Review Article

Big data and artificial intelligence (AI) methodologies for computer-aided drug design (CADD)

Jai Woo Lee, Miguel A. Maria-Solano, Thi Ngoc Lan Vu, Sanghee Yoon and Sun Choi

Global AI Drug Discovery Center, College of Pharmacy and Graduate School of Pharmaceutical Sciences, Ewha Womans University, Seoul 03760, Republic of Korea

There have been numerous advances in the development of computational and statistical methods and applications of big data and artificial intelligence (AI) techniques for computer-aided drug design (CADD). Drug design is a costly and laborious process considering the biological complexity of diseases. To effectively and efficiently design and develop a new drug, CADD can be used to apply cutting-edge techniques to various limitations in the drug design field. Data pre-processing approaches, which clean the raw data for consistent and reproducible applications of big data and AI methods are introduced. We include the current status of the applicability of big data and AI methods to drug design areas such as the identification of binding sites in target proteins, structure-based virtual screening (SBVS), and absorption, distribution, metabolism, excretion and toxicity (ADMET) property prediction. Data pre-processing and applications of big data and AI methods enable the accurate and comprehensive analysis of massive biomedical data and the development of predictive models in the field of drug design. Understanding and analyzing biological, chemical, or pharmaceutical architectures of biomedical entities related to drug design will provide beneficial information in the biomedical big data era.

Introduction

Drug design and discovery is a complicated, costly and laborious process considering the complexity of diseases. It involves the identification of potential targets and the development of therapeutically safe and effective drugs [1–3]. The process can benefit from computer-aided drug design (CADD), where various computational and statistical methods can be applied to effectively analyze biomedical entities for target identification and hit hunting [4,5]. CADD can further utilize the combined biochemical space to gain safety, efficacy and avoid toxicity for the completion of drug development. With the adoption of in silico techniques in academia, industry and government [6,7], significant progress has been made in drug design and discovery. Recently, with the growth of big data in biological, chemical and pharmaceutical medicine, various machine learning algorithms have been optimized and applied in the field of CADD. This integration offers significant improvement in the efficiency of drug design and discovery process. Successful applications in drug design, discovery and development can be achieved only when effective computational methods and tools are provided with accurate and reliable pre-processed data [8,9]. Hereafter, big data and artificial intelligence (AI) approaches to data pre-processing [10], modeling [11,12] and representative applications in drug design and discovery will be introduced.

Big data and AI methods in the drug discovery process

The limitations in the traditional drug discovery field caused by size and complexity of biomedical data can be computationally formulated and solved with the advent of computing and analysis techniques using big data and AI algorithms [13,14]. Big data and AI approaches covering pre-processing data, applications of AI algorithms and statistical methods help to build automated models to analyze protein three-dimensional structures, drug-receptor interactions, ADMET property prediction, etc. [15] (Figure 1).
Pre-processing and understanding data in CADD

The pre-processing steps are crucial for properly understanding and analyzing biochemical data and most importantly providing reliable data in the development of predictive models. For biomedical big data analysis, a data matrix with n samples and p biomedical features is considered. In the data matrix, p features can be any biomedical entities such as molecular descriptors, fingerprints, genes, sequence positions, protein structures, metabolites, etc. To statistically pre-process data, missing data imputation, outlier detection and redundant feature elimination are used (Figure 1). Implementing methods for data pre-processing requires effective algorithms for the accuracy of prediction and efficiency to run the program with optimized speed. The most relevant R packages for data pre-processing are listed in Table 1.

Missing data imputation

AI models, which learn the patterns and structures of sparse drug discovery data, often include insufficient information about the data. The size of experimental values in sparse drug discovery data may not be sufficient, and if a model is trained on sparse data, the prediction of an outcome using this model may lead to inaccurate or inconsistent prediction results. Filling missing values with an imputation model handling molecular descriptors will improve AI models to analyze drug discovery data.

