ELSEVIER

ADMET modeling approaches in drug discovery

Leonardo L.G. Ferreira and Adriano D. Andricopulo

Laboratory of Medicinal and Computational Chemistry, Center for Research and Innovation in Biodiversity and Drug Discovery, Physics Institute of Sao Carlos, University of Sao Paulo, Av. Joao Dagnone 1100, 13563-120, Sao Carlos, SP, Brazil

In silico **prediction of ADMET is an important component of pharmaceutical R&D. Last year, the FDA approved 59 new molecular entities, with small molecules comprising 64% of the therapies approved in 2018. Estimation of pharmacokinetic properties in the early phases of drug discovery has been central to guiding hit-to-lead and lead-optimization efforts. Given the outstanding complexity of the current R&D model, drug discovery players have intensely pursued molecular modeling strategies to identify patterns in ADMET data and convert them into knowledge. The field has advanced alongside the progress of chemoinformatics, which has evolved from traditional chemometrics to advanced machine learning methods.**

Introduction

Effective and safe drugs exhibit a finely tuned combination of pharmacodynamics (PD) and pharmacokinetics (PK), including high potency, affinity and selectivity against the molecular target, along with adequate absorption, distribution, metabolism, excretion and tolerable toxicity (ADMET). The coordinated optimization of these interdependent variables is a major hurdle in drug R&D [1]. To this end, unprecedented efforts have been put in place for the development of technologies aimed at predicting PD and PK endpoints during hit-to-lead and lead-optimization programs [2,3]. In this context, a wide diversity of tools is available today for the prediction of ADMET, such as QikProp, DataWarrior, MetaTox, MetaSite and StarDrop, to name a few [1,4–7]. Although loss of efficacy and safety concerns play an increasing part in drug R&D attrition rates, the impact of PK properties has decreased over the past decade [8]. This reduction stems from enhanced PK control programs and their increasingly earlier implementation in the research pipeline. By simultaneously targeting multiple PK parameters, fully integrated ADMET prediction platforms can readily exclude unsuitable compounds, reducing the number of synthesis-evaluation cycles and scaling down the number of more-expensive late-stage failures. Another important aspect to be considered is the prediction of pan-assay interference compounds (PAINS), which are promiscuous bioactive molecules that have led to assay artifacts or false-positive results. These 'frequent hitters' can show apparent activity against unrelated macromolecular targets and interfere with assay outcomes across different experimental methods. The follow-up of PAINS in hit-to-lead efforts represents a significant waste of valuable R&D resources [9]. The prediction of ADMET has become commonplace as the market's demand for innovative products has risen in pace with growing R&D risks and costs. The stringency of this situation is demonstrated by attrition rates of 90% between clinical trials and marketing authorization and estimated costs of US$ 2.6 billion per new chemical entity (NCE) [10]. In this context, continued development of efficient chemoinformatics tools to correlate molecular structure, physicochemical properties and ADMET endpoints is essential to the pharmaceutical sector. This review updates the drug discovery community with some of the newest developments in ADMET modeling, focusing on key advances made in the 2017–2018 period.

Chemoinformatics and drug discovery

Chemoinformatics emerged in the 1950s, when Ray and Kirsch published the first substructure searching routine [11]. In 1962, Hansch *et al.* paved the way for the development of quantitative

Corresponding author: Andricopulo, Adriano D..D. (aandrico@ifsc.usp.br)

https://doi.org/10.1016/j.drudis.2019.03.015

structure–activity and structure–property relationships (QSAR and QSPR, respectively) by correlating physicochemical properties to biological activity. This seminal study is referred to as the Hansch analysis [12]. In another landmark paper published in 1964, Hansch and Fujita advocated that lipophilicity (measured as the octanol–water partition coefficient) could be used to predict biological activity [13]. Following these groundbreaking ideas, other studies stood out in the 1960s and 1970s and contributed to the establishment of chemoinformatics as an active field of research: (i) Topliss and Costello's work on the statistical aspects of SAR [14]; (ii) Free and Wilson's analysis of the use of explicit structural variables in QSAR [15]; and (iii) the paper by Cramer *et al.* on the use of substructure analysis on structurally diverse datasets [16]. These foundations were then applied to PK investigations, among which the paper by Timmermans *et al.* published in 1977 reported on the correlation between lipophilicity and blood–brain barrier (BBB) permeability for 27 clonidine analogs [17]. In the 1980s, Hinderling *et al.* were pioneers in using chemometrics to correlate *in vivo* primary and secondary PK endpoints to octanol–buffer partition coefficients. This study was based on a series of β-adrenoceptor antagonists [18]. In the 1990s, Lipinski *et al.* were the first to correlate a series of physicochemical variables to PK properties for large datasets – an analysis that became the highly cited Rule of Five [19]. Following the drug-likeness concept, other relevant medicinal chemistry guidelines emerged in the late 1990s and early 2000s, including those proposed by Ghose *et al.*, Veber *et al.*, Egan *et al.* and Muegge *et al.* [20–23]. This period witnessed the birth of the so-called 'big data' age, which has been fostered by the increasing part played by automation and parallelization technologies, such as HTS and combinatorial chemistry. As a result, the computational counterpart has evolved into more-robust approaches that are capable of handling increasingly large, heterogeneous and multidimensional datasets. In this context, chemoinformatics has moved from the 'conventional' techniques explored in the studies mentioned above (e.g., multiple linear regression, MLR) to more-advanced machine learning (ML) methods [24].

