and industrial production. In this sense, prior to the synthesis and biological and pharmacological evaluation stages of the target molecule, AI driven analysis assists quickly and effectively in the identification, design, and development of bioactive prototypes against different types of diseases (Réda et al., 2020; Kokudeva et al., 2024).

As a rapidly evolving field, AI encompasses several domains, among which reasoning, knowledge representation, and machine learning (ML) stand out. Due to the large volume of data, machine learning has become a widely used tool in drug discovery. It employs various algorithms and techniques to recognize models and patterns within the provided datasets. Currently, its main application in drug design is to identify and explore the relationship between chemical structure and biological activity, known as the structure–activity relationship (SAR). The emergence of high-throughput sequencing approaches, such as next-generation sequencing (NGS), has led to an exponential growth in sequence data, enabling the identification of potential therapeutic targets (Gandwal & Lavecchia, 2024).

Machine Learning (ML) approaches have introduced questions about the understanding of intelligence for the design and development of algorithms capable of learning from data acquiring knowledge through experience in order to improve their learning behavior throughout the process. In general, ML has contributed to the prediction of pharmacological targets using large scale data sources and can be applied, for example, in healthcare, smart manufacturing, and everyday life (Peng et al., 2020).

Learning methods are classified into two subcategories: supervised learning and unsupervised learning methods. In the supervised method, algorithms are trained with a defined variable Y in an attempt to generate a mathematical function that generalizes this variable. These methods can be described in at least six types: random forest (RF), support vector machine (SVM), gradient boosting machine (GBM), elastic net regularization (EN), deep learning (DL), and deep neural networks (DNN) (Yang et al., 2019).

The growing increase in data and the limitations of ML approaches have led to the creation and development of the deep learning (DL) methodology, a subfield of machine learning that harnesses the power of artificial neural networks (ANNs). Computational methods for quantitative structure–activity/property relationships (QSAR/QSPR) are regression models used to predict biological activity as well as to design drugs based on chemical structure. The ANN model allows for the imitation of the action of electrical impulses generated by neurons through computational units referred to as "perceptrons". These units are commonly interconnected in a manner similar to neurons in the brain, enabling self-learning (Elton et al., 2019).

Artificial perceptrons in ANNs are part of a group of nodes essential for data input and output in solving problems at the biological and pharmacological levels. They are known to play a role in drug research, addressing challenges related to the complexity of chemical compound screening, as well as in estimating the pharmacokinetic and pharmacodynamic parameters of molecules. However, there are at least four other types of ANNs, among which the most notable are: multilayer perceptron networks (MLPs), recurrent neural networks (RNNs), convolutional neural networks (CNNs), and autoencoders, which employ supervised and/or unsupervised learning methods (Fleck et al., 2016).

The present review describes the role of big data and artificial intelligence in the modern era, focusing on molecular design, planning, and development of pharmacological prototypes. It also addresses the current "state of the art" in this field and the supervised and unsupervised methods involved in the process. Furthermore, it provides an overview of the implementation of big data resources using advanced AI algorithms and highlights how the current state of knowledge in machine learning and big data serves as an effective and essential tool in drug discovery.

THE IMPLEMENTATION AND EMERGENCE OF AI IN DRUG MOLECULAR DESIGN

From planning to the development of a drug, there are multiple and distinct stages, which are often complex and typically require significant time and high costs for the industry. Additionally, all work must be carried out by multidisciplinary teams. Advances in new drug development stemming from the Human Genome Project (HGP) have enabled more precise selection of specific chemical compounds as targets for a given disease. Compared to traditional approaches, in vitro and in silico methods offer the major advantage of reducing costs throughout the process. Moreover, the use of computational methods in the early stages of drug development also contributes to shortening the time required to identify a pharmacological prototype with specific therapeutic effects, except in cases where complex side effects arise (Kokudeva et al., 2024).

The use of modern pipelines in new drug discovery integrates hierarchical stages, primarily involving: target identification and validation, screening of potential candidates against the target, and optimization of the identified results to enhance affinity, selectivity, metabolic stability, and bioavailability. For the selected prototype to ultimately become a drug, it must demonstrate compatible activity results through preclinical and clinical assays (Tripathi et al., 2021).

Considering the development and the large amount of information obtained through computational chemistry, as well as high-throughput screening (HTS) methods and strategies, significant progress has been made in the rapid screening of millions of substances with potential biological and pharmacological activity against specific targets. In this way, the generation of massive data aimed at the discovery of new pharmacological prototypes has revolutionized modern methods and techniques for this purpose, marking the transition into the era of big data. Previously, big data analysis was almost exclusively limited to the field of information technology; however, today, many other areas such as engineering, healthcare, biological sciences, and exact sciences have benefited. Consequently, new computational tools and big data related algorithms have emerged for the management and handling of complex datasets, particularly those related to drug discovery, fostering studies and research projects across all fields of knowledge (Roy et al., 2010; Tripathi et al., 2021).

Advances in computer science and technology, along with the emergence of artificial intelligence (AI) and machine learning (ML) algorithms, have been essential in the search for optimal molecular designs as drug prototypes, enabling greater speed, lower costs, selectivity, and efficacy throughout the entire process (Figure 1) (Sliwoski et al., 2014).

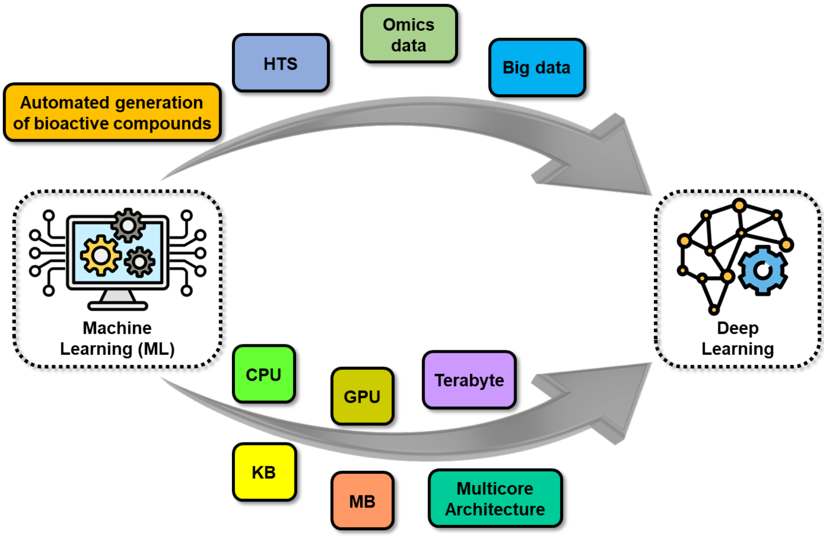

Figure 1. Overview of the growth of machine learning alongside the evolution of Big Data and computational power. HTS: High-Throughput Sequencing; CPU: Central Processing Unit; KB: Kilobyte; GPU: Graphics Processing Unit; MB: Megabyte.

