Published in final edited form as:

Drug Discov Today. 2020 September ; 25(9): 1624–1638. doi:10.1016/j.drudis.2020.07.005.

Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling

Linlin Zhao[1], **Heather L. Ciallella**[1], **Lauren M. Aleksunes**[2], **Hao Zhu**[1,3]

[1]The Rutgers Center for Computational and Integrative Biology, Camden, NJ 08102, USA

[2]Department of Pharmacology and Toxicology, Ernest Mario School of Pharmacy, Rutgers University, Piscataway, NJ 08854, USA

[3]Department of Chemistry, Rutgers University, Camden, NJ 08102, USA

Abstract

Advancing a new drug to market requires substantial investments in time as well as financial resources. Crucial bioactivities for drug candidates, including their efficacy, pharmacokinetics (PK), and adverse effects, need to be investigated during drug development. With advancements in chemical synthesis and biological screening technologies over the past decade, large amount of biological data points for millions of small molecules have been generated and are stored in various databases. These accumulated data, combined with new machine learning (ML) approaches, such as deep learning, have shown great potential to provide insights into relevant chemical structures to predict *in vitro, in vivo*, and clinical outcomes, thereby advancing drug discovery and development in the big data era.

Keywords

big data; computer-aided drug discovery; cheminformatics; public databases; machine learning; deep learning; ADME

Introduction

The development of new drugs is a lengthy and expensive venture. To advance through phases of preclinical and clinical development, drug candidates are extensively tested for their efficacy, PK, and adverse effects [1]. Over the past decades, innovations in combinatorial chemistry, robotics, and high-throughput screening (HTS) have accelerated the rapid screening of thousands to millions of compounds against specific drug targets [2,3]. For example, in 2006, Brandish *et al.* used a cell-based HTS to evaluate a library containing >1 million compounds for their ability to cross cell membranes and inhibit D-amino acid oxidase, an approach that took <12 weeks to complete [4]. Novel testing

Corresponding author: Zhu, H. (hao.zhu99@rutgers.edu).

approaches, such as 3D culture techniques, were integrated into HTS to construct a natural extracellular microenvironment [5]. More recent advances in microsystems technology and cell culture techniques accelerated the development of organ-on-chip microdevices to better understanding the drug effects on various functional units of organs [5–10]. In addition to HTS and novel testing approaches, the advent of pharmacogenomics, made possible by the completion of the Human Genome Project (HGP), has spurred major advances in new drug development, including precision medicines and targets for diseases [11]. For example, CRISPR/Cas gene editing can be used to determine the genes and proteins that cause or prevent disease by deliberately activating or inhibiting genes, thus showing potential to identify targets for potential drugs [12]. Furthermore, recent advancements in portable electronic technology made it possible to track physiological signals of patients, such as body temperature, body movements, blood pressure, metabolites, functional proteins, and oligonucleotides [13–16]. All these research efforts have generated enormous amounts of data for drugs and drug candidates and moved modern drug discovery into an era of ‘big data’.

The term ‘big data’ refers to massive data, which have large, varied and complex data structures, with associated difficulties of storing, analyzing, and visualizing them using traditional computational approaches [17–21]. Being mostly used in the information technology field, big data is now expanding in all science and engineering domains, including drug discovery [22,23]. There are ‘ten Vs’ characteristics that are intrinsic for big data in drug discovery [24–26] (Figure 1): include volume, velocity, variety, veracity, validity, vocabulary, venue, visualization, volatility, and value [27,28] (Box 1). Compilation of large amounts of data generated daily and shared through public databases, such as Enamine REAL Database [29], ChEMBL [30], PubChem [31], and so on, represent the volume and velocity of available data. Currently, most data depository portals (e.g., PubChem) gather data from diverse sources, which define the variety of data. Given the inconsistency in data quality, veracity reflects the degree of uncertainty inherent to data from different sources and requires novel technologies for data curation and management. The features vocabulary and venue always come along with the data variety. Data from different sources can be described using different formats, texts, or terms. The data vocabulary (i.e., terminology) needs to be normalized when obtained from the original data venue (i.e., platform). Furthermore, due to the high diversity of testing protocols, the available data needs be evaluated before acceptance, inchoate the importance of data *validity.* Given the complexity of big data, the visualization of various large data sets is also in high demand. Data management is also an important component for drug discovery and development and determines the duration of data usefulness. Thus, the volatility feature of big data requires an appropriate and efficient data management and data-sharing strategy. The value of data can be defined as the potential of data usefulness to reduce the cost of drug discovery and development. Given that the data-driven studies are normally predictive models of drug bioactivities, the value of data is strongly dependent on the other nine Vs.

When assessing the current landscape of accumulated big data that can be utilized for drug discovery, a variety of classification approaches can be observed. These include: (i) comprehensive databases of chemical collections, including drugs, drug derivatives (e.g., drug metabolites), lead compounds, and drug candidates; (ii) collections of drug targets, including receptor genomics and proteomics data; (iii) databases storing biological data obtained from assay screening, metabolism, and efficacy studies; and (iv) databases that assess liabilities and toxicities for drugs and chemicals. Together, these databases offer a wealth of information and big data sources for drug discovery and development.

The application of ML approaches in drug discovery and development, particularly during early stages, has proved valuable. For example, models based on quantitative structure–activity relationship (QSAR) approaches have been used to quickly predict large numbers of new compounds for various endpoints, including not only simple physicochemical properties, such as logP and solubility [32], but also various biological activities, such as ligand-bin cling activities [33], drug efficacy [34], and adverse effects [35]. These QSAR models were developed using classic ML algorithms, such as random forest [36], support vector machines (SVMs) [37], and k-nearest neighbors [38], and the molecular descriptors [39] describing the chemical structures. With increasing data size and computational power, a new generation of artificial intelligence, such as deep learning algorithms, was also successfully applied for drug bioactivity modeling. For example, two early studies used ~400 000 compounds [40] and −8000 compounds [41], respectively, as the training sets for neural network developments. In a later study, Eli Lilly used deep learning to model historical commercial data from 24 data sets comprising >1 million compounds [42]. All of these efforts showed potential to be used to prioritize drug candidates with desired therapeutic activities and to exclude unsuitable compounds with adverse effects in a virtual screen during drug discovery [43]. For example, Sprague *et al.* used several QSAR models to prioritize 148 novel chemopreventive compounds from >23 000 natural products [44]. Furthermore, Lipinski Rule of five [45], which was treated as a golden rule of identifying drug bioavailability, has been integrated into most predictive software [46–48]. These prediction tools are being used to virtually screen drug candidates and remove those being predicted to be nonbioavailable. By removing unsuitable compounds even before chemical synthesis, the computational models can greatly reduce the cost of drug discovery.

In the current big data scenario, merely having access to big data is not a guarantee of obtaining informative predictive models [1]. Given the multiple Vs characteristics of big data, successful ML methods require crucial support and improvement in data mining, curation, and management technologies [27,49]. It is necessary to develop novel approaches that systematically address the high volume, multidimensional, and high-sparse data sources needed to prechct drug efficacy and adverse effects in animals and/or humans [28,50]. Here, we summarize and explore current big data resources available for drug discovery and development. We highlight recent studies using classic ML and new deep learning technologies. We also address key challenges and important considerations for modern computational-aided drug discovery (CADD) in the current big data era.

Big data for drug discovery and development

Compared with applications in IT fields, such as social network analysis, data sets used for drug discovery research are relatively small [22]. However, with the developments of combinatorial chemistry synthesis, HTS techniques, and genomics/genetics knowledge, the databases for drugs and drug candidates have grown rapidly and new modeling approaches are needed to handle these larger data sets [28]. Current publicly available databases relevant to drug discovery and development are summarized in Table 1. Based on their application and relevance during different stages of drug discovery and development, these databases can be classified into seven categories: (i) comprehensive databases of chemical collections (e.g., Enamine REAL Database [29], PubChem [31], and ChEMBL [30]); (ii) chemical databases designed specifically for drug/drug-like compounds (e.g., DrugBank [51], AICD [52], and e-Drug3D [53]); (iii) collections of drug targets, including genomics and proteomics data (e.g., BindingDB [54], Supertarget [55], and Ligand Expo [56]); (iv) databases storing biological data obtained from assay screening, metabolism, and efficacy studies (e.g., HMDB [57], TTD [58], WOMBAT [59], and PKPB_DB [60]); (v) drug liabilities and toxicities (e.g., DrugMatrix [61], SIDER [62,63], and LTKB Benchmark Dataset [64]); and (vi) clinical databases [e.g., ClinicalTrials.gov [65], EORTC Clinical Trials Database (www.eortc.org/clinical-trials-database/), and PharmaGKB [66]]. These databases provide multidimensional data relating to drug candidates, such as chemical structure, physicochemical properties, and *in vitro, in vivo*, and clinical data. The number and size of the databases for drug-like compounds has expanded significantly, although some do not primarily focus on drug discovery and development. For example, PubChem [67] is a public repository for chemical structures and their biological properties. The number of PubChem compounds increased from 19 million in 2008 [68] to >1.1 million in 2019 [31]. During the same period, the number of bioassays deposited in PubChem increased from 1197 in 2008 to over 1.1 million in 2019, resulting in over five terabytes of data (Figure 2) [31,68]. Current statistics from PubChem inchoate that the repository contains 102.4 million compounds tested against 1.1 million bioassays (https://pubchem.ncbi.nlm.nih.gov, accessed 30 January 2020). The tremendous amount of PubChem bioassay data, with a total size of over five terabytes, constitutes a publicly accessible big data resource for all PubChem compounds, including most drugs and drug candidates, with a variety of target response information. Similar to PubChem, ChEMBL is a database containing protein binding, functional, ‘absorption, distribution, metabolism, and excretion’ (ADME), and toxicity data for numerous compounds [69]. Compared with PubChem, which is primarily updated directly by screening centers and other biological data-generating projects, ChEMBL contains a large amount of manually curated data from the published literature. Currently, the ChEMBL database comprises >1.8 million compounds tested against >12 000 targets, resulting in activity data for 15 million compound–target pairs (www.ebi.ac.uk/chembl/, accessed 30 January 2020). Several other data sources are specifically designed for drugs and drug candidates. For example, e-Drug3D monitors the current content of the *US Pharmacopeia of Small Drugs* (molecular weight ≤ 2000) from data released by the US Food and Drug Administration (FDA) [53]. It offers a public tool to explore FDA-approved drugs and active metabolites and can be used in a range of endeavors, including drug repurposing, drug design, privileged structure analyses, and SAR studies. The latest release (updated in June 2019) of e-Drug3D contains 1930 small-molecule drugs approved between 1939 and 2019 by FDA. The increasing availability of public data aims to reduce costs via increased outsourcing and engagement in precompetitive activities. These public data can have a significant impact on academic institutes, not-for-profit organizations, and industrial drug discovery by encouraging the development of new computational tools and predictive algorithms within the pubhc domain, benefiting the whole research community [70–72].

