drugs and drug candidates

MDPI

Review

Virtual Screening Algorithms in Drug Discovery: A Review Focused on Machine and Deep Learning Methods

Tiago Alves de Oliveira [1,2,*], Michel Pires da Silva [2], Eduardo Habib Bechelane Maia [2], Alisson Marques da Silva [2] and Alex Gutterres Taranto [1,*]

[1] Department of Bioengineering, Federal University of São João del-Rei, Praça Dom Helvécio, 74-Fábricas, São João del-Rei 36301-1601, Brazil

[2] Federal Center for Technological Education of Minas Gerais (CEFET-MG), Department of Informatics, Management and Design, Campus Divinópolis, Rua Álvares de Azevedo, 400-Bela Vista, Divinópolis 35503-822, Brazil

* Correspondence: tiago@cefetmg.br (T.A.d.O.); proftaranto@hotmail.com (A.G.T.); Tel.: +55-(37)-3229-1175 (T.A.d.O.); +55-(32)-3379-5134 (A.G.T.)

Abstract: Drug discovery and repositioning are important processes for the pharmaceutical industry. These processes demand a high investment in resources and are time-consuming. Several strategies have been used to address this problem, including computer-aided drug design (CADD). Among CADD approaches, it is essential to highlight virtual screening (VS), an in silico approach based on computer simulation that can select organic molecules toward the therapeutic targets of interest. The techniques applied by VS are based on the structure of ligands (LBVS), receptors (SBVS), or fragments (FBVS). Regardless of the type of VS to be applied, they can be divided into categories depending on the used algorithms: similarity-based, quantitative, machine learning, meta-heuristics, and other algorithms. Each category has its objectives, advantages, and disadvantages. This review presents an overview of the algorithms used in VS, describing them and showing their use in drug design and their contribution to the drug development process.

Keywords: drug discovery; virtual screening; machine learning; deep learning; CADD; algorithms; structural bioinformatics

check for updates

Citation: Oliveira, T.A.d.; Silva, M.P.d.; Maia, E.H.B.; Silva, A.M.d.; Taranto, A.G. Virtual Screening Algorithms in Drug Discovery: A Review Focused on Machine and Deep Learning Methods. *Drugs Drug Candidates* **2023**, *2*, 311–334. https://doi.org/10.3390/ddc2020017

Academic Editor: George Mihai Nitulescu

Received: 28 December 2022
Revised: 23 February 2023
Accepted: 6 March 2023
Published: 5 May 2023

1. Introduction

The discovery and manufacturing processes for new drugs have been in constant evolution in the last decade, with benefits and results directly related to the life quality improvement of the world population [1]. The effectiveness of the production and manufacturing strategies is increasing, although the costs involved remain a challenge [2]. The development of a new drug has an average cost between 1 and 2 billion USD and could take 10 to 17 years, from target discovery to drug registration [3]. These costs reflect a time-consuming, laborious, and expensive process, which implies drugs with high added value and, often, of restricted acquisition, at least for a third of the world's population [4,5].

In addition, part of these costs are related to a low success rate, with only 5% of phase I clinical trial drugs entering the market [6]. These impacts are even more significant in places such as Asia and Africa, where more than 50% of the population lives in poverty or with a lower average income than in the rest of the world [4,7]. The conditions that restrict the access of this fraction of the population to essential medicines contribute to the increase in the mortality of tens of thousands of people every year, 18 million of which could be avoided if these medicines became accessible [8].

One way to reduce drug discovery and manufacturing limitations, minimizing cost and time impacts, is by using computer-aided drug design (CADD), also known as molecular modeling. In such an approach, the design stages and analysis of drugs are carried out through a cyclic-assisted process entirely conducted through in silico simulations. These simulation techniques can evaluate essential factors in drug discovery, such as toxicity, activity, biological activity, and bioavailability, even before carrying out in vitro and in vivo clinical trials [9]. An initial step of the CADD approach is screening virtual compound libraries, known as virtual screening (VS). VS is a molecule classification method that exploits biological or chemical properties available in large datasets. The International Union of Pure and Applied Chemistry (IUPAC) defines VS as computational methods that classify molecules in a database according to their ability to present biological properties against a given molecular target [10].

Popular VS techniques originated in the 1980s, but the VS word first appeared in a 1997 publication [11]. In the early 1990s, the rapid evolution of computers created the hope that companies could accelerate the discovery process of new drugs. These computing advances have enabled several advances in combinatorial chemistry and high-throughput screening (HTS) technologies, allowing for vast libraries of compounds to be synthesized and screened in short periods. The VS process aims to choose an appropriate set of compounds, removing inappropriate structures and limiting the use of a significant number of resources. In this way, identifying hits with computational methods proved to be a promising approach because, even before carrying out the biological assays, computational simulations could indicate which compounds were most likely to be good hits.

Currently, VS represents a crucial step in the early stage of drug discovery due to it having been proven to be an excellent alternative to HTS, especially in terms of cost–effectiveness and probability of finding the appropriate result through a large virtual database [12,13]. In this context, we present a review of the main VS techniques. This paper aims to show an overview of the most recent algorithms used in VS, with a particular focus on machine learning (ML) and deep learning (DL).

The rest of this article is structured as follows: In Section 2, we describe the VS process, its implementation models, and the steps that compose it. Then, in Section 3, we describe the ML algorithms that are used with VS and the characteristics that make them suitable for the composition of CAAD models aimed at similarity evaluation. In Section 4, some guidelines for in silico models for CADD are presented and discussed. Finally, in Section 4, we present the conclusions of the exploited and presented strategies.

2. Virtual Screening

VS is an in silico technique used in the drug discovery process [12–16]. During VS, large databases of molecular structures are automatically evaluated using computational methods. With the use of VS, it is expected to identify molecules more susceptible to binding to the molecular target, typically a protein or enzyme receptor.

VS works like a filter (Figure 1), eliminating more molecules so that the number of final candidate molecules that may become a drug is much smaller than the initial number. In Figure 1, molecules that can become a good drug are initially selected because they have properties that favor their action or are similar to drugs with known functioning. Candidate ligands are then analyzed, and molecules in which pharmacophoric groups that indicate the likelihood of toxicity are identified are eliminated. In the next step, the best poses for candidate ligands are identified. During this VS process, the candidate ligands can have their compositions and structures readjusted to enhance their properties; mainly, their pharmacokinetic characteristics (absorption, distribution, metabolism, excretion, and toxicity (ADMET)). It may require new cycles of optimizations in the composition and structure of candidate ligands. This way, the ligands are prepared for biological assays after this process.

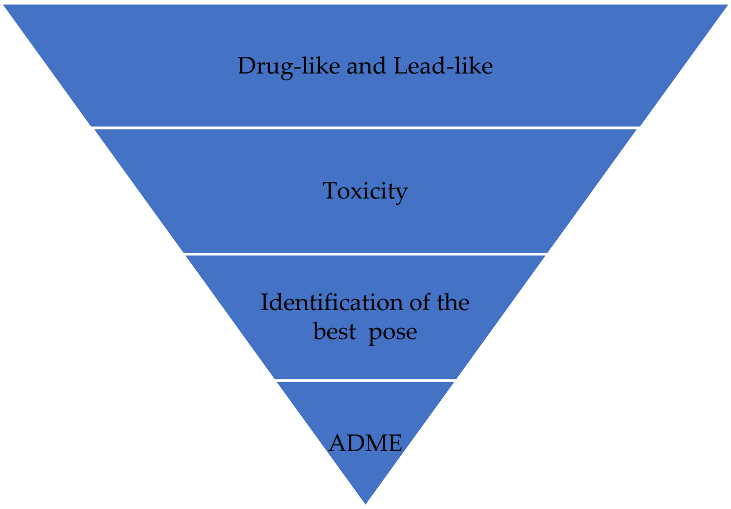

Figure 1. Virtual screening process.

Furthermore, the VS allows for the selection of compounds in a database of structures with a greater probability of presenting biological activity against a target of interest. VS can also eliminate compounds that may be toxic or that have unfavorable specific pharmacodynamic and pharmacokinetic properties. In this sense, biological assays should be performed only with the most promising molecules, leading to a lower cost and shorter development time [13].

There are essentially three main approaches of VS: structure-based VS (SBVS), ligand-based VS (LBVS), and fragment-based VS (FBVS). The following sections detail these approaches.

2.1. Structure-Based Virtual Screening

SBVS, also known as target-based VS (TBVS), attempts to predict the best binding orientation of two molecules to form a stable complex. The SBVS technique encompasses methods that explore the molecular target's 3D structure. SBVS is the preferred method when the 3D structure of the molecular target has been experimentally characterized [14]. SBVS tries to predict the likelihood of coupling between candidate ligands and target protein, considering the binding strength of the complex. The most used SBVS technique is molecular anchoring due to the low computational cost and the satisfactory results [15,16].

Molecular docking emerged in the 1980s when Kuntz et al. [17] developed an algorithm that explored the geometrically feasible docking of a ligand and target. They showed that structures close to the correct ones could be obtained after docking. This technique tries to predict the best position and orientation of a ligand in a binding site of a molecular target so that both form a stable complex. To this end, it uses the structural and chemical complementarity resulting from the interaction between a ligand and a molecular target through scoring functions, often complemented with pharmacophoric restrictions [18]. SBVS is often used with proteins and enzymes as receptors, but it can also be used in other structures, such as carbohydrates and DNA. However, although the approach was promising, it was only in the 1990s that it became widely used, when there was an improvement in the techniques used in conjunction with an increase in computational power and greater access to the structural data of target molecules.

The use of SBVS has advantages and disadvantages. Among the benefits are the following: (a) there is a decrease in the time and cost of screening millions of small molecules; (b) there is no need for the physical existence of the molecule, so it can be tested in silico before being synthesized, (c) there are several tools available to assist SBVS.

As a disadvantage of using SBVS, it can be highlighted that (a) some tools work best in specific cases but not in more general cases [19]; (b) it is difficult to accurately predict the correct binding position and classification of compounds due to the difficulty of parameterizing the complexity of ligand–receptor binding interactions; (c) it can generate false positives and false negatives.

Despite the disadvantages noted above, many studies using SBVS have been developed in recent years [12,20–25], which shows that although SBVS has weaknesses, it is still widely used for designing new drugs due to the reduction in time and cost.

2.2. Ligand-Based Virtual Screening

LBVS uses ligands with known biological activity aiming to identify molecules with similar structures using virtual libraries of compounds. In this approach, the molecular target structure is not considered. Instead, it follows the principle that structurally similar molecules may exhibit comparable biological activity. Thus, LBVS tries to identify compounds with similar molecular scaffolds or pharmacophore moieties, increasing the chance of finding biologically active compounds.

LBVS is performed by comparing the descriptors or characteristics of known molecules derived from reference compounds and compared with the descriptors of the molecules of the databases. For this purpose, similarity measures are used. There are several ways to verify the similarity between two sets of molecular characteristics, but the Tanimoto coefficient [26] is one of the most used to perform this task [27].

One of the most common LBVS classifications techniques is the number of dimensions of the represented descriptor (1D, 2D, or 3D):

- 1D descriptors cover molecular properties such as weight, number of hydrogen-bond donor and acceptor groups, number of rotatable bonds, number and atoms type, and computed physicochemical properties, such as logP and water solubility (log D), degree of ionization, and others;
- 2D descriptors are those based on molecular topology. Therefore, they are built based on the molecular connectivity of the compounds. Two-dimensional descriptors can exist as structural descriptors and topological indices [28]; structural descriptors are those characterizing the molecule by its chemical substructure. They can be represented in 2D graphs or binary vectors. For example, the number of aromatic rings in a molecule, the connectivity index, and the Carbo index (which calculates molecular similarity) are 2D descriptors. Topological indices define the molecule's structure according to its shape and size. Simpler indices characterize the molecules according to their size, shape, and degree of branching, and more complex indices consider both the properties of the atoms and their connectivity;
- 3D descriptors provide molecular information in the context of the spatial distribution of atoms, molecular properties, or chemical groups. 3D descriptors need to consider molecular conformations, meaning they need to consider the spatial distribution of particles, molecular properties, or chemical groups [29]. Studies have demonstrated that a single descriptor does not perform better than another for all VS [30]. Therefore, the molecules are, in most cases, described as a set of descriptors. Examples of 3D descriptors are the electrostatic potential and van der Walls.

