International Journal of
Molecular Sciences

MDPI

Review

A Structure-Based Drug Discovery Paradigm

Maria Batool, Bilal Ahmad and Sangdun Choi *

Department of Molecular Science and Technology, Ajou University, Suwon 16499, Korea; mariabatool.28@gmail.com (M.B.); bilalpharma77@gmail.com (B.A.)
* Correspondence: sangdunchoi@ajou.ac.kr; Tel.: +82-31-219-2600; Fax: +82-31-219-1615

Received: 10 May 2019; Accepted: 4 June 2019; Published: 6 June 2019

Abstract: Structure-based drug design is becoming an essential tool for faster and more cost-efficient lead discovery relative to the traditional method. Genomic, proteomic, and structural studies have provided hundreds of new targets and opportunities for future drug discovery. This situation poses a major problem: the necessity to handle the "big data" generated by combinatorial chemistry. Artificial intelligence (AI) and deep learning play a pivotal role in the analysis and systemization of larger data sets by statistical machine learning methods. Advanced AI-based sophisticated machine learning tools have a significant impact on the drug discovery process including medicinal chemistry. In this review, we focus on the currently available methods and algorithms for structure-based drug design including virtual screening and de novo drug design, with a special emphasis on AI- and deep-learning-based methods used for drug discovery.

Keywords: deep learning; artificial intelligence; neural network; structure-based drug discovery; virtual screening; scoring function

1. Introduction

In the drug discovery process, the development of novel drugs with potential interactions with therapeutic targets is of central importance. Conventionally, promising-lead identification is achieved by experimental high-throughput screening (HTS), but it is time consuming and expensive [1]. Completion of a typical drug discovery cycle from target identification to an FDA-approved drug takes up to 14 years [2] with the approximate cost of 800 million dollars [3]. Nonetheless, recently, a decrease in the number of new drugs on the market was noted due to failure in different phases of clinical trials [4]. In November 2018, a study was conducted to estimate the total cost of pivotal trials for the development of novel FDA-approved drugs. The median cost of efficacy trials for 59 new drugs approved by the FDA in the 2015–2016 period was $19 million [5]. Thus, it is important to overcome limitations of the conventional drug discovery methods with efficient, low-cost, and broad-spectrum computational alternatives.

In contrast to the traditional drug discovery method (classical or forward pharmacology), rational drug design is efficient and economical. The rational drug design method is also known as reverse pharmacology because the first step is to identify promising target proteins, which are then used for screening of small-molecule libraries [6]. Striking progresses have been made in structural and molecular biology along with advances in biomolecular spectroscopic structure determination methods. These methods have provided three-dimensional (3D) structures of more than 100,000 proteins [7]. In conjunction with the storage of (and organizing) such data, there has been much hype about the development of sophisticated and robust computational techniques. Completion of the Human Genome Project and advances in bioinformatics increased the pace of drug development because of the availability of a huge number of target proteins. The availability of 3D structures of therapeutically important proteins favors identification of binding cavities and has laid the foundation for structure-based drug design (SBDD). This is becoming a fundamental part of industrial drug

discovery projects and of academic researches [8]. SBDD is a more specific, efficient, and rapid process for lead discovery and optimization (Figure 1) because it deals with the 3D structure of a target protein and knowledge about the disease at the molecular level [9]. Among the relevant computational techniques, structure-based virtual screening (SBVS), molecular docking, and molecular dynamics (MD) simulations are the most common methods used in SBDD. These methods have numerous applications in the analysis of binding energetics, ligand–protein interactions, and evaluation of the conformational changes occurring during the docking process [10]. In recent years, developments in the software industry have been driven by a massive surge in software packages for efficient drug discovery processes. Nonetheless, it is important to choose outstanding packages for an efficient SBDD process [11]. Briefly, automation of all the steps in an SBDD process has shortened the SBDD timeline [8]. Moreover, the availability of supercomputers, computer clusters, and cloud computing has sped up lead identification and evaluation. In this review, we offer an overview of the SBDD process and the methods being used in the present era. Moreover, we provide an in-depth discussion about the machine learning (ML) methods intended to speed up this process and big-data handling.

2. An Overview of SBDD Process

In the entire drug discovery paradigm, SBDD is the most powerful and efficient process. Computational resources serve as an efficient technology for accelerating the drug discovery process, which includes various screening procedures, combinatorial chemistry, and calculations of such properties as absorption, distribution, metabolism, excretion and toxicity (ADMET) [12]. SBDD is an iterative process and it proceeds through multiple cycles leading an optimized drug candidate to clinical trials. Generally, a drug discovery process consists of four steps: the discovery phase, development phase, clinical trial phase, and registry phase. In the first phase, a potential therapeutic target and active ligands are identified. The fundamental step involves cloning of the target gene followed by the extraction, purification, and 3D structure determination of the protein. Many computer algorithms can be used to dock the huge databases of small molecules or fragments of compounds into the binding cavity of the target protein. These molecules are ranked according to a scoring system based on electrostatic and steric interactions with the binding site. Thorough investigation of electrostatic properties of the binding site, including the presence of cavities, clefts, and allosteric pockets can be carried out using a 3D structure of the target molecule. Current SBDD methods consider the key features of the binding cavity of the therapeutic target to design efficient ligands [13,14]. In the second phase, the top hits are synthesized and optimized [15]. Furthermore, the top-ranked compounds with high affinity for selective modulation of the target protein are tested in vitro in biochemical assays. These ligands interfere with crucial cellular pathways, thereby leading to the development of drugs with a desired therapeutic and pharmacological effect [16]. Biological properties like efficacy, affinity, and potency of the selected compounds are evaluated by experimental methods [17]. The next step is to determine the 3D structure of the target protein in complex with the promising ligand obtained in the first phase. The 3D structure provides detailed information about the intermolecular features that aid in the process of molecular recognition and binding of the ligand. Structural insights into the ligand–protein complex help with the analysis of various binding conformations, identification of unknown binding pockets, and ligand–protein interactions; elucidation of conformational changes resulting from ligand binding; and detailed mechanistic studies [7]. Subsequently, multiple iterations increase the efficacy and specificity of the lead. The third phase includes clinical trials of the lead compounds. Those compounds that pass the clinical trials proceed to the fourth phase in which the drug is distributed in the market for clinical use.

SBDD is a computational technique widely used by pharmaceutical companies and scientists. There are numerous drugs available on the market that have been identified by SBDD. Human immunodeficiency virus (HIV)-1-inhibiting FDA-approved drugs represent the foremost success story of SBDD [18]. Moreover, other drugs identified by the SBDD technique include a thymidylate synthase inhibitor, raltitrexed [8]; amprenavir, a potential inhibitor of HIV protease discovered by protein modeling and MD simulation [18,19]; and the antibiotic norfloxacin [20]. Other examples of success

cases of drug discovery via SBDD methods are listed in Table 1, whereas the interactions of these drugs with respective targets are shown in Figure 2. Some of the failure cases have also been documented; for example, RPX00023 has been reported as an antidepressant that was claimed to have an agonistic activity toward receptor 5-HT1A, but it inhibited the receptor [21]. These failure cases are the reason for limitations in SBDD strategies. Although SBDD workflow includes various efficient methods, they all have certain restrictions, which require further research work.