Since there are very few methods imputing missing values in drug discovery data analysis, a missing data imputation model such as Alchemite [16], a novel application of neural network, can be utilized to replace

---

**Table 1 Software programs for data pre-processing**

| Application                        | Method             | Software program | Link                                                        |
|------------------------------------|--------------------|------------------|-------------------------------------------------------------|
| Missing data imputation            | Neural Networks (NN)| Alchemite [23]   | https://intellegens.ai/products-services/alchemite-analytics/|
| Outlier detection                  | Neural Networks (NN)| Alchemite [23]   | https://intellegens.ai/products-services/alchemite-analytics/|
| Redundant feature elimination      | Random Forest (RF)  | RGIFE [22]       | http://ico2s.org/software/rgife.html                         |
missing values. In missing data imputation, the Alchemite method clearly outperforms random forest models, which present uncertainties. Alchemite deep learning imputation improving the prediction model has been proved to outperform collective matrix factorization, deep neural network or random forest when using sparse experimental ADMET data. Alchemite can estimate uncertainties on outcome prediction detecting assay activities [16].

**Outlier detection**

Data values in drug discovery datasets, such as the quantitative structure-activity relationship (QSAR) model, can be grouped by similarity using standard statistical methods. Identifying outlier compounds based on the standardization techniques can have a great impact on the QSAR model [17]. If data values that do not follow patterns in the data (i.e., outliers) are included or significant data values are excluded as outliers, the constructed model would lead to wrong predictions. For reliable prediction results, the molecular datasets, which are used to build the prediction model should cover the chemical space, and a new compound outside the applicability domain of molecule dataset should be detected [18,19]. Hence, outliers should be excluded before building the prediction model. There are very few QSAR models setting a reliable approach using dataset that includes potential outliers. Alchemite algorithm can also be used to detect potential outliers in the drug discovery data [16]. In this process, features following patterns in the data can be clustered and outliers excluded from the clustering procedure will be detected. Thus, Alchemite software program can conduct both missing data imputation and outlier detection to impute missing values and detect extreme values not following the patterns in data.

**Redundant feature elimination**

When the prediction model selects multiple significant features in the dataset, selecting redundant features such as highly correlated variables in statistical analysis and biological meaning can lead to a misinterpretation of the model analysis. It is essential to exclude redundant features for the appropriate comprehension of predictive models [20]. For instance, redundant feature elimination based on the information of target proteins in drug−protein interactions can avoid the class imbalance problem and remove the repeated features [21] to determine the best filtered significant molecular features. RGIFE, a ranked guided iterative feature elimination method [22], which iteratively removes the redundant features in the drug discovery data can be utilized. By removing redundant features and selecting relatively small set of relevant features, RGIFE helps machine learning classifiers to obtain a similar or better performance. RGIFE utilizing RF (Random Forest) algorithm can recursively select significant features by removing redundant features in the data. It is shown that over different biomedical datasets, RGIFE produces similar or better results compared with other feature selection algorithms such as correlation-based feature selection (CFS), Support Vector Machine Recursive Feature Elimination (SVM-RFE), ReliefF, Chi-Square and L1-based feature selection. The features selected by RGIFE were proven to produce relevant findings from a biological point of view [22].

**AI-based modeling methods**

In this section, AI methods for the construction of predictive models are introduced. In particular, we focus on regression methods, which build the models for the prediction of continuous outcomes; classification methods, which build the model for prediction of different classes; clustering methods, which group features based on similarity or distance between two features; and dimensionality reduction methods, which extract low-dimensional data consisting of significant features from high-dimensional data (Figure 1). Regression and classification belong to supervised learning, which estimates the outcome learning the structure of the input data. Clustering and dimensionality reduction belong to unsupervised learning, which investigates the interaction of features in the input data. Regarding regression and classification approaches, given input data and a target outcome, each model can be trained to learn the data to predict an outcome involving testing and possibly validation processes; training data is a portion of the input data used to build a model whereas testing data is a portion of the input data used to test and validate the performance of the model. It should be noted that most AI methods can be used for different categories of learning or analysis: As an example, neural networks can be used for clustering, regression and classification, and k-nearest neighbors can be utilized for missing data imputation to pre-process data and for classification of data values. It should be assumed that for each AI-based modeling method, drug discovery data with n rows of samples and p columns of features are used. In Tables 2 and 3, up-to-date software programs using AI-based modeling methods are listed including a brief description and their application in CADD.
Recent applications of big data & AI-driven technologies in CADD

There are several drug design areas where AI technologies have been successfully implemented in CADD. In this section, we focus on three relevant applications for the structure-based drug design processes: the identification of binding sites in target proteins, structure-based virtual screening (SBVS) and prediction of pharmacokinetic (ADME) and toxicity (T) properties. Furthermore, algorithms based on DT with molecular data can be utilized to analyze the effect of FDA-approved drugs, such as drug-induced liver injury and methods including NBC can be used to build frameworks to investigate exposures related to biologically, chemically or...
pharmaceutically diverse compound datasets [52]. AI-based software programs and tools used for these three applications are listed in Table 4. Each application includes different AI modeling methods.