During the early exploration stage of drug discovery, genomics and proteomics data are widely used for drug-target identification. The Binchng Database (BindingDB) is a public, web-accessible resource of drug–target binding data, including data of measured binchng affinities [54]. The targets included in BindingDB are proteins/enzymes that are considered as drug targets. BindingDB currently contains 1 756,093 binchng data, for 7371 protein targets and 780 240 small molecules (www.binchngclb.org/bincl/inclex.jsp, accessed 29 October 2019). During the drug development stage, databases storing biological data obtained from assay screening, metabolism information, and efficacy are widely used. The Human Metabolome Database (HMDB) is a freely available electronic database containing detailed information about small-molecule metabolites foundin the human body [73]. It currently contains 114 162 metabolite entries, including both water-soluble and lipid-soluble metabolites. WOMBAT is a bioactivity database for lead and drug discovery [59]. It currently contains 331 872 entries, representing 1966 unique targets, with bioactivity annotations. By contrast, DrugMatrix [61] focuses on the toxicogenomics data from ~600 drugs. The current DrugMatrix database contains large-scale rat gene expression data under drug treatment, mostly targeting several major organs (e.g., hver). Clinical data provide further drug adverse effect information. For example, AACT is a publicly available relational database that contains all information (protocol and result data elements) regarding every study registered in ClinicalTrials.gov [65]. It contains ~324 429 research studies in all 50 US states and in 209 countries. PharmGKB (/www.pharmgkb.org/) is a pharmacogenomics knowledge resource that encompasses clinical information of drug molecules, and contains 733 drugs with corresponding clinical information.

Challenges Multiple Vs in big data studies

Data-driven studies for CADD are required to solve challenges from the multiple Vs features, as described earlier. Most notably, these include a need for efficient handling of data sets generated from various sources (variety) at a rapid speed (veracity), and shared by different platforms (venue) with a specific time length of usefulness (volatility). The data sets from the public domain can be described using different terminologies (vocabulary), with certain qualities (veracity) and validity. The data volume used for drug discovery can be expansive because of the huge amount of data generated along with the long-term development procedure. The databases in Table 1, which are all relevant to drug discovery, can also be classified based on the associated stage of drug discovery: early exploration and discovery, hit identification, lead identification, lead optimization, and clinical studies (Figure 3). During early stages of discovery, various properties of drug candidates, such as physicochemical properties (data in the first category), protein-binding information (data in the second category), target information (genomic data in the third category), and adverse effect information (data in the fourth category) are generated. The high data variety makes it difficult to manage and incorporate heterogeneous data. In the meantime, these data are shared by different venues, which provide key information in different terminologies (vocabulary). For example, chemical identifiers chffer among different data platforms (e.g., CID for PubChem, and CAS for DrugBank and ChemIDplus, etc.) and drug structures are coded in different format (e.g., SMILES, InChi key, etc.) [74]. These features (venue, vocabulary, and variety) highlight the urgent need to develop universal criteria for data harmonization [75–78]. When moving from early stages of discovery to clinical trials, the data volume changes, that is, the size of data sets becomes considerably smaller because of limited data availability during the late stages of drug discovery. Most of the databases in the second, third, and fourth categories in Figure 3 comprises thousands to tens of thousands of compounds and serve to address specific needs, such as data regarding the ability of drug candidates to bind targets with specificity and strong affinity. Given that the clinical studies required for FDA approval of a new drug typically need progression through five testing phases (phase 0–IV), there is an enormous opportunity to enter accumulated data in clinical databases for individual drugs (Figure 3). Clinical databases often comprise thousands to hundreds of thousands data entries because one drug candidate typically undergoes extensive investigation and generates a large amount of data [79]. When comparing clinical databases to those that collect general information about chemicals, including property data (e.g., log P, solubility and etc.) and general biological activities (e.g., P450 inhibitions, cytotoxicity and etc.), have the largest size and always contain >1 million compounds (Figure 3). Given that the data are being collected from numerous sources, the variety and velocity of these databases are also the highest. These big data sources provide useful information for early drug discovery stages, but the multiple Vs features also bring new challenges.

Accelerated by the developments of novel testing technologies, data for drug discovery grow rapidly beyond our ability to use them [22]. Furthermore, a lack of quality control [80] is a common issue for public data sources. The ‘trash in, trash out’ principle [81] was introduced for all modeling studies to highlight the importance of quality control. When looking at available data for known drugs, there have been many testing results available for modeling purposes. For example, 1930 FDA-approved small-molecule drugs (molecular weight <2000) in e-Drug3D databases [18] were used to search against both ChEMBL [7] and PubChem [8] for their assay-testing results by using an in-house data-profiling tool [82]. There were 1114 ChEMBL assays with testing results for at least 25 of these drug molecules (Figure 4A). All these drugs were also tested against thousands of PubChem assays, and 299 assays had at least 25 active responses among these drug molecules (Figure 4B). There are >2 million data points in the response profile for ChEMBL and >500 000 data points in the PubChem response profile. Compared with PubChem, which is primarily shared directly by screening centers and data generation projects [68], ChEMBL contains a large amount of manually curated data from the published literature [30]. Currently, there is a significant overlap between ChEMBL and PubChem data because PubChem automatically acquires data from ChEMBL [83]. Studies have compared the data obtained from these two resources [84,85], with many responses in these profiles shown in gray as missing data (96% of the ChEMBL response profile in Figure 4A and 87% of the PubChem response profile in Figure 4B), because these drug compounds were not tested against all assays. In addition to the comparison of bioassay data in ChEMBL and PubChem, other tools and studies reported previously to compare the chemical space of these two databases [84] have the potential to handle the challenges arising from the data *variety.* Furthermore, the ratio of active responses in the PubChem data (e.g., 27% of all data in Figure 4B) is also biased. For example, acyclovir (CAS 59277-89-3) had 13 active and 204 inactive responses in these

PubChem assays. Given the nature of HTS techniques, general HTS data normally comprise fewer actives than inactives [86,87], especially for screening active hits against specific drug targets. In an early review of pharmacological data based on 4.8 million unique compounds, only ~5.7% of these compounds were found to show one (or more) active biological response [88], inchoating that most of the testing results were inactives. Notably, some drugs show high active responses in available data. For example, disulfiram (CAS 97-77-8) is used to deter alcohol consumption in patients with alcoholism. In PubChem, disulfiram has 163 active responses and 57 inactive results across multiple assays. Furthermore, clotrimazole (CAS 23593-75-1), an antifungal medicine, has 163 active responses and 42 inactive results across the assays in PubChem. Niclosamide (CAS 50-65-7) is used to treat tapeworm infestations, and has 157 active responses and 35 inactive results.

Besides veracity of data, the feature validity also determines data quality. Numerous testing approaches have been developed to evaluate drug candidates. However, hundreds and/or thousands of different protocols exist for the same testing purpose [89–92]. Thus, there is a need to understand the applicability of each protocol and the resulting data. For example, numerous protocols only use a single concentration (i.e., high concentration) of compounds for screening purposes, inchoating a potential flaw of the resulting classifications [93]. To address this issue, the National Center for Advancing Translational Sciences (NCATS) proposed quantitative HTS (qHTS), which utilizes multiple concentration testing to test drug molecules, generated more data for testing the same compounds [94]. Considered together with the feature veracity, it is important to manage and incorporate available big data meaningfully for drug discovery and development. For example, drug repurposing uses the preclinical, PK, pharmacodynamic (PD), and toxicity data of existing drugs, greatly reducing development costs [95]. Thus, computational modeling studies for drug repurposing are highly dependent on the data volatility. Finally, given the high velocity and variety, the visualization of big data in drug discovery also requires new tools [96–100].

For effective ML studies, steps can be taken to resolve issues induced by the multiple Vs features. For example, a common solution to fill the missing information of target compounds has been the development of ML models, such as QSAR models, for individual biological targets based on existing data. The resulting models are then used to predict compounds that were not tested in the relevant ‘training’ assays [101–104]. This strategy is applicable for physicochemical properties (e.g., logP) and/or target binding that utilize simple biological mechanisms (e.g., the binding target is rigid and specific). Recently, a new concept termed ‘read-across’, was developed based on a similar strategy [105,106]. Read-across is a method that fills in the missing data of target compounds based on similarities (largely structural) to the nearest neighbors among the compounds that have been tested. Read-across is becoming increasing used in the identification of chemical toxicity liabilities, including the prediction of adverse effects for drugs and in the evaluation of personal care products [78,107,108]. Currently, the European Chemical Agency (ECHA) (https://echa.europa.eu/home) is accepting read-across dossiers for chemical evaluation decision-making, and read-across strategy papers have been recently published [105,109]. However, using computational models (e.g., QSAR) or a read-across strategy to fill missing data can introduce extra uncertainty because of unanticipated prediction errors [103]. To deal with the biased nature of HTS data, more weighting should be given to active results, rather than inactive results, during modeling procedures [82]. ML modeling studies typically need the biased data sets to first be balanced by using various methods, such as down sampling [110–112]. Furthermore, the challenge of velocity can introduce potential experimental errors in public databases. This issue becomes crucial when data were gathered from different sources (i.e., the variety is high) because of unstandardized protocols, including data analysis, quality controls, and different experimentalists. Data sets with experimental errors will decrease the quality of the resulting computational models [113]. Although there is no universal solution to remove experimental errors, which are difficult to define in public data, researchers have proposed possible solutions to reduce potential experimental errors in the data set [114–116]. Cortes-Ciriano *et al.* [115] simulated experimental errors in QSAR modeling sets, and then compared the influence of different QSAR approaches on predictive accuracy. This study provided a practical reference for making a better decision about which modeling approach to use depending on the quality of modeling sets. Roy *et al.* [116] studied the relationship between systematic errors in the predictions and the applicability domain (AD) of QSAR modeling. They also exposed the flaw of using normal correlation coefficients to describe model predictivity. Zhao *et al.* [114] proposed a possible solution for identifying large experimental errors in the data sets, and investigated methods, including removing suspected data from the modeling set and applying AD for improving models, developed on questionable modeling sets.

Application of ML approaches for drug discovery and development using big data sources

Big data generated during drug discovery and shared via public databases have potentially significant value, which is the last of the ten Vs features, by reducing drug attrition during the development pipeline. Traditional ML applications, such as QSAR modeling, can be used to prioritize drug candidates with desired therapeutic activities and exclude unsuitable compounds with predicted adverse effects. Recently, remarkable improvements in computational power coupled with improvements in artificial intelligence (Al) technology, including the development of various deep learning algorithms [117–119], positioned CADD studies to progress to a new stage and capitalize on the rich infrastructure of big data studies [120].

AI, which is sometimes presented as machine intelligence, refers to the ability of computers to learn from existing data [121]. ML is a subfield of AI, and refers to methods that endow computers with learning ability, as defined by Arthur Samuel in 1959 [122]. QSAR is one of the classic applications of ML approaches in drug discovery. Since the QSAR approach was first developed by Hansch and Fujita in 1964 [123], it has remained an efficient method to find a statistically significant correlation between chemical structures and their properties and activities. During the early stages of QSAR application in drug discovery, QSAR modeling was limited to small data sets (e.g., <10 compounds) and based on simple linear regression methods [124]. Over the past few decades, QSAR has reached several milestones, including the development of novel chemical descriptors, including topological descriptors [125] and molecular fingerprints [126,127], and the application of new nonlinear modeling algorithms, such as random forest [36], SVMs [37], and k-nearest neighbors [38]. In the same period, model validation was emphasized and treated as a crucial component of modeling procedures [128]. In addition, the applicability domain, which defines the limitation of the applicability of a model, became standard practice for model development [129–132]. The application of QSAR modeling in drug discovery has created big value by saving resources through the virtual screening of drug candidates [133–135]. Drug candidates with desired biological activities (e.g., therapeutic activities) and fewer adverse effects can be prioritized before chemical synthesis. By applying QSAR models, the hit rates could be improved significantly by testing prioritized compounds selected from validated models [44,136–138].