Current ML approaches to ADMET modeling

ML has a wide range of applications in drug discovery and has been actively used in the prediction of PK profiles. Techniques such as *k*-nearest neighbor (*k*-NN), support vector machines (SVM), random forest (RF) and artificial neural networks (ANNs) are examples of ML methods that have recently excelled in the investigation of ADMET properties [25].

k-NN is a nonlinear method for pattern recognition that applies a distance metric to compare a new occurrence with a set of known events referred to as the training set [26]. The new occurrence is labeled according to the class to which most of the closest *k* neighbors in the training set belong. If the class is numeric, a distance-weighted average is used. *k*-NN can also be used for regression challenges. Shen *et al.* were among the first to develop *k*-NN-based QSPR models to investigate metabolic stability [27]. Test set compounds were correctly labeled as metabolically stable or unstable, with an accuracy >85%.

SVM is a linear ML method used in regression and classification analyses. SVM algorithms search for a combination of crucial borderline occurrences (the support vectors) from each class and construct a plane (a discriminant function) that maximizes the distance between the classes. By using a kernel function, the algorithm converts an *n*-dimensional descriptor space into a high-dimensional feature space and constructs a linear model referred to as the hyperplane [28]. In a pioneering study applying SVM to PK modeling, Doniger *et al.* investigated BBB penetration using a training set of 179 central nervous system (CNS) active compounds and 145 inactive molecules. Using 30 test sets, the model had an average performance of 81.5% [29].

RF is an ensemble method that can be used for classification and regression tasks [30]. Proposed by Breiman in 2001, RF combines multiple models (decision trees, DTs) that are independently derived and used as base learners [31]. To build the training set for each DT, a bootstrap sample from the whole dataset is used. Compounds not included in the training sets are assigned to the so-called out-of-bag (OOB) group. The best combinations of randomly selected descriptors are used to build the singular DTs. Next, each DT predicts the target endpoint for the OOB compounds. The outputs from each DT are then combined to produce a consensus model. Svetnik *et al.* were among the first to use a RF routine to investigate BBB penetration and P-glycoprotein (P-gp) binding. Both models had an accuracy of 80% [32].

ANNs are among the hottest topics in ML and have a wide range of applications in drug discovery – in regression and classification analyses [33]. ANNs seek to mimic the architecture of biological neural circuits so that they can process data similarly to the human brain. ANNs work by interconnecting many processing units (or neurons) that run in parallel. The neurons are sorted into layers: an input layer that receives the data, hidden layers that process the data using an activation function and an output layer that generates the answer to the problem at hand. Recently, deep ANNs have been introduced to drug discovery. Compared with conventional, shallow networks, deep nets have a higher number of hidden layers and algorithmic developments that enable them to better handle large datasets [33]. Gobburu and Shelver were pioneers in using ANNs to predict PK properties. They used a congeneric series of β-adrenoceptor antagonists to predict diverse properties, such as plasma protein binding (PPB), volume of distribution, renal and nonrenal clearance, and the mean residence time [34]. The ANNs outperformed MLR analyses carried out on the same dataset, yielding correlation coefficients >0.86 for all of the models. Following this perspective on the early developments of chemoinformatics and ML applied to ADMET, the next section focuses on current studies published in the 2017–2018 period.

Recent developments in ADMET modeling

Ensemble methods, also known as multiple classifier systems, are based on the combination of individual models and have been applied to handle the high-dimensionality issue and unbalanced datasets [35]. Unbalanced data compose datasets in which occurrences that represent a given class significantly outnumber occurrences that represent the other classes. This phenomenon often leads to biased ADMET datasets. Besides ensemble methods, other approaches have addressed this issue, such as those introduced by Norinder *et al.* and Svensson *et al.* [36,37]. Yang *et al.* developed a novel ensemble approach named the adaptive ensemble classification framework (AECF) (Fig. 1a) [38]. This method encompasses the steps of data balancing, generation of individual models,