LBVS is more suitable for use in the following situations: (a) whenever there is little information about the structure of the molecular target. In addition, it is used to enrich the database for SBVS experiments; (b) for targets with large amounts of experimental data available or where the drug-binding site is not well defined, LBVS methods are generally superior to SBVS methods [31]; (c) the simultaneous use of the LBVS and SBVS approaches can increase the accuracy of the VS, as the LBVS can eliminate some false-positive compounds identified as promising by the SBVS technique, increasing the chances of obtaining good results [11].

The LBVS technique has the following limitations: (a) the currently available LBVS techniques have not yet shown the desired performance, but the prediction accuracy is expected to increase rapidly in the coming years [14,20]; (b) minor chemical modifications in similar molecules can either potentiate an activity or make it inactive [32]. Wermuth [33] shows how relatively modest changes in a molecule with known biological activity can lead to compounds with very different activity profiles. This is known as activity cliffs [34]. In this way, a molecule identified by the LBVS technique as similar to another with known activity may be inactive, leading to a false-positive result.

2.3. Fragment-Based Virtual Screening

FBVS has also been a powerful approach to finding early hits that can lead development [35]. FBVS aims to test fragments of low-molecular-weight molecules against macromolecular targets of interest [36]. FBVS usually generates a candidate compound from a chemical fragment with low molecular weight (generally less than 300 Da), low binding affinity, and simple chemical structures [37]. These compounds are then used as starting points for drug development. However, due to their low molecular weight, fragment hits are generally weak binders and need to be made into larger molecules that bind more tightly to the target to be made into a lead. Therefore, the methods used in FBVS and the fragment binding modes are critical in fragment-based screening [35,37].

Over the last 20 years, FBVS using nuclear magnetic resonance (NMR) has been a prominent approach. FVBS has made it possible to obtain a high success rate with increased development speed and decreased cost of producing new drugs [36]. Moreover, there is no need for intervention by medicinal chemists in the early stages of increasing fragment molecules and their transformation into molecules with more significant biological activity. Several FBVS approaches have been used today, and the development of the ligand from a fragment using ML has gained prominence [38].

The main advantage of using FBVS is the low complexity of the fragments used in the simulation [35], which allows for the use of different techniques (which have pros and cons that must be analyzed according to the situation) for the development of new compounds and cost reduction in drug development [35].

3. Virtual Screening Algorithms

VS has become a widely used strategy for drug discovery, and as seen in the previous section, it can be classified into SBVS, FBVS, and LBVS. Moreover, VS techniques can be divided according to the algorithms used to perform its pipeline.

In the literature, several taxonomies are used to classify the algorithms, such as those presented in [39–44]. However, due to the diversity of the algorithms described in this work, we chose to use a new taxonomy, as shown in Figure 2, in which the algorithms are divided into similarity, quantitative, machine learning, meta-heuristic, and others. In the following sections, we explain the algorithms, describe how they are used, and how they contribute to the drug discovery process.

3.1. Similarity-Based Algorithms

In similarity-based approaches, algorithms assume that a molecule with a similar structure to another whose biological activity is known may share similar activity [16]. In a similarity search, a structure with a known biological activity relating to the target is used as a reference. These reference structures are then compared to a database of known molecules so that the similarity between them is determined. This similarity can be determined by a coefficient of similarity applied to all molecules in the database. However, two highly similar compounds may have different biological activities. Thus, small changes in a molecule can either generate an active or an inactive compound [27]. The methods based on similarity tend to only require a few pieces of reference information to create predictive models [45]. Therefore, their simplicity and effectiveness have made them widely accepted by the scientific community. Additionally, the similarity-based algorithms can be combined with other methods such as ML, meta-heuristics, and others to improve the results.

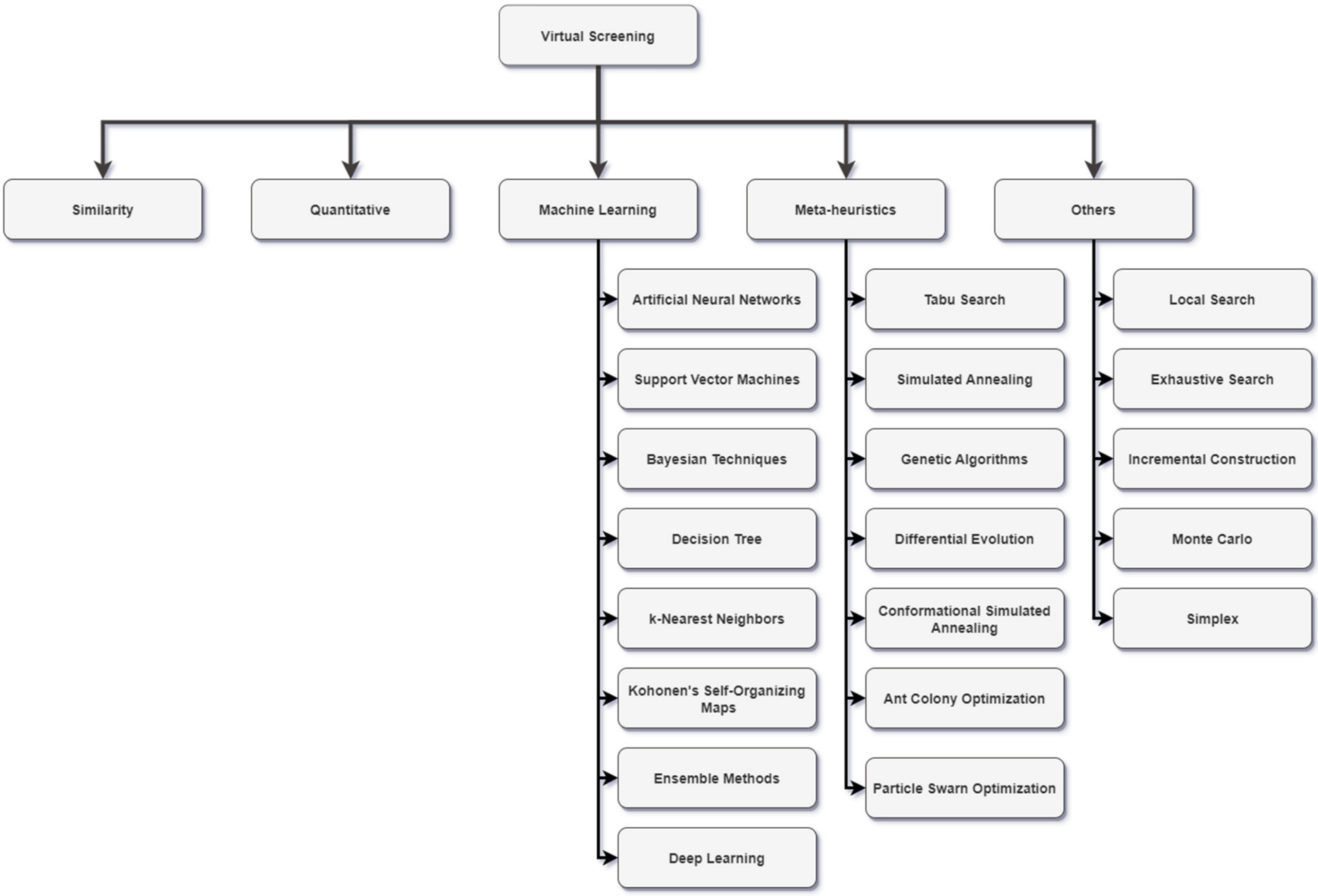

Figure 2. VS algorithms taxonomy.

The similarity-based algorithms can be divided into 2D, 3D, quantum, and hybrid approaches. The 2D approaches are the most intuitive VS technique based on the 2D similarity of substructures compared to various 2D descriptors using a similarity coefficient [46–48]. Then, in 3D approaches, the similarity search can also be performed through 3D molecular representations, such as in pharmacochemical [27,32,33,49], overlapping volumes [50,51], and molecular interaction fields (MIFs) [16]. Next, the quantum approach mainly focuses on quantum descriptors that produce good results but depend on conformation [52–55]. Finally, this could still be achieved with a hybrid approach [56–61] using more than one of the options above.

3.2. Quantitative Algorithms

QSAR modeling is one of the well-developed LBDD methods. QSAR models have attempted to predict the property and/or activity of interest of a given molecule.

A QSAR model is formed by a mathematical equation that compares the characteristics of the investigated molecules with their biological activities. Mathematically, according to Faulon and Bender [62], any QSAR method can usually be defined as the application of mathematical and statistical methods to the problem of finding empirical equations in the form Yi = F $(X_1, X_2, \ldots, X_n)$, where the variable Yi is the approximate biological activity (or another property of interest) of the molecules and $X_1, X_2, \ldots, X_n$ are structural experimental or calculated properties (molecular descriptors such as molar weight, logP, fragments, number of atoms and/or bonds). F is some mathematical relationship empirically determined that must be applied to descriptors to calculate property values for all molecules.

At first, QSAR modeling was only used with simple regression methods and limited congener compounds. However, QSAR models have grown, diversified, and evolved to modeling and VS in large databases using various ML techniques [63].

Benedetti and Fanelli [64] show several commonly used QSAR models (local, pharmacophore-based, and global QSAR models). Fan et al. [65] used a ligand-based 3D QSAR model based on in vitro antagonistic activity data against adenosine receptor 2A (A2A). The resulting models, obtained from 268 chemically diverse compounds, were used to test 1897 chemically distinct drugs, simulating the real-world challenge of a safety screening when presented with novel chemistry and a limited training set. Additionally, it presented an in-depth analysis of the appropriate use and interpretation of pharmacophore-based 3D QSAR models for safety.

The objective of a QSAR model is to establish a trend of molecular descriptor values that correlate with the values of biological activity [17]. Therefore, it is essential to minimize the prediction error, for example, by using the sum of squared differences or the sum of absolute values of the differences between the expected and observed activities in Yi for a training set. However, before constructing the model, the data need to be pre-processed and adequately prepared to build and validate the models [62].

3.3. Machine Learning-Based Algorithms

The availability of large databases of molecules storing information such as the structure and biological activity makes it possible to use predictive models based on ML, which require a large amount of data from active and inactive compounds to make the prediction. Therefore, ML techniques have been increasingly used in VS due to their accuracy, expansion of chemical libraries, new molecular descriptors, and similarity search techniques [45].

ML is a field of artificial intelligence and computer science devoted to understanding and building algorithms that learn based on data, analogy, and experience [37]. The ML algorithms use datasets to learn and, based on knowledge acquired in the learning process, make decisions, make previsions, and recognize patterns. In addition, it is expected that the overall performance of the ML system will improve over time and the system will adapt to changes [41].

ML techniques stand out due to their ability to learn from data (examples). Furthermore, the learning methods can be divided into supervised, unsupervised, or reinforcement, depending on the algorithm [66]. Supervised learning requires a training set in which each sample is composed of the input and its desired output, i.e., the learning algorithm uses labeled data [67]. In supervised learning, the system parameters are adjusted by the error between the system output and the desired output [39]. In unsupervised learning, the training set is composed only of the inputs. The desired output is not available or does not exist at all [67]. The objective of unsupervised learning is to construct the system knowledge based on the spatial organization of the inputs. In other words, unsupervised learning aims to discover patterns using unlabeled data. Reinforcement learning uses a mechanism based on trial-and-error and reward, in which the correct action awards the learner and the wrong actions punish them [68]. The learner decides on actions based on the current environmental state and the feedback from previous actions [69]. The learner has no prior knowledge of what action to take, and the aim is to find the actions that maximize long-term reinforcement. Reinforcement learning is often used in intelligent agents [41].

The power of VS is increased with ML because it makes it possible, instead of performing computationally costly simulations or exhaustive similarity searches, to track predicted hits much faster and more accurately [23,70,71]. ML is used to find new drugs, repositioning compounds, predict interactions between ligands and protein, discover drug efficacy, ensure safety biomarkers, and optimize the bioactivity of molecules, to mention a few [72]. We can see in Figure 2 that the ML algorithms can be divided into artificial neural networks (ANNs), support vector machines (SVMs), Bayesian techniques, decision trees (DTs), k-nearest neighbors (k-NN), Kohonen self-organized maps (SOMs), deep learning

(DL), and ensemble methods. The following sub-sections explain the ML algorithms, show how they can be used, and contribute to the drug discovery process.

3.3.1. Artificial Neural Networks

ANNs are mathematical models inspired by the processing capacity of the human brain. A neural network is composed of a set of computational units called neurons, which are a simple mathematical model of biological neurons [67]. The neurons are connected through synaptic weights, which simulate the strengths of synaptic connections of the biological brain [73]. The synaptic weights represent the knowledge in the ANNs, and the learning is performed by changing these weights. Generally, the ANNs use a supervised learning algorithm to update the weights based on errors obtained between the network output and the desired output. The neural network structure can be defined by the number of layers, the number and type of neurons, and the learning algorithm [74]. For example, Figure 3 illustrates the structure of a neural network with two inputs, four neurons in the first layer (hidden layer), one neuron in the second layer (output layer), and the synaptic weights [75].