Besides QSAR, there are many other ML applications in drug discovery and development [141]. For example, generative models were applied for *cle novo* drug design by applying the statistical framework to chemical pattern-matching studies [22,142–144]. These generative models can produce synthetically accessible molecules with desired properties and activities [145,146]. The advantages of these models include quick decision-making and providing an infinite virtual chemical space [147]. In another study, a multiple kernel learning algorithm was developed for drug-target interaction prediction [148] and allowed the integration of multiple heterogeneous information sources. A ML model developed for predicting blood–brain barrier (BBB) permeability combined public bioassay data with chemical information [101] for better predictivity. Similarly, another model developed for predicting the sensitivity of cancer cell to drugs integrated both genomic and chemical information [149]. ML applications were also used in drug repurposing studies, as discussed in a previous review [95]. Furthermore, ML was also applied to emerging ‘omics (genomics, transcriptomics, proteomics, and metabolomics) data, which could be used to expand our understanding of the complexities of human disease, to generate novel biomarkers for personalized medicine [150–152]. Thus, the application of ML modeling for various data, including public databases, can drive CADD to create big value in the big data era.

The advancement of computational power and the increasing availability of biological data stimulated the applications of new ML techniques, such as artificial neural network (ANN) modeling, to address the multiple Vs challenges brought by big data in drug discovery. Since the first reported application of the neural network modeling in drug discovery in 1989 [153], various neural network approaches have been developed and applied to drug discovery [154,155]. Deep learning, based on ANN, was originally presented during the 1980s [156]. However, it did not show significant advantages over other ML approaches during its infancy because the data used for model development were limited [157,158]. With increasing data size and computational power, deep learning has been applied to the life sciences and demonstrated its ability to contribute to drug discovery and development [156,159]. Deep learning applications in virtual screening were highlighted in the QSAR ML challenge supported by Merck in 2012, and the winning team used an ensemble of different ML methods including deep neural networks (DNN) and showed significantly better performance than other ML approaches in their following study [160]. The deep learning models in this study were based on a set of traditional molecular descriptors, such as atom pairs (AP) [161] and donor–acceptor pair (DP) [162]. Later in 2014, the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH) launched the Tox21 Challenge, in which participants were asked to model ~12 000 chemicals, including many drugs, for 12 different toxic effects [163]. In this competition, DeepTox, a computational toxicity model based on deep learning, exhibited the highest performance of all computational methods [164]. The chemical descriptors used in the DeepTox model were derived from a large number of molecular descriptors calculated using computational tools, including off-the-shelf software and JCompoundMapper [164]. Furthermore, there are other studies reporting the application of deep learning to the prediction of chemical properties, with the deep learning models showing improvements compared with classic ML models. Lusci *et al.* [165] reported deep learning models for the prediction of aqueous solubility for drug-like molecules. In this study, chemical structures were transformed into graphs and recursive neural network approaches were apphed for modeling. In another study, Xu *et al.* presented deep learning models using chemical 2D structure graphs as the input data for both liver toxicity [166] and acute oral toxicity [41]. Another study explored the performance of geometric deep learning methods in drug discovery, where deep learning methods were able to identify more useful chemical features compared with ML approaches [167]. Research has also compared model performance among multiple ML and deep learning approaches. For example, Lane *et al.* compared various ML models to identify hits for *Mycobacterium tuberculosis* [168]. These authors showed that SVM and DNN outperformed other ML algorithms on prediction accuracy. Overall, deep learning algorithms were the most efficacious in combining all the descriptors using all the metrics for training and cross-validation. However, in another study, Russo *et al.* [169] compared various ML and deep learning techniques for predicting estrogen receptor (ER) binchng agents. In this work, random forest outperformed the other algorithms, including deep learning, inchoating that there is no general advantage of deep learning to handle the modeling of all data sets.

Deep learning has been applied to other drug discovery studies, such as *cle novo* drug design [143,170]. For example, Gomez-Bombarelli *et al.* [171] reported exploring chemical space based on continuous encodings of molecules using DNN approaches. Another deep learning application for generating focused molecular libraries with the desired bioactivity was proposed [172]. In this study, recurrent neural networks (RNNs) were trained as generative models for drug molecular structures. Recently, another deep learning application, ReLeaSE [173], was reported for generating *cle novo* compounds with desired drug-related properties. ReLeaSE integrated two DNNs (generative and predictive) that were trained separately but were used jointly to generate novel targeted chemical libraries. These deep learning applications benefited from developments in ML applications of natural language processing and machine translation, with deep learning applications in *cle novo* drug design recently reviewed elsewhere [174]. Deep learning applications have also been developed that focus on using heterogeneous data. For example, a deep learning model was reported that could be used to prechct interactions between drugs and their biological targets based on 15 524 drug–target pairs from the DrugBank database, and using traditional molecular fingerprints [Extended Connectivity Fingerprints (ECFP)] [175]. Xie *et al.* [176] reported a deep learning study for the prediction of drug–target interactions using transcriptome data in the L1000 database from the Library of Integrated Network-Based Cellular Signatures program [44]. Reusing previous data by deep learning approaches is important for drug repurposing [95]. Donner *et al.* also used the L1000 database to develop a new method for measuring the compound functional similarity based on gene expression data for drug repurposing [177]. In this study, drugs with similar therapeutic and biological targets but dissimilar structures were identified to reveal previously unreported functions of compounds. Furthermore, multi-task learning based on DNN allows multiple related tasks to be modeled simultaneously. The multi-task learning approached demonstrated that DNN can reduce overfitting, solve issues of biased data, and identify variables from related tasks. Thus, multitask learning has a somewhat better performance compared with traditional models for some specific data sets [41,178–180]. However, there have also been arguments that ML models still can achieve better results compared with deep learning [42,169]. Given the complexity of biological systems and the multiple Vs features of big data for drug discovery, it remains chfficult to qualify a ML anchor deep learning method as universally superior to other approaches [42].

Concluding remarks and perspectives

In current era of big data, developments in computational tools and the rapid growth of public data sources have advanced the CADD. ML and deep learning approaches have been applied to the data generated along various stages of drug discovery and development and affirmed the *value* of big data by reducing the drug attrition. The challenges brought by the multiple Vs feature of big data require the development of appropriate computational approaches and algorithms. In addition to the progress in ML applications in drug discovery described earlier, the multiple Vs features, such as volume, velocity, variety, vocabulary, and volatility, require better database management data curation, and web portal design. The variety, veracity, validity, and venue features also require further refinements of experimental protocols, better quality controls, and more transparent data reporting. However, some clear limitations remain. For example, projects dealing with intellectual property (IP)-sensitive structures have no data-sharing authority [181]. Given the speed at which big data in drug discovery grow (velocity), it is difficult to update available CADD software with the newly generated data and recently developed algorithms and models as quickly. Most available prediction tools remain based on traditional QSAR approaches and have not changed for years. Although being used widely, the applications of data-driven ML modeling in drug discovery, especially deep learning, are still in the preliminary stage. In addition, applications of CADD tools in the industry is still questioned by the research community.

Regardless of the issues described earlier, when coupled with improvements in computer hardware and experimental screening techniques, ML modeling will continue to be crucial in illustrating the value of big data for drug discovery. New modeling algorithms and approaches will be key to addressing the multiple Vs challenges associated with big data.

Acknowledgment

This paper was partially supported by the National Institute of Environmental Health Sciences (grant numbers R01ES031080, R15ES023148, and P30ES005022)

Author biographies

Linlin Zhao

Linlin Zhao currently is a PhD student in the Center for Computational and Integrative Biology (CCIB) at Rutgers, The State University of New Jersey in Camden, under the mentorship of Hao Zhu. Her research focuses on read-across studies of computational toxicology by using various machine learning algorithms.

Heather L. Ciallella

Heather L. Ciallella currently is a PhD student in CCIB under the mentorship of Hao Zhu. Her research focuses on the applications of deep learning algorithms to chemical toxicity predictions.

Lauren M. Aleksunes

Lauren M. Aleksunes is a board-certified toxicologist and professor of pharmacology and toxicology at Rutgers University. She received her Pharm.D. and PhD from the University of Connecticut. She has authored or coauthored over 115 publications on mechanisms of xenobiotic disposition and toxicity.

Hao Zhu

Hao Zhu is a professor in the Chemistry Department and CCIB at Rutgers University. He received his PhD in computational chemistry from Case Western Reserve University in 2002. Dr Zhu has authored or coauthored over 75 peer-reviewed publications and book chapters in the applications of machine learning and big data modeling to chemical toxicity assessments, computer-aided drug discovery, and rational nanomaterial design.

References

1. Schneider G (2018) Automating drug discovery. Nat. Rev. Drug Discov 17, 97–113 [PubMed: 29242609]
2. Zheng XT et al. (2013) On-chip investigation of cell-drug interactions. Adv. Drug Deliv. Rev 65, 1556–1574 [PubMed: 23428898]
3. Hughes JP et al. (2011) Principles of early drug discovery. Br. J. Pharmacol 162, 1239–1249 [PubMed: 21091654]
4. Brandish PE et al. (2006) A cell-based ultra-high-throughput screening assay for identifying inhibitors of D-amino acid oxidase. J. Biomol. Screen 11, 481–487 [PubMed: 16760370]
5. Carvalho MR et al. (2020) Biomaterials and microfluidics for drug discovery and development. Adv. Exp. Med. Biol 1230, 121–135 [PubMed: 32285368]
6. Herland A et al. (2020) Quantitative prediction of human pharmacokinetic responses to drugs via fluidically coupled vascularized organ chips. Nat. Biomed. Eng 4, 421–436 [PubMed: 31988459]
7. Carney EF (2020) Pharmacokinetic modelling using linked organ chips. Nat. Rev. Nephrol 16, 188–188
8. Benam KH et al. (2020) Biomimetic smoking robot for in vitro inhalation exposure compatible with microfluidic organ chips. Nat. Protoc 15, 183–206 [PubMed: 31925401]
9. Torisawa YS and Tung YC (2020) Editorial for the Special Issue on Organs-on-Chips. Micromachines (Basel) 11, 369
10. Nguyen DT et al. (2018) Translational strategy: humanized mini-organs. Drug Discov. Today 23, 1812–1817 [PubMed: 29883729]
11. Hamburg MA and Collins FS (2010) The path to personalized medicine. N. Engl. J. Med 363, 301–304 [PubMed: 20551152]
12. Scott A (2018) A CRISPR path to drug discovery. Nature 555, S10–S11
13. Lou Z et al. (2020) Reviews of wearable healthcare systems: materials, devices and system integration. Mater. Sci. Eng. Rep 140, 100523
14. Liu QX et al. (2019) A high-performances flexible temperature sensor composed of polyethyleneimine/reduced graphene oxide bilayer for real-time monitoring. Adv. Mater. Technol 4, 1800594
15. Kim J et al. (2016) Noninvasive alcohol monitoring using a wearable tattoo-based iontophoretic-biosensing system. ACS Sensors 1, 1011–1019
16. Song CY et al. (2017) Ultrasensitive sliver nanorods array SERS sensor for mercury ions. Biosens. Bioelectron 87, 59–65 [PubMed: 27522013]
17. Sagiroglu S and Sinanc D (2013) Big data: a review. Proc. 2013 Int. Conf. Collab. Technol. Syst 2013, 42–47
18. Schadt EE et al. (2011) Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology. Nat. Rev. Genet 12, 224
19. Marx V (2013) Biology: the big challenges of big data. Nature 498, 255–260 [PubMed: 23765498]
20. Swarup V and Geschwind DH (2013) Alzheimer’s disease: from big data to mechanism. Nature 500, 34–35 [PubMed: 23883924]
21. Wu XD et al. (2014) Data mining with big data. IEEE Trans. Knowledge Data Eng 26, 97–107
22. Lusher SJ et al. (2014) Data-driven medicinal chemistry in the era of big data. Drug Discov. Today 19, 859–868 [PubMed: 24361338]
23. Lusher SJ et al. (2011) A molecular informatics view on best practice in multi-parameter compound optimization. Drug Discov. Today 16, 555–568 [PubMed: 21605698]