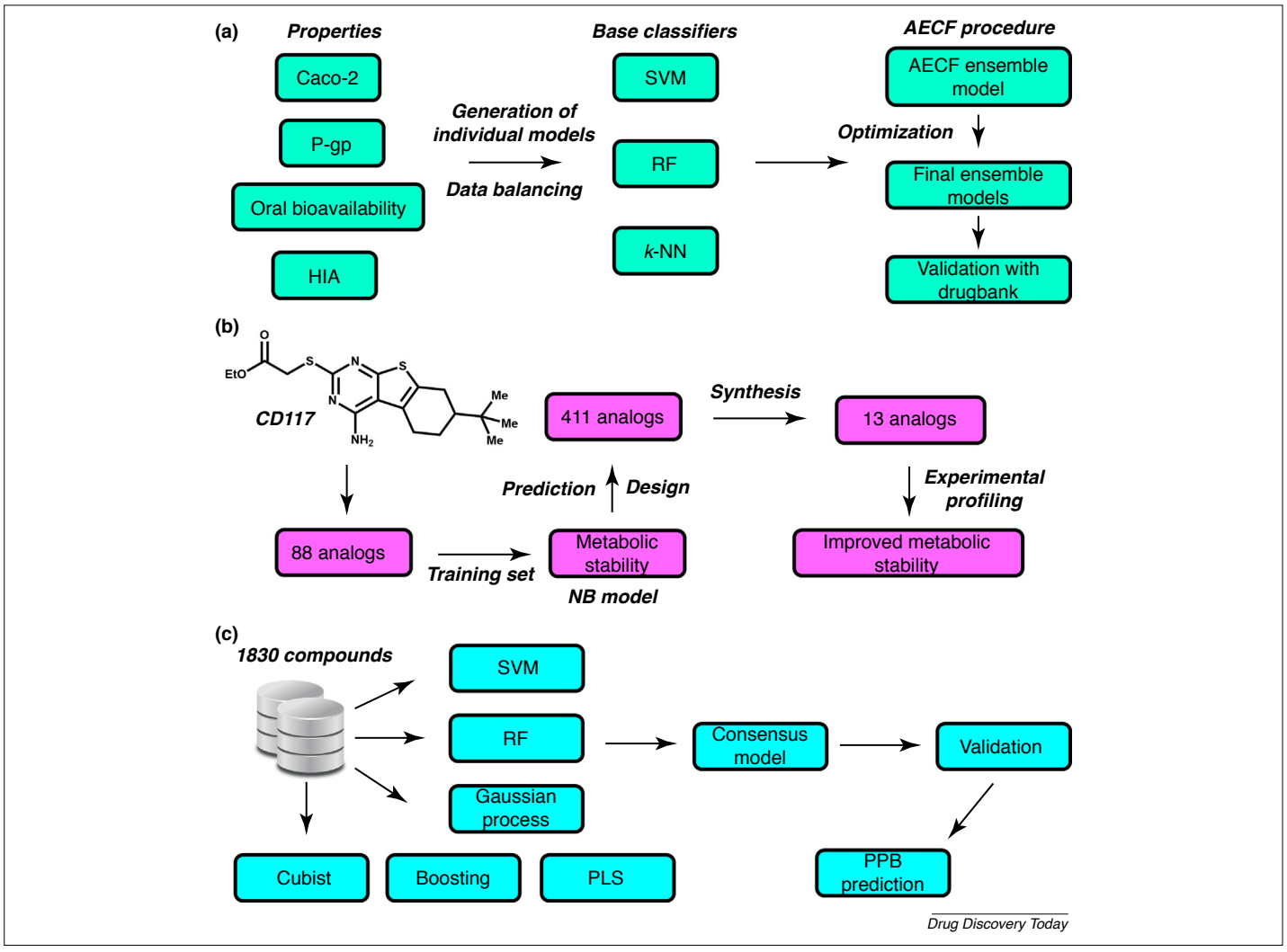

FIGURE 1

Recently developed machine learning models for the prediction of ADME properties. **(a)** The adaptive ensemble classification framework (AECF) for the prediction of Caco-2 permeability, human intestinal absorption (HIA), oral bioavailability and P-glycoprotein (P-gp) binding. The AECF models were built using support vector machine (SVM), Random Forest (RF) and *k*-nearest neighbors (*k*-NN). **(b)** A naive Bayes (NB) analysis was used to estimate metabolic stability for a series of analogs of the antitubercular agent CD117. **(c)** The development of quantitative structure–property relationship (QSPR) models based on partial least squares (PLS), RF, SVM, Cubist, Gaussian process (GP) and Boosting to predict plasma protein binding (PPB).

integration of models and optimization of the ensemble to simultaneously deal with the high-dimensionality and unbalance issues. The AECF method was used to build Caco-2 permeability, human intestinal absorption (HIA), oral bioavailability and P-gp binding models. Among the several ML techniques that were applied, the AECF routine selected the SVM, RF and *k*-NN models as the best base classifiers. Two independent datasets underwent the AECF procedure, and good average values for the area under the receiver operating characteristic curve (AUROC) resulted from all of the models (with the AUROC ranging from 0.78 to 0.91). Further studies using the DrugBank database [39] validated the AECF approach as a useful multiple classifier system for ADME predictions.

Stratton *et al.* developed a naive Bayes (NB) model to predict mouse liver microsomal (MLM) stability (Fig. 1b) [40]. A series of 88 analogs of the metabolically unstable antitubercular agent CD117 [MLM half-lifetime ($t_{1/2}$) <1 min] was used as the training set. Taking CD117 as a reference compound, 411 analogs were further designed. Thirteen molecules that were predicted to be metabolically stable were then synthesized and tested for MLM stability. All of the synthesized molecules outperformed the parent compound in terms of metabolic stability. These results demonstrated the robustness of the NB model for predicting the MLM stability of novel antitubercular agents and highlights its usefulness in lead-optimization campaigns focused on congeneric series.