In the current landscape, AI and ML combined with Big Data offer numerous applications. In pharmacology, these include protein folding prediction, protein–protein interactions, virtual screening, QSAR (Quantitative Structure–Activity Relationships), and the molecular design of novel drugs.

Several approaches are used to simulate pharmacological prototypes, particularly high-throughput virtual screening (HTVS), molecular docking, drug modeling, QSAR, and molecular dynamics. Chemoinformatics plays a key role in early-stage drug discovery by applying virtual screening (VS) to large chemical libraries, facilitating the identification of molecules with potential medicinal properties for specific targets. These methods are typically classified based on either the ligand structure (Ligand-Based Virtual Screening – LBVS) or the target structure (Structure-Based Virtual Screening – SBVS) (De Vivo et al., 2016; Alves et al., 2018).

In ligand-based virtual screening, ligands are docked to known protein targets to analyze protein–ligand interactions, with binding affinity assessed through scoring functions. These methods utilize molecular descriptors, physicochemical properties, and structural similarities to identify potential drug candidates based on reference compounds and database entries (Kadurin et al., 2017).

The extensive data on chemical structures and protein–ligand interactions has enabled AI driven inference, further advancing structure-based virtual screening (SBVS). Machine learning approaches, including support vector machines (SVM), random forests (RF), and reinforcement learning, have been instrumental in capturing the nonlinear dependencies governing ligand–target interactions (Lionta et al., 2014).

Deep learning (DL) approaches help overcome information loss during feature extraction in ML by generating higher-level hierarchical abstractions from Big Data, reducing reliance on manual feature engineering. Convolutional neural networks (CNNs) have been adapted for virtual screening, extracting features from small regions of input data (receptive fields). Tools such as DeepVS and PTPD (Predicting Therapeutic Peptides) implement CNN-based methods for screening active compounds and peptide-based ligands, respectively (Wu et al., 2019; Srivastava et al., 2023).

Ligand datasets are typically classified as active or inactive. By analyzing physicochemical and spatial similarities among active ligands, ML methods can predict the bioactivity of new compounds, even when target structures are unknown or imprecise. Consequently, ligand-based ML approaches improve the accuracy of drug design and activity prediction (Lionta et al., 2014).

Given the vast amount of data and numerous known bioactive compounds, ML algorithms are essential for analyzing datasets efficiently without compromising accuracy. Deep learning (DL), a subdivision of ML, enables handling large datasets and extracting multiple layers of abstraction, supporting both supervised and unsupervised learning (Elton et al., 2019).

Advances in computational power and open-source libraries such as TensorFlow and PyTorch have accelerated AI driven discovery of new bioactive molecules and drugs, significantly benefiting the pharmaceutical and medical fields (Chakraborty et al., 2024).

BIG DATA DATABASES IN MOLECULAR DESIGN ACROSS DIFFERENT FIELDS

Databases typically contain diverse data types, which can be raw or processed, standardized or not. Extracting meaningful insights from such heterogeneous datasets is highly challenging. Drug development requires integrating data from multiple interdisciplinary fields, including organic synthesis, structural elucidation, bioassays, pharmacology, and preclinical and clinical studies. AI is a crucial tool for managing the complexity and heterogeneity of these datasets (Chakraborty et al., 2024).

The growth of Big Data has driven the need for advanced computational resources, high-performance computing, cloud technologies, and GPUs. Data from

Big Data analyses in the search for new pharmacological prototypes can be organized into different categories or stored across multiple databases.

In the chemical field, there are databases such as ChemSpider (http://www.chemspider.com/) and Chemicalize (http://chemicalize.com/) that provide information on molecular chemical structures, chemical and commercial names, identifiers (e.g., CAS numbers), physical properties, and interactive spectra, among other data. SciFinder (https://scifinder.cas.org/) allows access to data on the structures and chemical reactions of over 100 million compounds registered with CAS (https://www.cas.org/) (Alves et al., 2018).

In the biological field, databases include information on chemical structures and their activities assessed through in vitro, in vivo, and high-throughput screening (HCS/HTS) assays. Among the most commonly used databases are ChEMBL (https://www.ebi.ac.uk/chembl/) and PubChem (http://pubchem.ncbi.nlm.nih.gov/), which provide biological data related to each compound. DrugBank (https://www.drugbank.ca/) stores information on approved drugs, including chemical structures, physicochemical properties, therapeutic uses, pharmacokinetics, toxicology, pharmacodynamics, and in some cases, molecular targets (Alves et al., 2018). The e-Drug3D platform contains a database with several collections of SD files featuring 3D structures of known drug molecules for drug screening. There are also collections containing genomic and proteomic data available from BindingDB and SuperTarget (Tripathi et al., 2021).

Macromolecule databases can also be included, with emphasis on PDB (Protein Data Bank, https://www.rcsb.org/) and BMRB (Biological Magnetic Resonance Data Bank, http://www.bmrb.wisc.edu/), which provide information on proteins, nucleic acids, and other complex biomacromolecules, serving as a foundation for research in health sciences, food science, drug design, and more (Alves et al., 2018).

MOLECULAR DESCRIPTORS AND THEIR CLASSIFICATIONS

Knowledge of molecular structures allows addressing important aspects such as physicochemical properties and biological activity, since the spatial geometry and nature of functional groups determine a compound's polarity, intermolecular forces, solubility, and reactivity. Additionally, these structures are crucial for fitting into specific enzyme and receptor sites (lock-and-key model), enabling recognition and interaction with other molecules for example, drugs that must bind to a target to produce the desired therapeutic effect. However, to predict structure–activity relationships using computational models, it is necessary to establish appropriate representations of the drug's molecular structure, which can be described as a unique numerical sequence (Hansch, 1990).