24. McAfee A and Brynjolfsson E (2012) Big data: the management revolution. Harv. Bus. Rev 90, 60–66, 68, 128 [PubMed: 23074865]

25. Arockia Panimalar S et al. (2017) The 17 V's of big data. Int. Res. J. Eng. Technol 4, 329–333

26. Oguntimilehin A and Ademola E (2014) A review of big data management, benefits and challenges. Rev. Big Data Manage. Benefits Challenges 5, 1–7

27. Ciallella HL and Zhu H (2019) Advancing computational toxicology in the big data era by artificial intelligence: data-driven and mechanism-driven modeling for chemical toxicity. Chem. Res. Toxicol 32, 536–547 [PubMed: 30907586]

28. Zhu H (2020) Big data and artificial intelligence modeling for drug discovery. Annu. Rev. Pharmacol. Toxicol 60, 573–589 [PubMed: 31518513]

29. Klingler FM et al. (2019) SAR by space: enriching hit sets from the chemical space. Molecules 24, 3096

30. Gaulton A et al. (2017) The ChEMBL database in 2017. Nucleic Acids Res. 45, D945–D954 [PubMed: 27899562]

31. Kim S et al. (2019) PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 47, D1102–D1109 [PubMed: 30371825]

32. Khan MT and Sydte I (2007) Predictive QSAR modeling for the successful predictions of the AD MET properties of candidate drug molecules. Curr. Drug Discov. Technol 4, 141–149 [PubMed: 17985997]

33. Tong W et al. (1997) QSAR models for binding of estrogenic compounds to estrogen receptor alpha and beta subtypes. Endocrinology 138, 4022–4025 [PubMed: 9275094]

34. Yoshida F and Topliss JG (2000) QSAR model for drug human oral bioavailability. J. Med. Chem 43, 2575–2585 [PubMed: 10891117]

35. Dearden JC (2003) In silico prediction of drug toxicity. J. Comput. Aided Mol. Des 17, 119–127 [PubMed: 13677480]

36. Breiman L (2001) Random forests. Machine Learn. 45, 5–32

37. Cortes C and Vapnik V (1995) Support-vector networks. Machine Learn. 20, 273–297

38. Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat 46, 175–185

39. Karelson M (2000) Molecular Descriptors in QSAR/QSPR, Wiley-Interscience

40. Korotcov A et al. (2017) Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. Mol. Pharm 14, 4462–4475 [PubMed: 29096442]

41. Xu Y et al. (2017) Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. J. Chem. Inf. Model 57, 2672–2685 [PubMed: 29019671]

42. Zhou YD et al. (2019) Exploring tunable hyperparameters for deep neural networks with industrial ADME data sets. J. Chem. Inf. Model 59, 1005–1016 [PubMed: 30586300]

43. Zhu H (2019) Big data and artificial intelligence modeling for drug discovery. Annu. Rev. Pharm. Toxicol 60, 573–589

44. Sprague B et al. (2014) Design, synthesis and experimental validation of novel potential chemopreventive agents using random forest and support vector machine binary classifiers. J. Comp-Aided Mol. Design 28, 631–646

45. Lipinski CA et al. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev 23, 3–25

46. Multi CASE Inc. (2017) CASE Ultra. Multi CASE

47. Chemical Computing Group Inc. (2016) Molecular Operating Environment (MOE), Chemical Computing Group Inc

48. Dimitrov SD et al. (2016) QSAR toolbox - workflow and major functionalities. SAR QSAR Env. Res 27, 203–219 [PubMed: 26892800]

49. Zhao L and Zhu H (2018) Big data in computational toxicology: challenges and opportunities In Computational Toxicology: Risk Assessment for Chemicals (XXXX, YYYY, ed.), pp. 291–312, Wiley

50. Zhang L et al. (2017) From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov. Today 22, 1680–1685 [PubMed: 28881183]

51. Wishart DS et al. (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 46, D1074–D1082 [PubMed: 29126136]

52. Wang K et al. (2019) AICD: an integrated anti-inflammatory compounds database for drug discovery. Sci. Rep 9, 7737 [PubMed: 31123286]

53. Douguet D (2018) Data sets representative of the structures and experimental properties of FDA-approved drugs. ACS Med. Chem. Lett 9, 204–209

54. Gilson MK et al. (2016) BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 44, D1045–D1053 [PubMed: 26481362]

55. Gunther S et al. (2008) SuperTarget and Matador: resources for exploring drug-target relationships. Nucleic Acids Res. 36, D919–D922 [PubMed: 17942422]

56. Feng ZK et al. (2004) Ligand Depot: a data warehouse for ligands bound to macromolecules. Bioinformatics 20, 2153–2155 [PubMed: 15059838]

57. Wishart DS et al. (2007) HMDB: the human metabolome database. Nucleic Acids Res. 35, D521–526 [PubMed: 17202168]

58. Zhu F et al. (2010) Update of TTD: therapeutic target database. Nucleic Acids Res. 38, D787–791 [PubMed: 19933260]

59. Olah M et al. (2007) WOMBAT and WOMBAT-FK: bioactivity databases for lead and drug discovery. Chem. Biol 1, 760–786

60. Thompson CM et al. (2009) Database for physiologically based pharmacokinetic (FBFK) modeling: physiological data for healthy and health-impaired elderly. J. Toxicol. Environ. Health B Crit. Rev 12, 1–24 [PubMed: 19117207]

61. Svoboda DL et al. (2019) An Overview of National Toxicology Program’s Toxicogenomic Applications: DrugMatrix and ToxFX. Adv. Comput. Toxicol 30, 141–157

62. Kuhn M et al. (2016) The SIDER database of drugs and side effects. Nucleic Acids Res. 44, D1075–D1079 [PubMed: 26481350]

63. Kuhn M et al. (2010) A side effect resource to capture phenotypic effects of drugs. Mol. Syst. Biol 6, XXX–YYY

64. Chen M et al. (2011) FDA-approved drug labeling for the study of drug-induced liver injury. Drug Discov. Today 16, 697–703 [PubMed: 21624500]

65. Zarin DA et al. (2011) The ClinicalTrials.gov results database - update and key issues. N. Engl. J. Med 364, 852–860 [PubMed: 21366476]

66. Thorn CF et al. (2010) Pharmacogenomics and bioinformatics: PharmGKB. Pharmacogenomics 11, 501–505 [PubMed: 20350130]

67. Kim S et al. (2019) PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 47, D1102–D1109 [PubMed: 30371825]

68. Bolton EE et al. (2008) PubChem: integrated platform of small molecules and biological activities. Annu. Rep. Comput. Chem 4, 217–241

69. Mendez D et al. (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. 47, D930–D940 [PubMed: 30398643]

70. Gaulton A and Overington JP (2010) Role of open chemical data in aiding drug discovery and design. Future Med. Chem 2, 903–907 [PubMed: 21426107]

71. Gonzalez-Medina M et al. (2017) Open chemoinformatic resources to explore the structure, properties and chemical space of molecules. RSC Adv. 7, 54153–54163

72. Tambuyser E et al. (2020) Therapies for rare diseases: therapeutic modalities, progress and challenges ahead. Nat. Rev. Drug Discov 19, 93–111 [PubMed: 31836861]

73. Wishart DS et al. (2018) HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 46), D608–D617 [PubMed: 29140435]

74. Fourches D et al. (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J. Chem. Inf. Model 50, 1189–1204 [PubMed: 20572635]

75. Fourches D et al. (2016) Trust, but Verify II: a practical guide to chemogenomics data Curation. J. Chem. Inf. Model 56, 1243–1252 [PubMed: 27280890]

76. Pezoulas VC et al. (2019) Medical data quality assessment: On the development of an automated framework for medical data curation. Computers Biol. Med 107, 270–283

77. Sansone SA et al. (2012) Toward interoperable bioscience data. Nat. Genet 44, 121–126 [PubMed: 22281772]

78. Zhao L et al. (2020) Mechanism-driven read-across of chemical hepatotoxicants based on chemical structures and biological data. Toxicol. Sci 174, 178–188 [PubMed: 32073637]

79. Cook JA and Collins GS (2015) The rise of big clinical databases. Br. J. Surg 102, e93–el01 [PubMed: 25627139]

80. Williams AJ and Ekins S (2011) A quality alert and call for improved curation of public chemistry databases. Drug Discov. Today 16, 747–750 [PubMed: 21871970]

81. Hartung T (2016) Making big sense from big data in toxicology by read-across. Altex 33, 83–93 [PubMed: 27032088]

82. Russo DP et al. (2017) CUPro: a new read-across portal to fill data gaps using public large-scale chemical and biological data. Bioinformatics 33, 464–466 [PubMed: 28172359]

83. Wang Y et al. (2017) PubChem BioAssay: 2017 update. Nucleic Acids Res. 45, D955–D963 [PubMed: 27899599]

84. Awale M et al. (2013) MQN-mapplet: visualization of chemical space with interactive maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. J. Chem. Inf. Model 53, 509–518 [PubMed: 23297797]

85. Capecchi A et al. (2019) PubChem and ChEMBL beyond Lipinski. Mol. Inform 38, el900016

86. Zhang J et al. (2014) Profiling animal toxicants by automatically mining public bioassay data: a big data approach for computational toxicology. PLoS ONE 9, XXX–YYY

87. Russo DP et al. (2019) Nonanimal models for acute toxicity evaluations: applying data-driven profiling and read-across. Env. Health Perspect 127, 047001

88. Paolini GV et al. (2006) Global mapping of pharmacological space. Nat. Biotechnol 24, 805–815 [PubMed: 16841068]

89. Di L et al. (2007) Comparison of cytochrome P450 inhibition assays for drug discovery using human liver microsomes with LC-MS, rhCYP450 isozymes with fluorescence, and double cocktail with LC-MS. Int. J. Pharm 335, 1–11 [PubMed: 17137735]

90. Chau CH et al. (2008) Validation of analytic methods for biomarkers used in drug development. Clin. Cancer Res 14, 5967–5976 [PubMed: 18829475]

91. Tiwari G and Tiwari R (2010) Bioanalytical method validation: an updated review. Pharm. Methods 1, 25–38 [PubMed: 23781413]

92. Buick AR et al. (1990) Method validation in the bioanalytical laboratory. J. Pharm. Biomed. Anal 8, 629–637 [PubMed: 2100599]

93. Michael S et al. (2008) A robotic platform for quantitative high-throughput screening. Assay Drug Dev. Technol 6, 637–657 [PubMed: 19035846]

94. Huang R (2016) A quantitative high-throughput screening data analysis pipeline for activity profiling. Methods Mol. Biol 1473, 111–122 [PubMed: 27518629]

95. Vanhaelen Q et al. (2017) Design of efficient computational workflows for in silico drug repurposing. Drug Discov. Today 22, 210–222 [PubMed: 27693712]