The ability to cross the BBB, measured by the blood–brain partitioning coefficient (logBB), was investigated for 287 compounds in a study by Zhu *et al.* [41]. QSPR models based on SVM, MLR, multivariate adaptive regression splines (MARS) and RF were built. The statistical robustness and predictive power of these models were assessed, and an analysis of their applicability domain was performed. The RF algorithm (predictive correlation coefficient, $r^2_{\text{pred}} = 0.84$) outperformed the other ML methods and identified the polar surface area (PSA) and octanol–water partition coefficient (logP) as the main determinants of BBB penetration. These results demonstrated the superiority of more-recent ML

techniques, such as RF, over traditional methods, such as MLR and partial least squares (PLS), for the prediction of logBB [42,43].

PPB was investigated by Wang *et al.* on a dataset of 1830 structurally diverse compounds [44]. Six predictive models were built using PLS, RF, SVM, Cubist, Gaussian process (GP) and Boosting (Fig. 1c). Individually, the PLS analysis showed the worst results, whereas the RF algorithm exhibited the best performance. A consensus model was then built by combining the five best models, yielding $r^2_{pred} = 0.78$. Using two external validation datasets, the consensus model produced r^2_{ext} values of 0.70 for both sets. Lipophilicity, the presence of aromatic rings and the partial charge distribution were identified as the key drivers of PPB. The study showed the superior performance of newer techniques, such as RF, over broadly used tools, such as PLS, in assessing the PPB profile of the training and validation sets.

Kumar *et al.* used SVM, ANN, *k*-NN, probabilistic neural network (PNN), PLS and linear discriminant analysis (LDA) for the prediction of HIA [45]. A dataset of 1242 drugs and drug-like compounds was used to construct the models. The SVM algorithm, which used a radial basis function kernel, outperformed the other methods. An accuracy of 90.38% was obtained for the training set (745 compounds), whereas 91.54% was obtained for the test set (497 compounds). The predictive power was observed to decrease according to the following sequence: SVM, ANN, PNN, *k*-NN, LDA and PLS. In another study, the authors used the same group of ML algorithms to develop QSPR models for the prediction of PPB [46]. Again, SVM outperformed the other methods in terms of its predictive power for the training (442 drug-like compounds, accuracy of 89.73%) and test sets (294 drug-like compounds, accuracy of 89.97%). The accuracy of the ANN model (86.91 and 88.12%; training and test set, respectively) was comparable to that of the SVM analysis. The LDA method had the poorest performance, yielding accuracy values of 78.12% and 79.51% for the training and test sets, respectively. Among the techniques used in the two studies, the SVM models performed very well, indicating their usefulness for the prediction of HIA and PPB of drug-like compounds.

Finkelmann *et al.* developed a method for the identification of sites of metabolism (SoMs) based on the atom's electronic and steric surroundings and relative location in the molecule [47]. A dataset consisting of 678 CYP substrates was used as the training set (Fig. 2a). Density functional theory (DFT) quantum mechanical calculations were run to derive partial charges. Steric environments and atom locations were attributed by using autocorrelation descriptors, radial distribution function descriptors, the topological distance and the 3D Euclidean distance. The models were trained using different techniques, such as ERTs, gradient boosting tree, SVM and RF, with the latter yielding the best results in terms of the Matthews correlation coefficient (MCC = 0.63). Additionally, a series of 25 compounds with known metabolic profiles was used as an external validation set (MCC = 0.55). This study represents an insightful combination of quantum mechanics, topological and distance-based metrics applied to the prediction of SoMs.

A set of RF models for the prediction of SoMs called fast metabolizer (FAME) was recently updated [48]. In the new version (FAME 2), Šícho *et al.* implemented a tree-based ensemble method termed extremely randomized trees (ERTs) to investigate the influence of topological and quantum molecular descriptors on prediction accuracy [49,50]. The best model produced an AUROC of 0.91 upon validation with an external test set. Additionally, models for predicting the regioselectivity of CYP3A4, CYP2D6 and CYP2C9 were developed and integrated into FAME 2. The package includes a graphical output that flags potential SoMs and indicates the probability of an atom participating in metabolic reactions (Fig. 2b). The software is available free of charge upon request.

Bocci *et al.* reported the development of ADME–Space, a tool that encloses 20 QSPR models for the prediction of ADMET properties, including metabolic stability, Caco-2 permeability, BBB permeability, P-gp binding and CYP450 inhibition [51]. The models were built on ~26 000 molecules and were based on the following algorithms: SVM, RF, ada boost (AB), LDA, gradient boosting (GB), DT, PLS and extra trees classifier (ETC) (Fig. 2c). In general, RF and SVM led to the best results. The best models were integrated to generate a 2D graphical output using the self-organizing map approach [52]. Self-organizing map analysis relies on an ANN that converts data from an *n*-dimensional matrix into a 2D map. The translation of multiple models into a single 2D fingerprint enables the simultaneous monitoring of different properties and the detection of likely ADMET drawbacks. Moreover, it is a straightforward way to evaluate how far newly designed compounds are from the optimal PK space.

