Biochemical Society Transactions (2022) **50** 241–252
https://doi.org/10.1042/BST20211240

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

Correspondence: Sun Choi (sunchoi@ewha.ac.kr)

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).

Received: 1 November 2021
Revised: 23 December 2021
Accepted: 23 December 2021

Version of Record published: 25 January 2022

Biochemical Society Transactions (2022) **50** 241–252
https://doi.org/10.1042/BST20211240

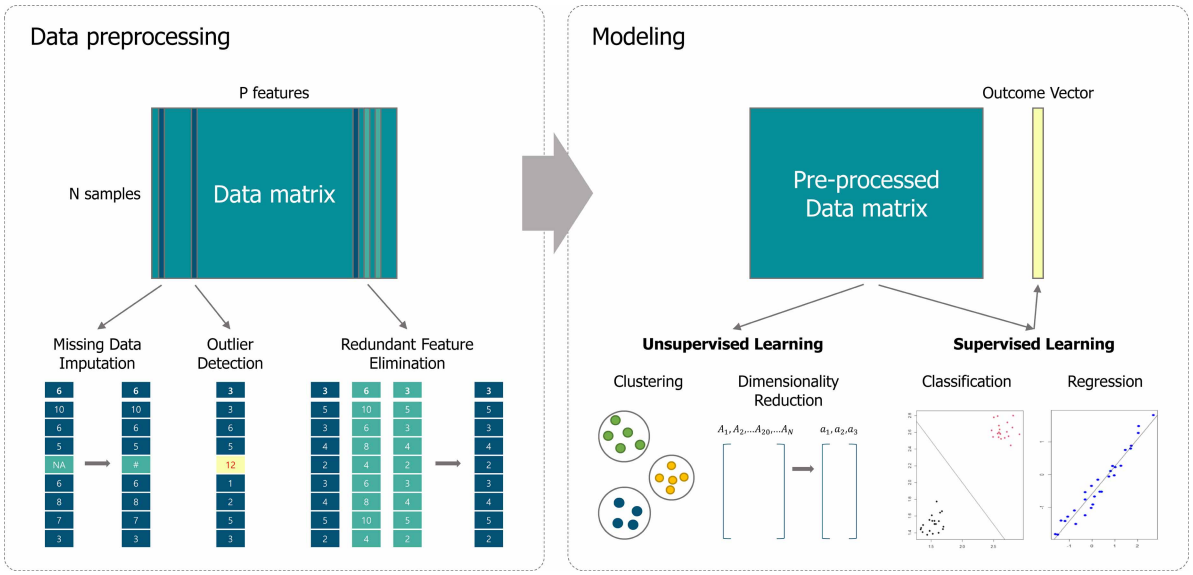

Figure 1. Data pre-processing and modeling.

Data pre-processing steps include missing data imputation, outlier detection, and redundant feature elimination. After the input data are pre-processed, predictive modeling including unsupervised learning (clustering and dimensionality reduction) and supervised learning (regression and classification) can be utilized.

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.

Biochemical Society Transactions (2022) **50** 241–252
https://doi.org/10.1042/BST20211240