96. Keim D et al. (2013) Big-data visualization. IEEE Comput. Graph Appl. 33, 20–21

97. Bolouri H et al. (2016) Big data visualization identifies the multidimensional molecular landscape of human gliomas. Proc. Natl. Acad. Sci. U. S. A 113, 5394–5399 [PubMed: 27118839]

98. Gaspar HA et al. (2015) Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. J. Chem. Inf. Model 55, 84–94 [PubMed: 25423612]

99. Iqbal U et al. (2016) Cancer-disease associations: A visualization and animation through medical big data. Comput. Methods Programs Biomed 127, 44–51 [PubMed: 27000288]

100. Kinjo S et al. (2018) Maser: one-stop platform for NGS big data from analysis to visualization. Database 2018, XXX–YYY

101. Wang WY et al. (2015) Developing enhanced blood-brain barrier permeability models: integrating external bio-assay data in QSAR modeling. Pharm. Res 32, 3055–3065 [PubMed: 25862462]

102. Kim MT et al. (2014) Critical evaluation of human oral bio availability for pharmaceutical drugs by using various cheminformatics approaches. Pharm. Res 31, 1002–1014 [PubMed: 24306326]

103. Kim MT et al. (2016) Mechanism profiling of hepatotoxicity caused by oxidative stress using antioxidant response element reporter gene assay models and big data. Environ. Health Perspect 124, 634–641 [PubMed: 26383846]

104. Ribay K et al. (2016) Predictive modeling of estrogen receptor binding agents using advanced cheminformatics tools and massive public data. Front. Env. Sci 4, 12 [PubMed: 27642585]

105. Ball N et al. (2016) t4 report: Toward Good Read-Across Practice (GRAP) Guidance. Altex 33, 149 [PubMed: 26863606]

106. Zhu H et al. (2016) t4 report: supporting read-across using biological data. Altex 33, 167 [PubMed: 26863516]

107. Guo YJ et al. (2019) Using a hybrid read-across method to evaluate chemical toxicity based on chemical structure and biological data. Ecotoxicol. Environ. Safety 178, 178–187 [PubMed: 31004930]

108. Zhu H et al. (2014) Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. Chem. Res. Toxicol 27, 1643–1651 [PubMed: 25195622]

109. Zhu H et al. (2016) Supporting read-across using biological data. Altex 33, 167–182 [PubMed: 26863516]

110. Zhang LY et al. (2013) Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. J. Chem. Inf. Model 53, 475–492 [PubMed: 23252936]

111. Ribay K et al. (2016) Predictive modeling of estrogen receptor binding agents using advanced cheminformatics tools and massive public data. Front. Environ. Sci 4, 12 [PubMed: 27642585]

112. Bharti DR et al. (2019) GCAC: galaxy workflow system for predictive model building for virtual screening. BMC Bioinformatics 19, 550 [PubMed: 30717669]

113. Wenlock MC and Carlsson LA (2015) How experimental errors influence drug metabolism and pharmacokinetic QSAR/QSPR models. J. Chem. Inf. Model 55, 125–134 [PubMed: 25406036]

114. Zhao LL et al. (2017) Experimental errors in QSAR modeling sets: what we can do and what we cannot do. ACS Omega 2, 2805–2812 [PubMed: 28691113]

115. Cortes-Ciriano I et al. (2015) Comparing the influence of simulated experimental errors on 12 machine learning algorithms in bioactivity modeling using 12 diverse data sets. J. Chem. Inf. Model 55, 1413–1425 [PubMed: 26038978]

116. Roy K et al. (2017) How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models? Chemometrics Intell. Lab. Syst 162, 44–54

117. Jing YK et al. (2018) Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. AAPS J. 20, 58 [PubMed: 29603063]

118. Chen HM et al. (2018) The rise of deep learning in drug discovery. Drug Discov. Today 23, 1241–1250 [PubMed: 29366762]

119. Clark E et a! (2019) Advances in deep learning and their applied utility toward chemical informatics & drug discovery. Abstr. Papers Am. Chem. Soc 257, XXX–YYY

120. Fleming N (2018) How artificial intelligence is changing drug discovery. Nature 557, S55–S55 [PubMed: 29849160]

121. Panch T et al. (2018) Artificial intelligence, machine learning and health systems. J. Global Health 8, XXX–YYY

122. Samuel AL (1959) Some studies in machine learning using the game of checkers. IBM J. Res. Dev.S, 210–229

123. Hansch C and Fujita T (1964) p-σ-π Analysis. A method for the correlation of biological activity and chemical structure. J. Am. Chem. Roc 86, 1616–1626

124. Martin YC (2010) Quantitative Drug Design: A Critical Introduction, CRC Press

125. Gozalbes R et al. (2002) Application of topological descriptors in QSAR and drug design: history and new trends. Curr. Drug Targets Infect. Disord 2, 93–102 [PubMed: 12462157]

126. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. Drug Discov. Today 11, 1046–1053 [PubMed: 17129822]

127. McGregor MJ and Muskal SM (1999) Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. J. Chem. Inf. Comput. Sci 39, 569–574 [PubMed: 10361729]

128. Golbraikh A and Tropsha A (2002) Beware of q2! J. Mol Graph Model 20, 269–276 [PubMed: 11858635]

129. Tetko IV et al. (2008) Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection. J. Chem. Inf. Model 48, 1733–1746 [PubMed: 18729318]

130. Zhu H et al. (2009) Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. Chem. Res. Toxicol 22, 1913–1921 [PubMed: 19845371]

131. Zhu H et al. (2008) Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. J. Chem. Inf. Model 48, 766–784 [PubMed: 18311912]

132. Tropsha A and Golbraikh A (2007) Predictive QSAR Modeling workflow, model applicability domains, and virtual screening. Curr. Pharm. Design 13, 3494–3504

133. Shoichet BK (2004) Virtual screening of chemical libraries. Nature 432, 862–865 [PubMed: 15602552]

134. Clark DE (2008) What has virtual screening ever done for drug discovery? Expert Opin. Drug Discov 3, 841–851 [PubMed: 23484962]

135. Schneider G (2010) Virtual screening: an endless staircase? Nat. Rev. Drug Discov 9, 273–276 [PubMed: 20357802]

136. Zhang L et al. (2013) Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. J. Chem. Inf. Model 53, 475–492 [PubMed: 23252936]

137. Prathipati P and Saxena AK (2006) Evaluation of binary QSAR models derived from LUDI and MOE scoring functions for structure based virtual screening. J. Chem. Inf. Model 46, 39–51 [PubMed: 16426038]

138. Tropsha A (2006) Variable selection QSAR modeling, model validation, and virtual screening. Annu. Rep. Comput. Chem 2, 113–126

139. Low Y et al. (2011) Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. Chem. Res. Toxicol 24, 1251–1262 [PubMed: 21699217]

140. Zakharov AV et al. (2016) QSAR modeling and prediction of drug-drug interactions. Mol. Pharm 13, 545–556 [PubMed: 26669717]

141. Mak KK and Pichika MR (2019) Artificial intelligence in drug development: present status and future prospects. Drug Discov. Today 24, 773–780 [PubMed: 30472429]

142. Wolber G et al. (2008) Molecule-pharmacophore superpositioning and pattern matching in computational drug design. Drug Discov. Today 13, 23–29 [PubMed: 18190860]

143. Schneider G and Fechner U (2005) Computer-based de novo design of drug-like molecules. Nat. Rev. Drug Discov 4, 649–663 [PubMed: 16056391]

144. Schneider G (2018) Generative models for artificially-intelligent molecular design. Mol. Informatics 37, XXX–YYY

145. Merk D et al. (2018) De novo design of bioactive small molecules by artificial intelligence. Mol. Informatics 37, XXX–YYY

146. Merk D et al. (2018) Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. Commun. Chem 1, 1–9

147. Schneider P et al. (2020) Rethinking drug design in the artificial intelligence era. Nat. Rev. Drug Discov 19, 353–364 [PubMed: 31801986]

148. Nascimento ACA et al. (2016) A multiple kernel learning algorithm for drug-target interaction prediction. BMC Bioinformatics 17, 46 [PubMed: 26801218]

149. Menden MP et al. (2013) Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS ONE 8, XXX–YYY

150. Matthews H et al. (2016) 'Omics'-informed drug and biomarker discovery: opportunities, challenges and future perspectives. Proteomes 4, 28

151. Hamet P and Tremblay J (2017) Artificial intelligence in medicine. Metabolism 69, S36–S40

152. Wishart DS (2016) Emerging applications of metabolomics in drug discovery and precision medicine. Nat. Rev. Drug Discov 15, 473–484 [PubMed: 26965202]

153. Aoyama T et al. (1989) Neural networks applied to pharmaceutical problems .1. Method and application to decision-making. Chem. Pharm. Bull 37, 2558–2560

154. Duch W et al. (2007) Artificial intelligence approaches for rational drug design and discovery. Curr. Pharm. Design 13, 1497–1508

155. Baskin II et al. (2016) A renaissance of neural networks in drug discovery. Expert Opin. Drug Discov 11, 785–795 [PubMed: 27295548]

156. Gawehn E et al. (2016) Deep learning in drug discovery. Mol. Informatics 35, 3–14

157. Roy K and Roy PP (2009) Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. Eur. J. Med. Chem 44, 2913–2922 [PubMed: 19128860]

158. Simmons K et al. (2008) Comparative study of machine-learning and chemometric tools for analysis of in-vivo high-throughput screening data. J. Chem. Inf. Model 48, 1663–1668 [PubMed: 18681397]

159. Xie L et al. (2017) Harnessing big data for systems pharmacology. Annu. Rev. Pharmacol. Toxicol 57, 245–262 [PubMed: 27814027]

160. Ma JS et al. (2015) Deep neural nets as a method for quantitative structure-activity relationships. J. Chem. Inf. Model 55, 263–274 [PubMed: 25635324]

161. Bjornsson ES et al. (2017) Azathioprine and 6-mereaptopurine induced liver injury: clinical features and outcomes. J. Clin. Gastroenterol 51, 63 [PubMed: 27648552]

162. Kearsley SK et al. (1996) Chemical similarity using physiochemical property descriptors. J. Chem. Inf. Comp. Sci 36, 118–127

163. Huang RL et al. (2016) Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. Front. Environ. Sci 3, 85

164. Mayr A et al. (2016) DeepTox: toxicity prediction using deep learning. Front. Environ. Sci 3, 80

165. Lusci A et al. (2013) Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. J. Chem. Inf. Model 53, 1563–1575 [PubMed: 23795551]

166. Xu YJ et al. (2015) Deep learning for drug-induced liver injury. J. Chem. Inf. Model 55, 2085–2093 [PubMed: 26437739]

167. Hop P et al. (2018) Geometric deep learning autonomously learns chemical features that outperform those engineered by domain experts. Mol. Pharm 15, 4371–4377 [PubMed: 29863875]

168. Lane T et al. (2018) Comparing and validating machine learning models for Mycobacterium tuberculosis drug discovery. Mol. Pharm 15, 4346–4360 [PubMed: 29672063]

169. Russo DP et al. (2018) Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction. Mol. Pharm 15, 4361–4370 [PubMed: 30114914]

170. Sanchez-Lengeling B et al. (2017) Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). Journal XX, XXX–YYY

171. Gomez-Bombarelli R et al. (2018) Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Sci. 4, 268–276

172. Segler MHS et al. (2018) Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Central Sci. 4, 120–131

173. Popova M et al. (2018) Deep reinforcement learning for de novo drug design. Sci. Adv 4, XXX–YYY

174. Zhavoronkov A (2018) Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry. Mol. Pharm 15, 4311–4313 [PubMed: 30269508]