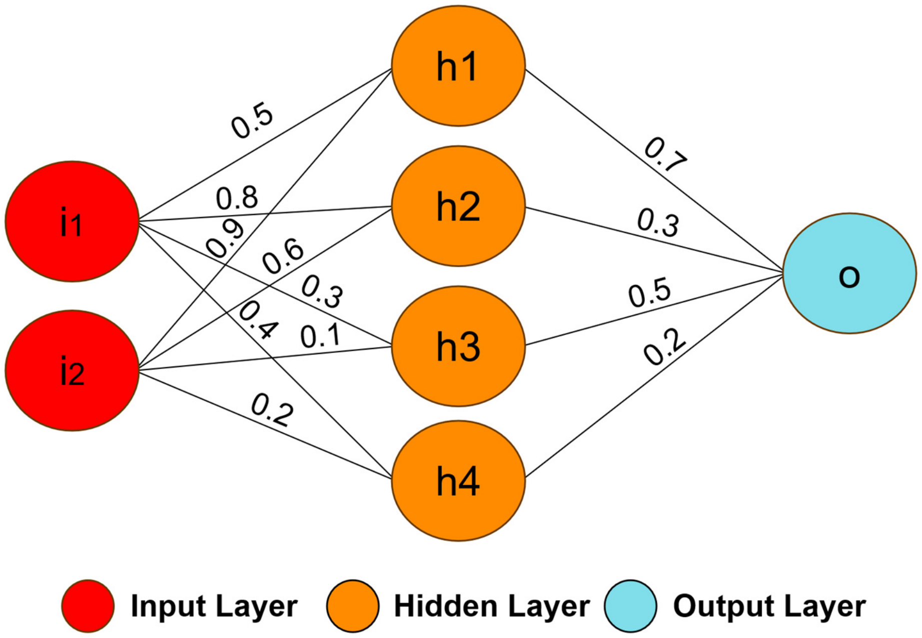

Figure 3. An example of a simple neural network model. The model was formed by two inputs (i1 and i2), four neurons (h1, h2, h3, and h4) in the first layer, one neuron (o) in the output layer, and the synaptic weights.

Different efforts introduce ANNs in medicinal chemistry to classify compounds, principally in studies with QSAR, to identify potential targets and the localization of structural and functional characteristics of biopolymers. In such a context, Lobanov [76] introduces a review to argue how ANNs can be used in the VS pipeline from combinatorial libraries, showing how selecting descriptors is essential to evaluate the ANNs. Furthermore, in such a review it is emphasized that descriptors can be found in one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) forms and can be used to create a model for ADMET properties profiles.

Tayarani et al. [77] propose an ANN model based on molecular descriptors to obtain the binding energy using the physical and chemical descriptions of the selected drugs. The authors showed a high correlation between the observed and calculated binding energies by the ANN since the average error ratio was less than 1%. The experimental results suggested that ANN is a powerful tool for predicting the binding energy in the drug design process. Recently, Mandlik, Bejugam, and Singh [78] wrote a book chapter to show the possibilities of using ANNs to aid in developing new drugs.

3.3.2. Support Vector Machine (SVM)-Based Techniques

SVM is a supervised ML algorithm for classification and regression tasks [41,79]. In a D-dimensional data space, an SVM creates a hyperplane (h) or a set of hyperplanes to cluster input data elements according to the similarity expressed in the descriptors' characteristics. Such hyperplanes employ linear and nonlinear functions to identify margins that separate a set of classes (x). In general, the greater the operating margin, the lower the classifier's generalization error. Intuitively, a good separation is achieved using the hyperplane with the most significant distance looked at in the closest training data of any class. For example, Figure 4 shows a classification instance where x1 and x2 denote a pair of classes, with h1, h2, and h3 as operating margins in which h1 does not separate the classes, h2 separates the classes, and h3 separates them with the highest operating margin.

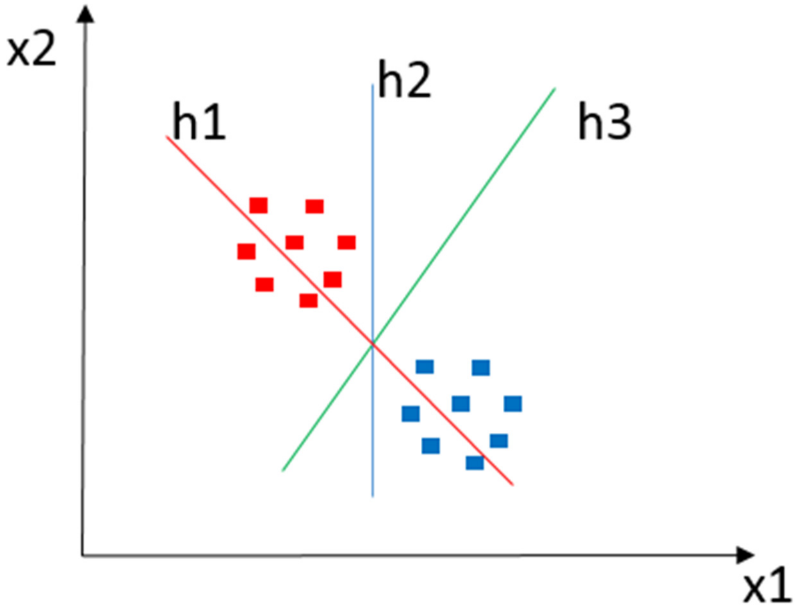

Figure 4. Classification example using SVM. The red and blue squares represent objects of each class to be classified. h1 does not separate classes. h2 does, but only by a small margin. h3 manages to separate them with the maximum margin.

SVMs have recently been cited as a promising technique for VS [79–82]. Rodrígues-Pérez et al. [77] describe how the algorithm can be used in VS and create a comparison with others in ML presenting its advantages and disadvantages. As an example, it is mentioned that SVM has evolved as one of the "premier" ML approaches and has been used as a approach of choice, given its typically high performance in compound classification and property predictions on the basis of limited training data. This is possible thanks to the adaptability and versatility of SVM for specialized applications.

Silva et al. [83] used Autodock Vina in the docking process, but they proposed an alternative SVM-based scoring function. They showed that while Autodock Vina offers a prediction of acceptable accuracy for most targets, classification using SVM was better, which illustrates the potential of using SVM-based protocols in VS. Finally, Li et al. [84] describe the ID-Score, a new scoring function based on a set of descriptors related to 2278 protein–ligand complexes extracted from Protein Data Bank (PDB). The results indicated that ID-Score performed consistently, implying that it can be applied across a wide range of biological target types.

3.3.3. Bayesian Techniques

Bayesian techniques are based on Bayes' theorem and describe the probability of an event occurring due to two or more causes. This approach combines previous beliefs with new evidence to compose a hypothesis and describe a model. The main implementation of this model was algorithms such as the naïve Bayes classifier and Bayesian networks [85].

In repositioning and drug discovery scenarios, the Bayesian algorithms identify the probability of a compound possessing biological activity for a specific target. Therefore, Bayesian techniques can be used, for instance, to find novel active scaffolds against the same target or to find compounds that are more active or that possess improved ADMET properties when compared to the main structure [86]. However, Bayesian networks were not explicitly superior to much simpler approaches, based, for example, on the Tanimoto coefficient [87].

3.3.4. Decision Trees

DTs are a supervised ML algorithm that describes functions to take data attributes as input and return a single output, sometimes called a "decision" by some authors [69]. DTs make decisions by performing a sequence of tests, starting in the first tree node called "root" and proceeding in a top–down strategy until reaching the last level of the structure, with nodes typically named leaves.

To compose the aforementioned structure as a tree, firstly, a test value that reports some relationship with each input attribute is associated with each internal node, except the root. Then, such nodes are labeled with the attribute's decision possibilities, and leaves specify the final results that the function should return according to the characterizations exploited. Figure 5 shows how a DT works. In Figure 5 diagram, we introduce a toy example to identify whether the weather is good for playing sports outdoors. The decision process starts in the root node where weather (i.e., sunny, cloudy, or rainy) is observed. Then, according to weather conditions, the decision function estimates a novel specific node (at a lower level), repeating this process until some leaf is achieved, which is used to extract and return the result.

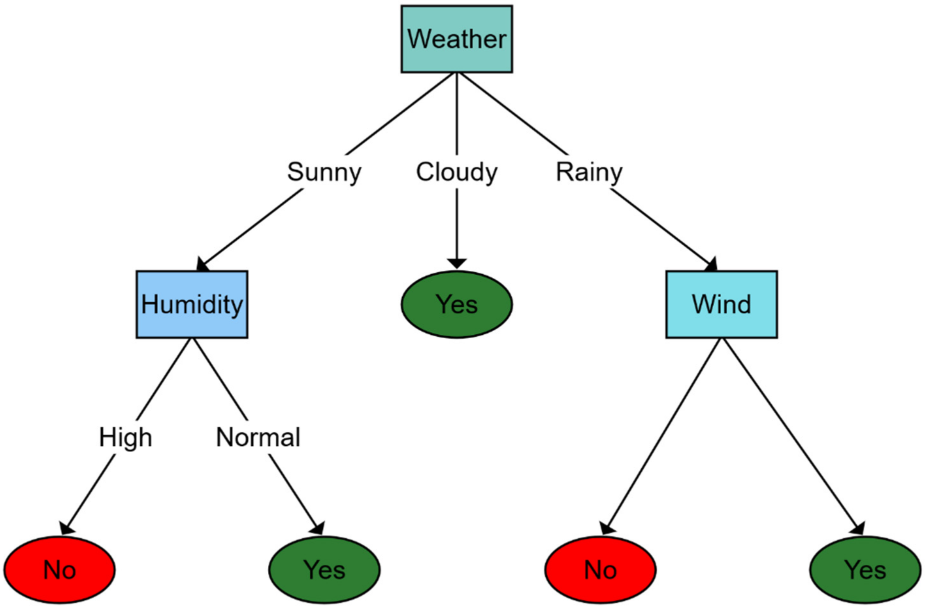

Figure 5. Classification example using a decision tree to determine whether the weather is good enough to go outside. The first level is determined by how the weather is. Depending on the answer, a new question is selected, or the answer is found. For example, if the weather is sunny, it is questioned about the humidity depending on whether it is high or normal to determine whether the final answer is yes or no.

Lavecchia [76], argue that DTs are a useful alternative to creating combinatorial libraries to predict the similarities between drugs or specific biological activities. However, to extract high performance using DTs, a single tree structure is not enough. In this context, some efforts, such as those presented in [88], propose applying a set of DTs to increase the forecast variation. These proposals are called random decision forests (RDFs). This generalization of DTs is employed by Lavecchia to demonstrate improving the LBVS performance, the predictive ability of quantitative models, and unique decision tree algorithms. Considering our paper's focus, decision-making forests can be considered as a choice to determine protein-binding affinity, in addition to being widely used in other coupling applications. In DT, molecules can be classified from the input data employed to design their structure during the training stage [70].

Considering VS, is possible to observe DTs employed as a support tool and classifier functions. An alternative method to construct DT-based classifiers with higher precision for repositioning and discovery of drugs, improving accuracy as its structure grows in complexity, has been described by Ho and collaborators. The proposed method uses a pseudorandom procedure to choose a member of a resource vector. According to the choice made, DTs are generated using only the components that have the selected characteristics [89]. DTs have been applied to investigate interactions among biological activities, chemical substructures, and phenotypic effects [88]. This study assumes chemical structure fingerprints provided by the PubChem system, divided across four assays according to biological activity data in a 10-fold cross-validation (CV) sensitivity, specificity, and Matthews correlation coefficient (MCC) with results of 57.2~80.5%, 97.3~99.0%, and 0.4~0.5, respectively.

3.3.5. K-Nearest Neighbors

KNN is the most popular unsupervised learning method used for classification and regression, whose strategy boils down to looking for, from some prefixed data input element, the k-closest neighbors. In scenarios where input data elements are fashioned by the same descriptors and drawn in the same n-Dimensional space, some metrics can be chosen to compare them, including those based on distances such as Euclidean, cosine, and Manhattan [90]. In a correlation analysis context, Pearson [91] and Kendall [92] may be sufficient. For input data elements represented by a particular structure such as trees, Resnik [93], Jiang [94], and Lin [95] should be considered.