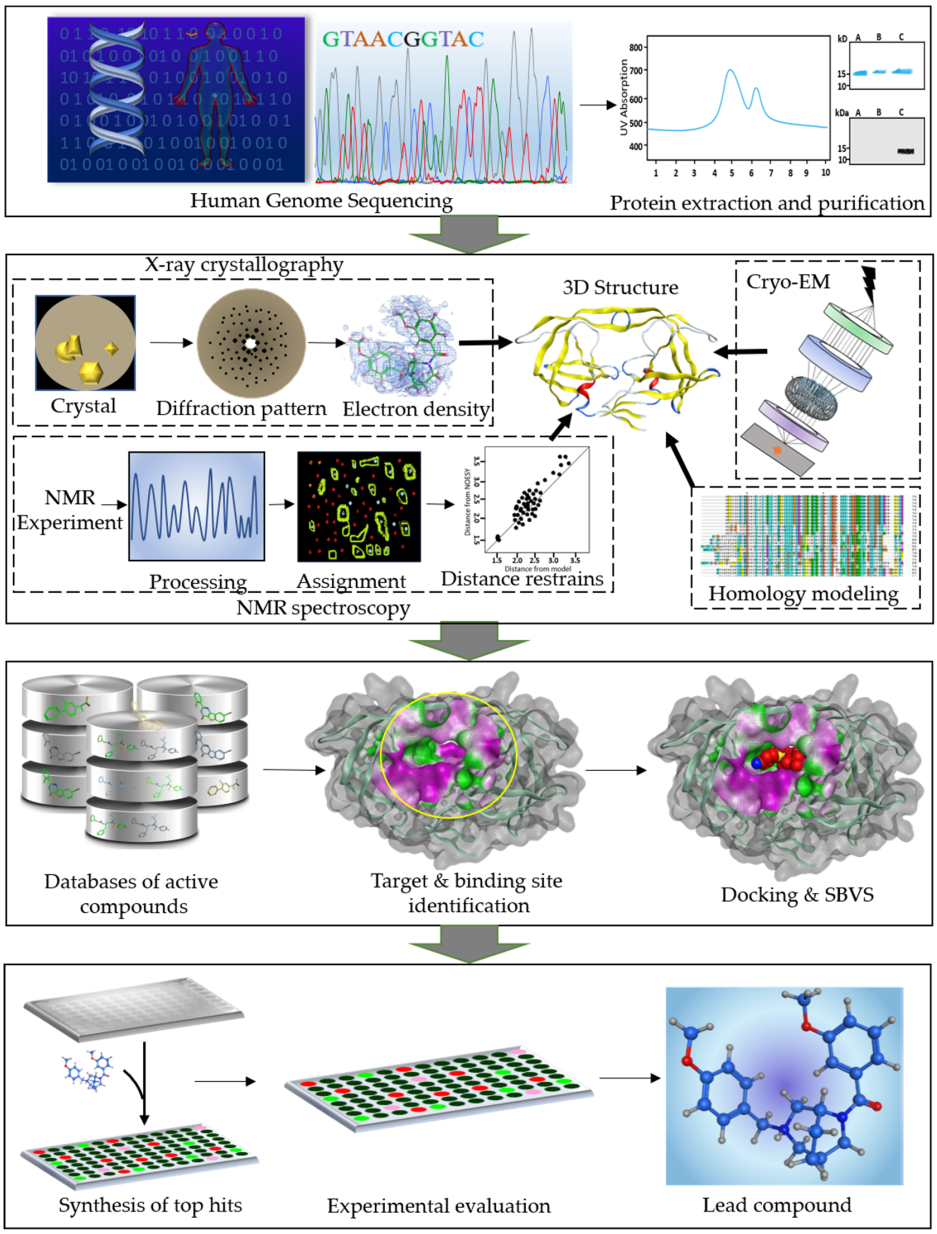

Figure 1. A workflow diagram of structure-based drug design (SBDD) process. The first panel shows the human genome sequencing followed by extraction and purification of the target proteins. Second panel represents the structure determination of the therapeutically important proteins using integrative structural biology approaches. Third panel represents the database preparation of the active compounds. The next step is identification of the druggable target protein and its binding site. Subsequently, the databases of active compounds are screened and docked into the binding cavity of the target protein. In the last panel, the identification of the potent lead compound is shown. The top hit compounds obtained as a result of virtual screening and docking are synthesized and tested in vitro. Further modifications can be done for optimization of the lead compound.

Table 1. The success cases of drug discovery by SBDD methods.

Drug	Drug Target	Target Disease	Technique	Ref.
Raltitrexed	Thymidylate synthase	Human immunodeficiency virus (HIV)	SBDD	[8]
Amprenavir	Antiretroviral protease	HIV	Protein modeling and molecular dynamics (MD)	[18, 19]
Isoniazid	InhA	Tuberculosis	Structure-based virtual screening (SBVS) and pharmacophore modeling	[22]
Pim-1 Kinase Inhibitors	Pim-1 Kinase	Cancer	Hierarchical multistage virtual screening (VS)	[23]
Epalrestat [2]	Aldose Reductase	Diabetic neuropathy	MD and SBVS	[24]
Flurbiprofen	Cyclooxygenase-2	Rheumatoid arthritis, Osteoarthritis	Molecular docking	[25, 26]
STX-0119	STAT3 [1]	Lymphoma	SBVS	[27]
Norfloxacin	Topoisomerase II, IV	Urinary tract infection	SBVS	
Dorzolamide	Carbonic anhydrase	Glaucoma, cystoid macular edema	Fragment-based screening	[28]

[1] Signal transducers and transcription activators (STATs). [2] Currently being sold in Japan under the brand name Kinedak®.

2.1. Target Protein and Binding Site Identification

The basic step in a typical SBDD process is target protein identification and validation [29]. The 3D structures of all therapeutically important proteins are determined experimentally by integrative structure biology techniques such as: NMR, X-ray crystallography, or cryo-electron microscopy but if a solution structure is not available, in silico methods are used to model the protein's 3D structure. There are three well-known structure prediction methods such as comparative modeling, threading, and ab initio modeling. Among them, homology modeling is one of the best and reliable approaches because it predicts the 3D structure of a target protein on the basis of the knowledge about the structure of homologous proteins with >40% similarity [2]. Once the 3D structure of the target is predicted, it is necessary to validate the model by checking the stereochemical properties in a Ramachandran plot. It shows the possible conformations of ψ and φ angles for all amino acid residues present in the protein structure [30]. There are many other methods for validation of the model [2,31,32].

After the structure of the target protein is resolved, the next step is to identify the binding pocket. This is a small cavity where ligands bind to the target to produce the desired effect. Therefore, it is necessary to identify the appropriate site on the target protein. In spite of the protein's dynamic nature, there are a few methods capable of spotting the potential binding residues. These methods consider the knowledge about interaction energy and van der Waals (vdW) forces for binding site mapping. Many methods have been developed for binding site mapping by interaction energy calculations specifically for SBDD. This method identifies particular sites on the target protein which interact favorably with important functional groups on drug-like molecules [33]. These methods identify energetically favorable interactions of specific probes with the proteins. Q-SiteFinder [33] is an energy-based method commonly used for binding site prediction. This method calculates vdW interaction energies of proteins with a methyl probe. Those with favorable energies are retained and clustered. These probe clusters are ranked based on their total interaction energies. In addition, interacting protein residues are functionally annotated to determine the binding site. The next step is

hit discovery, which is done by docking of compound libraries into the binding cavity of the target protein. In the initial phases of lead discovery, it is important to choose a specific set of ligands that play a key part in the lead identification and optimization [34]. For hit hunting, SBDD integrates two divergent methods (i.e., virtual screening (VS) and de novo design).