Table 2 Supervised AI modeling applications in drug design

Category	Name	Summary
Regression	Penalized Linear Regression	Penalized Linear Regression estimates significant interactions between features in an n-by-p data matrix and the continuous outcome [24]. It can be used for efficiently handling the data when the number of features including molecular descriptors, exceeds the number of compound samples [25].
	Partial Least Squares Regression (PLSR)	PLSR detects new significant features by combining the feature coordinates and extracts the optimal set of latent features by linearly combining them [26]. An extended version of PLS, a kernel-based PLS for pharmacophore mapping of QSAR methods provides types and environment effects of atoms [27].
Classification	Penalized Logistic Regression	Penalized Logistic Regression evaluates significant interactions between features in an n-by-p data matrix and the categorical outcome [28]. It can be used to efficiently identify the most influential descriptors to build a QSAR classification model with both high prediction accuracy and easy interpretability. [29].
	Support Vector Machine (SVM)	SVM builds a multidimensional hyperplane that separates data values in one category from data values in other categories by computing the largest possible distance between data values of different categories [30]. Biological or chemical structures with the optimal descriptors can be appropriately analyzed with SVM for QSAR predictions [31].
	K-Nearest Neighbors (kNN)	kNN defines a predicted category of an unknown sample based on the K closest data values in a training set [32]. Fuzzy kNN classification method was utilized to analyze drug compound data based on a 2D fingerprint via G protein-coupled receptors [33].
	Naïve Bayesian Classifier (NBC)	NBC calculates the set of probabilities by counting the frequency of categories for the feature to be predicted in the data [34]. One advantage utilizing a NBC with structural fingerprints, such as ECFP6, is to find important descriptor features frequently appearing in two classifying outcomes for the design of inhibitors [35].
	Decision Tree (DT)	DT expands subtrees and leaves to obtain a node labeled with a predicted outcome category [36]. Application of DT method can be used to prove that the outcome, the inhibition of InhA by ETH, is significantly related to specific residues determined by DT [37].
	Random Forest (RF)	RF, an ensemble of classification methods, efficiently analyzes high-dimensional data, merging and obtaining outcomes over individual decision trees [38]. RF method has been applied to meaningfully connect several drugs over cell lines using genomic information, drug targets and pharmacological information [39].
	Neural Networks (NN)	NN algorithm sets input features in an input layer, implements weighted transformations over hidden layers, and evaluates the outcome on an output layer [40]. Protein data are often treated as a grid of voxels. Grid-based approaches allows to project grid voxels into multi-channel protein descriptors, as for instance geometry and energy-based strategies [41]. Thus, each protein voxel contains the information of all descriptors. Protein multichannel grids have been successfully processed in 3D convolutional network (3D-CNN) models for the identification of protein binding sites and the prediction of good protein binders (see Section 3).

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

Table 3 Unsupervised AI modeling applications in drug design

Category	Name	Summary
Clustering	K-Means Clustering	K-means Clustering defines K clusters representing categories where the input data values are partitioned into [42]. In drug discovery studies, K-means clustering can generate proper molecular descriptors for each sample, compute the similarity between compound samples, and group compound features based on computed similarity [43].
	Hierarchical Clustering (HC)	In HC, the partitions of data values can be assigned with increasing cluster hierarchy. The partitioning process is finalized when a single cluster containing all n data values is formed or n clusters are assigned to n different data values each [44,45]. One of the most useful graphical representation of hierarchical cluster of compounds is a dendrogram, a tree diagram representing the distance between molecular features [43].
Dimensionality Reduction	Principal Component Analysis (PCA)	PCA transforms the original features into principal components, which are uncorrelated to each other but contain information from the original data [46,47]. PCA can be employed to build QSAR models with molecular descriptors, which explains how compound samples cause an impact on the biological, chemical, or pharmaceutical target. PCA model predicts biological activity when additional molecular descriptors are taken into the analysis of the same biological target, such as a protein affected by different receptors [48].
	Linear Discriminant Analysis (LDA)	LDA builds a prediction model, which classifies patterns in the data [49]. LDA detects features that better separate the categories of data by projecting the original data points on to these features. If two or more categories are estimated for given data points, LDA better separates them by applying the transformation mechanism [50]. An extended version of LDA, multi-label linear discriminant analysis, conducts feature dimension reduction of drug data features before constructing and predicting models. This dimensionality reduction step enhances the accuracy of the prediction models and decreases the computing time of training in the prediction model to analyze drug discovery data [51].

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

Biochemical Society Transactions (2022) **50** 241–252
https://doi.org/10.1042/BST20211240

Table 4 Software and tools for AI modeling applications in structure-based drug design

Application	Method	Software program	Link
Identification of binding sites in target proteins	NBC	ENRI [56]	Source code: https://github.com/fibonaccirabbits/enri
	DT RF	P2Rank [53]	Source code: http://github.com/rdk/p2rank
	NN	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/imolecule/tools/bitenet/
	K-Means Clustering HC PCA	iFeature [79]	Web server: https://ifeature.erc.monash.edu/
	LDA	SpotOn [80]	Web server: https://alcazar.science.uu.nl/cgi/services/SPOTON/spoton/
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/azevedolab/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 NBC	SwissADME [73]	Web server: http://www.swissadme.ch/
	RF SVM kNN	admetSAR2.0 [83]	Web server: http://lmmd.ecust.edu.cn/admetsar2/
	kNN	vNN-ADMET [76]	Web server: https://vnnadmet.bhsai.org/vnnadmet/
	RF NN	AMPL [77]	Source code: https://github.com/ATOMconsortium/AMPL
	RF SVM PLSR NBC DT	ADMETlab [84]	Web server: https://admet.scbdd.com/home/index/