**Identification of binding sites in target proteins**

Protein binding sites are structural elements whereupon drug-like molecules bind and trigger a therapeutic response. The large-scale identification of such binding sites still remains challenging [53,54]. This is in part attributed to the dynamic nature of proteins, which sample a wide range of conformations in solution and often only a fraction of them harbor binding sites. The increasing number of available conformations together with the complexity of protein conformational landscapes make protein data analysis more challenging [55,56].

To search for those pharmaceutically rich protein conformations, several tools have been developed using classical approaches including Fpocket [57], SiteHound [58] and MetaPocket [59]. These tools can predict binding sites considering geometric and potential energy factors of protein surfaces. Different AI methods such as oversampling and binary classification (ENRI) [56], random forest (P2Rank) [53] and most recently deep learning approaches (DeepSite) [41,60] have emerged as potential strategies to enhance the binding pocket identification performance.

In this new scenario, Kozlovskii and Popov [54] developed BiteNet (Binding site neural Network), a rapid and accurate deep learning approach. After a curation procedure, 5,946 protein–ligand complexes containing 11,949 binding sites from Protein data bank [61] were used as training data set for the construction of a neural network model. Protein–ligand complexes structures offer a wide coverage of protein binding pockets, which usually are not detectable in ligand-free structures [62]. In the BiteNet fashion, protein ensembles are treated as 3D videos, protein structures as 3D images and binding sites as objects. This is performed by processing the data in a 3D-CNN as protein multi-channel grids of voxels in which the channels only consider atomic densities. It proves that in absence of geometry and energy descriptors (i.e., by essentially treating proteins as 3D...
images), binding sites can be successfully predicted. In fact, BiteNet significantly outperforms classical binding site prediction methods and state-of-the-art AI methods in terms of predictive power and computational efficiency. The thoughtful curation of the training set and preparation of the training process were found to be the key for the outperformance of BiteNet [54]. Thus, deep learning-based tools can be successfully applied to identify druggable conformations along molecular dynamics (MD) trajectories as input (Figure 2). The detected druggable conformations are advantageous to the following structure-based drug design procedures.

In practice, there are some aspects that need special attention for users when using a deep learning software for binding site prediction. First, it is inevitable that such training sets contain false negatives because protein–ligand complexes may also encompass empty binding sites, as a consequence, the prediction of some binding sites and especially novel allosteric sites could be omitted. Thus, the combination of classical and deep learning approaches could be beneficial by yielding complementary outcomes. Second, it is advisable to consider the applicability domain derived from the training data set. It has been shown that the performance of different

| Application                                  | Method                          | Software program           | Link                                           |
|----------------------------------------------|--------------------------------|-----------------------------|------------------------------------------------|
| Identification of binding sites in target proteins | NBC                            | ENRI [56]                   | Source code: https://github.com/fibonaccirabbits/enri |
|                                              | DT                             | P2Rank [53]                 | Source code: http://github.com/rdk/p2rank       |
|                                              | RF                             | DeepSite [60]               | Web server: www.playmolecule.org/deepsite/      |
|                                              | NN                             | BiteNet [54]                | Data set: https://doi.org/10.5281/zenodo.4043664 |
|                                              |                                |                             | Source code: https://github.com/i-Molecule/bitenet |
|                                              |                                |                             | Web server: https://sites.skoltech.ru/molecule/tools/bitenet/ |
| Structure-based Virtual Screening (SBVS)     | NN                             | DeepBSP [68]                | Source code: https://github.com/BaoJingxiao/DeepBSP |
|                                              | Penalized linear regression/Penalized logistic regression | SAnDReS [81] | Source code: https://github.com/aizedolabol/sandres |
|                                              | RF                             | RF-Score-v3 [82]            | Software: http://istar.cse.cuhk.edu.hk/rf-score-3.tgz  |
|                                              |                                |                             | http://crcm.marseille.inserm.fr/fileadmin/rf-score-3.tgz |
| Prediction of Pharmacokinetics (ADME) and Toxicity (T) | SVM                            | SwissADME [73]              | Web server: http://www.swissadme.ch/             |
|                                              | NBC                            |                             | Web server: http://lmmd.ecust.edu.cn/admetsar2/ |
|                                              | RF                             | admetSAR2.0 [83]            | Web server: http://binadrug.mumc.org/svm/         |
|                                              | SVM                            | vNN-ADMET [76]              | Web server: https://vnnadmet.bhslai.org/vnnadmet/  |
|                                              | kNN                            | AMPL [77]                   | Source code: https://github.com/ATOMconsortium/AMPL |
|                                              | RF                             | ADMETlab [84]               | Web server: https://admet.scbdd.com/home/index/  |
|                                              | NN                             |                             |                                                |
|                                              | RF                             |                             |                                                |
|                                              | SVM                            |                             |                                                |
|                                              | PLSR                           |                             |                                                |
|                                              | NBC                            |                             |                                                |
|                                              | DT                             |                             |                                                |
Figure 2. Overview of the applications workflow and their interconnection.