Once the adequate metric is found, the param K in the KNN algorithm should be defined before execution, which matches the number of neighborhood points extracted as results. For example, if k = 1, only the training element closest to the instance to be classified is selected. If k = 5, the five elements closest to the instance are selected, and based on the classes of the five elements, the class of the test element is inferred. The idea of this algorithm is that the class to which a compound belongs can be defined by the class to which the closest neighbors belong (more similar compounds, for example), considering weighted similarities between the compound and its closest neighbors [96].

In VS, Shen et al. [97] created a new drug design strategy that uses KNN to validate developed QSAR models. They identified and synthesized nine potential anticonvulsant drugs by evaluating a library of more than 250,000 molecules. Of these, seven were confirmed as bioactive (a success rate of 77%). Peterson et al. [98] used KNN in QSAR and VS studies and found 47 types I geranyl transferase inhibitors in a database with 9,500,000 chemicals. Seven were successfully tested in vitro and presented activity at the micromolar level. These new hits could not be identified using traditional similarity search methods that demonstrate that KNN models are promising and can be used as an efficient VS method.

3.3.6. Kohonen's SOMs

Kohonen's self-organizing maps (SOMs) [99] are unsupervised algorithms that use competitive networks formed by a two-layered structure of neurons. The first layer is called the input layer, with its neurons completely interconnected to the neurons of the second layer, which is organized in a dependent arrangement of the object to be mapped. Kohonen maps are therefore formed by connected nodes with an associated vector corresponding to the map's input data (e.g., molecular descriptors). Kohonen algorithms can analyze grouped data to discover multidimensional structures and patterns. The Kohonen networks can also be considered an unsupervised neural network since having a known target vector is not necessary. It is considered a self-organized algorithm since it can decrease the size of a data group while maintaining the real representation concerning the relevant properties of the input vectors, resulting in a set of characteristics of the input space.

The Kohonen self-organizing maps are defined by Ferreira et al. [100] as competitive neural networks with a high degree of interconnection between their neurons that can form mappings, preserving the topology between the input and output spaces. For example, in VS, Kohonen's SOMs could be employed extensively in drug repositioning and scaffold hopping [101]. Noeske et al. [102] used Kohonen's SOMs to discover new targets for metabotropic glutamate receptor antagonists (mGluR). Experiments have revealed distinct subclusters of mGluR antagonists, and localization overlaps with ligands known to bind to histamine (H1R), dopamine (D2R), and various other targets. These interactions were later confirmed experimentally and exhibited significant binding affinities between the predicted mGluR antagonists for the targets. This led to a mGluR1 antagonist and selective subtype (Ki = 24 nM) based on a coumarin scaffold. This compound was later developed into a series of leads.

In another study, Noeske et al. [103] used Kohonen's SOMs to map the drug-like chemical space using pharmacophore descriptors. The experiments demonstrated that other G protein-coupled receptors (GPCRs) could interact with mGluR ligands (mGluR1: dopamine D2 and D3 receptors, histamine H1 receptor, mACh receptor; mGluR5: histamine H1 receptor). Lastly, the results were experimentally confirmed, and their IC50 ranged from 5 to 100 µM. Hristozov et al. [104] use Kohonen's SOMs in LBVS to rule out compounds with a low probability of having biological activity. The proposed idea can be used, according to the authors, (1) to improve the recovery of potentially active compounds; (2) to discard compounds that are unlikely to have a specific biological activity; and (3) to select potentially active compounds from a large dataset. In a recent paper, Palos et al. [105] applied Kohonen's SOMs in ligand clustering to perform drug repositioning in FDA-approved drugs. This research suggests that four FDA drugs could be used for Trypanosoma cruzi infections.

3.3.7. Ensemble Methods

Ensemble methods combine multiple models rather than employing a single model to increase model accuracy. The results are significantly more accurate when the combined models are used. This has increased the popularity of ensemble methods in ML [106].

Sequential and parallel ensemble techniques are the two main types of ensemble methods. Adaptive boosting (AdaBoost) is an example of a sequential ensemble technique that generates base learners in a sequence. The sequential generation of base learners facilitates the base learners' dependence on one another. The model's performance is then improved by giving previously misrepresented learners more weight.

Base learners are generated in parallel ensemble techniques, such as a random forest. To foster independence among base learners, parallel methods use the parallel generation of base learners. The application of averages results in a significant reduction in error due to the independence of base learners.

Most ensemble methods employ a single algorithm in base learning, resulting in uniformity among all base learners. Homogeneous base learners are base learners of the same type and share many of the same characteristics. Heterogeneous ensembles are created when heterogeneous base learners are used in other approaches. Different types of learners make up heterogeneous base learners.

In a recently published paper, Helguera [107] reported that using ensemble techniques, such as bagging [108] and boosting [109], provides better predictions than any of the above techniques in isolation in ML. Therefore, Helguera proposed an ensemble method based on genetic algorithms (GAs). To investigate potential Parkinson's disease therapeutics, Helguera applied his proposed method for the LBVS of dual-target A2A adenosine receptor antagonists and MAO-B inhibitors. Finally, he showed that the ensemble method outperformed individual models.

A combination of VS and MD with Ensemble methods to evaluate the ΔT_m is described in [110], which used Ensemble methods to estimate the ΔT_m experimental values of DNA intercalating agents. These authors also evaluated three docking methodologies, and the best five molecules were submitted to MD simulations.

3.3.8. Deep Learning

DL is a large class of ML techniques based on neural networks in which inputs are successively transformed into alternative representations that better allow for pattern extraction [39,41–44,67,69,73,111]. The word "deep" refers to the fact that circuits are typically organized in many layers, meaning that the computation paths from inputs to outputs have many steps. DL is currently the most widely used approach for applications such as visual object recognition, machine translation, speech recognition, speech synthesis, and image synthesis; it also plays a significant role in reinforcement learning applications. Figure 6 shows the main difference between ANNs and the DL approach, namely the number of hidden layers. In ANNs, one or two hidden layers are usually used, while in DL, more than five layers are used, leading to more meaningful processing and better results.

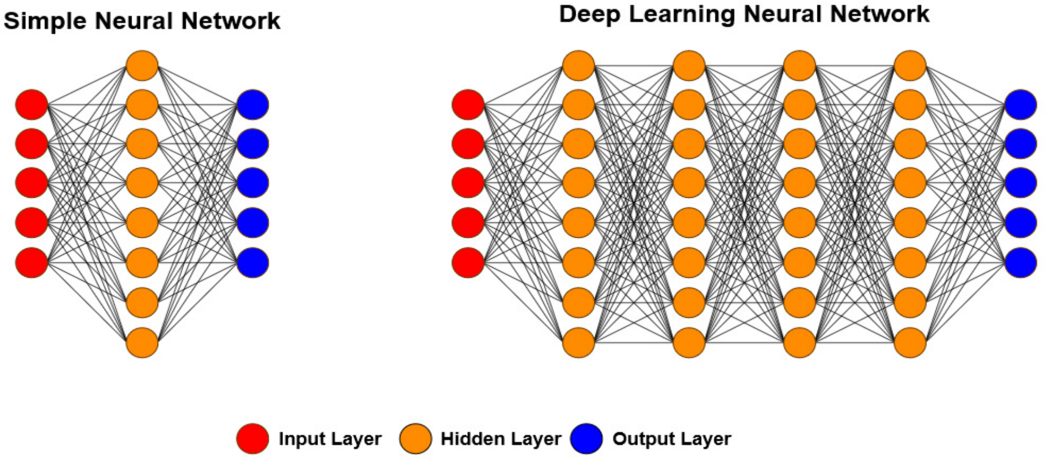

Figure 6. Differences between ANN and DL architectures. In a simple neural network, one or two hidden layers are used, while in DL, more than two layers are always used. With this, the required processing is more significant; however, in most cases, the results are better.

Several studies with DL have been applied in VS. Multiple reviews [63,112–115] have presented a detailed explanation of DL and how it can be used in VS and described methods and techniques along with problems and challenges in the area. Bahi et al. [116] present a compound classification method based on a deep neural network for VS (DNN-VS) using the Spark-H2O platform to label small molecules from large databases. Experimental results have shown that the proposed approach outperforms state-of-the-art ML techniques with an overall accuracy of more than 99%.

Joshi et al. [117] used DL to create a predictive model for VS, and the molecular dynamics (MD) of natural compounds against the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) were deployed on the Selleck database containing 1611 natural compounds for VS. From 1611 compounds, only four compounds were found with drug-like properties, and three were non-toxic. The results of MD simulations showed that two compounds, Palmatine and Sauchinone, formed a stable complex with the molecular target. This study suggests that the identified natural compounds may be considered for therapeutic development against SARS-CoV-2.

More recently, Schneider et al. [118] presented DL for VS in antibodies to predict antibody–antigen binding for antigens with no known antibody binders. As a result, the DL approach for antibody screening (DLAB) can improve antibody–antigen docking and structure-based VS of antibody drug candidates. Additionally, DLAB enables improved pose ranking for antibody docking experiments and the selection of antibody–antigen pairings for which accurate poses are generated and correctly ranked.

3.4. Meta-Heuristic Algorithms

Meta-heuristic approaches are widely applied in searching for and finding approximate optimum solutions. Meta-heuristic can be classified as single-solution or population-based [119]. Single-solution meta-heuristics aim to change and improve the performance of a single solution. Tabu search (Section 3.4.1.) and simulated annealing (Section 3.4.2.) are examples of single-solution meta-heuristics used in drug discovery. On the other hand, population-based methods maintain and improve multiple solutions, each corresponding to a unique point in the search space [120]. Population-based meta-heuristics applied in drug discovery include genetic algorithms (Section 3.4.3.), differential evolution (Section 3.4.4.), conformational space annealing (Section 3.4.5.), ant colony optimization (Section 3.4.6.), and particle swarm optimization (Section 3.4.7.).

3.4.1. Tabu Search

The TS algorithm performs the search using a memory structure that accepts the investigation of movements that do not improve the current solution [121]. In TS, the memory is used to create a Tabu list consisting of solutions that will not be revisited, preventing cycles, making it possible to explore unvisited regions, and improving the decision-making process through past experiences.

PRO_LEADS [122] is one of the most popular algorithms using TS in VS. It proposes using a TS algorithm to explore the large search space and uses an empirical scoring function that estimates the binding affinities to order the possible solutions. PRO_LEADS was tested on 50 protein–ligand complexes and accurately predicted the binding mode of 86% of the complexes.

3.4.2. Simulated Annealing (SA)

Simulated annealing, a general probabilistic meta-heuristic, is a method that can approximate the global optimum of a given function given ample search space [123]. In summary, SA is an approach for resolving bound-constrained and unconstrained optimization issues. This method simulates the actual physical procedure of heating a material and gradually lowering the temperature to reduce defects and reduce system energy.

A new point is generated randomly during each iteration of the simulated annealing algorithm. A probability distribution with a temperature-dependent scale is the foundation for the new point's distance from the current point or the scope of the search. All new points that lower the objective are accepted by the algorithm, as are points that raise the objective with a certain probability. The algorithm can look globally for more potential solutions rather than becoming stuck in local minima by accepting points that raise the goal. An annealing schedule is chosen to lower the temperature as the algorithm progresses. The algorithm reduces the scope of its search to a minimum to converge as the temperature drops.

Within the MD context, simulated annealing molecular dynamics (SAMD) involves heating a particular crystallographic (ligand–receptor) complex to high temperature, followed by gradual slow cooling to unveil conformers of reasonable energy minima [124]. The process is usually repeated over multiple cycles (10–100) [124,125]. One use of SAMD is calculating the optimal docked pose/conformation for a particular ligand within a specific binding pocket, as in the CDOCKER docking engine. Hatmal and Taha [126] use a methodology implemented to develop pharmacophores for acetylcholinesterase and protein kinase C-θ. The resulting models were validated by receiver-operating characteristic analysis and in vitro bioassay. Thus, in this research, four new protein kinase C-θ inhibitors among captured hits, two of which exhibited nanomolar potencies, were identified.

3.4.3. Genetic Algorithms

GAs are a technique for tackling both obliged and unconstrained improvement issues that depend on regular choice, the interaction that drives natural development [127]. They are implemented as a computer simulation where a population undergoes mutations to find better solutions. Each new individual represents a possible solution to the problem. For each new generation, the most suitable individuals are evaluated. This process selects some individuals for the next generation, and they are recombined and mutated to generate new individuals. Although random, they are targeted to a better solution by exploring available information to find new search points where better performance is expected. This process is repeated until it is finalized. The problem with GAs is that they typically have a poor processing time. Therefore, they are mostly used in challenging problems.