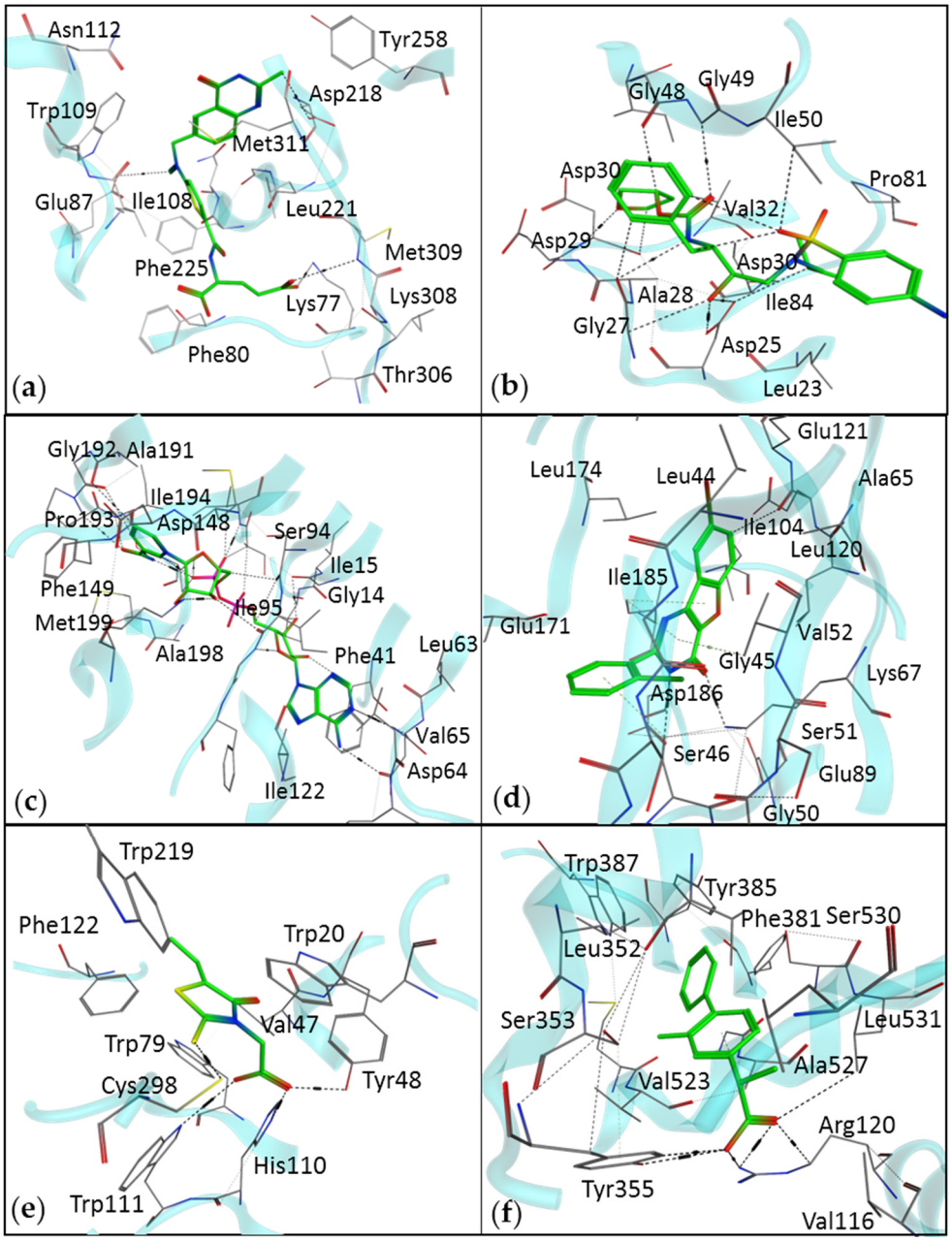

Figure 2. The interaction diagram of drugs identified by SBDD methods, with their respective therapeutic targets. (**a**) An interaction of raltitrexed with thymidylate synthase (Protein Data Bank (PDB) ID: 5X5Q). (**b**) An interaction of amprenavir with HIV protease (PDB ID: 3EKV). (**c**) Isoniazid, a drug for tuberculosis, identified by the SBVS method (PDB ID: 1ENY). (**d**) Pim-1 kinase inhibitor, benzofuropyrimidine, for the treatment of various types of cancers (PDB ID: 4ALU). (**e**) Epalrestat is an aldose reductase inhibitor (PDB ID: 4JIR). (**f**) Flurbiprofen is a cyclooxygenase 2 inhibitor (PDB ID: 3PGH).

2.2. *Virtual Screening: A Lead Identification Approach*

In medicinal chemistry, VS is a robust approach to lead identification [3]. In VS, databases of millions of drug-like or lead-like compounds are screened computationally against the target proteins

with well-known 3D structures. The screening of compound libraries is accomplished by docking, where ligands are filtered based on their binding affinity [35,36]. The top hits of the computational screening are then tested in vitro [3,37]. VS is classified into two major types: ligand-based VS (LBVS) and SBVS. In LBVS, biological data are analyzed to separate inactive compounds from the active compounds. This information is then employed to identify highly active scaffolds on the basis of consensus pharmacophores [38], similarity, or various descriptors. In SBVS, the knowledge about the 3D structure of the target protein is necessary. The target protein is docked with the huge libraries of drug-like compounds, available commercially, via computer algorithms. A scoring function is executed to evaluate the binding force of the docked complex followed by experimental assays to validate the binding. The scoring of ligands is a critical step in SBVS. Unlike ligand-based methods, structure-based approaches do not rely on already available experimental data.

2.3. De Novo Drug Design

De novo drug design is a method of building novel chemical compounds starting from molecular units. The gist of this approach is to develop chemical structures of the small molecules that bind to the target binding cavity with good affinity [39]. Generally, a stochastic approach is used for de novo design, and it is important to take the search space knowledge into consideration in the design algorithm. The two designs, positive and negative, are being used. In the former design, a search is restricted to the specific regions of chemical space with higher probability of finding hits having required features. In contrast, the search criteria are predefined in the negative mode, to prevent the selection of false positives [40]. The chemical compound designing by computational techniques can be related to imitation of synthetic chemistry, while scoring functions perform binding assays [41]. Critical assessment of candidates is crucial for the design process, and the scoring function is one of the assessment tools. Multiple scoring functions can be employed parallelly for multi-objective drug design [42], which considers multiple features at once.

Two methods—(i) ligand-based and (ii) receptor-based de novo drug design—can be used. The latter approach is more prevalent. The quality of target protein structures and accurate knowledge about its binding site are important for receptor-based design because suitable small molecules are designed by fitting the fragments into the binding cavities of the receptors. This could be either done by means of a computational program or by cocrystallization of the ligand with the receptor [43]. There are two techniques for receptor-based design: building blocks, either atoms or fragments such as single rings, amines, and hydrocarbons are linked together to form a complete chemical compound or simply by growing a ligand from a single unit. In the fragment-linking method, the binding site is identified to map the probable interacting points for different functional groups present in the fragments [44]. These functional groups are attached together to build an absolute compound. In the fragment-growing technique, the growth of fragments is accomplished within the binding site monitored by suitable search algorithms [45]. These search algorithms involve scoring functions to assess the probability of growth. Fragment-based de novo design uses the whole chemical space to generate novel compounds. In case of the linking approach, the selection of linkers is critical. Fragment anchoring in the binding site can be performed by (i) the outside-in approach and (ii) the inside-out approach. In the former approach, the building blocks are primarily arranged at the periphery of the binding site, and it grows inward. In the course of the inside-out approach, building blocks are casually fitted into the binding site and built outward [10].

2.4. Molecular Docking

Docking is a technique of virtual simulation of molecular interactions [46]. Molecular docking predicts the conformation and binding of ligands within a target active site with high accuracy; therefore, it is the most popular technique in SBDD [47,48]. This method can be applied to study important molecular phenomena such as a ligand-binding pose and intermolecular interactions for stability of a complex [49]. Moreover, docking algorithms predict binding energies and rank the ligands

by means of various scoring functions [49,50]. The appropriate ligand-binding conformation depends on two factors: (i) large conformational space defining possible binding poses and (ii) explicit prediction of binding energy correlating with each conformation [51]. Multiple iterations are performed, until the minimum energy state is attained, in which ligand-binding is assessed by various scoring functions [7].

There are two types of molecular docking: flexible-ligand search docking and flexible-protein docking. In flexible-ligand search docking, three types of algorithms are designed to deal with the ligand flexibility. These algorithms are the stochastic method, systematic method, and simulation method [52]. The systematic algorithms are aimed at analyzing degrees of freedom. This task can be accomplished by the fragmentation method, one of the frequently used techniques. In this method, a ligand grows gradually inward in a binding cavity [52,53]. In the conformational search technique, rotatable bonds of the molecule are rotated 360° systematically at a fixed-increment rate, or in the database approach, pregenerated libraries of conformational ensembles are utilized for ligand flexibility. In the stochastic algorithms, random modifications are applied to a single ligand or a group. These modifications are accepted or rejected depending upon probability functions such as genetic algorithm methods [52,54] and the Monte Carlo (MC) method. Lastly, MD simulation is a comprehensive technique for studying the dynamic behavior of macromolecules. Energy minimization is implemented as integration with simulations to achieve local minima. The algorithms available for energy minimization are the Newton–Raphson method, steepest descent, least squares methods, and a conjugate gradient [52]. Many biological systems show movements upon ligand binding; thus, in the flexible-protein docking method, the receptor remains flexible during the docking procedure to mimic the natural biological environment. In addition to the full protein movement, in a few cases, small motions are also noticed such as side chain rearrangement or movement of highly flexible loops. MD and MC methods are suitable for flexible-protein docking [55,56].

2.5. Scoring Functions

A scoring function helps a docking program to delve into the ligand-binding site. Once a significant binding conformation is identified, the scoring function calculates binding affinity. Accordingly, scoring functions are thought to have a substantial impact on docking. Scoring functions are trained on a training dataset of a similar class of compounds for which their experimental binding affinity is available. Scoring functions are divided into four general classes: force field, empirical, knowledge-based, and machine learning (ML) [57–59]. The force field is calculated by estimating the intermolecular interactions such as electrostatic and vdW forces between the binding partners. Empirical scoring functions are calculated based on the atom numbers in the ligand and target protein and are used for affinity and pose prediction [60]. The latter includes hydrophobic forces, hydrophilic forces, hydrogen bonding, and entropy. A statistical method called multiple linear regression is employed to fit scoring-function coefficients. A knowledge-based scoring function depends on statistical potentials of intermolecular interactions. This method is based purely on the assumption that frequently occurring functional groups or a certain type of atoms are energetically favorable and contribute to binding affinity [61]. In contrast to classical scoring functions, ML methods do not constrain analysis to a predefined functional form among structural features and binding affinity values [62]. ML methods are dynamic techniques for construction and optimization of models to predict a binding pose and affinity. Lately, the development of novel scoring functions by ML is becoming popular [63]. These methods implicitly take into consideration the interactions between a ligand and target while ignoring error-prone interactions. Furthermore, different methods of the ML technique such as random forest (RF), support vector machine (SVM), and neural networks (NN) work with nonlinear dependence among binding interactions. Thus, ML-based scoring functions perform better than others do in case of binding energy calculations [1]. Another scoring function known as consensus scoring employs collective scores to minimize the error rate in individual scores and to increase the possibility of true positive selection [52].

The efficiency of various scoring functions has been compared in many studies [64–68], regarding binding affinity prediction, reproducibility of a known binding conformation, and ranking of

a library. All modern scoring functions have different accuracy rates under different conditions. Thus, none of the scoring functions can outperform the others. However, consensus scoring function can perform better than single-scoring approach and is widely used in various bioinformatics applications. Consensus scoring function compensates the limitations of single-scoring functions. It improves the hit rate by combining multiple scoring functions based on a simple cause: the true value tends to be closer to the mean value of replicated experiments [69]. In case of single-scoring functions, a binding pose can be predicted accurately, but in terms of binding energy calculations, there is still a need to improve the performance of current scoring functions. Hence, a lot of efforts have been made to upgrade the abilities of the currently available scoring functions. Prevalent methods include the addition of certain features for calculation of entropic and solvation effects [70], development of a consensus scoring function to overcome the limitations of others [69], and calculations of quantum-energy terms [71]. Targeted scoring functions are known to significantly enhance VS performance and might be a solution to the limitations of other scoring functions [72]. Such scoring functions generate output with higher probability of true hits and a decreased rate of false positives.