175. Wen M et al. (2017) Deep-learning-based drug-target interaction prediction. J. Proteome Res 16, 1401–1409 [PubMed: 28264154]

176. Xie L et al. (2018) Deep learning-based transcriptome data classification for drug-target interaction prediction. BMC Genomics 19, 667 [PubMed: 30255785]

177. Donner Y et al. (2018) Drug repurposing using deep embeddings of gene expression profiles. Mol. Pharm 15, 4314–4325 [PubMed: 30001141]

178. Cai C et al. (2019) Deep learning-based prediction of drug-induced cardiotoxicity. J. Chem. Inf. Model 59, 1073–1084 [PubMed: 30715873]

179. Wenzel J et al. (2019) Predictive multitask deep neural network models for ADME-Tox properties: learning from large data sets. J. Chem. Inf. Model 59, 1253–1268 [PubMed: 30615828]

180. Li X et al. (2018) Prediction of human cytochrome P450 inhibition using a multitask deep autoencoder neural network. Mol. Pharm 15, 4336–4345 [PubMed: 29775322]

181. Griffen EJ et al. (2018) Can we accelerate medicinal chemistry by augmenting the chemist with Big Data and artificial intelligence? Drug Discov. Today 23, 1373–1384 [PubMed: 29577971]

182. Sterling T and Irwin JJ (2015) ZINC 15-ligand discovery for everyone. J. Chem. Inf. Model 55, 2324–2337 [PubMed: 26479676]

183. Pence HE and Williams A (2010) ChemSpider: an online chemical information resource. J. Chem. Educ 87, 1123–1124

184. Chevillard F and Kolb P (2015) SCUBIDOO: a large yet screenable and easily searchable database of computationally created chemical compounds optimized toward high likelihood of synthetic tractability. J. Chem. Inf. Model 55, 1824–1835 [PubMed: 26282054]

185. Zhang RZ et al. (2017) TCM-Mesh: the database and analytical system for network pharmacology analysis for TCM preparations. Sci. Rep 7, 2821 [PubMed: 28588237]

186. Banerjee P et al. (2015) Super Natural II--a database of natural products. Nucleic Acids Res. 43 (Database issue), D935–939 [PubMed: 25300487]

187. Singla D et al. (2010) BIAdb: a curated database of benzylisoquinoline alkaloids. BMC Pharmacol. 10, 4 [PubMed: 20205728]

188. Janes J et al. (2018) The ReFRAME library as a comprehensive drug repurposing library and its application to the treatment of cryptosporidiosis. Proc. Natl. Acad. Sci. U. S. A 115, 10750–10755 [PubMed: 30282735]

189. Siramshetty VB et al. (2018) SuperDRUG2: a one stop resource for approved/marketed drugs. Nucleic Acids Res. 46, D1137–D1143 [PubMed: 29140469]

190. Dimitropoulos D et al. (2006) Using MSDchem to search the PDB ligand dictionary. Curr. Protoc. Bioinformatics 15, 14.13.11–14.13.21

191. Wang R et al. (2005) The PDBbind database: methodologies and updates. J. Med. Chem 48, 4111–4119 [PubMed: 15943484]

192. Kuhn M et al. (2008) STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res. 36, D684–688 [PubMed: 18084021]

193. Chatr-Aryamontri A et al. (2017) The BioGRID interaction database: 2017 update. Nucleic Acids Res. 45, D369–D379 [PubMed: 27980099]

194. Hu LG et al. (2005) Binding MOAD (Mother of All Databases). Proteins 60, 333–340 [PubMed: 15971202]

195. Pandy-Szekeres G et al. (2018) GPCRdb in 2018: adding GPCR structure models and ligands. Nucleic Acids Res. 46, D440–D446 [PubMed: 29155946]

196. Armstrong JF et al. (2020) The IUPHAR/BPS Guide to Pharmacology in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to Malaria Pharmacology. Nucleic Acids Res. 48, D1006–D1021 [PubMed: 31691834]

197. Frolkis A et al. (2010) SMPDB: The Small Molecule Pathway Database. Nucleic Acids Res. 38, D480–487 [PubMed: 19948758]

198. Karp PD et al. (2019) The BioCyc collection of microbial genomes and metabolic pathways. Brief. Bioinform 20, 1085–1093 [PubMed: 29447345]

199. King ZA et al. (2016) BiGG models: a platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 44, D515–522 [PubMed: 26476456]

200. Schomburg I et al. (2002) BRENDA: a resource for enzyme data and metabolic information. Trends Biochem. Sci 27, 54–56 [PubMed: 11796225]

201. Croft D et al. (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 39, D691–697 [PubMed: 21067998]

202. Chelliah V et al. (2015) BioModels: ten-year anniversary. Nucleic Acids Res. 43, D542–548 [PubMed: 25414348]

203. Mathias SL et al. (2013) The CARLSBAD database: a confederated database of chemical bio activities. Database 2013, bat044 [PubMed: 23794735]

204. Voigt JH et al. (2001) Comparison of the NCI open database with seven large chemical structural databases. J. Chem. Inf. Comput. Sci 41, 702–712 [PubMed: 11410049]

205. Ihlenfeldt WD et al. (2002) Enhanced CACTVS browser of the Open NCI Database. J. Chem. Inf. Comput. Sci 42, 46–57 [PubMed: 11855965]

206. Mangal M et al. (2013) NPACT: Naturally Occurring Plant-based Anti-cancer Compound-Activity-Target database. Nucleic Acids Res. 41, D1124–1129 [PubMed: 23203877]

207. Wishart D et al. (2015) T3DB: the toxic exposome database. Nucleic Acids Res. 43, D928–934 [PubMed: 25378312]

208. Lim E et al. (2010) T3DB: a comprehensively annotated database of common toxins and their targets. Nucleic Acids Res. 38, D781–D786 [PubMed: 19897546]

209. Judson R et al. (2008) ACTOR - Aggregated Computational Toxicology Resource. Toxicol. Appl. Pharmacol 233, 7–13 [PubMed: 18671997]

210. Wang CC et al. (2017) SkinSensDB: a curated database for skin sensitization assays. J. Cheminformatics 9, 1–6

211. Davis AP et al. (2019) The Comparative Toxicogenomics Database: update 2019. Nucleic Acids Res. 47, D948–D954 [PubMed: 30247620]

212. Neveu V et al. (2020) Exposome-Explorer 2.0: an update incorporating candidate dietary biomarkers and dietary associations with cancer risk. Nucleic Acids Res. 48, D908–D912 [PubMed: 31724701]

Highlights

- Drug discovery has been advanced to a big data era with a large amount of public data sources available.
- Ten V features (*volume, velocity, variety, veracity, validity, vocabulary, venue, visualization, volatility*, and *value*) bring new challenges to machine learning modeling.
- Recent progress of machine learning to deep learning and the development of new algorithms answers the big data challenges.

Box 1.

Ten Vs features of big data in drug discovery

Volume: size of data.

Velocity: speed of new data generation.

Variety: various formats of data.

Veracity: quality of data.

Validity: authenticity of data.

Vocabulary: terminology of data.

Venue: platform of data generation.

Visualization: view of data.

Volatility: duration of data usefulness.

Value: potential of data usefulness to reduce the cost of drug discovery and development.

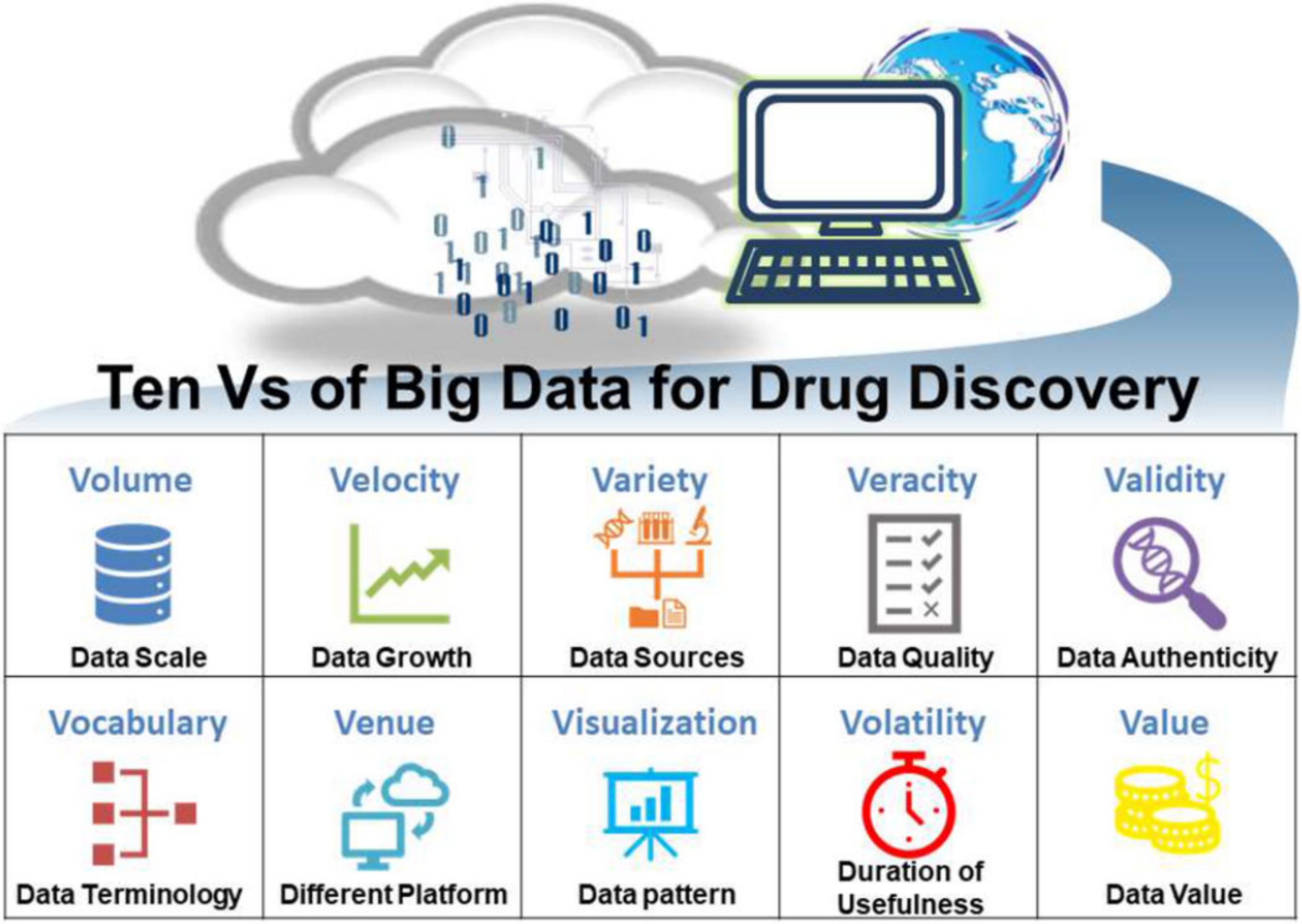

Figure 1.

Ten Vs scheme of ‘big data’ to assist the drug discovery and development.