GAs are widely used in VS. For example, Bhaskar et al. [128] used a genetic algorithm in their docking process to uncover 14 lead candidates that selectively inhibit the NorA efflux pump competing with its substrates. In addition, Rohilla, Khare, and Tyagi [129] evaluated, using a genetic algorithm, the ability of the IdeR transcription factor (essential in regulating the intracellular levels of iron) to bind to the DNA after being exposed to some compounds. The most potent inhibitors showed IC50 values of 9.4 mM and 6.9 mM. After analyzing the inhibition results, it was possible to find an important scaffold in inhibiting the protein (benzo-thiazol benzene sulfonic).

Xia et al. [130] applied a genetic algorithm to investigate a potential therapy to treat neurodegenerative disorders (e.g., Parkinson's disease) and stroke. They performed a VS experiment based on the X-ray structure of the Pgk1/terazosin complex and selected 13 potential anti-apoptotic agents selected from the Specs chemical library. In vitro experiments performed by the authors demonstrated a high chance of three of the selected compounds binding to hPgk1-like terazosin. These results indicate that they can act by activating hPgk1 and as apoptosis inhibitors.

3.4.4. Differential Evolution

DE is a heuristic strategy for globally optimizing nonlinear and non-differentiable continuous space functions. A broader group of evolutionary computing algorithms includes the DE algorithm. The DE algorithm begins with an initial population of candidate solutions, just like other popular direct search strategies such as GAs and evolution strategies. Then, by introducing mutations into the population, these candidate solutions are iteratively improved, keeping the fittest candidate solutions with lower objective function values. The basic idea of this operation is to add a randomly selected individual to the difference between two other randomly selected individuals [131].

MolDock [132] combines DE with a cavity prediction algorithm. In a VS test of ligands against target proteins, MolDock identified the correct binding mode of 87% of the complexes. This result was better than Glide [133], Gold [134], Surflex [135], and Flexx [136] in similar databases.

3.4.5. Conformational Space Annealing (CSA)

CSA is a stochastic global search optimization algorithm that uses randomness in the search process. Due to this, this method works well for nonlinear objective functions where other local search algorithms do not. It utilizes global optimization to find the optimal solution for a given objective function satisfying for any. CSA can be considered a type of GA because it performs genetic operations while maintaining a conformational population [137]. Here, the terminologies of "conformation" and "solution" are used interchangeably. CSA contains two essential ingredients of Monte Carlo and SA.

As in Monte Carlo, CSA only works with locally minimized conformations. However, in SA, annealing is performed in terms of temperature, while in CSA, annealing is performed in an abstract conformational space while the diversity of the population is actively controlled.

The CSA performance is described in Shin [138], which evaluates two existing docking programs, GOLD [139] and AutoDock [140], on the Astex diverse set.

3.4.6. Ant Colony Optimization

ACO is an algorithm inspired by the behavior of ants in the search of food. One of the central ideas of ACO, proposed by Dorigo et al. [141], is the indirect communication based on trails (or paths) of pheromones between a colony of agents called ants. In nature, ants search for food by walking on random tracks until they find it. Then, they return to the colony and leave behind a trail of pheromones. When other ants find this trail, they stop following random tracks and follow the trail to obtain food and return to the colony. Over time, the pheromone evaporates, reducing the ants' attraction to a certain path. The more ants go through a path, the more time it takes for the pheromone to evaporate. Pheromone evaporation is important because it prevents an optimal local solution from becoming excessively attractive over time.

The Protein–Ligand ANT System (PLANTS) [142] is used in protein–ligand docking. In this algorithm, an artificial colony of ants is used to find the conformation of the least energized ligand at the binding site. In this artificial colony, the behavior of real ants is imitated by marking the low-energy conformations with pheromone trails. This information of the pheromone track is then modified in the subsequent iterations, aiming to increase the probability of generating conformations with lower energy. According to Reddy et al. [143], this algorithm behaves well when tested with GOLD in experimentally developed structures.

3.4.7. Particle Swarm Optimization

Particle swarm optimization (PSO) optimizes a problem iteratively by improving a candidate solution respecting a certain quality measure. It has many similarities to GAs. In PSO, the system is initialized with an individual (population) representing random solutions, and it searches for the best solution by updating the generations. In the PSO, the possible solutions are called particles and fly through the problem space following the current optimal particles. Comparing it with GA, PSO is easier to implement because it has fewer parameters to calibrate. Although classified as an evolutionary algorithm, PSO does not have the survival characteristic of the fittest or the use of genetic operators, such as crossing and mutation [144].

Liu, Li, and Ma [145] compared a PSO algorithm against four state-of-the-art docking tools (GOLD, DOCK, FlexX, and AutoDock with Lamarckian GA) and it performed better in the evaluated database. Gowthaman, Lyskov, Karanicolas [146], and PSOVina [147] are other examples of particle swarming applied in VS.

3.5. Other Techniques-Based Algorithms

In the literature, algorithms are also found based on other techniques applied in the drug development process. Among these, we can highlight Local Search (Section 3.5.1), Exhaustive search (Section 3.5.2), Simplex (Section 3.5.3), Incremental Construction (Section 3.5.4), and Monte Carlo (Section 3.5.5).

3.5.1. Local Search

Local search is a method to solve computationally complex optimization problems. Local search algorithms test all solutions in the candidate solution space, applying only local changes until a solution considered optimal is found or a specific time limit has elapsed. These algorithms look for reasonable solutions in large or infinite-state spaces. Autodock Vina [148], SwissDock/EADock [149], and GlamDock [150] are examples of tools that use the local search for VS.

3.5.2. Exhaustive Search

For problems with no efficient solution method, it could be interesting to test each possibility sequentially to determine the best solution. This exhaustive examination of all possibilities is known as exhaustive search, direct research, or the "brute force" method. An exhaustive search is typically used when the problem's size is limited, or problem-specific heuristics can be used to reduce the number of possible solutions to a manageable level. This method is also used when implementation simplicity is more important than speed.

In VS, eHiTS [151] software offers the first genuinely exhaustive systematic search algorithm that considers all poses without a severe steric clash of all ligands.

3.5.3. Simplex Algorithm

The Simplex algorithm is an iterative procedure for solving linear problems in a finite number of steps [152]. It consists of (i) knowing an initial basic feasible solution, (ii) testing whether the solution is optimal, (iii) improving the solution from a set of rules, and repeating the process until an optimal solution is obtained. For example, the Simplex method was used with GA to create rDock [153].

3.5.4. Incremental Construction

This is an algorithm for incremental construction in which the ligand is gradually incorporated into the binding site [58–60]. The chemical structure is initially broken up into several pieces with this strategy. Next, one of these pieces is chosen to serve as an anchor fragment, and it is docked in a complementary area of the binding site while the other pieces are added one at a time. The procedure goes on until the entire ligand is made. Finally, the calculation executes the conformational search just for the additional pieces, lessening the levels of opportunity to be investigated, and consequently staying away from the combinatorial blast.

3.5.5. Monte Carlo

A Monte Carlo method uses a statistical methodology based on a large set of random samples to obtain results that approximate reality [154]. Thus, Monte Carlo methods perform a sufficiently high number of successive simulations to allow for probabilities to be calculated heuristically.

When used as docking methods, Monte Carlo methods randomly generate an initial conformation of the ligand and calculate its binding energy. Based on this initial conformation, a new configuration is generated. Let us suppose that the binding energy for the new configuration is lesser (i.e., more negative) than for the initial conformation. In that case, it is automatically accepted as the reference for the next iteration. Otherwise, another evaluation is performed to verify whether it should be used as the reference. This process is repeated until the desired number of iterations is reached.

4. Guidelines for In Silico Models for CADD

Although some in silico models have been created and used for years to evaluate chemicals in some countries, without a transparent validation process and an objective determination of the reliability of the models it is essential to increase their regulatory acceptance.

In 2004, the Organization for Economic Co-operation and Development (OECD) showed an initiative for (Q)SAR models [155] and created principles for the validation of such models based on the following concepts: (i) (Q)SRAR models should be fashioned toward a well-defined aim, (ii) aid algorithms should be focused only on the aims of (Q)SAR models, (iii) a well-defined application domain should be presented, and (iv) appropriate measures of goodness of fit, robustness and predictivity, and a mechanistic interpretation, should be applied if it is possible. Later in 2004 was published a full report describing all these points [156].

In 2007, the "Guidance Document on the Validation of (Q)SAR Models" was published by the OECD to provide advice on how to evaluate specific (Q)SAR models in light of the OECD's principles [157]. Attached as annexes are a validation checklist, a validation reporting format, and validation case studies.

In November 2011, the Government of Canada published the third edition of the PSPC National CADD Standard. A concerted effort was made to simplify the standard, and PSPC is aware of the emerging technology and processes related to building information modeling (BIM). However, as BIM represents a significant change, a new BIM standard, by necessity, will be created, facilitating the transition in the architecture, engineering, and construction (AEC) industries. In addition, some regions have developed a regional CADD standard, which is to be used as a complement to this national standard. Additionally, the Government of the USA and the European Chemicals Agency have already published their standards. Therefore, the recommendation is to create or follow a protocol with the best recommendations of some standards and updates them as needed.

Recently, Czub et al. [158] published a paper that evaluated whether an ML-based model could meet the OECD principles' regulation for the 5-HT1A receptor box. The model was developed based on a database with close to 9500 molecules by using an automatic ML tool (AutoML). First, the model selection was based on the Akaike information criterion value and 10-fold cross-validation routine. Later, good predictive ability was confirmed with an additional external validation dataset with over 700 molecules. Moreover, the multi-start technique was applied to test whether an automatic model development procedure results in reliable results. The information provided indicates that our final model leads to affinity predictions within the error range indicated by the FDA.

5. Final Considerations

Several in silico methods and techniques have been used to improve drug development. CADD allowed for the development of new compounds with a decrease in time and cost. As a result, VS has emerged as one of the most promising in silico drug design methods. This review focused on the most used algorithms in the CADD process and described how they can be used to contribute to the various stages of drug development, focusing primarily on the use of VS.

VS techniques can be divided according to algorithms based on similarity, quantitative, ML, meta-heuristics, and others. Each category was explained, and use cases were demonstrated. Among these techniques, the one that has been highlighted most recently is ML. ML methods have successfully been used to screen employees and aided in drug discovery. In addition, DL has been producing even more accurate models in the last 5 years. As a result, several studies and innovations benefit from the application of DL in CADD.

However, CADD tools have required a wide variety of knowledge from researchers, which is necessary for their integration in computation, chemistry, and biology to select and prepare the targets and ligands, generate the models, and analyze the results. As a result, the formation of multidisciplinary teams and researcher training is increasingly important for the selection of new hits and for optimizing HTS experiments.