A molecular descriptor is the result of values derived from logical-mathematical operations that transform encoded chemical information into a symbolic representation of a compound (Consonni et al., 2002). Descriptors are typically structured in a matrix or bit vector (bit vector or STD logic vector).

Furthermore, descriptors can be classified according to their level to make them suitable for analyses in machine learning, including: one-dimensional (1D), which considers physicochemical properties and molecular formula (e.g., molecular weight); two-dimensional (2D), based on properties estimable from a 2D representation (e.g., molecular fingerprints, atom counts, connectivity indices); and three-dimensional (3D), associated with and dependent on the molecule's 3D spatial conformation (e.g., volume) (Table 1) (Alves et al., 2018).

Levels	Descriptor class	Particular properties of each descriptor class
1	0D or count descriptor	Atom and bond count, molecular weight
2	1D or fingerprint	Molecular weight
3	2D or topological descriptor	Atom and bond count, atomic connectivity, drug characteristics, adjacency and distance matrices, molecular weight
4	3D or geometric (spatial) descriptor	Potential energy, surface area, shape and volume, spatial conformation

Table 1. Types of molecular descriptor classes and their levels.

Among the most commonly used molecular representations are the Simplified Molecular Input Line Entry System (SMILES) and strings. The increase in dimensionality of descriptor classes (0D/1D/2D/3D), considering the SMILES format or the two-dimensional structure of compounds, also proportionally reflects the information content of the descriptors. Currently, various software tools are available, including Open Babel, PaDEL, Dragon, MOE, PeptiDesCalculator, AlvaDes, QuBiLS-MAS, VolSurf, and MLR-MobyDigs, among others (Cruciani et al., 2000; Todeschini et al., 2003; Tripathi et al., 2021).

Spatial occupancy measures of molecules in the training set, obtained through conformation sampling and alignment spaces, are classified as 4D descriptors. In this case, 4D-QSAR analysis allows the incorporation of spatial conformational freedom and alignment in extending 3D-QSAR models using training sets related to structure–activity relationships. 5D descriptors are an extension of 4D, also incorporating conformational freedom to enable comprehensive characterization of ligand topology within the active site. Furthermore, 6D descriptors have been developed, which take into account various solvation models (Vedani & Dobler, 2002; Hopfinger et al., 2003).

MACHINE LEARNING: SUPERVISED AND UNSUPERVISED

The machine learning (ML) method focuses on understanding the intelligence of a design and developing a set of algorithms that can learn from data without human intervention or explicit instructions. It consists of a process based on three main steps: data representation, hypothesis optimization, and generalization. This method is part of a rapidly evolving technical field, involving multiple application domains, with particular relevance to intelligent industry and healthcare, and can be applied in everyday life, such as recommendation systems, speech recognition, autonomous driving, and more (Holzinger, 2019; Lin et al., 2020).

In this method, an equation must be deduced to establish the relationship between descriptors and activity, defined iteratively according to the chosen function and algorithm. Subsequently, the hypothesis is evaluated based on its generalization capability, i.e., the ability of the generated equation to predict biological activity or properties.

Learning methods can be classified as supervised or unsupervised. Supervised learning allows training algorithms with a defined target variable (Y) in order to establish a mathematical function that generalizes this variable. Currently, several algorithms are used in these cases, including neural networks (NN), random forest (RF), support vector machines (SVM), among others (Alves et al., 2018).

Furthermore, big data has created many opportunities for the advancement of machine learning methods, particularly in cases dealing with volume, variety, velocity, and veracity. Regarding volume, traditional ML algorithms face several challenges, such as processing time and memory requirements. In terms of variety, data can exist in different forms/structures, categorized as structured, semi-structured, and unstructured. Velocity relates to the speed or frequency at which incoming data is processed. Finally, veracity is directly linked to the reliability and trustworthiness of the available data (Tripathi et al., 2021).

ML algorithms are commonly used for classification and regression tasks. In classification tasks, the main focus is on discriminating problems among two or more classes, whereas regression is concerned with predicting a quantity or a real-valued variable (Sarker, 2021).

In the operational workflow for implementing machine learning (ML) prediction methods in the development and discovery of new drugs, three central steps must be considered (Figure 2). The first step involves data preprocessing, which requires the selection and preparation of data for the vast majority of ML algorithms, including discretization and standardization. The second step, referred to as model learning, entails the actual implementation of the algorithms in a concrete manner. The final step, designated as evaluation and validation, is based on performance evaluation methods and metrics, aiming to monitor and validate the various ML models (Tripathi et al., 2021).

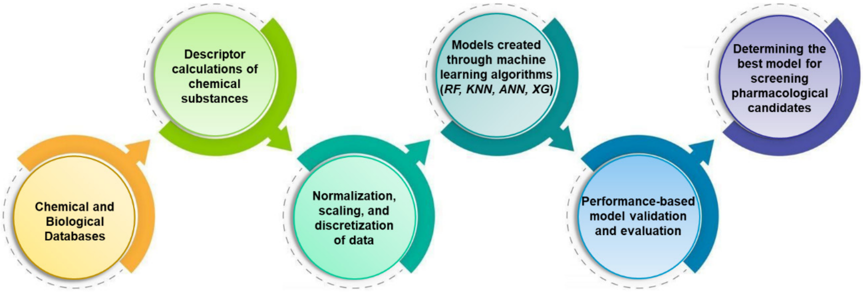
Chemical and Biological Databases
Descriptor calculations of chemical substances
Normalization, scaling, and discretization of data
Models created through machine learning algorithms (*RF, KNN, ANN, XG*)
Performance-based model validation and evaluation
Determining the best model for screening pharmacological candidates

Figure 2. Operational workflow of machine learning (ML) in big data for the development and discovery of new pharmacological candidates.

Unsupervised learning methods are applied in cases where input labels are unknown, and learning occurs by detecting patterns within the features of high-dimensional input data. The main objectives of this approach are the grouping of data subsets based on feature similarity and the quantitative identification of clusters (hierarchical clustering) present within the data. These methods can also be used to identify patterns in datasets considering only the descriptors, since the target variable Y is undefined (Bal et al., 2014).