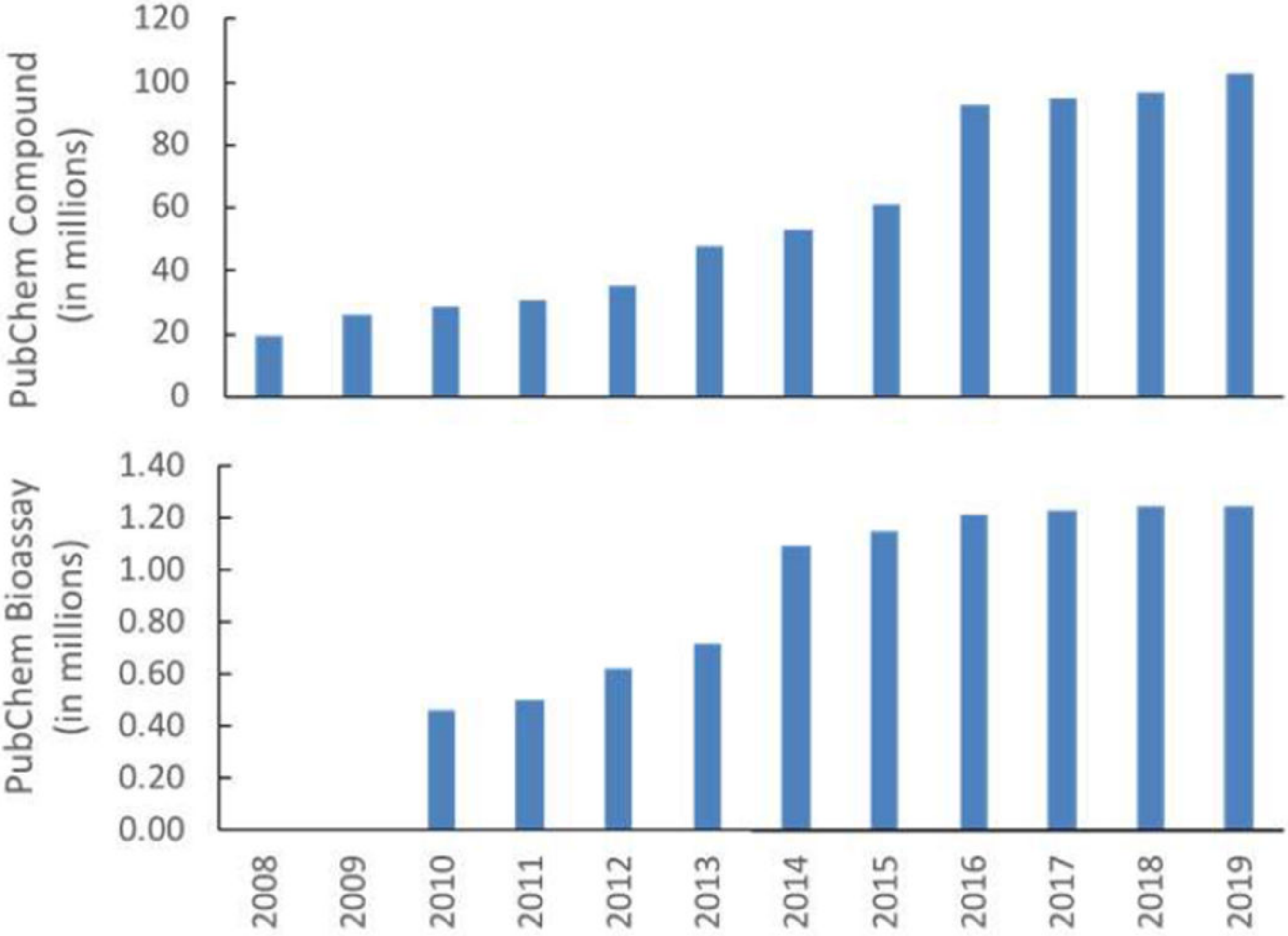

Figure 2.

The number of compounds and bioassays increase in PubChem within 12 years.

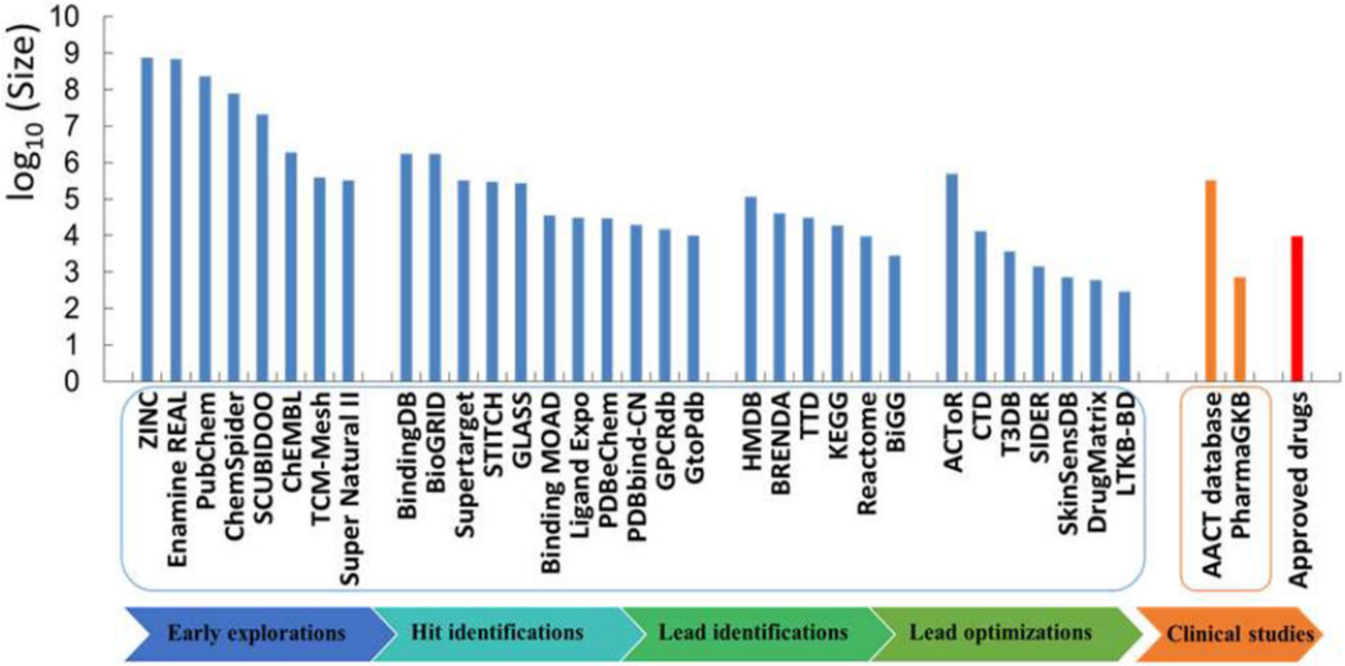

Figure 3.

Size of available databases at different stages of drug discovery and development. The definition of the size of these databases was based primarily on the number of molecules stored in the database. The sizes of BindingDB, Supertarget, Binding MOAD, PDBbind-CN, AACT database, PharmaGKB and Approved drugs were defined by the data entries provided by the databases.

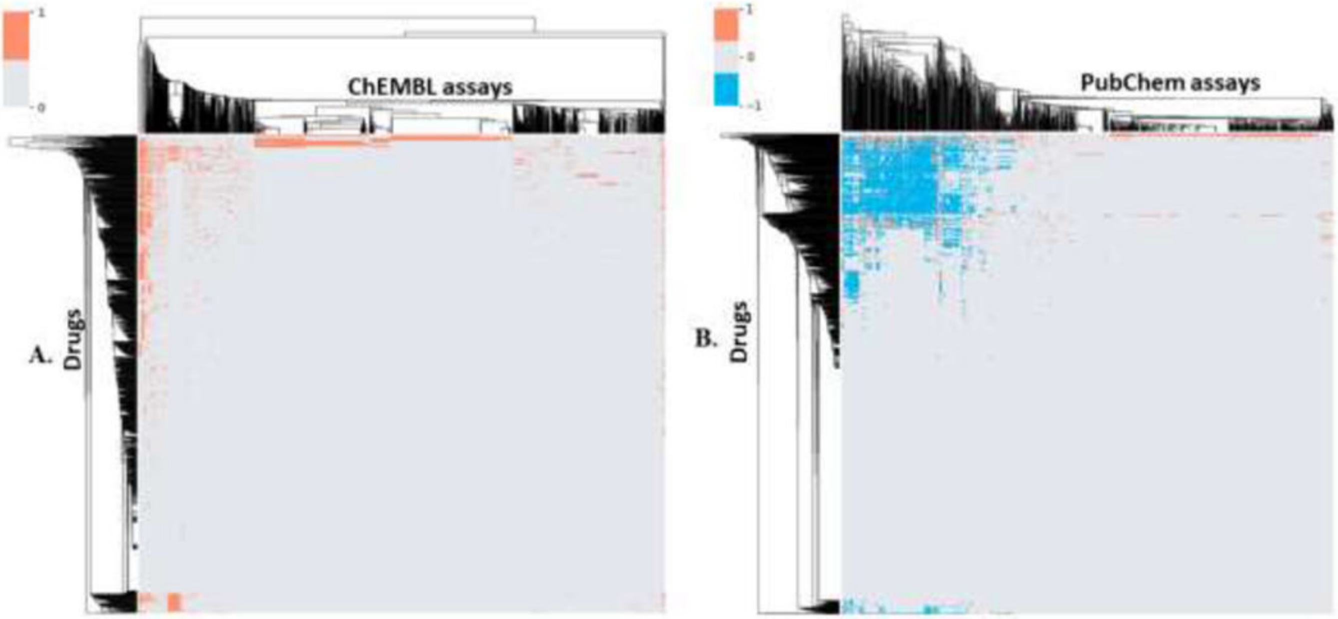

Figure 4.

Biological data profiles of 1930 US Food and Drug Administration (FDA)-approved drugs represented by data from ChEMBL and PubChem. **(a)** Data obtained from 1114 ChEMBL assays, which have at least 25 testing results (red spots) among these compounds; **(b)** Data obtained from 299 PubChem assays, which have at least 25 active responses (red spots) among these compounds. The gray spots indicate missing data (no data or 'inconclusive' results) and the blue spots indicate inactives. The 'active', 'inactive', and 'inconclusive' results for a specific assay were defined individually in PubChem, which could be found using its AID. For example, for PubChem AID 928, the active compounds have a PUBCHEM_ACTIVITY_SCORE between 40 and 100, inconclusive compounds have a PUBCHEM_ACTIVITY_SCORE between 1 and 39, and all inactive compounds have a PUBCHEM_ACTIVITY_SCORE of 0.

Table 1.