Author Contributions: Conceptualization, A.G.T. and A.M.d.S.; methodology, A.G.T.; validation, A.G.T. and A.M.d.S.; formal analysis, T.A.d.O., M.P.d.S. and E.H.B.M.; investigation, T.A.d.O., M.P.d.S. and E.H.B.M.; resources, A.G.T. and A.M.d.S.; data curation, A.G.T. and A.M.d.S.; writing—original draft preparation, T.A.d.O.; writing—review and editing, A.G.T., A.M.d.S., E.H.B.M. and M.P.d.S.; visualization, T.A.d.O.; supervision, A.G.T.; project administration, A.G.T.; funding acquisition, A.G.T. and A.M.d.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by PPBE and PPGCF/UFSJ; Fundação de Amparo à Pesquisa do Estado de Minas Gerais—FAPEMIG, grants APQ-02741-17, APQ-00855-19, APQ-01733-21, and APQ-04559-22; and Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPq-Brazil, grants 305117/2017-3, 426261/2018-6; and by the fellowship of 2021 (grant 310108/2020-9).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the Federal University of São João del-Rei (UFSJ), Programa de Pós-graduação em Bioengenharia (PPBE/UFSJ), and the Federal Center for Technological Education of Minas Gerais (CEFET-MG) for providing the physical infrastructure and human resources.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Maia, E.H.B.; Medaglia, L.R.; da Silva, A.M.; Taranto, A.G. Molecular Architect: A User-Friendly Workflow for Virtual Screening. *ACS Omega* **2020**, *5*, 6628–6640. [CrossRef] [PubMed]
2. Rodrigues, R.P.; Mantoani, S.P.; de Almeida, J.R.; Pinsetta, F.R.; Semighini, E.P.; da Silva, V.B.; da Silva, C.H.T.P. Virtual Screening Strategies in Drug Design. *Rev. Virtual De Química* **2012**, *4*, 739–776. [CrossRef]
3. Leelananda, S.P.; Lindert, S. Computational Methods in Drug Discovery. *Beilstein J. Org. Chem.* **2016**, *12*, 2694–2718. [CrossRef] [PubMed]
4. Leisinger, K.M.; Garabedian, L.F.; Wagner, A.K. Improving Access to Medicines in Low and Middle Income Countries: Corporate Responsibilities in Context. *South Med. Rev.* **2012**, *5*, 3–8.
5. Chan, M. *Ten Years in Public Health 2007–2017*; Report by Dr Margaret Chan Director-General World Health Organization; World Health Organization: Geneve, Switzerland, 2017; ISBN 9789241512442.
6. Arrowsmith, J. A Decade of Change. *Nat. Rev. Drug Discov.* **2012**, *11*, 17–18. [CrossRef]
7. Stevens, H.; Huys, I. Innovative Approaches to Increase Access to Medicines in Developing Countries. *Front. Med.* **2017**, *4*, 1–6. [CrossRef]
8. Sridhar, D. Improving Access to Essential Medicines: How Health Concerns Can Be Prioritised in the Global Governance System. *Public Health Ethics* **2008**, *1*, 83–88. [CrossRef]
9. Ferreira, R.S.; Glaucius, O.; Andricopulo, A.D. Integrating Virtual and High-Throughput Screening: Opportunities and Challenges in Drug Research and Development. *Quim. Nova.* **2011**, *34*, 1770–1778. [CrossRef]
10. Martin, Y.C.; Abagyan, R.; Ferenczy, G.G.; Gillet, V.J.; Oprea, T.I.; Ulander, J.; Winkler, D.; Zefirov, N.S. Glossary of Terms Used in Computational Drug Design, Part II (IUPAC Recommendations 2015). *Pure Appl. Chem.* **2016**, *88*, 239–264. [CrossRef]
11. Horvath, D. A Virtual Screening Approach Applied to the Search for Trypanothione. *J. Med. Chem.* **1997**, *2623*, 2412–2423. [CrossRef]
12. Surabhi, S.; Singh, B. Computer aided drug design: An overview. *J. Drug Deliv. Ther.* **2018**, *8*, 504–509. [CrossRef]
13. Zhang, G.; Guo, S.; Cui, H.; Qi, J. Virtual Screening of Small Molecular Inhibitors against DprE1. *Molecules* **2018**, *23*, 524. [CrossRef] [PubMed]
14. Banegas-Luna, A.-J.; Cerón-Carrasco, J.P.; Pérez-Sánchez, H. A Review of Ligand-Based Virtual Screening Web Tools and Screening Algorithms in Large Molecular Databases in the Age of Big Data. *Future Med. Chem.* **2018**, *10*, 2641–2658. [CrossRef] [PubMed]
15. Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput. Aided Drug Des.* **2011**, *7*, 146–157. [CrossRef]
16. Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discov. Today* **2006**, *11*, 1046–1053. [CrossRef]
17. Kuntz, I.D.; Blaney, J.M.; Oatley, S.J.; Langridge, R.; Ferrin, T.E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269–288. [CrossRef]
18. Vázquez, J.; López, M.; Gibert, E.; Herrero, E.; Javier Luque, F. Merging Ligand-Based and Structure-Based Methods in Drug Discovery: An Overview of Combined Virtual Screening Approaches. *Molecules* **2020**, *25*, 4723. [CrossRef]
19. Lionta, E.; Spyrou, G.; Vassilatis, D.; Cournia, Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr. Top. Med. Chem.* **2014**, *14*, 1923–1938. [CrossRef]
20. Carregal, A.P.; Maciel, F.V.; Carregal, J.B.; dos Reis Santos, B.; da Silva, A.M.; Taranto, A.G. Docking-Based Virtual Screening of Brazilian Natural Compounds Using the OOMT as the Pharmacological Target Database. *J. Mol. Model.* **2017**, *23*, 111. [CrossRef]
21. Mugumbate, G.; Mendes, V.; Blaszczyk, M.; Sabbah, M.; Papadatos, G.; Lelievre, J.; Ballell, L.; Barros, D.; Abell, C.; Blundell, T.L.; et al. Target Identification of Mycobacterium Tuberculosis Phenotypic Hits Using a Concerted Chemogenomic, Biophysical, and Structural Approach. *Front. Pharmacol.* **2017**, *8*, 1–13. [CrossRef]
22. Wójcikowski, M.; Ballester, P.J.; Siedlecki, P. Performance of Machine-Learning Scoring Functions in Structure-Based Virtual Screening. *Sci. Rep.* **2017**, *7*, 46710. [CrossRef] [PubMed]

23. Carpenter, K.A.; Huang, X. Machine Learning-Based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review. *Curr. Pharm. Des.* **2018**, *24*, 3347–3358. [CrossRef] [PubMed]
24. Dutkiewicz, Z.; Mikstacka, R. Structure-Based Drug Design for Cytochrome P450 Family 1 Inhibitors. *Bioinorg. Chem. Appl.* **2018**, *2018*, 1–21. [CrossRef]
25. Nunes, R.R.; da Fonseca, A.L.; Pinto, A.C.D.S.; Maia, E.H.B.; da Silva, A.M.; de Varotti, F.P.; Taranto, A.G. Brazilian Malaria Molecular Targets (BraMMT): Selected Receptors for Virtual High-Throughput Screening Experiments. *Mem. Inst. Oswaldo Cruz* **2019**, *114*, 1–10. [CrossRef] [PubMed]
26. Kumar, A. Chemical Similarity Methods- A Tutorial Review. *Chem. Educ.* **2011**, *16*, 46–50.
27. Kristensen, T.G.; Nielsen, J.; Pedersen, C.N.S. Methods for similarity-based virtual screening. *Comput. Struct. Biotechnol. J.* **2013**, *5*, e201302009. [CrossRef] [PubMed]
28. Gozalbes, R.; Doucet, J.P.; Derouin, F. Application of Topological Descriptors in QSAR and Drug Design: History and New Trends Introduction: The Problem Of Obtaining New Drugs And The Qsar Approach. *Curr. Drug Targets Infect. Disord.* **2002**, *2*, 93–102. [CrossRef]
29. Maldonado, A.G.; Doucet, J.P.; Petitjean, M.; Fan, B.T. Molecular Similarity and Diversity in Chemoinformatics: From Theory to Applications. *Mol. Divers.* **2006**, *10*, 39–79. [CrossRef]
30. Leach, A.R.; Gillet, V.J.; Lewis, R.A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010**, *53*, 539–558. [CrossRef]
31. Liu, S.; Alnammi, M.; Ericksen, S.S.; Voter, A.F.; Ananiev, G.E.; Keck, J.L.; Hoffmann, F.M.; Wildman, S.A.; Gitter, A. Practical Model Selection for Prospective Virtual Screening. *J. Chem. Inf. Model.* **2018**, *59*, acs.jcim.8b00363. [CrossRef]
32. Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discov.* **2002**, *1*, 882–894. [CrossRef] [PubMed]
33. Wermuth, C.G. Similarity in Drugs: Reflections on Analogue Design. *Drug Discov. Today* **2006**, *11*, 348–354. [CrossRef] [PubMed]
34. Cruz-Monteagudo, M.; Medina-Franco, J.L.; Pérez-Castillo, Y.; Nicolotti, O.; Cordeiro, M.N.D.S.; Borges, F. Activity Cliffs in Drug Discovery: Dr Jekyll or Mr Hyde? *Drug Discov. Today* **2014**, *19*, 1069–1080. [CrossRef]
35. Li, Q. Application of Fragment-Based Drug Discovery to Versatile Targets. *Front. Mol. Biosci.* **2020**, *7*, 180. [CrossRef]
36. Singh, M.; Tam, B.; Akabayov, B. NMR-Fragment Based Virtual Screening: A Brief Overview. *Molecules* **2018**, *23*, 233. [CrossRef]
37. Kirsch, P.; Hartman, A.M.; Hirsch, A.K.H.; Empting, M. Concepts and Core Principles of Fragment-Based Drug Design. *Molecules* **2019**, *24*, 4309. [CrossRef] [PubMed]
38. Zhou, H.; Cao, H.; Skolnick, J. FRAGSITE: A Fragment-Based Approach for Virtual Ligand Screening. *J. Chem. Inf. Model.* **2021**, *61*, 2074–2089. [CrossRef] [PubMed]
39. Kubat, M. *An Introduction to Machine Learning*; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; ISBN 9783319639130.
40. Maia, E.H.B.; Assis, L.C.; de Oliveira, T.A.; da Silva, A.M.; Taranto, A.G. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Front. Chem.* **2020**, *8*, 343. [CrossRef]
41. Kordon, A.K. *Applying Computational Intelligence: How to Create Value*; Springer: Berlin/Heidelberg, Germany, 2010; ISBN 9783540699101.
42. Konar, A. *Computational Intelligence: Principles, Techniques and Applications*; Springer: Berlin/Heidelberg, Germany, 2008; ISBN 3540208984.
43. Smith, J. *Machine Learning Systems—Design That Scale*; Manning: Shelter Island, NY, USA, 2018; ISBN 9781617293337.
44. Engelbrecht, A.P. *Computational Intelligence: An Introduction*; John Wiley & Sons: Hoboken, NJ, USA, 2007; ISBN 978-0-470-51250-0.
45. Hönig, S.M.N.; Lemmen, C.; Rarey, M. Small Molecule Superposition: A Comprehensive Overview on Pose Scoring of the Latest Methods. *WIREs Comput. Mol. Sci.* **2023**, *13*, e1640. [CrossRef]
46. Tresadern, G.; Bemporad, D.; Howe, T. A Comparison of Ligand Based Virtual Screening Methods and Application to Corticotropin Releasing Factor 1 Receptor. *J. Mol. Graph. Model.* **2009**, *27*, 860–870. [CrossRef]
47. Richmond, N.J.; Abrams, C.A.; Wolohan, P.R.N.; Abrahamian, E.; Willett, P.; Clark, R.D. GALAHAD: 1. Pharmacophore Identification by Hypermolecular Alignment of Ligands in 3D. *J. Comput. Aided Mol. Des.* **2006**, *20*, 567–587. [CrossRef] [PubMed]
48. Jones, G.; Willett, P.; Glen, R.C. A Genetic Algorithm for Flexible Molecular Overlay and Pharmacophore Elucidation. *J. Comput. Aided Mol. Des.* **1995**, *9*, 532–549. [CrossRef]
49. Ostrowska, K.; Grzeszczuk, D.; Głuch-Lutwin, M.; Gryboś, A.; Siwek, A.; Leśniak, A.; Sacharczuk, M.; Trzaskowski, B. 5-HT1A and 5-HT2A Receptors Affinity, Docking Studies and Pharmacological Evaluation of a Series of 8-Acetyl-7-Hydroxy-4-Methylcoumarin Derivatives. *Bioorg. Med. Chem.* **2018**, *26*, 527–535. [CrossRef] [PubMed]
50. Kumar, A.; Zhang, K.Y.J. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Front. Chem.* **2018**, *6*, 315. [CrossRef] [PubMed]
51. Puertas-Martín, S.; Redondo, J.L.; Ortigosa, P.M.; Pérez-Sánchez, H. OptiPharm: An Evolutionary Algorithm to Compare Shape Similarity. *Sci. Rep.* **2019**, *9*, 1398. [CrossRef]
52. Lam, Y.H.; Abramov, Y.; Ananthula, R.S.; Elward, J.M.; Hilden, L.R.; Nilsson Lill, S.O.; Norrby, P.O.; Ramirez, A.; Sherer, E.C.; Mustakis, J.; et al. Applications of Quantum Chemistry in Pharmaceutical Process Development: Current State and Opportunities. *Org. Process. Res. Dev.* **2020**, *24*, 1496–1507. [CrossRef]
53. Fei, J.; Mao, Q.; Peng, L.; Ye, T.; Yang, Y.; Luo, S. The Internal Relation between Quantum Chemical Descriptors and Empirical Constants of Polychlorinated Compounds. *Molecules* **2018**, *23*, 2935. [CrossRef]