On the left, the training data set sources and curation process are shown. The data type and the relevant information for the training process of each application are described in the horizontal arrows. On the right, the different neural network models built using the training data are represented. Application i uses protein ensembles as input and provides druggable conformations encompassing binding sites as output. Application ii is fed with the docked complexes between the druggable conformations and drug candidates as input, yielding hit compounds as output. Application iii takes the hit compounds as input to finally obtain the lead compounds.
methods depend on the protein family under study. Third, training sets overlook protein flexibility since they are usually constructed from X-ray rigid structures. This can be addressed with data augmentation techniques by computationally generating ensembles of protein–ligand conformations [63].

**Structure-based virtual screening (SBVS)**

Once druggable protein conformations are identified, one may proceed to obtain good binders from a chemical space of drug candidates that can cause the desired therapeutic effects. Those potent binders are referred to as hit compounds. Molecular interactions between drug candidates and protein binding sites can be virtually simulated using docking techniques. Specifically, in SBVS, a vast number of ligands from chemical libraries are ranked according to their binding affinity, which is predicted by a regression model, known as scoring function (SF) [64].

Recently, a new generation of SFs has been developed, which apply AI to utilize the ever-growing biological and structural data [65]. AI-based SFs continue to show their outperformance over classical SFs, with their ability to learn from low-level features in protein–ligand complexes [66]. In addition, unlike traditional SFs, the flexible nature of AI-based SFs allows customization of training datasets to focus on protein families of interest [67], including additional information to improve predictive performance [66] or diversify outcomes. For instance, instead of binding affinity, AI-based SFs developed in the DeepBSP tool [68] can directly predict the root mean square deviation (RMSD) between the docked and the native binding poses. In DeepBSP, a thoroughly curated dataset of 11,925 native protein–ligand complexes from PDBbind database [69] and more than 165,000 docked poses were represented using 3D voxel grids. These volumetric data, together with respective RMSD values calculated using DockRMSD program were used to train the model with a 3D-CNN structure (Figure 2). This model does not generate ligand–protein poses but re-rank ensembles of docked poses used as inputs by predicting their hypothetical RMSD values. The AI-based SF shows significantly improved docking power compared with that produced by the native SF of the baseline docking program (Autodock Vina) [68]. This model can support one in selecting good binders with correct binding structures from a pool of generated docking poses to eventually identify hit compounds.

However, there is current controversy over the applications of AI-based SFs in SBVS due to the lack of eligible validation experiments [66,70]. In retrospective validation on frequently used benchmark datasets, AI-based SFs stably showed good performance both when trained with protein–ligand complex information and when trained with ligand information alone [66,70]. These results indicate that the protein structural information does not significantly affect the prediction tasks. Bias-controlled validation considering the lack of interpretability of AI-based algorithms in general is required in order to ensure the reliability of the methods in the field [70].