3. Big Data in Drug Discovery

The "big data" approach influences our daily life, and drug discovery is not an exception. By current computational techniques, molecular characteristics can be studied in a logical and systematic manner. The data collected from each compound can be subjected to analyses from different perspectives [73]. In the modern era of technology, there has probably been an increase in the size of data generation. According to a recent estimate, the total size of stored data is approximately two zettabytes (10^{21}) with expected doubling every two years [74]. Hence, excavation of massively produced digital information offers a multitude of opportunities to increase productivity. Nevertheless, apart from the volume and production rate of big data, the variety and complexity of big data pose challenges for effective analysis [75]. Furthermore, sometimes generated data contain inconsistencies, such as missing or incomplete information, errors, and duplications, thereby affecting the outcomes of accurate simulation and analytical activities. Therefore, preliminary analysis and curation are required as advanced measures to ensure fairness, accuracy, and experimental efficacy [76]. On the other hand, precollection and curation measures vary among research communities, depending on preceding observations and experimental records. Yet, there is high demand for a simple, unified, and well-established curation protocol that ensures the quality of generated simulation and analytical datasets.

Several studies examined the impact of quality on research activities [77]. Several others recommend conducting a fair evaluation of the quality and impact of a particular work [78]. Hence, the existing standard of research continues to adhere to the "less-is-more" principle. Big data have played a vital role in medicinal and combinatorial chemistry, whereas HTS contributes to the generation of a huge amount of data over a short span of time. Big data dependency will likely increase as the perception of personalized medicine improves. Earlier, big data have been regarded as the beginning of computation-oriented medicinal chemistry (i.e., processing stacks of generated data, resulting in shortening of the time taken to complete a drug development cycle). For instance, a well-known global pandemic spanning more than 40 years, HIV, has infected more than 37 million people, where only 57% are being treated with antiviral agents (World Health Organization (WHO), 2018). In the past few years, many studies have addressed the inhibition of viral reverse transcriptase and/or integrase [79,80]. Although this technique has proven effective enough, it comes with several shortcomings such as viral resistance and poor bioavailability.

In the early 1990s, the roles of chemokines and $CD4^+$ cells were described. Chemokine activity is associated with their G-protein-coupled receptors (GPCRs); in the CCR5 case, it is a "C-C" receptor with 75% homology to CCR2 [81]. With the emergence of CCR5 as an interesting and a druggable novel target to combat HIV, numerous pharmaceutical firms turned to their GPCR inhibitor libraries in search of a putative ligand for this protein. A strong lead, an imidazopyridine (UK107543) was identified by Pfizer, a well-known pharmaceutical company, using HTS [82]. Maraviroc (Selzentry),

an antiretroviral drug, classified as an entry inhibitor was later declared as an approved drug for HIV-1 treatment by the FDA [83]. Such real-world use cases spotlight the significance of big data resources in medicinal chemistry. Therefore, among medicinal chemists, we are seeing a major demand for rational awareness of data-driven processes and for information-handling skills [84].

From this standpoint, the scientific communities started investing in the development of applications, tools, and software to handle massively generated and already stored data. Nevertheless, a major concern limiting the usability of these computational platforms includes security and privacy concerns for the users [85]. Aside from these factors, freely and publicly accessible resources provide a versatile collection, which can be manipulated beyond the pharmaceutical scope [86].

4. Artificial Intelligence and Machine Learning in Drug Discovery

Artificial intelligence (AI) mimics human behavior by simulating human intelligence by computer techniques [87]. ML, a subfield of AI, uses statistical methods for learning with or without being programmed [88]. In the drug development process, AI has shifted the mood from hype to hope [87]. Computational technologies and ML algorithms have revolutionized drug discovery in the pharmaceutical industry. Integration of ML algorithms in an automatic manner–to discover new compounds by analyzing, learning, and explaining pharmaceutical big data–is the application of AI to drug design [89]. Big Pharma is increasing investment in AI; this situation shows the truth behind the use of ML algorithms to identify and screen potential drug candidates. For instance, SYNSIGHT has introduced an AI-based integrated platform in combination with VS and molecular modeling to create huge biological models for drug development [90]. Many leading biopharmaceutical companies are collaborating to integrate AI and ML methods with their drug discovery pipelines. Pfizer has been collaborating with IBM since December 2016 to take advantage of their multicloud platform Watson [91] for immuno-oncology drug discovery [92]. Similarly, Exscientia Ltd., a UK-based world class AI-driven drug design company [93] is collaborating with Sanofi to find a cure for metabolic disorders [94], and Clegene, another leading pharmaceutical company, aims to accelerate drug discovery in the areas of autoimmunity and oncology [95]. Recently, Exscientia announced a success story in collaboration with GlaxoSmithKline (GSK), where they claimed the discovery of a highly potent lead molecule for the treatment of chronic obstructive pulmonary disease by means of AI-based drug discovery workflow [96].

ML success has been repeatedly demonstrated in classification, generative modeling, and reinforcement learning (RL). Different categories of ML are supervised learning, unsupervised learning, and RL. The subcategory of supervised learning, classification, and regression methods predicts the model on the basis of input and output data sources. Supervised ML is applicable to a disease in diagnostic methods, ADMET in a classification method's output, and to drug efficacy in regression methods [97]. SVMs with supervised ML algorithms use binary activity prediction to distinguish between a drug and nondrug [98,99] or between specific and nonspecific compounds [100,101]. SVM classification is performed in LBVS to rank the database compounds by decreasing activity probability. To minimize error in SVM ranking, optimized special ranking functions are used [101]. The clustering method for an unsupervised learning category can discover a disease subtype as outputs, while a feature-finding method can identify a target in a disease [102,103]. Decision-making RL maximizes its performance in de novo drug design via modeling and quantum chemistry. RL is less dependent on dataset learning. With RL, the desired physical and biological properties of newly generated chemical structures can be biased [104]. ML exploits the relationship between a biological activity and chemical structure during drug design. Structure prediction of biological targets (protein structure, binding pocket, transmembrane regions, and phosphorylation and glycosylation sites) and quantitative structure–activity relationship (QSAR) models, pharmacophore models, molecular docking analysis, and ranking/scoring functions in similarity searches–can be implemented and statistically validated by ML techniques [105]. Classifying a pharmacokinetic and toxicological (ADMET) profile, discovery or optimization of biologically active hit compounds, and the constructed model or biological activity of a new ligand can aid with a drug discovery process at several steps by ML techniques [106]. Multiple ML models can be used to drive

multiparameter optimization. The output of ML methods depends on multiple parameters like diversity of the training dataset, an ability to handle imbalanced datasets of active and inactive compounds in the library and defining precise parameters to cover full chemical space including active and inactive molecules [107]. Proficient ML models can be developed to screen huge libraries which generate few false positives and a good number of active compounds in the output. This goal can be attained using versatile training datasets comprising predicted inactive compounds [108,109].

5. The Role of Deep Learning in Drug Design

NN represent a supervised neurology-inspired ML technique that is employed routinely and successfully to address such issues as speech and image recognition. Artificial neural networks (ANNs) are ML algorithms that operate as neurons in the brain: they receive numerous input signals and generate an activation response by calculating a weighted sum of the inputs through a nonlinear activation function and pass the output signal to subsequent connected neurons [110]. The basic structure of an ANN consists of an input layer, hidden layer, and the output layer (Figure 3).