54. Bultinck, P.; Carbó-Dorca, R. Molecular Quantum Similarity Using Conceptual DFT Descriptors. *J. Chem. Sci.* **2005**, *117*, 425–435. [CrossRef]
55. Gugler, S.; Reiher, M. Quantum Chemical Roots of Machine-Learning Molecular Similarity Descriptors. *J. Chem. Theory Comput.* **2022**, *18*, 6670–6689. [CrossRef]
56. Elokely, K.M.; Doerksen, R.J. Docking Challenge: Protein Sampling and Molecular Docking Performance. *J. Chem. Inf. Model.* **2013**, *53*, 1934–1945. [CrossRef]
57. Huang, S.-Y.; Grinter, S.Z.; Zou, X. Scoring Functions and Their Evaluation Methods for Protein-Ligand Docking: Recent Advances and Future Directions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12899–12908. [CrossRef] [PubMed]
58. Ferreira, L.G.; dos Santos, R.N.; Oliva, G.; Andricopulo, A.D. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **2015**, *20*, 13384–13421. [CrossRef]
59. Haga, J.H.; Ichikawa, K.; Date, S. Virtual Screening Techniques and Current Computational Infrastructures. *Curr. Pharm. Des.* **2016**, *22*, 3576–3584. [CrossRef] [PubMed]
60. Breda, A.; Basso, L.A.; Santos, D.S.; De Azevedo, J.R.; Walter, F. Virtual Screening of Drugs: Score Functions, Docking, and Drug Design. *Curr. Comput. Aided Drug Des.* **2008**, *4*, 265–272. [CrossRef]
61. Liu, J.; Wang, R. On Classification of Current Scoring Functions. *J. Chem. Inf. Model.* **2015**, *55*, 475–482. [CrossRef] [PubMed]
62. Faulon, J.; Bender, A. *Handbook of Chemoinformatics Algorithms*; CRC Press: Boca Raton, FL, USA, 2010.
63. Neves, B.J.; Braga, R.C.; Melo-Filho, C.C.; Moreira-Filho, J.T.; Muratov, E.N.; Andrade, C.H. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. *Front. Pharmacol.* **2018**, *9*, 1275. [CrossRef]
64. de Benedetti, P.G.; Fanelli, F. Computational Quantum Chemistry and Adaptive Ligand Modeling in Mechanistic QSAR. *Drug Discov. Today* **2010**, *15*, 859–866. [CrossRef]
65. Fan, F.; Hamadeh, H.; Warshaviak, D.T.; Dunn, R. The Utilization of Pharmacophore-Based 3D QSAR Modeling and Virtual Screening in Safety Profiling: A Case Study to Identify Antagonistic Activities against Adenosince Receptor, A2aR, Using 1, 897 Known Drugs. *bioRxiv* **2018**, *14*, 413385. [CrossRef]
66. Morales, E.F.; Escalante, H.J. A Brief Introduction to Supervised, Unsupervised, and Reinforcement Learning. In *Biosignal Processing and Classification Using Computational Learning and Intelligence*; Academic Press: Cambridge, MA, USA, 2022; pp. 111–129. [CrossRef]
67. Haykin, S. *Neural Networks and Learning Machines*, 3rd ed.; Prentice Hall: New York, NY, USA, 2007; Volume 3, ISBN 9780131471399.
68. Ertel, W. *Introduction to Artificial Intelligence*; Springer: New York, NY, USA, 2017. [CrossRef]
69. Russel, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 4th ed.; Pearson: London, UK, 2021; ISBN 9780137505135.
70. Lavecchia, A. Machine-Learning Approaches in Drug Discovery: Methods and Applications. *Drug Discov. Today* **2015**, *20*, 318–331. [CrossRef]
71. Rifaioglu, A.S.; Atas, H.; Martin, M.J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. Recent Applications of Deep Learning and Machine Intelligence on in Silico Drug Discovery: Methods, Tools and Databases. *Brief. Bioinform.* **2018**, *20*, 1–36. [CrossRef]
72. Dara, S.; Dhamercherla, S.; Jadav, S.S.; Babu, C.M.; Ahsan, M.J. Machine Learning in Drug Discovery: A Review. *Artif. Intell. Rev.* **2022**, *55*, 1947. [CrossRef] [PubMed]
73. Aggarwal, C.C. *Neural Networks and Deep Learning*; Springer International Publishing: Berlin/Heidelberg, Germany, 2018.
74. Eberhart, R.C.; Dobbins, R.W. Early Neural Network Development History: The Age of Camelot. *IEEE Eng. Med. Biol. Mag.* **1990**, *9*, 15–18. [CrossRef]
75. Chan, H.C.S.; Shan, H.; Dahoun, T.; Vogel, H.; Yuan, S. Advancing Drug Discovery via Artificial Intelligence. *Trends Pharmacol. Sci.* **2019**, *40*, 592–604. [CrossRef] [PubMed]
76. Lobanov, V. Using Artificial Neural Networks to Drive Virtual Screening of Combinatorial Libraries. *Drug Discov. Today Biosilico* **2004**, *2*, 149–156. [CrossRef]
77. Tayarani, A.; Baratian, A.; Naghibi Sistani, M.-B.; Saberi, M.R.; Tehranizadeh, Z. Artificial Neural Networks Analysis Used to Evaluate the Molecular Interactions between Selected Drugs and Human Cyclooxygenase2 Receptor. *Iran. J. Basic Med. Sci.* **2016**, *16*, 1196–1202. [CrossRef]
78. Mandlik, V.; Bejugam, P.R.; Singh, S. *Application of Artificial Neural Networks in Modern Drug Discovery*; Elsevier Inc.: Amsterdam, The Netherlands, 2016; ISBN 9780128015599.
79. Sengupta, S.; Bandyopadhyay, S. Application of Support Vector Machines in Virtual Screening. *Int. J. Comput. Biol.* **2014**, *1*, 56–62. [CrossRef]
80. Han, B.; Ma, X.; Zhao, R.; Zhang, J.; Wei, X.; Liu, X.; Liu, X.; Zhang, C.; Tan, C.; Jiang, Y.; et al. Development and Experimental Test of Support Vector Machines Virtual Screening Method for Searching Src Inhibitors from Large Compound Libraries. *Chem. Cent. J.* **2012**, *6*, 139. [CrossRef]
81. Deshmukh, A.L.; Chandra, S.; Singh, D.K.; Siddiqi, M.I.; Banerjee, D. Identification of Human Flap Endonuclease 1 (FEN1) Inhibitors Using a Machine Learning Based Consensus Virtual Screening. *Mol. Biosyst.* **2017**, *13*, 1630–1639. [CrossRef]
82. Rodríguez-Pérez, R.; Bajorath, J. Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. *J. Comput. Aided Mol. Des.* **2022**, *36*, 355–362. [CrossRef]
83. Silva, C.; Simoes, C.; Carreiras, P.; Brito, R. Enhancing Scoring Performance of Docking-Based Virtual Screening Through Machine Learning. *Curr. Bioinform.* **2016**, *11*, 408–420. [CrossRef]

84. Li, G.B.; Yang, L.L.; Wang, W.J.; Li, L.L.; Yang, S.Y. ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein-Ligand Interactions. *J. Chem. Inf. Model.* **2013**, *53*, 592–600. [CrossRef] [PubMed]
85. Abdullah, A.A.; Hassan, M.M.; Mustafa, Y.T. A Review on Bayesian Deep Learning in Healthcare: Applications and Challenges. *IEEE Access* **2022**, *10*, 36538–36562. [CrossRef]
86. Zhou, W.-N.; Zhang, Y.-M.; Qiao, X.; Pan, J.; Yin, L.-F.; Zhu, L.; Zhao, J.-N.; Lu, S.; Lu, T.; Chen, Y.-D.; et al. Virtual Screening Strategy Combined Bayesian Classification Model, Molecular Docking for Acetyl-CoA Carboxylases Inhibitors. *Curr. Comput. Aided Drug Des.* **2018**, *15*, 193–205. [CrossRef] [PubMed]
87. Abdo, A.; Chen, B.; Mueller, C.; Salim, N.; Willett, P. Ligand-Based Virtual Screening Using Bayesian Networks. *J. Chem. Inf. Model.* **2010**, *50*, 1012–1020. [CrossRef]
88. Han, L.; Wang, Y.; Bryant, S.H. Developing and Validating Predictive Decision Tree Models from Mining Chemical Structural Fingerprints and High–Throughput Screening Data in PubChem. *BMC Bioinform.* **2008**, *9*, 401. [CrossRef]
89. Ho, T.K. The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. Pattern Anal. Mach Intell.* **1998**, *20*, 832–844. [CrossRef]
90. Homayouni, R.; Heinrich, K.; Wei, L.; Berry, M.W. Gene Clustering by Latent Semantic Indexing of MEDLINE Abstracts. *Bioinformatics* **2005**, *21*, 104–115. [CrossRef]
91. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson Correlation Coefficient. *Springer Top. Signal Process.* **2009**, *2*, 1–4.
92. Kendall, M.G. A New Measure of Rank Correlation. *Biometrika* **1938**, *30*, 81. [CrossRef]
93. Resnik, P. Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language. *J. Artif. Intell. Res.* **1999**, *11*, 95–130. [CrossRef]
94. Jiang, J.J.; Conrath, D.W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *arXiv* **1997**. [CrossRef]
95. Lin, D. An Information-Theoretic Definition of Similarity. *ICML* **1998**, *388*, 296–304.
96. Priya, S.; Tripathi, G.; Singh, D.B.; Jain, P.; Kumar, A. Machine Learning Approaches and Their Applications in Drug Discovery and Design. *Chem. Biol. Drug Des.* **2022**, *100*, 136–153. [CrossRef]
97. Shen, M.; Béguin, C.; Golbraikh, A.; Stables, J.P.; Kohn, H.; Tropsha, A. Application of Predictive QSAR Models to Database Mining: Identification and Experimental Validation of Novel Anticonvulsant Compounds. *J. Med. Chem.* **2004**, *47*, 2356–2364. [CrossRef]
98. Peterson, Y.K.; Wang, X.S.; Casey, P.J.; Tropsha, A. The Discovery of Geranylgeranyltransferase-I Inhibitors with Novel Scaffolds by the Means of Quantitative Structure-Activity Relationship Modeling, Virtual Screening, and Experimental Validation. *J. Med. Chem.* **2009**, *52*, 83–88. [CrossRef]
99. Vracko, M. Kohonen Artificial Neural Network and Counter Propagation Neural Network in Molecular Structure-Toxicity Studies. *Curr. Comput. Aided Drug Des.* **2005**, *1*, 73–78. [CrossRef]
100. Ferreira, L.L.G.; Andricopulo, A.D. ADMET Modeling Approaches in Drug Discovery. *Drug Discov. Today* **2019**, *24*, 1157–1165. [CrossRef] [PubMed]
101. Schneider, P.; Tanrikulu, Y.; Schneider, G. Self-Organizing Maps in Drug Discovery: Compound Library Design, Scaffold-Hopping, Repurposing. *Curr. Med. Chem.* **2009**, *16*, 258–266. [CrossRef] [PubMed]
102. Noeske, T.; Sasse, B.C.; Stark, H.; Parsons, C.G.; Weil, T.; Schneider, G. Predicting Compound Selectivity by Self-Organizing Maps: Cross-Activities of Metabotropic Glutamate Receptor Antagonists. *Chem. Med. Chem.* **2006**, *1*, 1066–1068. [CrossRef]
103. Noeske, T.; Jirgensons, A.; Starchenkovs, I.; Renner, S.; Jaunzeme, I.; Trifanova, D.; Hechenberger, M.; Bauer, T.; Kauss, V.; Parsons, C.G.; et al. Virtual Screening for Selective Allosteric MGluR1 Antagonists and Structure-Activity Relationship Investigations for Coumarine Derivatives. *Chem. Med. Chem.* **2007**, *2*, 1763–1773. [CrossRef]
104. Hristozov, D.; Oprea, T.I.; Gasteiger, J. Ligand-Based Virtual Screening by Novelty Detection with Self-Organizing Maps. *J. Chem. Inf. Model.* **2007**, *47*, 2044–2062. [CrossRef]
105. Palos, I.; Lara-Ramirez, E.E.; Lopez-Cedillo, J.C.; Garcia-Perez, C.; Kashif, M.; Bocanegra-Garcia, V.; Nogueda-Torres, B.; Rivera, G. Repositioning FDA Drugs as Potential Cruzain Inhibitors from Trypanosoma Cruzi: Virtual Screening, in Vitro and in Vivo Studies. *Molecules* **2017**, *22*, 1015. [CrossRef] [PubMed]
106. Ardabili, S.; Mosavi, A.; Várkonyi-Kóczy, A.R. Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods. *Lect. Notes Netw. Syst.* **2020**, *101*, 215–227.
107. Helguera, A.M.; Perez-Castillo, Y.D.S.; Cordeiro, M.N.; Tejera, E.; Paz-Y-Miño, C.; Sánchez-Rodríguez, A.; Teijeira, M.; Ancede-Gallardo, E.; Cagide, F.; Borges, F.; et al. Ligand-Based Virtual Screening Using Tailored Ensembles: A Prioritization Tool for Dual A2AAdenosine Receptor Antagonists/Monoamine Oxidase B Inhibitors. *Curr. Pharm. Des.* **2016**, *22*, 3082–3096. [CrossRef] [PubMed]
108. Korkmaz, S.; Zararsiz, G.; Goksuluk, D. MLViS: A Web Tool for Machine Learning-Based Virtual Screening in Early-Phase of Drug Discovery and Development. *PLoS ONE* **2015**, *10*, 1–15. [CrossRef] [PubMed]
109. Melville, J.L.; Burke, E.K.; Hirst, J.D. Machine Learning in Virtual Screening. *Curr. Top. Med. Chem.* **2009**, *12*, 332–343. [CrossRef]
110. de Oliveira, T.A.; Medaglia, L.R.; Maia, E.H.B.; Assis, L.C.; de Carvalho, P.B.; da Silva, A.M.; Taranto, A.G. Evaluation of Docking Machine Learning and Molecular Dynamics Methodologies for DNA-Ligand Systems. *Pharmaceuticals* **2022**, *15*, 132. [CrossRef]
111. Dong, S.; Wang, P.; Abbas, K. A Survey on Deep Learning and Its Applications. *Comput. Sci. Rev.* **2021**, *40*, 100379. [CrossRef]
112. Carpenter, K.A.; Cohen, D.S.; Jarrell, J.T.; Huang, X. Deep Learning and Virtual Drug Screening. *Future Med. Chem.* **2018**, *10*, 2557–2567. [CrossRef]