In the field of chemistry, this approach can identify homogeneous subgroups even within heterogeneous datasets and can be applied in cases where the analysis involves assessing the consistency of different datasets and how interferences may influence activity, potentially revealing new structure–activity relationship (SAR) rules. In such cases, algorithms such as Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA), and Self-Organizing Maps (SOM) can be employed (Alves et al., 2018).

Supervised methods are used in situations where a predictive model learns from an input dataset based on label knowledge. In this way, the labels can train the machine learning (ML) model to recognize predictive patterns. These methods are fundamentally linked to ML applications, as they involve a predictive model; therefore, it is possible to generate predictions directly from the trained model using new input data. This approach relies on labeled training data to estimate a function and is also applied in cases where objectives need to be achieved through the establishment of a dataset. The most commonly used tasks are classification and regression, which focus on separating and fitting the data (Badillo et al., 2020; Sarker, 2021).

OVERVIEW OF DEEP LEARNING NEURAL NETWORKS

Big data has undergone significant transformations and became revolutionary with the advent of deep learning neural networks (DLNNs). These networks gained prominence with the introduction of the ReLU (Rectified Linear Unit) activation function, which effectively mitigated issues related to the vanishing gradient problem that could otherwise hinder neural network training from the outset. Architecturally, DLNNs typically comprise an input layer, an output layer, and multiple hidden layers. The network's ability to extract features is closely associated with the number of hidden layers, such that the complexity of the learned features increases proportionally with the number of hidden layers (Bui et al., 2020).

Effective training of DLNNs generally requires large datasets, and careful consideration of various hyperparameters is essential to achieve optimal performance [86]. These hyperparameters, also referred to as tuning parameters, play a critical role in shaping the network's training process and can significantly impact its performance. They are typically optimized using specific algorithms. Key hyperparameters in DLNNs include the number of layers, the number of neurons per layer, and the choice of activation function (Nath & Karthikeyan, 2018).

In this context, an equation (**1**) was established relating the loss function with L1 regularization.

$$\text{loss} = (y, \hat{y}) + \lambda \sum_{i=1}^{n} |\beta_i| \qquad (1)$$

Where: y = true value; $\hat{y}$ = predicted value; λ = parameter controlling the magnitude of the penalty applied to the model; n = number of features; β_i = model coefficient.

However, in L2-regularized loss, a squared magnitude of the feature coefficients is used, as shown in equation (**2**), resulting in uniform shrinkage of the coefficients. This is particularly important in cases where features are collinear.

$$\text{loss} = (y, \hat{y}) + \lambda \sum_{i=1}^{n} \beta_i^2 \qquad (2)$$

Where: y = true value; $\hat{y}$ = predicted value; λ = parameter controlling the magnitude of the penalty applied to the model; n = number of features; β_i = model coefficient.

Deep neural networks, which have a considerable number of parameters, can be attributed to machine learning systems with high computational power. However, overfitting is one of the main challenges in DLNNs, and to address this issue, the Dropout technique is applied. The primary goal of Dropout is to randomly remove units from the network during the training process, preventing the units from co-adapting excessively. During training, Dropout effectively samples from multiple thinned networks and can approximate the averages of their predictions using just one non-thinned network with reduced weights. This approach drastically reduces overfitting, leading to significant improvements compared to other methods, while also making the network more efficient in memorization and enhancing generalization (Figure 3) (Srivastava et al., 2014).

Dropout has also proven to be an important technique for mitigating overfitting. It involves the random exclusion of a specific percentage of neurons and their connections across different deep layers of the network. This makes the network more robust to memorization and improves generalization (Figure 3) (Tripathi et al., 2021).

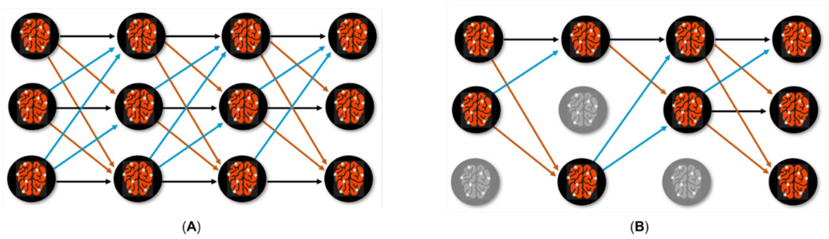

Figure 3. (**A**): Deep learning neural network (DLNN) without Dropout. (**B**) Deep learning neural network (DLNN) with Dropout.

A BRIEF OVERVIEW OF GENERATIVE ADVERSARIAL NETWORK ARCHITECTURES

Generative adversarial networks (GANs) are deep neural network architectures composed of two networks: a generative network and a discriminative network, that compete against each other. The discriminative network focuses on classifying and distinguishing between real and fake data, while the generative network produces fake data based on feedback from the discriminator. In this process, the discriminator is trained on labeled real data, such as information regarding the class of a compound. A Nash equilibrium must be achieved during the handling and optimization of fake data, so that the generated data closely matches the real data, considering both the generative and discriminative networks, ensuring that neither the generator's nor the discriminator's cost decreases (Man et al., 2025).

Currently, in chemometrics, this tool, along with variations such as conditional GANs and Wasserstein GANs, has been applied in numerous situations and applications, particularly in the development and search for new pharmacological prototypes (Figure 4).

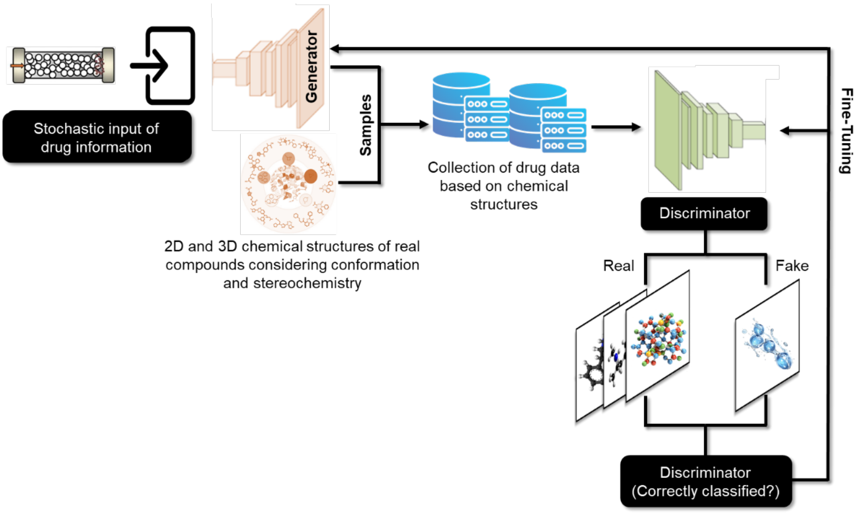
Stochastic input of drug information
Generator
Samples
2D and 3D chemical structures of real compounds considering conformation and stereochemistry
Collection of drug data based on chemical structures
Discriminator
Fine-Tuning
Real
Fake
Discriminator (Correctly classified?)