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

Biochemical Society Transactions (2022) **50** 241–252
https://doi.org/10.1042/BST20211240

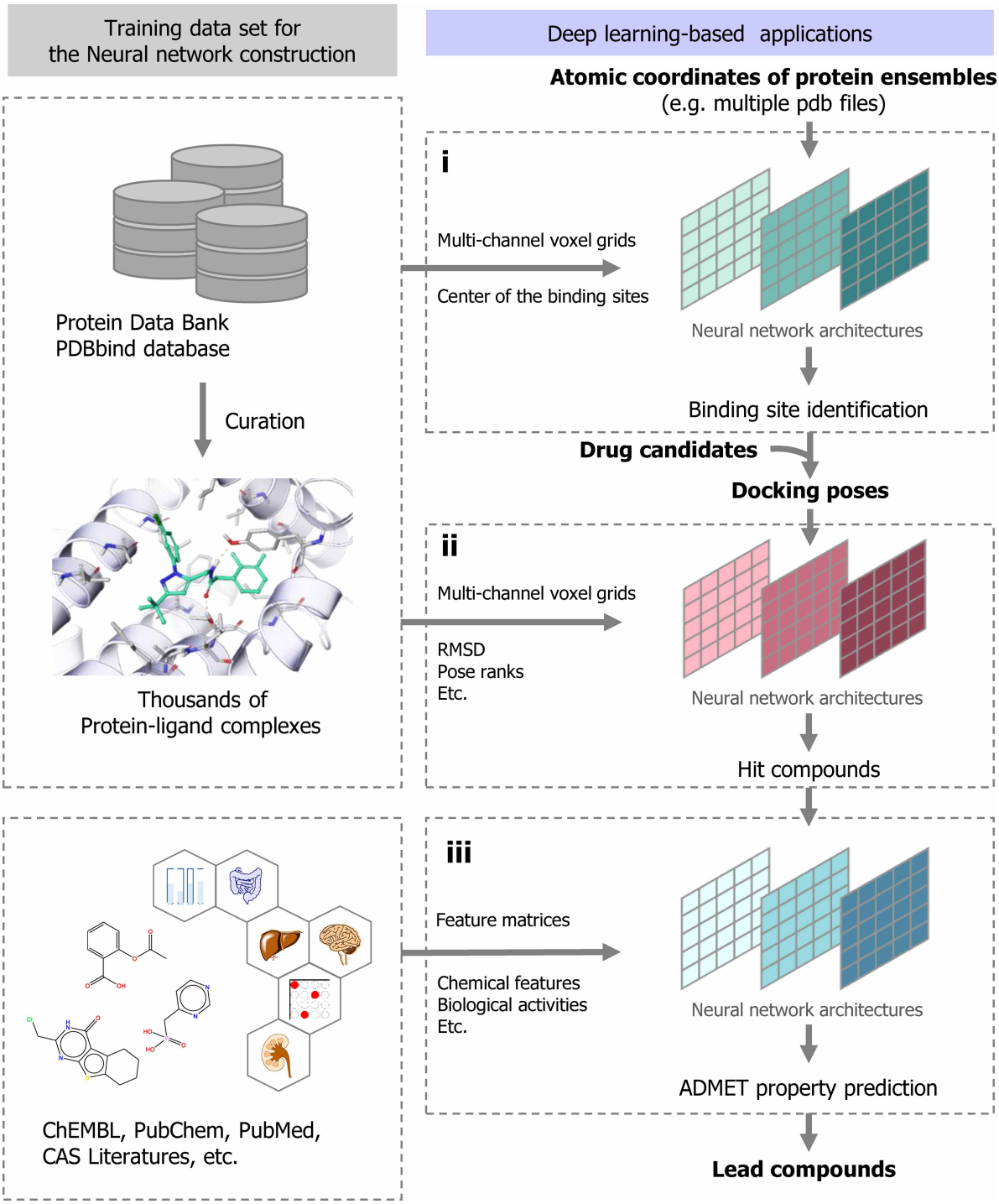

Figure 2. Overview of the applications workflow and their interconnection.

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.