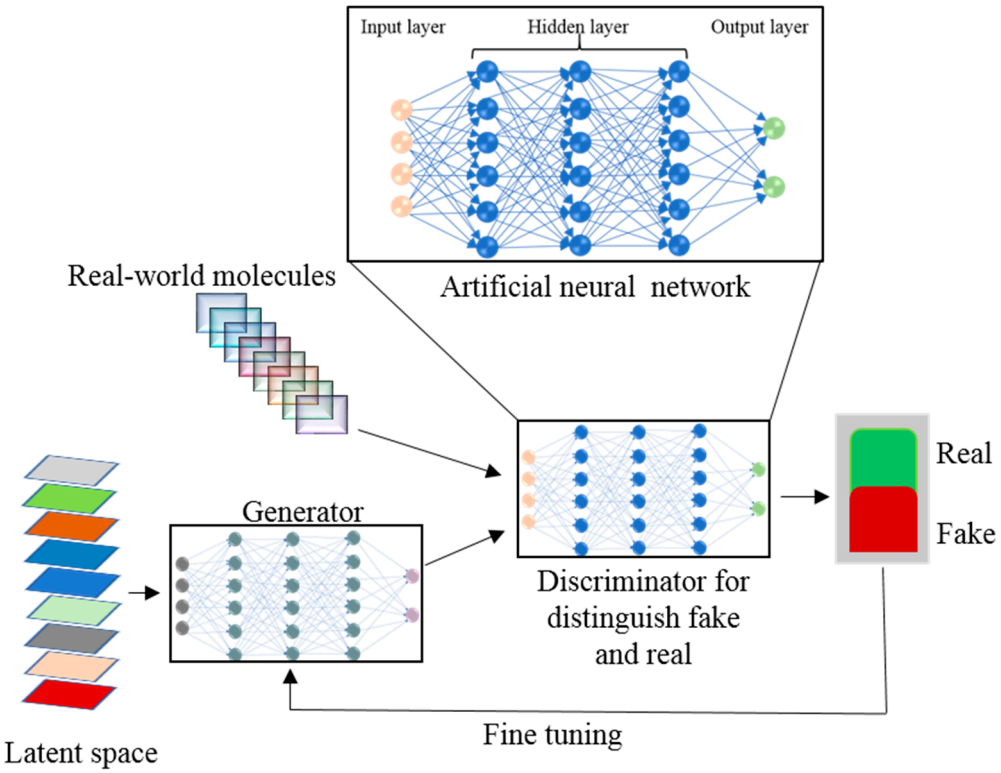

Figure 3. A workflow of the generative adversarial network approach with an artificial neural networks (ANN) for new molecule design.

In the ANN, the processing nodes are either fully or partially connected. From input nodes, the input variables are taken and are transformed through hidden nodes into the output nodes where output values are calculated. By back-propagation methods, the ANN training is done in an iterative fashion to train the network [111]. Due to overfitting, a diminishing gradient, and other problems, the traditional ANN methods have not performed well and have been replaced by other ML algorithms like RF [112] and SVM [113]. The deep learning (DL) concept has originated from ANN's feedforward NNs with many hidden layers [114]. DL's recent development has given the ANN a renaissance. DL is changing our everyday life and has achieved huge success in self-driving cars, computer games, speech recognition, natural language processing, and other applications [115]. With the rapid explosion of chemical "big data" from combinatorial synthesis and HTS, ML techniques have become an indispensable tool for drug designers to retrieve chemical information from large compound

databases to design drugs rationally. Big data volume, velocity, variety, and veracity characterization are not possible via traditional QSAR approaches. ML techniques are more efficient than the physical model for scaling big datasets. DL, being the data-hungry ML algorithm for analyzing and exploring big data, is in high demand. As compared to other ML methods, the DL architecture is flexible [116]. Atomwise, the first DL-based technology for structure-based small-molecule drug discovery has helped to design new potential drugs for 27 disease targets with accuracy and precision [117]. A straightforward method with a fully connected deep neural network (DNN) is used for model building of compounds having the same number of molecular descriptors. To the Merck Kaggle challenge dataset, Dahl et al. [118] applied a DNN and showed better performance as compared to RF on 13 of the total 15 targets. DNNs can handle thousands of descriptors without overfitting and feature selection problems as in the traditional ANN, in an optimized manner, owing to the number of nodes and hidden layers. Mayr et al.'s multitasking DNN method won the Tox21 dataset challenge consisting of 12,000 compounds for 12 high-throughput toxicity assays. In this challenge the computational toxicity prediction of chemicals and drugs was given. The chemical structures and assay measurements from stress and nuclear receptor signaling pathway assays for 12 different toxic effects were available to the participants to check structure-activity relationships. Mayer et al. developed a DeepTox pipeline for toxicity prediction which uses deep learning algorithms. DeepTox normalizes the chemical structures followed by computation of the chemical descriptors. The computed descriptors are used in DL methods to predict the toxicity of chemicals. Later, these models are combined to ensembles [119]. Statistically, a DNN outperforms other ML models such as SVM [120], RF, and others when applied to seven datasets selected from ChEMBL database [121]. In variational autoencoder (VAE), an encoder NN generates a chemical structure via unsupervised learning to map chemical structures from a database onto a latent space. The trained VAE from the latent vector in the latent space transforms the molecular structure into a simplified molecular-input line-entry system (SMILES) string. Kadurin et al. [122] have generated new structures having specific anticancer properties by coupling the generative adversarial network (GAN) with VAE. In a GAN (Figure 3), two ANN models—the generator and discriminator—are trained simultaneously and generate a new molecule from scratch by optimizing a different and opposing objective function in a zero-sum game [123]. A reinforced adversarial neural computer (RANC) with DL architecture, based on the GAN paradigm and RL, generates unique and adequate structures [124]. The RANC uses the SMILES string dataset with key distribution of chemical features like molecular weight, log P, and topological polar surface area for de novo design of small molecules against different biological targets and pathways. Relevant to drug discovery, RANC trained on SMILES string representation outperforms other methods on several metrics [124]. Segler et al. [125] and Yuan et al. [126] have used a recurrent neural network (RNN) for new structure generation acknowledging its success in natural language processing. RNN generates molecular structures by using the probability distribution learning on the SMILES string training set. Target specific libraries were generated by Segler et al. [125] while exploring the RNNs. RNN together with deep Q-learning the RL technology generates SMILES with desirable properties like quantitative estimate of drug-likeness (QED) [127] and clogP [128]. Olivecrona et al. overcame the incorporation of handwritten rules for undesirable structure penalties by tuning the pretrained RNN using the policy based RL approach [129]. Pereira et al. reported deep-learning-based virtual screening method where they compared 95,316 decoys with 2950 ligands docked on 40 receptors and those ranked by the deep convolutional neural network showed better performance than other docking programs [130]. New molecular fingerprints or focused molecule libraries with modeled pharmacokinetic properties of potential drugs can be generated using DL [131].

6. Challenges and Emerging Problems

Drug discovery still faces a lot of challenges, such as (i) upgrading the efficacy of virtual screening methods, (ii) improving computational chemogenomic studies, (iii) boosting the quality and number of computational web sources, (iv) improving the structure of multitarget drugs, (v) enhancing the

algorithms for toxicity prediction, and (vi) collaborating with other related fields of study for better lead identification and optimization.

Computer-aided structure-based drug discovery is an integral part of multidisciplinary work. Computer-aided drug discovery can be used in combination with combinatorial chemistry or HTS, by means of various algorithms to prepare combinatorial libraries for HTS, including chemical space characterization [50]. VS is known to shorten the time and cost of HTS methods. The major drawback of VS is that while generating screening libraries, it ignores the protonation and tautomerism effect as well as ionization states of compounds, thereby missing out on significant hits. Availability of limited experimental data and reliable output of computational methods cause researchers to ignore tautomerization, but they are still irresistible [10,132]. In the drug discovery process, ADMET prediction remains a hurdle. Nonetheless, availability of various computational methods for prediction of these values has reduced the time and the number of tests on animals. Further development of informatics toxicology is needed [133].

In the de novo lead generation method, though this process seems to be efficient and acceptable, there are limitations of the linking procedure. The first limitation is that the linking fragments should be placed accurately in the cavity for appropriate linking. Moreover, de novo design is thought to be fully automated, but still there is some work to be done manually, which is quite laborious. Furthermore, compounds designed by this technique are not always easy to synthesize in the laboratory. Thus, new software is needed that considers the synthesis factors while including de novo designing of compounds [10].

In the case of molecular docking, a variety of docking algorithms and scoring functions are available, but it is important to choose an appropriate scoring function, which requires deep knowledge about such software. The limitations of the scoring functions are a major drawback among docking programs because this software provides an efficient evaluation of ligand binding energy but ignores accuracy [52]. Several molecular determinants such as electrostatic interactions and entropy calculations are entirely ignored during ligand-binding energy calculations [48]. No single software package is suitable for work with all types of proteins and ligands. Similarly, accurate binding affinity calculation is still debated [10]. Despite a lot of improvements and currents developments in SBDD, a consistent solution is yet to be developed. To overcome fundamental issues such as considering water molecules and flexibility of a target molecule, revolutionary innovations are still needed.

Author Contributions: M.B. and S.C. conceived the idea. M.B. and B.A. collected the data and literature. M.B. and B.A. wrote the manuscript. S.C. coordinated the project and wrote the manuscript. All authors have given their approval to the final version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (2019R1H1A2039674) and the Commercialization Promotion Agency for R&D Outcomes funded by the Ministry of Science and ICT (2018K000369).