Figure 4. Generative adversarial network (GAN) architecture for chemical compounds and pharmaceuticals use.

One of the constitutive forms of DLNNs is the convolutional neural network (CNN), which is a subset of deep learning widely used for processing grid-structured data. The applications and tasks of this network are primarily focused on computer vision and object recognition. Inspired by biological aspects, this type of network consists of three main components: the convolutional layer, the pooling layer, and the fully connected layer (Ayene, 2022; Zhao et al., 2024).

The first layer extracts features from the image, such as color, textures, shapes, and edges, allowing the production of feature or activation maps. The second layer is responsible for the spatial dimensionality reduction of these feature maps, which helps decrease memory usage and mitigate overfitting. The extracted features are then fed into fully connected (dense) layers, which combine the data for the final image classification. Among state-of-the-art CNNs for computation and classification are architectures such as Inception and ResNet. Moreover, the emergence of high-performance and high-precision CNN models has been applied in computer vision, autonomous vehicles, content creation, and as an aid in medical diagnosis of diseases, such as cancer (Zhao et al., 2024).

With the advancement of CNNs, it is now possible to train them to predict protein-ligand interactions as well as to estimate compound toxicity based on graphical images (Wang et al., 2023).

AN OVERVIEW OF AUTOENCODERS

Autoencoders (AEs) are defined as a type of neural network architecture that utilizes unsupervised machine learning and is trained to efficiently encode input data, followed by the reconstruction or decoding of that data at the output nodes. They have been widely used for learning from datasets in the design of new bioactive compounds. Among hidden or latent variables, autoencoders have the ability to transform inputs into hierarchical representations, generate realistic synthetic data, or predict anomalies (Figure 5) (Berahmand et al., 2024).

When a completely independent set of input data is provided, it becomes significantly more challenging for the model to retain the information and provide an accurate dimensional representation without losses. Due to the distinct and varying characteristics of the data, it is not possible to anticipate which attribute will provide the most effective training for a given machine learning algorithm. Over time, various types of autoencoders have been developed, including conditional, adversarial, denoising, convolutional, and variational autoencoders (VAEs) (Liu et al., 2022).

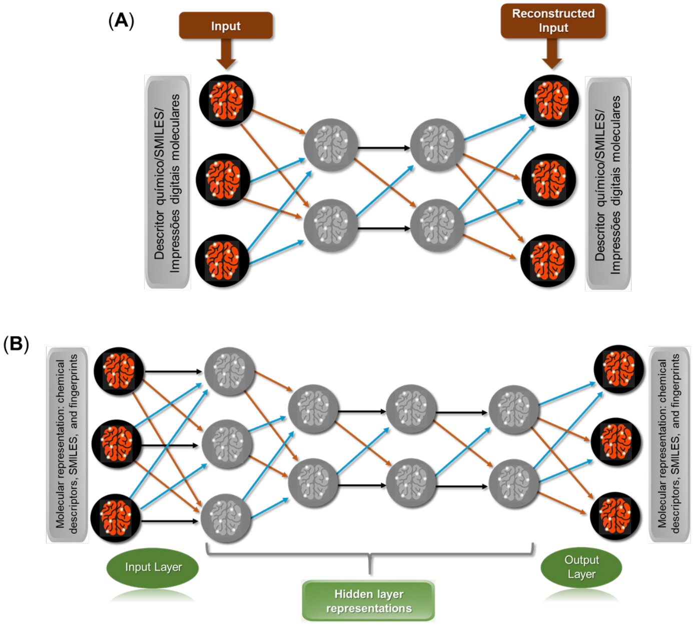
(A)
Input
Reconstructed Input
Descritor químico/SMILES/ Impressões digitais moleculares
Descritor químico/SMILES/ Impressões digitais moleculares
(B)
Molecular representation: chemical descriptors, SMILES, and fingerprints
Molecular representation: chemical descriptors, SMILES, and fingerprints
Input Layer
Hidden layer representations
Output Layer

Figure 5. (**A**) Illustration of an autoencoder, where the gray circle represents the hidden layer. (**B**) Illustration of a deep autoencoder with hidden layers, which can be used in training learning algorithms.

The variational autoencoder (VAE) has proven useful in addressing issues of overfitting and discontinuities in standard autoencoders, as it employs techniques to regularize actions in the latent space. In addition, individual and separate points in the latent space are replaced through a probability distribution. VAEs have become a valuable tool in molecule construction, offering new perspectives and approaches for designing novel structures in drug development (Gómez-Bombarelli et al., 2018).

The conditional variational autoencoder (CVAE), on the other hand, allows compound properties to be incorporated as information during the encoding process, which can then be manipulated. This enables the generation of drug-like chemical structures with specific properties and characteristics, such as hydrogen bond donor and acceptor centers, molecular weight, logP, and regions with high electron density and polarity. One of the main advantages of this methodology is its ability to control the properties and features of each molecule without affecting the others (Lim et al., 2018).

USE OF DEEP GENERATIVE MODELS (DGMS) IN THE DESIGN OF NEW BIOACTIVE COMPOUNDS

In computational terms, a generative model can be regarded as a machine learning model capable of producing new data that closely resembles the training data. Generative models using AI are designed to learn the patterns and distributions of training data, enabling the generation of new content based on the input data.