113. Kimber, T.B.; Chen, Y.; Volkamer, A. Deep Learning in Virtual Screening: Recent Applications and Developments. *Int. J. Mol. Sci.* **2021**, *22*, 4435. [CrossRef]
114. Lin, X.; Li, X.; Lin, X. A Review on Applications of Computational Methods in Drug Screening and Design. *Molecules* **2020**, *25*, 1375. [CrossRef] [PubMed]
115. Jarada, T.N.; Rokne, J.G.; Alhajj, R. A Review of Computational Drug Repositioning: Strategies, Approaches, Opportunities, Challenges, and Directions. *J. Chem.* **2020**, *12*, 1–23. [CrossRef] [PubMed]
116. Bahi, M.; Batouche, M. Deep Learning for Ligand-Based Virtual Screening in Drug Discovery. In Proceedings of the PAIS 2018: International Conference on Pattern Analysis and Intelligent Systems, Tebessa, Algeria, 24–25 October 2018. [CrossRef]
117. Joshi, T.; Joshi, T.; Pundir, H.; Sharma, P.; Mathpal, S.; Chandra, S. Predictive Modeling by Deep Learning, Virtual Screening and Molecular Dynamics Study of Natural Compounds against SARS-CoV-2 Main Protease. *Sci. Rep.* **2020**, *39*, 6728–6746. [CrossRef] [PubMed]
118. Schneider, C.; Buchanan, A.; Taddese, B.; Deane, C.M. DLAB: Deep Learning Methods for Structure-Based Virtual Screening of Antibodies. *Bioinformatics* **2022**, *38*, 377–383. [CrossRef]
119. Blum, C.; Roli, A. Metaheuristics in Combinatorial Optimization. *ACM Comput. Surv.* **2003**, *35*, 268–308. [CrossRef]
120. Eiben, A.E.; Smith, J.E. *Introduction to Evolutionary Computing*; Springer: Berlin/Heidelberg, Germany, 2015. [CrossRef]
121. Glover, F. Tabu Search: A Tutorial. *Interfaces* **1990**, *20*, 74–94. [CrossRef]
122. Baxter, C.A.; Murray, C.W.; Clark, D.E.; Westhead, D.R.; Eldridge, M.D. Flexible Docking Using Tabu Search and an Empirical Estimate of Binding Affinity. *Proteins Struct. Funct. Genet.* **1998**, *33*, 367–382. [CrossRef]
123. Suman, B.; Kumar, P. A Survey of Simulated Annealing as a Tool for Single and Multiobjective Optimization. *J. Oper. Res. Soc.* **2005**, *57*, 1143–1160. [CrossRef]
124. Doucet, N.; Pelletier, J.N. Simulated Annealing Exploration of an Active-Site Tyrosine in TEM-1β-Lactamase Suggests the Existence of Alternate Conformations. *Proteins Struct. Funct. Bioinform.* **2007**, *69*, 340–348. [CrossRef]
125. Jao, C.C.; Hegde, B.G.; Chen, J.; Haworth, I.S.; Langen, R. Structure of Membrane-Bound—Synuclein from Site-Directed Spin Labeling and Computational Refinement. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 19666–19671. [CrossRef] [PubMed]
126. Hatmal, M.M.; Taha, M.O. Simulated Annealing Molecular Dynamics and Ligand–Receptor Contacts Analysis for Pharmacophore Modeling. *Future Med. Chem.* **2017**, *9*, 1141–1159. [CrossRef] [PubMed]
127. Katoch, S.; Chauhan, S.S.; Kumar, V. A Review on Genetic Algorithm: Past, Present, and Future. *Multimed. Tools Appl.* **2021**, *80*, 8091–8126. [CrossRef] [PubMed]
128. Bhaskar, B.V.; Babu, T.M.C.; Reddy, N.V.; Rajendra, W. Homology Modeling, Molecular Dynamics, and Virtual Screening of NorA Efflux Pump Inhibitors of Staphylococcus Aureus. *Drug Des. Dev. Ther.* **2016**, *10*, 3237–3252. [CrossRef] [PubMed]
129. Rohilla, A.; Khare, G.; Tyagi, A.K. Virtual Screening, Pharmacophore Development and Structure Based Similarity Search to Identify Inhibitors against IdeR, a Transcription Factor of Mycobacterium Tuberculosis. *Sci. Rep.* **2017**, *7*, 1–14. [CrossRef]
130. Xia, J.; Feng, B.; Shao, Q.; Yuan, Y.; Wang, X.S.; Chen, N.; Wu, S. Virtual Screening against Phosphoglycerate Kinase 1 in Quest of Novel Apoptosis Inhibitors. *Molecules* **2017**, *22*, 1029. [CrossRef] [PubMed]
131. Bilal; Pant, M.; Zaheer, H.; Garcia-Hernandez, L.; Abraham, A. Differential Evolution: A Review of More than Two Decades of Research. *Eng. Appl. Artif. Intell.* **2020**, *90*, 103479. [CrossRef]
132. Thomsen, R.; Christensen, M.H. MolDock: A New Technique for High Accuracy Molecular Docking. *J. Med. Chem.* **2006**, *49*, 3315–3321. [CrossRef]
133. Friesner, R.A.; Banks, J.L.; Murphy, R.B.; Halgren, T.A.; Klicic, J.J.; Mainz, D.T.; Repasky, M.P.; Knoll, E.H.; Shelley, M.; Perry, J.K.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749. [CrossRef]
134. Verdonk, M.L.; Cole, J.C.; Hartshorn, M.J.; Murray, C.W.; Taylor, R.D. Improved Protein-Ligand Docking Using GOLD. *Proteins Struct. Funct. Genet.* **2003**, *52*, 609–623. [CrossRef]
135. Spitzer, R.; Jain, A.N. Surflex-Dock: Docking Benchmarks and Real-World Application. *J. Comput. Aided Mol. Des.* **2012**, *26*, 687–699. [CrossRef]
136. Hui-fang, L.; Qing, S.; Jian, Z.; Wei, F. Evaluation of Various Inverse Docking Schemes in Multiple Targets Identification. *J. Mol. Graph. Model.* **2010**, *29*, 326–330. [CrossRef] [PubMed]
137. Joung, I.S.; Kim, J.Y.; Gross, S.P.; Joo, K.; Lee, J. Conformational Space Annealing Explained: A General Optimization Algorithm, with Diverse Applications. *Comput. Phys. Commun.* **2018**, *223*, 28–33. [CrossRef]
138. Shin, W.H.; Heo, L.; Lee, J.; Ko, J.; Seok, C.; Lee, J. LigDockCSA: Protein-Ligand Docking Using Conformational Space Annealing. *J. Comput. Chem.* **2011**, *32*, 3226–3232. [CrossRef]
139. Jones, G.; Willett, P.; Glen, R.C.; Leach, A.R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727–748. [CrossRef] [PubMed]
140. Morris, G.M.; Goodsell, D.S.; Halliday, R.S.; Huey, R.; Hart, W.E.; Belew, R.K.; Olson, A.J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639–1662. [CrossRef]
141. Dorigo, M.; Maniezzo, V.; Colorni, A. Ant System: Optimization by a Colony of Cooperating Agents. *IEEE Trans. Syst. Man. Cybern.* **1996**, *26*, 29–41. [CrossRef]
142. Korb, O.; Stutzle, T.; Exner, T.E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. *Theor. Comput. Sci.* **2009**, *49*, 84–96. [CrossRef]

143. Reddy, R.H.; Kim, H.; Cha, S.; Lee, B.; Kim, Y.J. Structure-Based Virtual Screening of Protein Tyrosine Phosphatase Inhibitors: Significance, Challenges, and Solutions. *J. Microbiol. Biotechnol.* **2017**, *27*, 878–895.
144. Pervaiz, S.; Ul-Qayyum, Z.; Bangyal, W.H.; Gao, L.; Ahmad, J. A Systematic Literature Review on Particle Swarm Optimization Techniques for Medical Diseases Detection. *Comput. Math. Methods Med.* **2021**, *2021*, 5990999. [CrossRef]
145. LIU, Y.; LI, W.; MA, R. Particle Swarm Optimization on Flexible Docking. *Int. J. Biomath.* **2012**, *5*, 1250044. [CrossRef]
146. Gowthaman, R.; Lyskov, S.; Karanicolas, J. DARC 2.0: Improved Docking and Virtual Screening at Protein Interaction Sites. *PLoS ONE* **2015**, *10*, 1–24. [CrossRef] [PubMed]
147. Ng, M.C.K.; Fong, S.; Siu, S.W.I. PSOVina: The Hybrid Particle Swarm Optimization Algorithm for Protein-Ligand Docking. *J. Bioinform. Comput. Biol.* **2015**, *13*, 1541007. [CrossRef]
148. Trott, O.; Olson, A.J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2009**, *31*, 455–461. [CrossRef] [PubMed]
149. Grosdidier, A.; Zoete, V.; Michielin, O. SwissDock, a Protein-Small Molecule Docking Web Service Based on EADock DSS. *Nucleic Acids Res.* **2011**, *39*, 270–277. [CrossRef] [PubMed]
150. Tietze, S.; Apostolakis, J. GlamDock: Development and Validation of a New Docking Tool on Several Thousand Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2007**, *47*, 1657–1672. [CrossRef] [PubMed]
151. Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S.B.; Johnson, A.P. EHiTS: A New Fast, Exhaustive Flexible Ligand Docking System. *J. Mol. Graph. Model.* **2007**, *26*, 198–212. [CrossRef]
152. Lalami, M.E.; El-Baz, D.; Boyer, V. Multi GPU Implementation of the Simplex Algorithm. In Proceedings of the 2011 IEEE International Conference on High Performance Computing and Communications, Banff, AB, Canada, 2–4 September 2011; pp. 179–186. [CrossRef]
153. Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A.B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R.E.; Morley, S.D. RDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.* **2014**, *10*, 1–7. [CrossRef]
154. Harrison, R.L. Introduction to Monte Carlo Simulation. *AIP Conf. Proc.* **2010**, *1204*, 17–21.
155. Ocde Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models. 2004. Available online: https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf (accessed on 5 February 2023).
156. Ocde Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology Oecd Series on Testing and Assessment Number 49 the Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(q)sars] on the Principles for the Validation of (q)Sars. 2004. Available online: https://one.oecd.org/document/env/jm/mono(2004)24/en/pdf (accessed on 6 February 2023).
157. Ocde env/jm/mono(2007)2 2 Oecd Environment Health and Safety Publications Series on Testing and Assessment no. 69 Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(q)sar] Models. 2007. Available online: https://one.oecd.org/document/env/jm/mono(2007)2/en/pdf (accessed on 8 February 2023).
158. Czub, N.; Pacławski, A.; Szlęk, J.; Mendyk, A. Do AutoML-Based QSAR Models Fulfill OECD Principles for Regulatory Assessment? A 5-HT1A Receptor Case. *Pharmaceutics* **2022**, *14*, 1415. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.