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

Abbreviations

AI	Artificial intelligence
DL	Deep learning
HTS	High throughput screening
vdW	van der Waals
VS	Virtual screening
SBVS	Structure-based virtual screening
HIV	Human Immunodeficiency Virus
GA	Genetic algorithm
MC	Monte Carlo
SVM	Support vector machine
VAE	Variational autoencoder
RF	Random forest
ANN	Artificial neural network

DNN	Deep neural network
GAN	Generative adversarial network
ADMET	Absorption, distribution, metabolism, excretion and toxicity
GSK	GlaxoSmithKline
RANC	Reinforced adversarial neural computer
RL	Reinforcement learning
MD	Molecular dynamics
GPCRs	G-protein-coupled receptors
STATs	Signal transducers and transcription activators
3D	Three-dimensional
RNN	Recurrent neural network
ML	Machine learning
SBDD	Structure-based drug design
PDB	Protein data bank
NN	Neural Network
QSAR	Quantitative structure–activity relationship
QED	Quantitative estimate of drug-likeness
SMILES	Simplified molecular-input line-entry system

References

1. Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S.H. Structure-based virtual screening for drug discovery: A problem-centric review. *AAPS J.* **2012**, *14*, 133–141. [CrossRef] [PubMed]
2. Song, C.M.; Lim, S.J.; Tong, J.C. Recent advances in computer-aided drug design. *Brief. Bioinform.* **2009**, *10*, 579–591. [CrossRef] [PubMed]
3. Lavecchia, A.; di Giovanni, C. Virtual screening strategies in drug discovery: A critical review. *Curr. Med. Chem.* **2013**, *20*, 2839–2860. [CrossRef] [PubMed]
4. Lavecchia, A.; Cerchia, C. In silico methods to address polypharmacology: Current status, applications and future perspectives. *Drug Discov. Today* **2016**, *21*, 288–298. [CrossRef] [PubMed]
5. Moore, T.J.; Zhang, H.; Anderson, G.; Alexander, G.C. Estimated costs of pivotal trials for novel therapeutic agents approved by the us food and drug administration, 2015–2016. *JAMA Intern. Med.* **2018**, *178*, 1451–1457. [CrossRef] [PubMed]
6. Swinney, D.C.; Anthony, J. How were new medicines discovered? *Nat. Rev. Drug Discov.* **2011**, *10*, 507–519. [CrossRef]
7. 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] [PubMed]
8. Anderson, A.C. The process of structure-based drug design. *Chem. Biol.* **2003**, *10*, 787–797. [CrossRef]
9. Lionta, E.; Spyrou, G.; Vassilatis, D.K.; Cournia, Z. Structure-based virtual screening for drug discovery: Principles, applications and recent advances. *Curr. Top. Med. Chem.* **2014**, *14*, 1923–1938. [CrossRef]
10. Kalyaanamoorthy, S.; Chen, Y.P. Structure-based drug design to augment hit discovery. *Drug Discov. Today* **2011**, *16*, 831–839. [CrossRef]
11. Searls, D.B. Data integration: Challenges for drug discovery. *Nat. Rev. Drug Discov.* **2005**, *4*, 45–58. [CrossRef] [PubMed]
12. Batool, M.; Choi, S. Identification of druggable genome in staphylococcus aureus multidrug resistant strain. In Proceedings of the 2017 IEEE Life Sciences Conference (LSC), Sydney, NSW, Australia, 13–15 December 2017; pp. 270–273.
13. Blaney, J. A very short history of structure-based design: How did we get here and where do we need to go? *J. Comput. Aided Mol. Des.* **2012**, *26*, 13–14. [CrossRef] [PubMed]
14. Mandal, S.; Moudgil, M.; Mandal, S.K. Rational drug design. *Eur. J. Pharm.* **2009**, *625*, 90–100. [CrossRef] [PubMed]
15. Wilson, G.L.; Lill, M.A. Integrating structure-based and ligand-based approaches for computational drug design. *Future Med. Chem.* **2011**, *3*, 735–750. [CrossRef] [PubMed]
16. Urwyler, S. Allosteric modulation of family c g-protein-coupled receptors: From molecular insights to therapeutic perspectives. *Pharm. Rev.* **2011**, *63*, 59–126. [CrossRef] [PubMed]

17. Fang, Y. Ligand-receptor interaction platforms and their applications for drug discovery. *Expert Opin. Drug Discov.* **2012**, *7*, 969–988. [CrossRef] [PubMed]
18. Wlodawer, A.; Vondrasek, J. Inhibitors of HIV-1 protease: A major success of structure-assisted drug design. *Annu Rev. Biophys Biomol. Struct.* **1998**, *27*, 249–284. [CrossRef]
19. Clark, D.E. What has computer-aided molecular design ever done for drug discovery? *Expert Opin. Drug Discov.* **2006**, *1*, 103–110. [CrossRef]
20. Rutenber, E.E.; Stroud, R.M. Binding of the anticancer drug zd1694 to *E. Coli* thymidylate synthase: Assessing specificity and affinity. *Structure* **1996**, *4*, 1317–1324. [CrossRef]
21. De Paulis, T. Drug evaluation: Prx-00023, a selective 5-ht1a receptor agonist for depression. *Curr. Opin. Investig. Drugs* **2007**, *8*, 78–86.
22. Marrakchi, H.; Laneelle, G.; Quemard, A. Inha, a target of the antituberculous drug isoniazid, is involved in a mycobacterial fatty acid elongation system, fas-ii. *Microbiology* **2000**, *146*, 289–296. [CrossRef] [PubMed]
23. Ren, J.X.; Li, L.L.; Zheng, R.L.; Xie, H.Z.; Cao, Z.X.; Feng, S.; Pan, Y.L.; Chen, X.; Wei, Y.Q.; Yang, S.Y. Discovery of novel pim-1 kinase inhibitors by a hierarchical multistage virtual screening approach based on svm model, pharmacophore, and molecular docking. *J. Chem. Inf. Model.* **2011**, *51*, 1364–1375. [CrossRef] [PubMed]
24. Wang, L.; Gu, Q.; Zheng, X.; Ye, J.; Liu, Z.; Li, J.; Hu, X.; Hagler, A.; Xu, J. Discovery of new selective human aldose reductase inhibitors through virtual screening multiple binding pocket conformations. *J. Chem. Inf. Model.* **2013**, *53*, 2409–2422. [CrossRef] [PubMed]
25. Dadashpour, S.; Tuylu Kucukkilinc, T.; Unsal Tan, O.; Ozadali, K.; Irannejad, H.; Emami, S. Design, synthesis and in vitro study of 5,6-diaryl-1,2,4-triazine-3-ylthioacetate derivatives as cox-2 and beta-amyloid aggregation inhibitors. *Arch. Pharm.* **2015**, *348*, 179–187. [CrossRef] [PubMed]
26. Miller, Z.; Kim, K.S.; Lee, D.M.; Kasam, V.; Baek, S.E.; Lee, K.H.; Zhang, Y.Y.; Ao, L.; Carmony, K.; Lee, N.R.; et al. Proteasome inhibitors with pyrazole scaffolds from structure-based virtual screening. *J. Med. Chem.* **2015**, *58*, 2036–2041. [CrossRef] [PubMed]
27. Matsuno, K.; Masuda, Y.; Uehara, Y.; Sato, H.; Muroya, A.; Takahashi, O.; Yokotagawa, T.; Furuya, T.; Okawara, T.; Otsuka, M.; et al. Identification of a new series of stat3 inhibitors by virtual screening. *ACS Med. Chem. Lett.* **2010**, *1*, 371–375. [CrossRef]
28. Grover, S.; Apushkin, M.A.; Fishman, G.A. Topical dorzolamide for the treatment of cystoid macular edema in patients with retinitis pigmentosa. *Am. J. Ophthalmol.* **2006**, *141*, 850–858. [CrossRef]
29. Grant, M.A. Protein structure prediction in structure-based ligand design and virtual screening. *Comb. Chem. High Throughput Screen.* **2009**, *12*, 940–960. [CrossRef]
30. Krieger, E.; Joo, K.; Lee, J.; Lee, J.; Raman, S.; Thompson, J.; Tyka, M.; Baker, D.; Karplus, K. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in casp8. *Proteins* **2009**, *77*, 114–122. [CrossRef]
31. Bordoli, L.; Kiefer, F.; Arnold, K.; Benkert, P.; Battey, J.; Schwede, T. Protein structure homology modeling using swiss-model workspace. *Nat. Protoc.* **2009**, *4*, 1–13. [CrossRef]
32. Potapov, V.; Cohen, M.; Inbar, Y.; Schreiber, G. Protein structure modelling and evaluation based on a 4-distance description of side-chain interactions. *BMC Bioinform.* **2010**, *11*, 374. [CrossRef] [PubMed]
33. Laurie, A.T.; Jackson, R.M. Q-sitefinder: An energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **2005**, *21*, 1908–1916. [CrossRef]
34. Wunberg, T.; Hendrix, M.; Hillisch, A.; Lobell, M.; Meier, H.; Schmeck, C.; Wild, H.; Hinzen, B. Improving the hit-to-lead process: Data-driven assessment of drug-like and lead-like screening hits. *Drug Discov. Today* **2006**, *11*, 175–180. [CrossRef]
35. Shoichet, B.K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865. [CrossRef] [PubMed]
36. Phatak, S.S.; Stephan, C.C.; Cavasotto, C.N. High-throughput and in silico screenings in drug discovery. *Expert. Opin. Drug Discov.* **2009**, *4*, 947–959. [CrossRef]
37. Reddy, A.S.; Pati, S.P.; Kumar, P.P.; Pradeep, H.N.; Sastry, G.N. Virtual screening in drug discovery—A computational perspective. *Curr. Protein Pept. Sci* **2007**, *8*, 329–351. [CrossRef]
38. Pedretti, A.; Mazzolari, A.; Gervasoni, S.; Vistoli, G. Rescoring and linearly combining: A highly effective consensus strategy for virtual screening campaigns. *Int. J. Mol. Sci* **2019**, *20*, 2060. [CrossRef]
39. Hartenfeller, M.; Schneider, G. De novo drug design. *Methods Mol. Biol.* **2011**, *672*, 299–323.
40. Richardson, J.S.; Richardson, D.C. The de novo design of protein structures. *Trends Biochem. Sci* **1989**, *14*, 304–309. [CrossRef]