In the field of new drug design, deep generative models (DGMs) have represented a significant advancement, particularly for chemical compounds with specific and required properties. Among these properties, a strong binding affinity to selected protein targets is particularly notable. DGMs allow the adjustment and optimization of drug structural features, such as solubility, through fine-tuning of molecular spectra. Additionally, DGMs are valuable for practical aspects, such as predicting synthetic routes, and are effective in pharmacology, drug repositioning, and the design of multi-target drugs. They also enable the early identification of potential side effects during drug development and provide insights into mechanisms of action, allowing the generation of molecules for high-throughput screening (HTS) and the prediction of certain molecular properties during ADMET (absorption, distribution, metabolism, excretion, and toxicity) analysis (Gangwal & Lavecchia, 2024).

Numerous studies have employed deep generative models (DGMs) to identify new bioactive compounds using deep learning, particularly during the SARS-CoV-2 pandemic, demonstrating significant advances and success. A notable example is the work by Bung et al. (2021), which improved a stacked RNN model with transfer learning (TL) and trained it on approximately 1.5 million ChEMBL compounds to identify ligands targeting the SARS-CoV-2 viral protease. Using reinforcement learning (RL), the model quantitatively evaluated features such as molecular similarity, molecular weight, synthetic accessibility (SA) score, and logP, incorporating the QED drug-likeness metric. Docking simulations subsequently identified 31 compounds as potential pharmacological candidates.

RNNs trained on SMILES sequences have also been applied to assess antimicrobial potential against *Staphylococcus aureus* and *Plasmodium falciparum*, producing promising results in the discovery of new antibacterial agents. Another study combined deep Q-learning with RNNs to generate SMILES strings with specific molecular properties. More recently, a hybrid approach combining LSTM networks with transformer architectures, termed AMPTrans-LSTM, was developed to generate peptides with diverse antimicrobial activities. While this model has shown multiple advantages in the antimicrobial domain, further refinement and validation are required to address challenges such as transformer training stability and to confirm its efficacy against other microorganisms (Gangwal & Lavecchia, 2024).

Blaschke et al. (2017) employed a variational autoencoder (VAE) to identify antagonists of the dopamine type 2 receptor. In a subsequent study, a VAE with a graph-based latent space incorporating a Gaussian mixture, termed GraphGMVAE, was developed for scaffold hopping, enabling the generation of compounds with high precision. This approach also facilitated molecular classification to enhance method validation, with upadacitinib, a human Janus kinase 1 (JAK1) inhibitor, serving as a reference. Synthesis and biochemical testing of seven compounds demonstrated that GraphGMVAE is effective in designing compounds for medicinal chemistry, producing results comparable to those of human experts. The strategy of structural modification to improve a drug's therapeutic potential is a well-established and widely used tool in the literature. Even when molecular generators perform as expected, their effectiveness and efficiency must be evaluated according to the core principles and requirements of medicinal chemistry (Yu et al., 2021).

The initial version of DrugEx, based on reinforcement learning with RNNs (RL-RNN), was developed and trained to identify compounds targeting G protein-coupled receptors (GPCRs), which are implicated in cardiovascular diseases and inflammatory processes, including the adenosine A2A receptor. During training, DrugEx generates SMILES sequences derived from the ZINC 15 database, incorporating a stochastic element to enhance diversity. The results demonstrated that the RNN could produce a wide range of compounds, with the machine-generated bioactives encompassing those identified using fingerprints of adenosine A2A receptor ligands. Subsequent updates to DrugEx introduced new encoding strategies, allowing the evaluation of specific molecular substructures and further refinement of potential compounds (Gangwal & Lavecchia, 2024).

Currently, despite the significant results achieved with DGM models, there remains a vast field to explore, both in pharmaceutical and medicinal chemistry as well as in organic synthesis. Reports in the literature evaluating the efficacy of these methods in vitro tests have generated considerable interest within the scientific community, as this is still a relatively unexplored and rapidly growing area, particularly in the medical field.

DGM AND QSAR MODELS AS REFERENCE TOOLS IN DRUG DISCOVERY

Deep generative models (DGMs) are primarily designed to construct chemical structures with pharmacological potential capable of targeting various diseases, leveraging training datasets. Key databases commonly employed for this purpose include PubChem and ChEMBL (Hu et al., 2017).

The performance of deep learning models during training is highly dependent on data quality, which can be affected by dataset size, coverage and properties of chemical space, diversity, and potential errors. Integrating publicly available data with proprietary datasets may introduce redundancies and inaccuracies, potentially compromising model performance. Standardized assay protocols, particularly those from the pharmaceutical industry, produce more uniform and homogeneous datasets. Nonetheless, merging multiple data sources remains complex, highlighting the importance of careful curation and dataset harmonization to achieve optimal results when applying DGMs (Yonchev et al., 2018).

Accessible compound activity data often lack negative examples, resulting in imbalances relative to high-throughput screening (HTS). This limitation can be mitigated by incorporating negative or decoy data to enhance model training (Cáceres et al., 2020).

Benchmark platforms suggest several evaluation metrics for deep generative models (DGMs), including validity, novelty, uniqueness, and controllability. Tools such as Molecular Sets (MOSES) and GuacaMol further enable benchmarking by assessing structural similarity to reference drugs, synthetic feasibility, and target specificity. These platforms provide a valuable framework for facilitating the early-stage discovery of new pharmacological prototypes (Jhanwar et al., 2011; Polykovskiy et al., 2020).

QSAR predictive models are capable of establishing quantitative relationships between chemical structures and biological activities or chemical properties, employing mathematical and biostatistical approaches to predict new pharmacological prototypes and chemical entities. These models can utilize both linear and nonlinear regression methods (Jhanwar et al., 2011).

They rely on datasets encompassing molecular descriptors, physicochemical and structural characteristics (including hydrophobic, electronic, conformational, and steric effects), as well as activity data sourced from databases such as ChEMBL and PubChem (Cáceres et al., 2020).

Software tools such as RDKit and the Cheminformatics Toolkit play a crucial role in evaluating model performance, providing predictions of solubility, toxicity, mutagenicity, carcinogenicity, drug design, and ADMET parameters across various chemical structures. Nonetheless, QSAR model benchmarking platforms face notable challenges, particularly in dataset selection, bias mitigation, and validation of chemical group predictions. These platforms facilitate virtual screening (VS), compound optimization, toxicity prediction, and structure-activity relationship (SAR) analyses, serving as reliable computational tools in chemoinformatics. They enable the integration of chemical data with corresponding in vitro and in vivo biological activities based on experimental evidence (Gangwal & Lavecchia, 2024).