Biochemical Society Transactions (2022) 50 241–252
https://doi.org/10.1042/BST20211240

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 PBDbind 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 (ADME) 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 web 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. Biomed. Sci.* **27**, 1 https://doi.org/10.1186/s12929-019-0592-z

Biochemical Society Transactions (2022) **50** 241–252
https://doi.org/10.1042/BST20211240

3 Emmerich, C.H., Gamboa, L.M., Hofmann, M.C.J., Bonin-Andresen, M., Arbach, O., Schendel, P. et al. (2021) Improving target assessment in biomedical research: the GOT-IT recommendations. *Nat. Rev. Drug Discov.* **20**, 64–81 https://doi.org/10.1038/s41573-020-0087-3
4 Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R.K. and Kumar, P. (2021) Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol. Divers.* **25**, 1315–1360 https://doi.org/10.1007/s11030-021-10217-3
5 Bellis, L.J., Akhtar, R., Al-Lazikani, B., Atkinson, F., Bento, A.P., Chambers, J. et al. (2011) Collation and data-mining of literature bioactivity data for drug discovery. *Biochem. Soc. Trans.* **39**, 1365–1370 https://doi.org/10.1042/BST0391365
6 Schaduangrat, N., Lampa, S., Simeon, S., Gleeson, M.P., Spjuth, O. and Nantasenamat, C. (2020) Towards reproducible computational drug discovery. *J. Cheminform.* **12**, 9 https://doi.org/10.1186/s13321-020-0408-x
7 Yang, X., Wang, Y.F., Byrne, R., Schneider, G. and Yang, S.Y. (2019) Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev.* **119**, 10520–10594 https://doi.org/10.1021/acs.chemrev.8b00728
8 Katsila, T., Spyroulias, G.A., Patrinos, G.P. and Matsoukas, M.T. (2016) Computational approaches in target identification and drug discovery. *Comput. Struct. Biotechnol. J.* **14**, 177–184 https://doi.org/10.1016/j.csbj.2016.04.004
9 Macalino, S.J.Y., Gosu, V., Hong, S.H. and Choi, S. (2015) Role of computer-aided drug design in modern drug discovery. *Arch. Pharm. Res.* **38**, 1686–1701 https://doi.org/10.1007/s12272-015-0640-5
10 Hu, Y.H., Lin, W.C., Tsai, C.F., Ke, S.W. and Chen, C.W. (2015) An efficient data preprocessing approach for large scale medical data mining. *Technol. Health Care* **23**, 153–160 https://doi.org/10.3233/THC-140887
11 Car, J., Sheikh, A., Wicks, P. and Williams, M.S. (2019) Beyond the hype of big data and artificial intelligence: building foundations for knowledge and wisdom. *BMC Med.* **17**, 143 https://doi.org/10.1186/s12916-019-1382-x
12 Saez, C. and Garcia-Gomez, J.M. (2018) Kinematics of big biomedical data to characterize temporal variability and seasonality of data repositories: functional data analysis of data temporal evolution over non-parametric statistical manifolds. *Int. J. Med. Inform.* **119**, 109–124 https://doi.org/10.1016/j.ijmedinf.2018.09.015
13 Miller, J.B. (2019) Big data and biomedical informatics: preparing for the modernization of clinical neuropsychology. *Clin. Neuropsychol.* **33**, 287–304 https://doi.org/10.1080/13854046.2018.1523466
14 Suh, D., Lee, J.W., Choi, S. and Lee, Y. (2021) Recent applications of deep learning methods on evolution-and contact-based protein structure prediction. *Int. J. Mol. Sci.* **22**, 6032 https://doi.org/10.3390/ijms22116032
15 Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N.C. and Ping, P.P. (2019) Machine learning and integrative analysis of biomedical big data. *Genes* **10**, 87 https://doi.org/10.3390/genes10020087
16 Irwin, B., Whitehead, T.M., Rowland, S., Mahmoud, S.Y., Conduit, G.J. and Segall, M.D. (2021) Deep imputation on large-scale drug discovery data. *Appl. AI Lett.* **2**, e31 https://doi.org/10.1002/ail2.31
17 Tropsha, A. (2010) Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.* **29**, 476–488 https://doi.org/10.1002/minf.201000061