Web-accessible tools

Dong *et al.* recently announced a freely accessible web-based resource for PK and toxicity prediction called ADMETlab (http://admet.scbdd.com) [53]. The QSPR models used in the platform were constructed using ~289 000 compounds and six algorithms: recursive partitioning (RP), DT, RF, SVM, PLS and NB. The regression models were built with RF, SVM, RP and PLS, and the classification models were generated using RF, SVM, NB and DT (Fig. 3a). In general, RF produced the best results for classification and regression analyses. ADMETlab allows the user to conduct drug-likeness analyses based on five rules: Lipinski, Ghose, Veber, Varma and Oprea [19,20,21,54,55]. Moreover, 31 predictive models are available for physicochemical properties (three models), absorption (six), distribution (three), metabolism (ten), elimination (two) and toxicity (seven). Given the amount of data collected and the number of QSPR models, ADMETlab is one of the most thorough platforms that has recently been made available for ADMET prediction.

Tian *et al.* reported on the development of a freely available tool to predict CYP450 reactants [56]. CypReact (https://bitbucket.org/Leon_Ti/cypreact) predicts whether query molecules are likely to undergo metabolism by any of the CYP450 isoforms: 1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1 and 3A4 (Fig. 3b). The algorithm was trained with a dataset containing 1632 reactants and decoys. SVM, logistic regression (LR), DT, RF and an ensemble method were used to build the classification models to distinguish between CYP450 reactants and non-reactants. The best model for each isoform produced AUROC scores that varied from 83 to 92%. The RF model was the most accurate for the 1A2, 2A6, 2B6, 2C8, 2C19, 2E1 and 3A4 isoforms, whereas the ensemble method was the best for 2C9 and 2D6.

Quinone metabolites produced by CYP450 enzymes and peroxidases are highly reactive Michael acceptors and are responsible for a