Current publicly available databases that can be used for drug discovery and development

Database	Description	Size (as of 29 October 2019)	Link	Refs
Chemical collections				
Enamine REAL Database	Tool used to find new hit molecules using large-scale virtual screening and for searching analogs to hit molecules	>700 million compounds that comply with 'rule of 5' and Verber criteria	https://enamine.net/hit-finding/compound-collections/real-database	[29]
ZINC	Contains compound information including 2D/3D structure, purchasability, target, and biology-related information	>230 million compounds in 3D formats and >750 million compounds for analog-searching	http://zinc.docking.org/	[182]
PubChem	Contains chemical molecule (mostly small molecule) information, including chemical structures, identifiers, chemical and physical properties, biological activities, safety and toxicity data	97 million compounds, 236 million substances, 268 million bioactivities	https://pubchem.ncbi.nlm.nih.gov/	[31]
ChemSpider	Free chemical structure database providing fast access to >67 million structures, properties, and associated information	>78 million compound structures	www.chemspider.com/	[183]
SCUBIDOO	Freely accessible database that currently holds 21 million virtual products originating from small library of building blocks and collection of robust organic reactions	21 million virtual products	http://kolblab.org/scubidoo/index.php	[184]
ChEMBL	Manually curated database of bioactive molecules with drug-like properties; brings together chemical, bioactivity, and genomic data to aid translation of genomic information into effective new drugs.	>1.9 million compounds, 1.1 million pieces of assay information	www.ebi.ac.uk/chembl/	[30]
TCM-Mesh	Integration of a database and a data-mining system for network pharmacology analysis of all respects of traditional Chinese medicine, including herbs, herbal ingredients, targets, related diseases, adverse effect, and toxicity	383 840 compounds, 6235 herbs	http://mesh.tcm.microbioinformatics.org/	[185]
Super Natural II	Contains natural compounds, including information about corresponding 2D structures, physicochemical properties, predicted toxicity class and potential vendors	325 508 natural compounds	http://bioinf-applied.charite.de/supernatural_new/index.php	[186]
BIAdb	Comprehensive database of benzylisoquinoline alkaloids, containing information about ~846 unique benzylisoquinoline alkaloids	~846 unique benzylisoquinoline alkaloids	https://webs.iiitd.edu.in/raghava/biadb/index.html	[187]
Drug/drug-like compounds				
AICD	Anti-Inflammatory Compounds Database (AICD) deposits compounds with potential anti-inflammation activities	79 781 small molecules	http://956023.ichengyun.net/AICD/index.php	[52]
Drug Bank	Unique bioinformatics and cheminformatics resource that combines detailed drug data with	13 441 drug entries	www.drugbank.ca/	[51]

Database	Description	Size (as of 29 October 2019)	Link	Refs
	comprehensive drug target information			
ReFRAME	Screening library of 12 000 molecules assembled by combining three databases (Clarivate Integrity, GVK Excelra GoStar, and Citeline Pharmaprojects) to facilitate drug repurposing	12 000 molecules	https://reframedb.org/	[188]
SuperDRUG2	Contains approved/marketed drugs with regulatory details, chemical structures (2D and 3D), dosage, biological targets, physicochemical properties, external identifiers, adverse effects, and PK data	>4600 active pharmaceutical ingredients	http://cheminfo.charite.de/superdrug2/	[189]
Drugs@FDA database	Information about drugs from FDA	~23 391 drug application records	www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files	
e-Drug3D	Contains 1930 molecular structures approved by FDA between 1939 and 2019 with a molecular weight <2000	1930 drugs	https://chemoinfo.ipmc.cnrs.fr/MOLDB/index.php	[53]
Drug targets, including genomics and proteomics data				
BindingDB	Public, web-accessible database of measured binding affinities, focusing chiefly on interactions of proteins considered to be candidate drug targets with ligands that are small, drug-like molecules	1 756 093 binding data, for 7371 protein targets and 780 240 small molecules	/www.bindingdb.org/bind/index.jsp	[54]
Supertarget	It is an extensive web resource for analysing drug-target interactions.	332 828 drug-target interactions	http://insilico.charite.de/supertarget/index.php?site=home	[55]
Ligand Expo	Provides chemical and structural information about small molecules within structure entries of Protein Data Bank.	30 440 entries of ligand	http://ligand-expo.rutgers.edu/index.html	[56]
PDBeChem	Consistent and enriched library of ligands, small molecules, and monomers referenced as residues and hetgroups in PDB entries	>29 922 ligands	www.ebi.ac.uk/pdbe-srv/pdbechem/	[190]
PDBbind-CN	Provides essential linkage between energetic and structural information of biomolecular complexes, which is helpful for various computational and statistical studies on molecular recognition in biological systems	19 588 biomolecular complexes	www.pdbbind-cn.org/	[191]
STITCH	Database integrating information about interactions from metabolic pathways, crystal structures, binding experiments, and drug–target relationships	Interactions between 300 000 small molecules and 2.6 million proteins from 1133 organisms	http://stitch.embl.de/	[192]
BioGRID	The Biological General Repository for Interaction Datasets is an open-access database on protein, genetic, and chemical interactions for humans and all major model organisms	1 753 686 protein and genetic interactions, 28 093 chemical associations and 874 796 post-translational modifications from major model organisms	https://thebiogrid.org/	[193]
Binding MOAD	Created from a subset of Protein Data Bank (PDB), containing every high-quality example of ligand–protein binding.	36 047 protein–ligand structures, and 13 353 binding data	http://bindingmoad.org/	[194]
GPCRdb	Contains data from GPCRs, including crystal structures, sequence alignments, and receptor mutations;	15 149 proteins, and 144 917 ligands	www.gpcrdb.org	[195]

Database	Description	Size (as of 29 October 2019)	Link	Refs
	can be visualized in interactive diagrams; provides online analysis tools			
Guide to Pharmacology	IUPHAR/BPS Guide to PHARMACOLOGY is an open-access, expert-curated database of molecular interactions between ligands and their targets.	2937 targets, and 9859 ligands	www.guidetopharmacology.org/	[196]
GLASS	GPCR-Ligand Association (GLASS) database is a manually curated repository for experimentally validated GPCR–ligand interactions; along with relevant GPCR and chemical information, GPCR–ligand association data are extracted and integrated into GLASS from literature and public databases	~277 651 unique ligands and 3048 GPCRs	https://zhanglab.ccmb.med.umich.edu/GLASS/	
Biological data from assay screening, metabolism, and efficacy studies				
HMDB	Freely available electronic database containing detailed information about small-molecule metabolites found in human body	114 162 metabolite entries	www.hmdb.ca/about	[57]
SMPDB	Small Molecule Pathway Database (SMPDB) is an interactive, visual database containing >30 000 small-molecule pathways found in humans only	>30 000 small-molecule pathways	http://smpdb.ca/	[197]
TTD	Therapeutic Target Database (TTD) is a database providing information about known and explored therapeutic protein and nucleic acid targets, targeted disease, pathway information and corresponding drugs directed at each of these targets	2589 targets, and 31 614 drugs	http://db.idrblab.net/ttd/	[58]
BioCyc	Collection of 7615 pathway/genome databases; each database in BioCyc collection describes genome and metabolic pathways of a single organism	7615 pathway/genome databases	https://biocyc.org/	[198]
BiGG	Metabolic reconstruction of human metabolism designed for systems biology simulation and metabolic flux balance modeling	2004 proteins, 2766 metabolites, and 3311 metabolic and transport reactions	http://bigg.ucsd.edu/	[199]
BRENDA	Main collection of enzyme functional data available to scientific community	At least 40 000 different enzymes from >6900 different organisms	www.brenda-enzymes.org/	[200]
Reactome	Curated, peer-reviewed knowledgebase of biological pathways, including metabolic pathways, and protein trafficking and signaling pathways	>9600 proteins, 9800 reactions, and 2000 pathways for humans	https://reactome.org/	[201]
BioModels Database	Repository of computational models of biological processes; models described from literature are manually curated and enriched with cross-references	6753 patient-derived genome-scale metabolic models, 112 898 metabolic models etc.	www.ebi.ac.uk/biomodels-main/	[202]
KEGG	Database resource that integrates genomic, chemical, and systemic functional information	18 652 metabolites	https://www.genome.jp/kegg/	[202]
CARLSBAD	Database and knowledge inference system that integrates multiple	932 852 CARLSBAD activities, 890 323	http://carlsbad.health.unm.edu/carlsbad/?mode=home	[203]

Database	Description	Size (as of 29 October 2019)	Link	Refs
	bioactivity data sets to provide researchers with novel capabilities for mining and exploration of available structure activity relationships (SAR) throughout chemical biology space.	unique structure–target pairs, 3542 targets, 435 343 unique structures		
WOMBAT	Contains 331 872 entries, representing 1966 unique targets, with bioactivity annotations	268 246 unique structures	http://dud.docking.org/wombat/	[59]
Open NCI Database	Maintained by the National Cancer Institute; contains small-molecule information such as names, biological activities, structures; useful resource for researchers working in cancer/ AIDS fields	>250 000 compounds	https://cactus.nci.nih.goV/ncidb2.2/	[204,205]
NPACT	Provides information on plant-derived natural compound, including structure, properties (physical, elemental, and topological), cancer type, cell lines, inhibitory values (IC50, ED50, EC50, GI50), molecular targets, commercial suppliers, and drug likeness of compounds	1574 entries	http://crdd.osdd.net/raghava/npact/	[206]
PKPB_DB	Contains physiological parameter values for humans from early childhood through senescence; intended to be used in physiologically based (PB)PK modeling; also contains similar data for animals (primarily rodents)	N/A	https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=204443	[60]
Drug liabilities and toxicities				
T3DB	Unique bioinformatics resource that combines detailed toxin data with comprehensive toxin target information	3678 toxins	www.t3db.ca/	[207,208]
DrugMatrix	One of world's largest toxicogenomic reference resources	~600 drug molecules and 10 000 genes	https://ntp.niehs.nih.gov/data/drugmatrix/	[61]
ACToR	Includes computational toxicology information about compounds, including HTS, chemical exposure, sustainable chemistry (chemical structures and physicochemical properties) and virtual tissue data	>500 000 chemicals	https://actor.epa.gov/actor/home.xhtml	[209]
SkinSensDB	Contains curated data from published AOP-related skin sensitization assays	710 unique chemicals	https://cwtung.kmu.edu.tw/skinsensdb/	[210]
SIDER	Contains information on marketed medicines and their recorded adverse drug reactions, including frequency, drug and adverse effect classifications	1430 drugs with 5868 side effect information	http://sideeffects.embl.de/download/	[62,63]
LTKB Benchmark Dataset	Contains drugs with potential to cause drug-induced liver injury in humans; established using FDA-approved prescription drug labels	287 prescription drugs	www.fda.gov/science-research/liver-toxicity-knowledge-base-ltkb/ltkb-benchmark-dataset	[64]
CTD	Comparative Toxicogenomics Database (CTD) is a premier public resource for literature-based, manually curated associations between chemicals, gene products, phenotypes, diseases, and environmental exposures	13 378 unique chemicals and related information	http://ctdbase.org/	[211]

Database	Description	Size (as of 29 October 2019)	Link	Refs
Clinical databases				
ClinicalThals.gov	Database of privately and publicly funded clinical studies conducted around the world	~324 429 research studies in all 50 US states and 209 countries	https://clinicaltrials.gov/	[65]
AACT database	Publicly available relational database that contains all information (protocol and result data elements) about every study registered in ClinicalTrials.gov. Content is downloaded from ClinicalTrials.gov daily and loaded into AACT	~324 429 research studies in all 50 US states and 209 countries	https://aact.ctti-clinicaltrials.org/	[65]
EORTC Clinical Trials Database	Contains information about EORTC clinical trials and clinical trials from other organizations with EORTC participation	N/A	www.eortc.org/clinical-trials/	
Exposome-Explorer	Contains detailed information on nature of biomarkers, populations and subjects where measured, samples analyzed, methods used for biomarker analyses, concentrations in biospecimens, correlations with external exposure measurements, and biological reproducibility over time	908 biomarkers	http://exposome-explorer.iarc.fr/	[212]
PharmaGKB	A pharmacogenomics knowledge resource that encompasses clinical information about drug molecules	733 drugs with their clinical information	www.pharmgkb.org/	[66]