18 Perez-Villanueva, J., Santos, R., Hernandez-Campos, A., Giulianotti, M.A., Castillo, R. and Medina-Franco, J.L. (2010) Towards a systematic characterization of the antiprotozoal activity landscape of benzimidazole derivatives. *Bioorgan. Med. Chem.* **18**, 7380–7391 https://doi.org/10.1016/j.bmc.2010.09.019
19 Yosipof, A. and Senderowitz, H. (2014) Optimization of molecular representativeness. *J. Chem. Inform. Model.* **54**, 1567–1577 https://doi.org/10.1021/ci400715n
20 Zhang, B.T. and Cao, P. (2019) Classification of high dimensional biomedical data based on feature selection using redundant removal. *PLos ONE* **14**, e0214406 https://doi.org/10.1371/journal.pone.0214406
21 Radovic, M., Ghalwash, M., Filipovic, N. and Obradovic, Z. (2017) Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics* **18**, 9 https://doi.org/10.1186/s12859-016-1423-9
22 Lazzarini, N. and Bacardit, J. (2017) RGIFE: a ranked guided iterative feature elimination heuristic for the identification of biomarkers. *BMC Bioinformatics.* **18**, 322 https://doi.org/10.1186/s12859-017-1729-2
23 Alchemite™ Analytics 2021 [Available from: https://intellegens.ai/products-services/alchemite-analytics/
24 Jozaghi, A., Shen, H.J., Ghazvinian, M., Seo, D.J., Zhang, Y., Welles, E. et al. (2021) Multi-model streamflow prediction using conditional bias-penalized multiple linear regression. *Stoch. Environ. Res. Risk A* **35**, 2355–2373 https://doi.org/10.1007/s00477-021-02048-3
25 Algamal, Z.Y., Lee, M.H., Al-Fakih, A.M. and Aziz, M. (2016) High-dimensional QSAR modelling using penalized linear regression model with L-1/2-norm. *Sar. Qsar. Environ. Res.* **27**, 703–719 https://doi.org/10.1080/1062936X.2016.1228696
26 Fowler, S.M., Wheeler, D., Morris, S., Mortimer, S.I. and Hopkins, D.L. (2021) Partial least squares and machine learning for the prediction of intramuscular fat content of lamb loin. *Meat Sci.* **177**, 108505 https://doi.org/10.1016/j.meatsci.2021.108505
27 Falchi, F., Bertozzi, S.M., Ottonello, G., Ruda, G.F., Colombano, G., Fiorelli, C. et al. (2016) Kernel-based, partial least squares quantitative structure-retention relationship model for UPLC retention time prediction: a useful tool for metabolite identification. *Anal. Chem.* **88**, 9510–9517 https://doi.org/10.1021/acs.analchem.6b02075
28 Zhou, Z.M., Huang, H.H. and Liang, Y. (2021) Cancer classification and biomarker selection via a penalized logsum network-based logistic regression model. *Technol. Health Care* **29**, S287–SS95 https://doi.org/10.3233/THC-218026
29 Algamal, Z.Y. and Lee, M.H. (2017) A novel molecular descriptor selection method in QSAR classification model based on weighted penalized logistic regression. *J. Chemometr.* **31**, e2915 https://doi.org/10.1002/cem.2915
30 Yucelbas, S. and Yucelbas, C. (2021) Autism spectrum disorder detection using sequential minimal optimization-support vector machine hybrid classifier according to history of jaundice and family autism in children. *Concurr. Comp.-Pract. E* **34**, e6498 https://doi.org/10.1002/cpe.6498
31 Alvarsson, J., Lampa, S., Schaal, W., Andersson, C., Wikberg, J.E.S. and Spjuth, O. (2016) Large-scale ligand-based predictive modelling using support vector machines. *J. Cheminform.* **8**, 39 https://doi.org/10.1186/s13321-016-0151-5
32 Miranda-Vega, J.E., Rivas-Lopez, M. and Fuentes, W.F. (2021) k-nearest neighbor classification for pattern recognition of a reference source light for machine vision system. *IEEE Sens. J.* **21**, 11514–11521 https://doi.org/10.1109/JSEN.2020.3024094
33 Bagherian, M., Sabeti, E., Wang, K., Sartor, M.A., Nikolovska-Coleska, Z. and Najarian, K. (2021) Machine learning approaches and databases for prediction of drug-target interaction: a survey paper. *Brief. Bioinform.* **22**, 606 https://doi.org/10.1093/bib/bbaa020