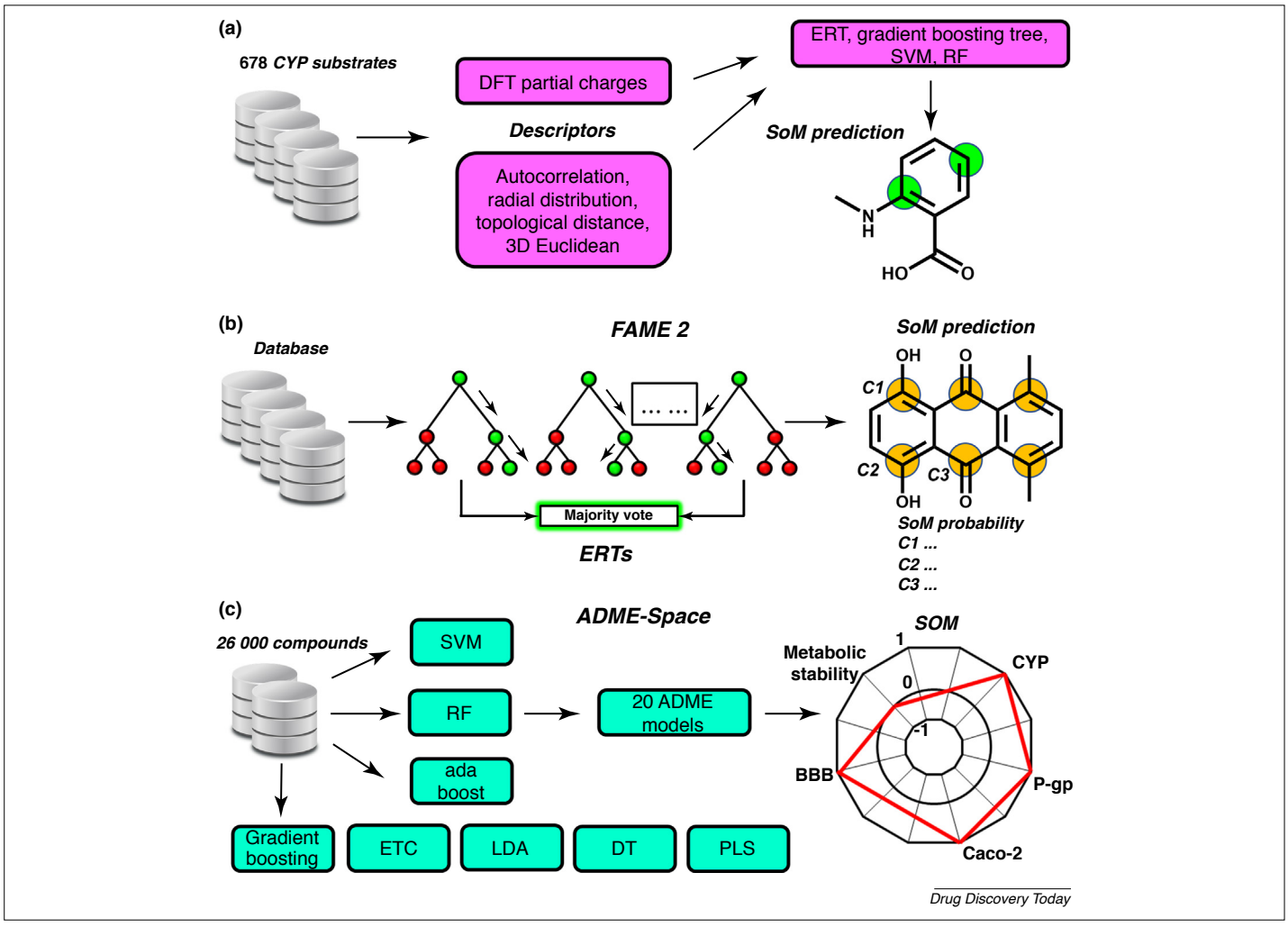

FIGURE 2

Regression and classification approaches for the prediction of pharmacokinetics (PK) and toxicity properties. **(a)** The use of extremely randomized trees (ERTs), gradient boosting tree, support vector machines (SVM) and Random Forest (RF) for the identification of sites of metabolism (SoMs). Density functional theory (DFT)-derived partial charges, topological and distance-based metrics were used as descriptors to develop the models. **(b)** The incorporation of ERTs into FAME 2 for the prediction of SoMs. **(c)** ADME–Space: a set of 20 quantitative structure–property relationship (QSPR) models for the prediction of metabolic stability, Caco-2 permeability, blood–brain barrier (BBB) permeability, P-glycoprotein (P-gp) binding and cytochrome P450 (CYP450) inhibition. The models rely on SVM, RF, ada boost (AB), linear discriminant analysis (LDA), gradient boosting (GB), decision trees (DT), partial least squares (PLS) and extra trees classifier (ETC). Abbreviation: SOM, self-organizing map.

variety of adverse effects [57]. Using a deep ANN on 718 training set compounds, Hughes and Swamidass were the first to develop an ML model to predict quinone formation by metabolic oxidation (Fig. 3c) [58]. The model succeeded in distinguishing quinone-forming molecules from non-quinone-forming molecules (AUROC = 88.2%). Moreover, the deep ANN identified atom pairs that are likely to originate quinones (AUROC = 97.6%). This functionality includes a color-coded graphical output that highlights and attributes a score for the quinone-generating atoms. The so-called XenoSite model is freely available (http://swami.wustl.edu/xenosite/p/quinone) and can be used as a practical resource for identifying molecules that are susceptible to producing quinone toxic metabolites.

A new and improved version of the web server FAF-Drugs was recently announced [59,60]. The new version, called FAF-Drugs4, includes several features focused on PK issues and can be freely accessed (http://fafdrugs4.mti.univ-paris-diderot.fr/) [60]. FAF-Drugs4 can filter large compound collections for profiling ADMET, flagging toxic groups, predicting physicochemical properties, identifying PAINS and providing structural alerts. Moreover, the new version includes a quantitative estimate for drug-likeness that was benchmarked using a dataset of 771 oral drugs.

Podlewska and Kafel communicated the development of MetStabOn, a freely available online server for metabolic stability prediction [61]. Separate training sets for metabolic stability on human (5234 compounds), rat (2829) and mouse (1136) liver microsomes and plasma were used to derive regression and classification models (Fig. 4a). The regression model was generated using sequential minimal optimization (SMO), which is an adaptation of SVM [62], and five classification systems were generated using SMO, RF, NB, *k*-NN and DT. The SMO, *k*-NN and RF models outperformed those based on DT and NB in terms of their predictive power for the training sets. When external test sets were used, SMO, NB and RF outperformed the other ML techniques. In addition to predicting clearance, $t_{1/2}$ and metabolic stability, for each query compound, the platform provides a list of similar structures that are part of the training set.