41. Lameijer, E.W.; Tromp, R.A.; Spanjersberg, R.F.; Brussee, J.; Ijzerman, A.P. Designing active template molecules by combining computational de novo design and human chemist's expertise. *J. Med. Chem* **2007**, *50*, 1925–1932. [CrossRef]
42. Gillet, V.J. New directions in library design and analysis. *Curr. Opin. Chem. Biol.* **2008**, *12*, 372–378. [CrossRef] [PubMed]
43. Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.* **2005**, *4*, 649–663. [CrossRef] [PubMed]
44. Keseru, G.M.; Makara, G.M. Hit discovery and hit-to-lead approaches. *Drug Discov. Today* **2006**, *11*, 741–748. [CrossRef] [PubMed]
45. Tang, Y.; Zhu, W.; Chen, K.; Jiang, H. New technologies in computer-aided drug design: Toward target identification and new chemical entity discovery. *Drug Discov. Today Technol.* **2006**, *3*, 307–313. [CrossRef] [PubMed]
46. Prada-Gracia, D.; Huerta-Yepez, S.; Moreno-Vargas, L.M. Application of computational methods for anticancer drug discovery, design, and optimization. *Bol. Med. Hosp. Infan.t Mex.* **2016**, *73*, 411–423.
47. 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]
48. Kitchen, D.B.; Decornez, H.; Furr, J.R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949. [CrossRef] [PubMed]
49. Huang, S.Y.; Zou, X. Advances and challenges in protein-ligand docking. *Int. J. Mol. Sci* **2010**, *11*, 3016–3034. [CrossRef]
50. Lopez-Vallejo, F.; Caulfield, T.; Martinez-Mayorga, K.; Giulianotti, M.A.; Nefzi, A.; Houghten, R.A.; Medina-Franco, J.L. Integrating virtual screening and combinatorial chemistry for accelerated drug discovery. *Comb. Chem. High. Throughput Screen.* **2011**, *14*, 475–487. [CrossRef]
51. Kapetanovic, I.M. Computer-aided drug discovery and development (caddd): In silico-chemico-biological approach. *Chem. Biol. Interact.* **2008**, *171*, 165–176. [CrossRef]
52. Sousa, S.F.; Fernandes, P.A.; Ramos, M.J. Protein-ligand docking: Current status and future challenges. *Proteins* **2006**, *65*, 15–26. [CrossRef] [PubMed]
53. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489. [CrossRef] [PubMed]
54. Taylor, R.D.; Jewsbury, P.J.; Essex, J.W. A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.* **2002**, *16*, 151–166. [CrossRef] [PubMed]
55. Oshiro, C.M.; Kuntz, I.D.; Dixon, J.S. Flexible ligand docking using a genetic algorithm. *J. Comput. Aided Mol. Des.* **1995**, *9*, 113–130. [CrossRef] [PubMed]
56. Hart, T.N.; Read, R.J. A multiple-start monte carlo docking method. *Proteins* **1992**, *13*, 206–222. [CrossRef] [PubMed]
57. Ain, Q.U.; Aleksandrova, A.; Roessler, F.D.; Ballester, P.J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput Mol. Sci* **2015**, *5*, 405–424. [CrossRef] [PubMed]
58. Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C.R. Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharm.* **2008**, *153*, 7–26. [CrossRef]
59. 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]
60. Guedes, I.A.; Pereira, F.S.S.; Dardenne, L.E. Empirical scoring functions for structure-based virtual screening: Applications, critical aspects, and challenges. *Front. Pharm.* **2018**, *9*, 1089. [CrossRef]
61. Muegge, I. Pmf scoring revisited. *J. Med. Chem.* **2006**, *49*, 5895–5902. [CrossRef]
62. Li, H.; Peng, J.; Leung, Y.; Leung, K.S.; Wong, M.H.; Lu, G.; Ballester, P.J. The impact of protein structure and sequence similarity on the accuracy of machine-learning scoring functions for binding affinity prediction. *Biomolecules* **2018**, *8*, 12. [CrossRef] [PubMed]
63. David, H.; Gary, B.F. Computational intelligence methods for docking scores. *Curr. Comput. Aided Drug Des.* **2009**, *5*, 56–68.
64. Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **2009**, *49*, 1079–1093. [CrossRef] [PubMed]

65. Warren, G.L.; Andrews, C.W.; Capelli, A.M.; Clarke, B.; LaLonde, J.; Lambert, M.H.; Lindvall, M.; Nevins, N.; Semus, S.F.; Senger, S.; et al. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931. [CrossRef] [PubMed]
66. Ferrara, P.; Gohlke, H.; Price, D.J.; Klebe, G.; Brooks, C.L., 3rd. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032–3047. [CrossRef] [PubMed]
67. Wang, R.; Lu, Y.; Fang, X.; Wang, S. An extensive test of 14 scoring functions using the pdbbind refined set of 800 protein-ligand complexes. *J. Chem. Inf. Comput. Sci* **2004**, *44*, 2114–2125. [CrossRef] [PubMed]
68. Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287–2303. [CrossRef] [PubMed]
69. Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci* **2001**, *41*, 1422–1426. [CrossRef]
70. Huang, S.Y.; Zou, X. Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. *J. Chem. Inf. Model.* **2010**, *50*, 262–273. [CrossRef]
71. Raub, S.; Steffen, A.; Kamper, A.; Marian, C.M. Aiscore chemically diverse empirical scoring function employing quantum chemical binding energies of hydrogen-bonded complexes. *J. Chem. Inf. Model.* **2008**, *48*, 1492–1510. [CrossRef]
72. Seifert, M.H. Targeted scoring functions for virtual screening. *Drug Discov. Today* **2009**, *14*, 562–569. [CrossRef] [PubMed]
73. Prieto-Martínez, F.D.; López-López, E.; Eurídice Juárez-Mercado, K.; Medina-Franco, J.L. Chapter 2—computational drug design methods—current and future perspectives. In *In silico drug design*; Roy, K., Ed.; Academic Press: Cambridge, MA, USA, 2019; pp. 19–44.
74. Akoka, J.; Comyn-Wattiau, I.; Laoufi, N. Research on big data—A systematic mapping study. *Comput. Stand. Interfaces* **2017**, *54*, 105–115. [CrossRef]
75. Secchi, P. On the role of statistics in the era of big data: A call for a debate. *Stat. Probab. Lett.* **2018**, *136*, 10–14. [CrossRef]
76. Cox, D.R.; Kartsonaki, C.; Keogh, R.H. Big data: Some statistical issues. *Stat. Probab. Lett.* **2018**, *136*, 111–115. [CrossRef] [PubMed]
77. Bornmann, L. Measuring the societal impact of research. *EMBO Rep.* **2012**, *13*, 673. [CrossRef] [PubMed]
78. Mårtensson, P.; Fors, U.; Wallin, S.-B.; Zander, U.; Nilsson, G.H. Evaluating research: A multidisciplinary approach to assessing research practice and quality. *Res. Policy* **2016**, *45*, 593–603. [CrossRef]
79. Cabrera, M.T.; Brewer, E.M.; Grant, L.; Tarczy-Hornoch, K. Exudative retinal detachment documented by handheld spectral domain optical coherence tomography after retinal laser photocoagulation for retinopathy of prematurity. *Retin. Cases Brief. Rep.* **2018**. [CrossRef] [PubMed]
80. Ghosh, A.K.; Osswald, H.L.; Prato, G. Recent progress in the development of HIV-1 protease inhibitors for the treatment of hiv/aids. *J. Med. Chem.* **2016**, *59*, 5172–5208. [CrossRef]
81. Barmania, F.; Pepper, M.S. C-c chemokine receptor type five (ccr5): An emerging target for the control of hiv infection. *Appl. Transl. Genom* **2013**, *2*, 3–16. [CrossRef]
82. MacArthur, R.D.; Novak, R.M. Reviews of anti-infective agents: Maraviroc: The first of a new class of antiretroviral agents. *Clin. Infect. Dis.* **2008**, *47*, 236–241. [CrossRef]
83. Kuritzkes, D.; Kar, S.; Kirkpatrick, P. Maraviroc. *Nat. Rev. Drug Discov.* **2008**, *7*, 15. [CrossRef]
84. Lusher, S.J.; McGuire, R.; van Schaik, R.C.; Nicholson, C.D.; de Vlieg, J. Data-driven medicinal chemistry in the era of big data. *Drug Discov. Today* **2014**, *19*, 859–868. [CrossRef] [PubMed]
85. Ebejer, J.P.; Fulle, S.; Morris, G.M.; Finn, P.W. The emerging role of cloud computing in molecular modelling. *J. Mol. Graph. Model.* **2013**, *44*, 177–187. [CrossRef] [PubMed]
86. Kissin, I. What can big data on academic interest reveal about a drug? Reflections in three major us databases. *Trends Pharm. Sci* **2018**, *39*, 248–257. [CrossRef] [PubMed]
87. Mak, K.K.; Pichika, M.R. Artificial intelligence in drug development: Present status and future prospects. *Drug Discov. Today* **2019**, *24*, 773–780. [CrossRef] [PubMed]
88. Bishop, C.M. Model-based machine learning. Philos Trans. *A Math. Phys. Eng. Sci* **2013**, *371*, 20120222. [CrossRef] [PubMed]
89. Duch, W.; Swaminathan, K.; Meller, J. Artificial intelligence approaches for rational drug design and discovery. *Curr. Pharm. Des.* **2007**, *13*, 1497–1508. [CrossRef] [PubMed]