34 Gao, H.Y., Zeng, X. and Yao, C.H. (2019) Application of improved distributed naive Bayesian algorithms in text classification. *J. Supercomput.* **75**, 5831–5847 https://doi.org/10.1007/s11227-019-02862-1
35 Kang, D., Pang, X.C., Lian, W.W., Xu, L.J., Wang, J.H., Jia, H. et al. (2018) Discovery of VEGFR2 inhibitors by integrating naive Bayesian classification, molecular docking and drug screening approaches. *RSC Adv.* **8**, 5286–5297 https://doi.org/10.1039/C7RA12259D
36 Mao, Y.X., He, Y.H., Liu, L.M. and Chen, X.S. (2020) Disease classification based on Eye movement features with decision tree and random forest. *Front. Neurosci.* **14**, 798 https://doi.org/10.3389/fnins.2020.00798
37 Barros, R.C., Winck, A.T., Machado, K.S., Basgalupp, M.P., de Carvalho, A.C.P.L.F., Ruiz, D.D. et al. (2012) Automatic design of decision-tree induction algorithms tailored to flexible-receptor docking data. *BMC Bioinformatics* **13**, 310 https://doi.org/10.1186/1471-2105-13-310
38 Chen, W., Li, Y., Xue, W.F., Shahabi, H., Li, S.J., Hong, H.Y. et al. (2020) Modeling flood susceptibility using data-driven approaches of naive Bayes tree, alternating decision tree, and random forest methods. *Sci. Total Environ.* **701**, 134979 https://doi.org/10.1016/j.scitotenv.2019.134979
39 Lind, A.P. and Anderson, P.C. (2019) Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PLos ONE* **14**, e0219774 https://doi.org/10.1371/journal.pone.0219774
40 Lahmiri, S., Tadj, C. and Gargour, C. (2021) Biomedical diagnosis of infant Cry signal based on analysis of cepstrum by deep feedforward artificial neural networks. *IEEE Instrum. Meas. Mag.* **24**, 24–29 https://doi.org/10.1109/MIM.2021.9400952
41 Jiang, M.J., Li, Z., Bian, Y.J. and Wei, Z.Q. (2019) A novel protein descriptor for the prediction of drug binding sites. *BMC Bioinformatics* **20**, 478 https://doi.org/10.1186/s12859-019-3058-0
42 Khan, I., Luo, Z.W., Shaikh, A.K. and Hedjam, R. (2021) Ensemble clustering using extended fuzzy k-means for cancer data analysis. *Expert Syst. Appl.* **172**, 114622 https://doi.org/10.1016/j.eswa.2021.114622
43 Voicu, A., Duteanu, N., Voicu, M., Vlad, D. and Dumitrascu, V. (2020) The rcdk and cluster R packages applied to drug candidate selection. *J. Cheminform.* **12**, 3 https://doi.org/10.1186/s13321-019-0405-0
44 Liu, X.Y., Chen, Y.Y., Cheng, A.S.K., Zeng, Y.C., Ullah, S. and Feuerstein, M. (2021) Conceptualizing problems with symptoms, function, health behavior, health-seeking skills, and financial strain in breast cancer survivors using hierarchical clustering. *J. Cancer Surviv.* https://doi.org/10.1007/s11764-021-01068-w
45 Wang, H.C., Chou, M.C., Wu, C.C., Chan, L.P., Moi, S.H., Pan, M.R. et al. (2021) Application of the interaction between tissue immunohistochemistry staining and clinicopathological factors for evaluating the risk of oral cancer progression by hierarchical clustering analysis: a case-control study in a Taiwanese population. *Diagnostics* **11**, 925 https://doi.org/10.3390/diagnostics11060925
46 Xie, H.B., Zhou, P., Guo, T.R., Sivakumar, B., Zhang, X. and Dokos, S. (2016) Multiscale two-Directional two-dimensional principal component analysis and Its application to high-dimensional biomedical signal classification. *IEEE T Bio-Med. Eng.* **63**, 1416–1425 https://doi.org/10.1109/TBME.2015.2436375