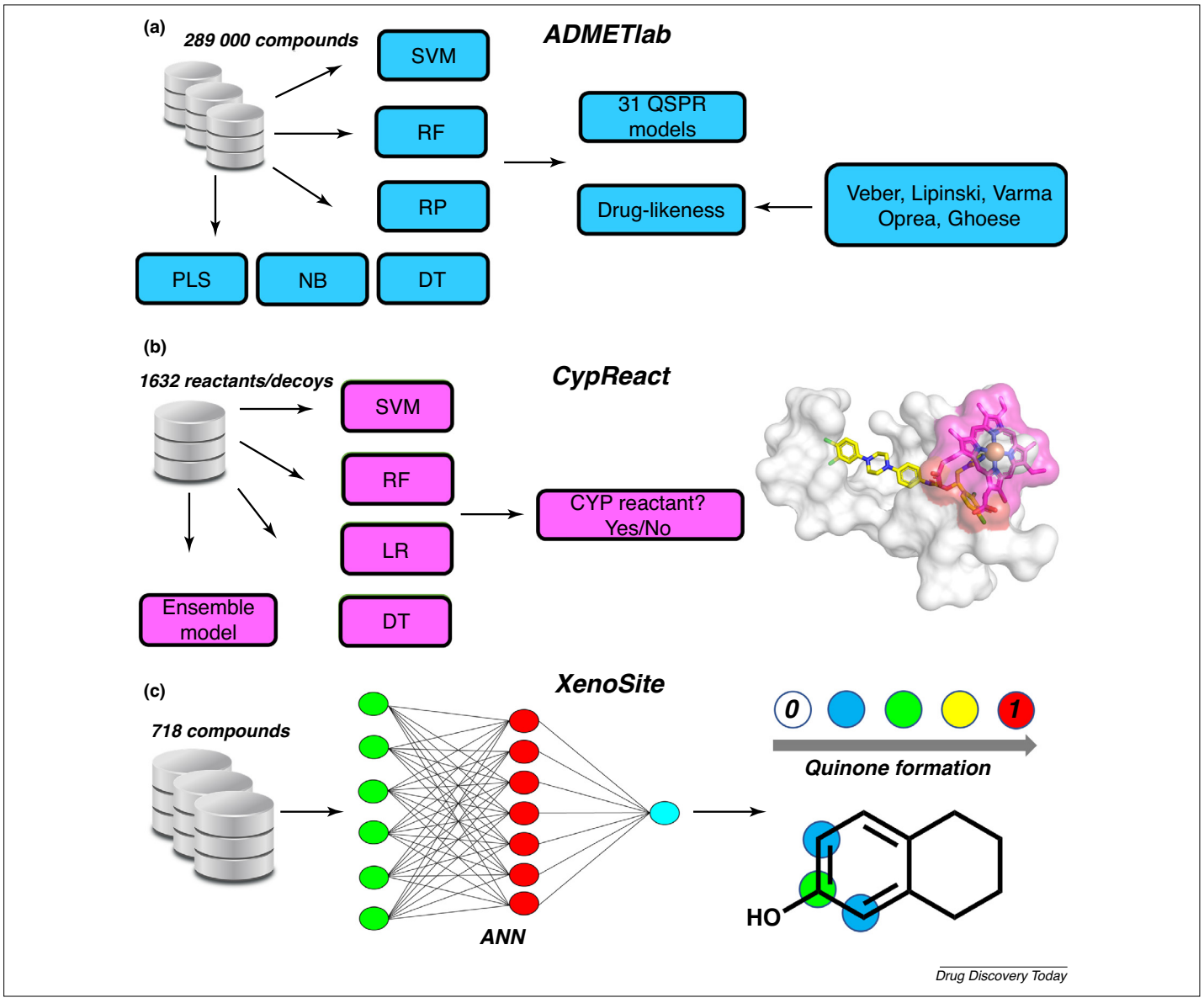

FIGURE 3

Design and implementation of freely accessible resources for pharmacokinetics (PK) profiling. **(a)** ADMETlab: a web server for the prediction of PK and toxicity endpoints. The models were constructed using recursive partitioning (RP), decision trees (DT), Random Forest (RF), support vector machines (SVM), partial least squares (PLS) and naive Bayes (NB). **(b)** The CypReact tool to predict CYP450 reactants. A dataset of 1632 reactants and decoys was used to build the SVM, logistic regression (LR), DT, RF and ensemble classification models. **(c)** XenoSite: an artificial neural network (ANN) to predict quinone metabolites produced by metabolic oxidation via cytochrome P450 (CYP450) enzymes.

SwissADME was recently introduced as a web-based tool for ADMET modeling (http://www.swissadme.ch/index.php) [63]. SwissADME enables the prediction of ten physicochemical properties, in addition to comprising six lipophilicity (logP) and three water solubility (logS) models (Fig. 4b). Moreover, nine PK models are available: gastrointestinal absorption, BBB permeability, P-gp binding, skin permeation (logKp) and CYP450 inhibition for five of the isoforms. Additionally, drug-likeness can be assessed by applying different filters: Lipinski, Ghose, Veber, Egan and Muegge. This functionality enables a straightforward identification of the number of rule violations and which properties disagree with these rules. Additionally, medicinal chemistry alerts are given, such as PAINS, Brenk structural alert [64], lead-likeness [65] and synthetic accessibility. SwissADME models were built using SVM on large datasets, and molecular and physicochemical features as descriptors. The skin-permeation model relies on a previously published MLR analysis [66]. The SwissADME output includes the so-called 'BOILED-Egg' plot, which allows evaluation on how gastrointestinal absorption and BBB penetration vary with logP and the topological polar surface area [67]. A 'bioavailability radar' is also provided. This drug-likeness plot encloses a colored zone that represents the ideal physicochemical landscape and graphically indicates which properties fall outside this zone. SwissADME offers a rapid way to input query molecules and obtain results, which can be easily saved in table format. Each of the predicted endpoints includes a readily visible summary of the enclosed model. These and other features make SwissADME one of the most practical tools recently developed for ADMET prediction.