90. Probst, C.; Schneider, S.; Loskill, P. High-throughput Organ-on-a-chip systems: Current status and remaining challenges. *Curr. Opin. Biomed. Eng.* **2018**, *6*, 33–41. [CrossRef]
91. IBM. Ibm Watson. Available online: https://www.ibm.com/watson (accessed on 1 May 2019).
92. Smalley, E. Ai-powered drug discovery captures pharma interest. *Nat. Biotechnol.* **2017**, *35*, 604–605. [CrossRef]
93. Exscientia. At the forefront of small molecule drug discovery. Available online: https://www.exscientia.co.uk/ (accessed on 1 May 2019).
94. Fleming, N. How artificial intelligence is changing drug discovery. *Nature* **2018**, *557*, 55–57. [CrossRef]
95. Exscientia. Celgene and exscientia enter 3-year ai drug discovery collaboration focused on accelerating drug discovery in oncology and autoimmunity. Available online: https://www.exscientia.co.uk/news (accessed on 1 May 2019).
96. Exscientia. Exscientia achieves molecule discovery milestone as part of gsk collaboration. Available online: https://www.exscientia.co.uk/news (accessed on 1 May 2019).
97. Guncar, G.; Kukar, M.; Notar, M.; Brvar, M.; Cernelc, P.; Notar, M.; Notar, M. An application of machine learning to haematological diagnosis. *Sci Rep.* **2018**, *8*, 411. [CrossRef] [PubMed]
98. Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci* **2003**, *43*, 1882–1889. [CrossRef] [PubMed]
99. Zernov, V.V.; Balakin, K.V.; Ivaschenko, A.A.; Savchuk, N.P.; Pletnev, I.V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci* **2003**, *43*, 2048–2056. [CrossRef] [PubMed]
100. Warmuth, M.K.; Liao, J.; Ratsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci* **2003**, *43*, 667–673. [CrossRef] [PubMed]
101. Jorissen, R.N.; Gilson, M.K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561. [CrossRef] [PubMed]
102. Koohy, H. The rise and fall of machine learning methods in biomedical research. *F1000Res* **2017**, *6*, 2012. [CrossRef] [PubMed]
103. Young, J.D.; Cai, C.; Lu, X. Unsupervised deep learning reveals prognostically relevant subtypes of glioblastoma. *BMC Bioinform.* **2017**, *18*, 381. [CrossRef] [PubMed]
104. Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci Adv.* **2018**, *4*, eaap7885. [CrossRef]
105. Lo, Y.C.; Rensi, S.E.; Torng, W.; Altman, R.B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **2018**, *23*, 1538–1546. [CrossRef]
106. Lima, A.N.; Philot, E.A.; Trossini, G.H.; Scott, L.P.; Maltarollo, V.G.; Honorio, K.M. Use of machine learning approaches for novel drug discovery. *Expert Opin. Drug Discov.* **2016**, *11*, 225–239. [CrossRef]
107. Ma, X.H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z.R.; Chen, Y.Z. Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. *COMB Chem. High Throughput Screen.* **2009**, *12*, 344–357. [CrossRef] [PubMed]
108. Han, L.Y.; Ma, X.H.; Lin, H.H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z.R.; Cao, Z.W.; Ji, Z.L.; Chen, Y.Z. A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor. *J. Mol. Graph. Model.* **2008**, *26*, 1276–1286. [CrossRef] [PubMed]
109. Liu, X.H.; Ma, X.H.; Tan, C.Y.; Jiang, Y.Y.; Go, M.L.; Low, B.C.; Chen, Y.Z. Virtual screening of abl inhibitors from large compound libraries by support vector machines. *J. Chem. Inf. Model.* **2009**, *49*, 2101–2110. [CrossRef] [PubMed]
110. Van Gerven, M.; Bohte, S. *Artificial Neural Networks as Models of Neural Information Processing*; Frontiers Media SA: Lausanne, Switzerland, 2018.
111. Dreyfus, S. The computational solution of optimal control problems with time lag. *IEEE Trans. Autom. Control.* **1973**, *18*, 383–385. [CrossRef]
112. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
113. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
114. Mamoshina, P.; Vieira, A.; Putin, E.; Zhavoronkov, A. Applications of deep learning in biomedicine. *Mol. Pharm.* **2016**, *13*, 1445–1454. [CrossRef] [PubMed]

115. Howard, J. The business impact of deep learning. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; ACM: New York, NY, USA, 2013; p. 1135.
116. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
117. Atomwise. Artificial intelligence for drug discovery. Available online: https://www.atomwise.com/ (accessed on 24 April 2019).
118. Ma, J.; Sheridan, R.P.; Liaw, A.; Dahl, G.E.; Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263–274. [CrossRef]
119. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. Deeptox: Toxicity prediction using deep learning. *Front. Env. Sci.* **2016**, *3*, 80. [CrossRef]
120. Suthaharan, S. Support vector machine. In *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*; Springer US: Boston, MA, USA, 2016; pp. 207–235.
121. Bento, A.P.; Gaulton, A.; Hersey, A.; Bellis, L.J.; Chambers, J.; Davies, M.; Kruger, F.A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The chembl bioactivity database: An update. *Nucleic Acids Res.* **2014**, *42*, 1083–1090. [CrossRef] [PubMed]
122. Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. Drugan: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol. Pharm.* **2017**, *14*, 3098–3104. [CrossRef] [PubMed]
123. Butler, K.T.; Davies, D.W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555. [CrossRef] [PubMed]
124. Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced adversarial neural computer for de novo molecular design. *J. Chem. Inf. Model.* **2018**, *58*, 1194–1204. [CrossRef] [PubMed]
125. Segler, M.H.S.; Kogej, T.; Tyrchan, C.; Waller, M.P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci* **2018**, *4*, 120–131. [CrossRef] [PubMed]
126. Yuan, W.; Jiang, D.; Nambiar, D.K.; Liew, L.P.; Hay, M.P.; Bloomstein, J.; Lu, P.; Turner, B.; Le, Q.T.; Tibshirani, R.; et al. Chemical space mimicry for drug discovery. *J. Chem. Inf. Model.* **2017**, *57*, 875–882. [CrossRef] [PubMed]
127. Bickerton, G.R.; Paolini, G.V.; Besnard, J.; Muresan, S.; Hopkins, A.L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90–98. [CrossRef]
128. Nieto-Draghi, C.; Fayet, G.; Creton, B.; Rozanska, X.; Rotureau, P.; de Hemptinne, J.C.; Ungerer, P.; Rousseau, B.; Adamo, C. A general guidebook for the theoretical prediction of physicochemical properties of chemicals for regulatory purposes. *Chem. Rev.* **2015**, *115*, 13093–13164. [CrossRef]
129. Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017**, *9*, 48. [CrossRef]
130. Pereira, J.C.; Caffarena, E.R.; dos Santos, C.N. Boosting docking-based virtual screening with deep learning. *J. Chem. Inf. Model.* **2016**, *56*, 2495–2506. [CrossRef]
131. Hughes, T.B.; Miller, G.P.; Swamidass, S.J. Modeling epoxidation of drug-like molecules with a deep machine learning network. *ACS Cent. Sci* **2015**, *1*, 168–180. [CrossRef] [PubMed]
132. Martin, Y.C. Let's not forget tautomers. *J. Comput. Aided Mol. Des.* **2009**, *23*, 693–704. [CrossRef] [PubMed]
133. Mangiatordi, G.F.; Alberga, D.; Altomare, C.D.; Carotti, A.; Catto, M.; Cellamare, S.; Gadaleta, D.; Lattanzi, G.; Leonetti, F.; Pisani, L.; et al. Mind the gap! A journey towards computational toxicology. *Mol. Inf.* **2016**, *35*, 294–308. [CrossRef] [PubMed]