47 Uguz, H. (2012) A biomedical system based on artificial neural network and principal component analysis for diagnosis of the heart valve diseases. *J. Med. Syst.* **36**, 61–72 https://doi.org/10.1007/s10916-010-9446-7
48 Yoo, C. and Shahlaei, M. (2018) The applications of PCA in QSAR studies: a case study on CCR5 antagonists. *Chem. Biol. Drug Des.* **91**, 137–152 https://doi.org/10.1111/cbdd.13064
49 Wang, F., Wang, Q., Nie, F.P., Li, Z.H., Yu, W.Z. and Wang, R. (2021) Unsupervised linear discriminant analysis for jointly clustering and subspace learning. *IEEE T Knowl. Data En.* **33**, 1276–1290 https://doi.org/10.1109/TKDE.2019.2939524
50 Lin, W.M., Gao, Q.Q., Du, M., Chen, W.S. and Tong, T. (2021) Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data. *Comput. Biol. Med.* **134**, 104478 https://doi.org/10.1016/j.compbiomed.2021.104478
51 Pirhadi, S., Shiri, F. and Ghasemi, J.B. (2015) Multivariate statistical analysis methods in QSAR. *RSC Adv.* **5**, 104635–65 https://doi.org/10.1039/C5RA10729F
52 Terranova, N., Venkatakrishnan, K. and Benincosa, L.J. (2021) Application of machine learning in translational medicine: current status and future opportunities. *AAPS J.* **23**, 74 https://doi.org/10.1208/s12248-021-00593-x
53 Krivak, R. and Hoksza, D. (2018) P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. Cheminform.* **10**, 39 https://doi.org/10.1186/s13321-018-0285-8
54 Kozlovskii, I. and Popov, P. (2020) Spatiotemporal identification of druggable binding sites using deep learning. *Commun. Biol.* **3**, 618 https://doi.org/10.1038/s42003-020-01350-0
55 Amaro, R.E., Baudry, J., Chodera, J., Demir, O., McCammon, J.A., Miao, Y.L. et al. (2018) Ensemble docking in drug discovery. *Biophys. J.* **114**, 2271–2278 https://doi.org/10.1016/j.bpj.2018.02.038
56 Akbar, R., Jusoh, S.A., Amaro, R.E. and Helms, V. (2017) ENRI: a tool for selecting structure-based virtual screening target conformations. *Chem. Biol. Drug Design* **89**, 762–771 https://doi.org/10.1111/cbdd.12900
57 Le Guilloux, V., Schmidtke, P. and Tuffery, P. (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 https://doi.org/10.1186/1471-2105-10-168
58 Hernandez, M., Ghersi, D. and Sanchez, R. (2009) SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res.* **37**, W413–W416 https://doi.org/10.1093/nar/gkp281
59 Zhang, Z.M., Li, Y., Lin, B.Y., Schroeder, M. and Huang, B.D. (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* **27**, 2083–2088 https://doi.org/10.1093/bioinformatics/btr331
60 Jimenez, J., Doerr, S., Martinez-Rosell, G., Rose, A.S. and De Fabritiis, G. (2017) Deepsite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **33**, 3036–3042 https://doi.org/10.1093/bioinformatics/btx350
61 Westbrook, J., Feng, Z.K., Chen, L., Yang, H.W. and Berman, H.M. (2003) The protein data bank and structural genomics. *Nucleic Acids Res.* **31**, 489–491 https://doi.org/10.1093/nar/gkg068
62 Cimermancic, P., Weinkam, P., Rettenmaier, T.J., Bichmann, L., Keedy, D.A., Woldeyes, R.A. et al. (2016) Cryptosite: expanding the druggable proteome by characterization and prediction of cryptic binding sites. *J. Mol. Biol.* **428**, 709–719 https://doi.org/10.1016/j.jmb.2016.01.029
63 Son, J. and Kim, D. (2021) Development of a graph convolutional neural network model for efficient prediction of protein-ligand binding affinities. *PLos ONE* **16**, e0249404 https://doi.org/10.1371/journal.pone.0249404