**Prediction of pharmacokinetic properties and toxicity**

The prediction of absorption, distribution, metabolism, and excretion (ADMET) properties and toxicity (T) helps the selection of good drug candidates [71,72] and fosters drug-likeness in the process of drug development. Recent studies have employed a wide range of AI-based methods to predict ADMET properties to reduce a preclinical failure in the drug discovery industry (Figure 2). The SwissADME web tool provides the prediction of physicochemical properties, descriptors, and drug-likeness with the ADMET properties, which are built by SVM or Bayesian methods [73]. AdmetSAR server with 27 predictive models [74] and admetSAR2.0 with 47 predictive models were developed. The collected dataset was represented as molecular fingerprints and constructed using RF, SVM, and kNN models. In another study, variable nearest neighbor (vNN) method was developed as a complement to the kNN method [75]. Fifteen prediction models were constructed using vNN and implemented in the vNN-ADMET web server [76]. ATOM Modeling PipeLine, an open-source software pipeline was built to construct prediction models [77]. It covers data curation, model training and tuning, visualization and analysis. Regarding data curation, RDKit and MolVS packages were provided, and DeepChem, Mordred, and Molecular Operating Environment were included in the module. Diverse datasets can be used and supported by RF, XGBoost, NN, GCNN methods to construct new models.

As prediction model quality depends on input data, large high-quality data are required to obtain accurate prediction results. With numerous efforts to build comprehensive databases and benchmarks [78] together with algorithmic development, a better prediction of ADMET properties will be obtained in the field of drug discovery using AI-based modeling.
Perspectives

- With the utilization of big data and AI methods, CADD enables a better understanding of health and disease. Effective and efficient approaches for the analysis of biomedical big data help to identify significant targets or define features strongly related to specific health outcomes.

- Recent development and applications of big data and AI techniques to build computational and statistical models to solve various problems in drug discovery requires high-quality data as essential parts of research. We have discussed different sections of big data pre-processing, AI modeling methods and AI-based applications in drug design, including the identification of binding sites in target proteins, SBVS and ADMET property prediction.

- Despite the present success, there is still a big room for improvement in terms of method accuracy. Furthermore, the increase in high-dimensional data arising from structural and dynamic elements of sophisticated biochemical entities, will push the drug design field to the innovation of big data and AI tools based on theories and methodologies in statistics. Combined approaches of different data pre-processing and AI methods that learn core patterns from the structures of biomedical big data could significantly improve the predictive models for the drug design, discovery and development.

Competing Interests
The authors declare that there are no competing interests associated with the manuscript.

Funding
This work was supported by the Mid-career Researcher Program (NRF-2020R1A2C2101636), Medical Research Center (MRC) grant (2018R1A5A2025286), Brain Pool Program (NRF-2021H1D3A2A02038434), and Bio & Medical Technology Development Program (NRF-2019M3E5D4065251) funded by the Ministry of Science and ICT (MSIT) and the Ministry of Health and Welfare (MOHW) through the National Research Foundation of Korea (NRF). It was also supported by the Ewha Womans University Research Grant of 2021.

Author Contributions
Conceptualization, J.W.L. and S.C.; Writing — original draft preparation, J.W.L.; Writing — review and editing, J.W.L., M.A.M.S., T.N.L.V., S.Y. and S.C.; Visualization, J.W.L., M.A.M.S., T.N.L.V. and S.Y.; Supervision and Funding acquisition, S.C.

Abbreviations
ADMET, absorption, distribution, metabolism, excretion and toxicity; AI, artificial intelligence; CADD, computer-aided drug design; DT, decision tree; kNN, k-nearest neighbors; LDA, linear discriminant analysis; NN, neural networks; PCA, principal component analysis; PLSR, partial least squares regression; QSAR, quantitative structure-activity relationship; RF, random forest; RGIFE, ranked guided iterative feature elimination; RMSD, root mean square deviation; SBVS, structure-based virtual screening; SF, scoring function; SVM, support vector machine; vNN, variable nearest neighbor.

References
1 Grechishnikova, D. (2021) Transformer neural network for protein-specific de novo drug generation as a machine translation problem. Sci. Rep. 11, 321 https://doi.org/10.1038/s41598-020-79682-4
2 Lu, R.M., Hwang, Y.C., Liu, I.J., Lee, C.C., Tsai, H.Z., Li, H.J. et al. (2020) Development of therapeutic antibodies for the treatment of diseases. J. Biomol. Sci. 27, 1 https://doi.org/10.1186/s12392-019-0592-z