REFERENCE

ALVES, V. M.; BRAGA, R. C.; MURATOV, E. N.; ANDRADE, C. H. Cheminformatics: an introduction. **Química Nova**, v. 41, n. 2, p. 202-212, 2018.

AYENI, J. A. Convolutional Neural Network (CNN): The architecture and applications. **Applied Journal of Physical Science**, v. 4, n. 4, p. 42-50, 2022.

BADILLO, S.; BANFAI, B.; BIRZELE, F.; DAVYDOV, I. I.; HUTCHINSON, L.; KAM-THONG, T.; SIEBOURG-POLSTER, J.; STEIERT, B.; ZHANG, J. D. An Introduction tom achine learning. **Clinical Pharmacology and Therapeutics**, v. 107, p. 871-885, 2020.

BAL, M.; AMASYALI, M. F.; SEVER, H.; KOSE, G.; DEMIRHAN, A. Performance evaluation of the machine learning algorithms used in inference mechanism of a medical decision support system. **The Scientific World Journal**, 2014:137896, 2014.

BERAHMAND, K.; DANESHFAR, F.; SALEHI, E. S.; LI, Y.; XU, Y. Autoencoders and their applications in machine learning: a survey. **Artificial Intelligence Review**, v. 57, n. 28, p. 1-52, 2024.

BLASCHKE, T.; OLIVECRONA, M.; ENGKVIST, O.; BAJORATH, J.; CHEN, H. Application generative autoencoder in de novo molecular design. **Molecular Informatics**, v. 37, p. 1-13, 2017.

BUI, D. T.; TSANGARATOS, P.; NGUYEN, V-T.; LIEM, N. V.; TRINH, P. T. Comparing the prediction performance of a deep learning neural network model with conventional machine learning models in landslide susceptibility assessment. **Catena**, v. 188, 104426, 2020.

BUNG, N.; KRISHNAN, S. R.; BULUSU, G.; ROY, A. De novo design of new chemcial entities for SARS-CoV-2 using artificial intelligence. **Future Medicinal Chemistry**, v. 13, p. 575-585, 2021.

CÁCERES, E. L.; MEW, N. C.; KEISER, M. J. Adding stochastic negative examples into machine learning improves molecular bioactivity prediction. **Journal of Chemical Information and Modeling**, v. 60, p. 5957-5970, 2020.

CHAKRABORTY, C.; BHATTACHARYA, M.; LEE, S-S.; WEN, Z-H.; LO, Y-H. The changing scenario of drug discovery using AI to deep learning: Recent advancement, success stories, collaborations, and challenges. **Molecular Therapy: Nucleic Acid**, v. 35, p. 1-21, 2024.

CONSONNI, V.; TODESCHINI, R.; PAVAN, M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. **Journal of Chemical Information and Computer Sciences**, v. 42, n.3, p. 682-692, 2002.

CRUCIANI, G.; PASTOR, M.; GUBA, W. VolSurf: A new tool for the pharmacokinetic optimization of lead compounds. **European Journal of Pharmaceutical Sciences**, v. 11, Suppl. 2:S29-39, 2000.

DASH, S.; SHAKYAWAR, S. K.; SHARMA, M.; KAUSHIK, S. Big data in healthcare: management, analysis and future prospects. **Jounal of Big Data**, v. 6, n. 54, p. 1-25, 2019.

DE MAURO, A.; GRECO, M.; GRIMALDI, M. A formal definition of Big Data based on its essential features. **Library Review**, v. 65, n. 3, p. 122-135, 2016.

DE VIVO, M.; MASETTI, M.; BOTTEGONI, G.; CAVALLI, A. Role of molecular dynamics and related methods in drug discovery. **Journal of Medicinal Chemistry**, v. 59, p. 4035-4061, 2016.

ELTON, D. C.; BOUKOUVALAS, Z.; FUGE, M. D.; CHUNG, P. W. Deep learning for molecular design–a review of the state of the art. **Molecular Systems Design & Engineering**, v. 4, p. 828-849, 2019.

FLECK, L.; TAVARES, M. H. F.; EYNG, E.; HELMANN, A. C.; ANDRADE, M. A. M. Redes neurais artificiais: princípios básicos. **Revista Eletrônica Científica Inovação e Tecnologia**, v. 1, n. 13, p. 47-57, 2016.

GANGWAL, A.; LAVECCHIA, A. Unleashing the power of generative AI in drug discovery. **Drug Discovery Today**, v. 29, n. 6, p. 1-24, 2024.

GÓMEZ-BOMBARELLI, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. **ACS Central Science**, v. 4, n. 2, p. 268-276, 2018.

HANSCH, C. Comprehensive medicinal chemistry: the rational design, mechanistic study & therapeutic application of chemical compounds; New York: Pergamon Press, 1990.

HOLZINGER, A. Introduction to machine learning & knowledge extraction (MAKE). **Machine Learning & Knowledge Extraction**, v. 1, n. 1, p. 1-20, 2019.

HOPFINGER, A. J.; WANG, S.; TOKARSKI, J. S.; JIN, B.; ALBUQUERQUE, M.; MADHAV, P. J.; DURAISWAMI, C. Construction of 3D-QSAR models using the 4D-QSSAR analysis formalism. **Journal of the American Chemical Society**, v. 119, n. 43, p. 10509-10524, 1997.

HU, Y.; STUMPFE, D.; BAJORATH, J. Recent advances in scaffold hopping. **Journal of Medicinal Chemistry**, v. 60, p. 1238-1246, 2017.

JHANWAR, B.; SHARMA, V.; SINGLA, R. K.; SHRIVASTAVA, B. QSAR - Hansch analysis and related approaches in drug design. **Pharmacologyonline**, v. 1, p. 306-344, 2011.

KADURIN, A.; NIKOLENKO, S.; KHRABROV, K.; ALIPER, A.; ZHAVORONKOV, A. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in sílico. **Molecular Pharmaceutics**, v. 14, n. 9, p. 3098-3104, 2017.