Biochemical Society Transactions (2022) **50** 241–252
https://doi.org/10.1042/BST20211240

64 Ghislat, G., Rahman, T. and Ballester, P.J. (2021) Recent progress on the prospective application of machine learning to structure-based virtual screening. *Curr. Opin. Chem. Biol.* **65**, 28–34 https://doi.org/10.1016/j.cbpa.2021.04.009
65 Liu, J. and Wang, R. (2015) Classification of current scoring functions. *J. Chem. Inf. Model.* **55**, 475–482 https://doi.org/10.1021/ci500731a
66 Morrone, J.A., Weber, J.K., Huynh, T., Luo, H. and Cornell, W.D. (2020) Combining docking pose rank and structure with deep learning improves protein–ligand binding mode prediction over a baseline docking approach. *J. Chem. Inf. Model.* **60**, 4170–4179 https://doi.org/10.1021/acs.jcim.9b00927
67 Bitencourt-Ferreira, G., Duarte da Silva, A. and Filgueira,. (2021) Application of machine learning techniques to predict binding affinity for drug targets: a study of cyclin-dependent kinase 2. *Curr. Med. Chem.* **28**, 253–265 https://doi.org/10.2174/2213275912666191102162959
68 Bao, J., He, X. and Zhang, J.Z. (2021) DeepBSP—a machine learning method for accurate prediction of protein–ligand docking structures. *J. Chem. Inf. Model.* **61**, 2231–2240 https://doi.org/10.1021/acs.jcim.1c00334
69 Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y. et al. (2017) Forging the basis for developing protein–ligand interaction scoring functions. *Acc. Chem. Res.* **50**, 302–309 https://doi.org/10.1021/acs.accounts.6b00491
70 Sieg, J., Flachsenberg, F. and Rarey, M. (2019) In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J. Chem. Inf. Model.* **59**, 947–961 https://doi.org/10.1021/acs.jcim.8b00712
71 Kaitin, K. (2008) Obstacles and opportunities in new drug development. *Clin. Pharmacol. Ther.* **83**, 210–212 https://doi.org/10.1038/sj.clpt.6100462
72 Ghosh, J., Lawless, M.S., Waldman, M., Gombar, V. and Fraczkiewicz, R. (2016) Modeling ADMET. In *Silico Methods for Predicting Drug Toxicity* (Emilio, B., ed.), pp. 63–83, Springer, Humana Press, New York https://doi.org/10.1007/978-1-4939-3609-0
73 Daina, A., Michielin, O. and Zoete, V. (2017) SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **7**, 42717 https://doi.org/10.1038/srep42717
74 Cheng, F.X., Li, W.H., Zhou, Y.D., Shen, J., Wu, Z.R., Liu, G.X. et al. (2012) admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J. Chem. Inf. Model.* **52**, 3099–3105 https://doi.org/10.1021/ci300367a
75 Liu, R., Tawa, G. and Wallqvist, A. (2012) Locally weighted learning methods for predicting dose-dependent toxicity with application to the human maximum recommended daily dose. *Chem. Res. Toxicol.* **25**, 2216–2226 https://doi.org/10.1021/tx300279f
76 Schyman, P., Liu, R.F., Desai, V. and Wallqvist, A. (2017) vNN web server for ADMET predictions. *Front. Pharmacol.* **8**, 889 https://doi.org/10.3389/fphar.2017.00889
77 Minnich, A.J., McLoughlin, K., Tse, M., Deng, J., Weber, A., Murad, N. et al. (2020) AMPL: a data-driven modeling pipeline for drug discovery. *J. Chem. Inf. Model.* **60**, 1955–1968 https://doi.org/10.1021/acs.jcim.9b01053
78 Wu, Z.Q., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S. et al. (2018) Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 https://doi.org/10.1039/C7SC02664A
79 Chen, Z., Zhao, P., Li, F.Y., Leier, A., Marquez-Lago, T.T., Wang, Y.N. et al. (2018) Ifeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **34**, 2499–2502 https://doi.org/10.1093/bioinformatics/bty140
80 Moreira, I.S., Koukos, P.I., Melo, R., Almeida, J.G., Preto, A.J., Schaarschmidt, J. et al. (2017) Spoton: high accuracy identification of protein-protein interface hot-spots. *Sci. Rep.* **7**, 8007 https://doi.org/10.1038/s41598-017-08321-2
81 Bitencourt-Ferreira, G., Rizzotto, C. and de Azevedo Junior, W.F. (2021) Machine learning-based scoring functions, development and applications with SAnDReS. *Curr. Med. Chem.* **28**, 1746–1756 https://doi.org/10.2174/0929867327666200515101820
82 Li, H., Leung, K.S., Wong, M.H. and Ballester, P.J. (2015) Improving AutoDock Vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Mol. Inform.* **34**, 115–126 https://doi.org/10.1002/minf.201400132
83 Yang, H.B., Lou, C.F., Sun, L.X., Li, J., Cai, Y.C., Wang, Z. et al. (2019) admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. *Bioinformatics* **35**, 1067–1069 https://doi.org/10.1093/bioinformatics/bty707
84 Dong, J., Wang, N.N., Yao, Z.J., Zhang, L., Cheng, Y., Ouyang, D.F. et al. (2018) ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. *J. Cheminform.* **10**, 29 https://doi.org/10.1186/s13321-018-0283-x