KOKUDEVA, M.; VICHEV, M.; NASEVA, E.; MITEVA, D. G.; VELIKOVA, T. Artificial intelligence as a tool in drug discovery and development. **World Journal of Experimental Medicine**, v. 14, n. 3, p. 1-9, 2024.

LIM, J.; RYU, S.; KIM, J. W.; KIM, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. **Journal of Cheminformatics**, v. 10, n. 31, p. 1-9, 2018.

LIN, E.; LIN, C-H.; LANE, H-Y. Relevant applications of generative adversarial networks in drug design and Discovery: molecular de novo design, dimensionality reduction, and de novo peptide and protein design. **Molecules**, v. 25, n. 14, p. 1-25, 2020.

LIONTA, E.; SPYROU, G.; VASSILATIS, D.; COURNIA, Z. Structure-base virtual screening for drug Discovery: principles, applications and recente advances. **Current Topics in Medicinal Chemistry**, v. 14, n. 23, p. 1923–1938, 2014.

LIU, Y.; JIANG, H.; WANG, Y.; WU, Z.; LIU, S. A conditional variational autoencoding generative adversarial networks with self-modulation for rolling bearing fault diagnosis. **Measurement**, v. 192, 110888, 2022.

MAN, W.; XU, L.; HE, C. Evolutionary architecture search for generative adversarial networks using an aging mechanism-based strategy. **Neural Networks**, v. 181, n. 1, 106877, 2025.

NATH, A.; KARTHIKEYAN, S. Enhanced prediction of recombination hotspots using input features extracted by class specific autoenconders. **Journal of Theoretical Biology**, v. 444, p. 73-82, 2018.

PENG, J.; LI, J.; SHANG, X. A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. **BMC Bioinformatics**, v. 21, supl. 13:394, p. 1-13, 2020.

POLYKOVSKIY, D. et al. Molecular sets (MOSES): a benchmarking platform for molecular generation models. **Frontiers in Pharmacology**, v. 11, 2020.

QIAN, T.; ZHU, S.; HOSHIDA, Y. Use of bid data in drug development for precision medicine: an update. Expert Review of Precision Medicine and Drug Development, v. 4, n. 3, p. 189, 200, 2019.

RÉDA, C.; KAUFMANN, E.; DELAHAYE-DURIEZ, A. Machine learning applications in drug development. **Computational and Structural Biotechnology Journal**, v. 18, p. 241-252, 2020.

ROY, A.; McDONALD, P. R.; SITTAMPALAM, S.; CHAGUTURU, R. Open access high throughput drug discovery in the public domain: a mount Everest in the making. **Current Pharmaceutical Biotechnology**, v. 11, p. 764-778, 2010.

SARKER, I. H. Machine learning: algorithms, realworld applications and research directions. **SN Computer Science**, 2:160, p. 1-21, 2021.

SCHNEIDER, P. et al. Rethinking drug design in the artificial intelligence era. **Natural Reviews Drug Discovery**, v. 19, p. 353-364, 2020.

SIVARAJAH, U.; KAMAL, M. M.; IRANI, Z.; WEERAKKODY, V. Critical analysis of Big Data challenges and analytical methods. **Journal of Business Research**, v. 70, p. 263-286, 2017.

SLIWOSKI, G.; KOTHIWALE, S.; MEILER, J.; LOWE, E. W. Computational methods in drug discovery. **Pharmacological Reviews**, v. 66, p. 334-395, 2014.

SRIVASTAVA, R.; AVASTHI, V.; PRIYA, K. Deep convolutional neural network for partial discharge monitoring system. Advances in Engineering Software, 103407, 2023.

TODESCHINI, R.; MAURI, A.; PAVAN, M. MobyDigs: software for regression and classification models by genetic algorithms. **Nature-Inspired methods in chemometrics; genetic algorithms and artificial neural networks**. Lerdi, R., cap. 5, Ed.:Elsevier, 2003.

TRIPATHI, M. K.; NATH, A.; SINGH, T. P.; ETHAYATHULLA, A. S.; KAUR, P. Evolving scenario of big data and artificial intelligence (AI) in drug discovery. **Molecular Diversity**, v. 25, p. 1439-1460, 2021.

VEDANI, A.; DOBLER, M. 5D-QSAR: the key for simulating induced fit? **Journal of Medicinal Chemistry**, v. 45, n. 11, p. 2139-2149, 2002.

YANG, X.; WANG, Y.; BYRNE, R.; SCHNEIDER, G.; YANG, S. Concepts of artificial intelligence for computer-assisted drug Discovery. **Chemical Reviews**, v. 119, n. 18, p. 10520-10594, 2019.

YONCHEV, D.; DIMOVA, D.; STUMPFE, D.; VOGT, M.; BAJORATH, J. Redundancy in two major compounds databases. **Drug Discovery Today**, v. 23, p. 1183-1186, 2018.

YU, Y. et al. A novel scalarized scaffold hopping algorithm with graph-based variational autoencoder for discovery of JAK1 inhibitors. **ACS Omega**, v. 6, p. 22945-22954, 2021.

WANG, Y.; JIAO, Q.; WANG, J.; CAI, X.; ZHAO, W.; CUI, X. Prediction of protein-ligand binding affinity with deep learning. **Computational and Structural Biotechnology Journal**, v. 21, p. 5796-5806, 2023.

WISHART, D. S. Introduction to Cheminformatics. **Current Protocols in Bioinformatics**, sup. 53:14.1.1-14.1.21, 2016.

WU, C.; GAO, R.; ZHANG, Y.; DE MARINIS, Y. PTPD: predicting therapeutic peptides by deep learning and word2vec. **BMC Bioinformatics**, v. 20, n. 456, p. 1-8, 2019.

ZHAO, L.; CIALLELLA, H. L.; ALEKSUNES, L. M.; ZHU, H. Advancing computer-aided drug discovery (CADD) by bid data and data-driven machine learning modeling. **Drug Discovery Today**, v. 25, n. 9, p. 1624-1638, 2020.

ZHAO, X.; WANG, L.; ZHANG, Y.; HAN, X.; DEVECI, M.; PARMAR, M. A review of convolutional reural networks in computer vision. **Artificial Intelligence Review**, v. 57, n. 99, p. 1-43, 2024.