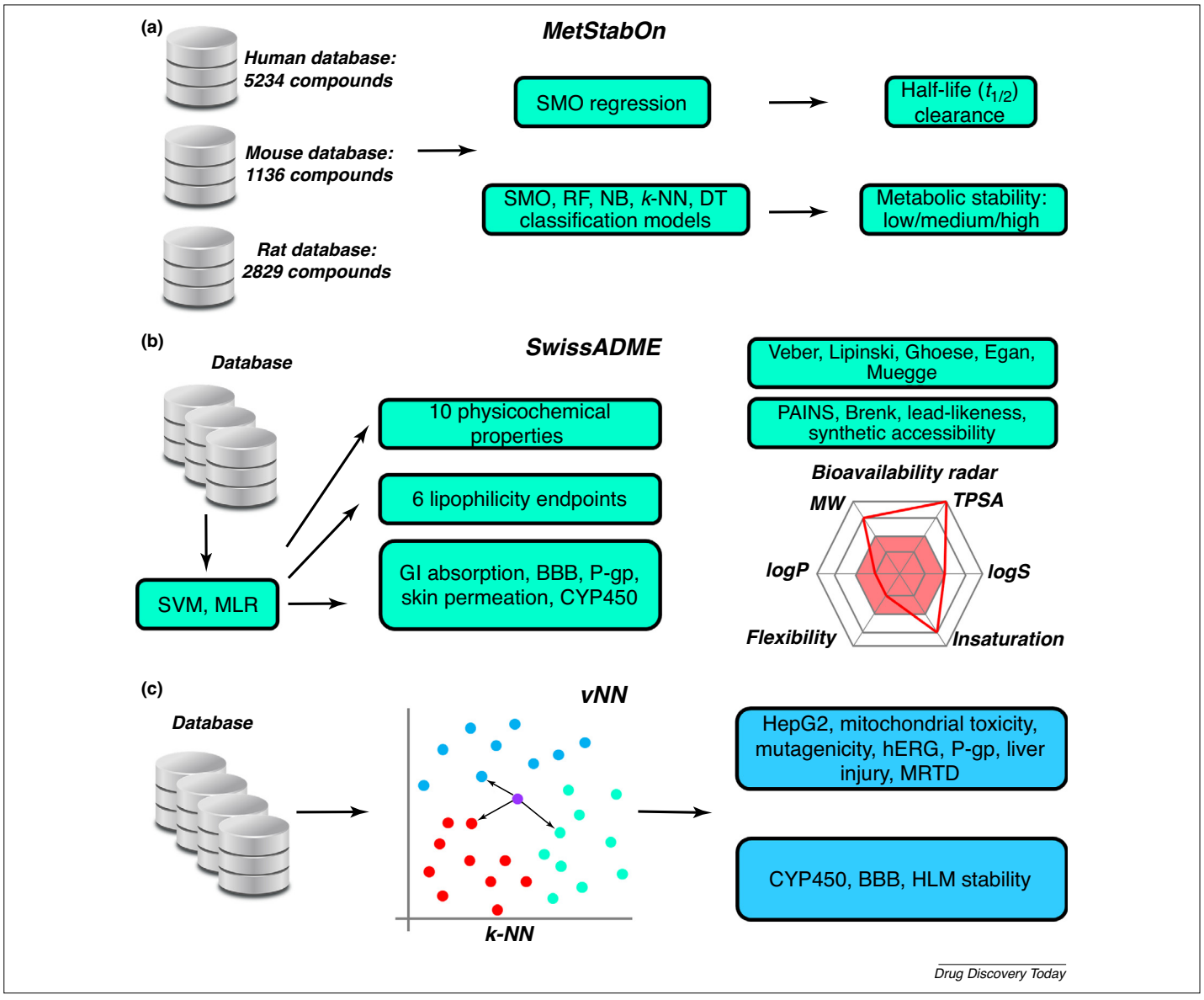

FIGURE 4

Recently implemented web servers for pharmacokinetics (PK) investigations. **(a)** MetStabOn: an online server for metabolic stability prediction. Regression and classification models were generated by applying sequential minimal optimization (SMO), Random Forest (RF), decision trees (DT), naive Bayes (NB) and *k*-nearest neighbors (*k*-NN). **(b)** SwissADME: a web-based tool for PK modeling. The platform enables the prediction of physicochemical properties, gastrointestinal (GI) absorption, blood–brain barrier (BBB) penetration, P-glycoprotein (P-gp) binding, cytochrome P450 (CYP450) inhibition and skin permeation. In addition, the identification of pan assay interference compounds (PAINS) and evaluation of drug-likeness, lead-likeness, synthetic accessibility and medicinal chemistry suitability are provided. The models were built using support vector machines (SVM), except for the skin-permeation model, which is based on multiple linear regression (MLR). **(c)** vNN: a platform that features 15 models for the prediction of PK and toxicity parameters. The models were built using *k*-NN and provide predictions for human hepatoblastoma cell (HepG2) toxicity, mitochondrial toxicity, mutagenicity, ether-a-go-go-related protein 2 (hERG) blocking, P-gp binding, liver injury, maximum recommended therapeutic dose (MRTD), CYP450 inhibition, BBB permeability and human liver microsomal (HLM) stability. *Abbreviations*: MW, molecular weight; logP, octanol–water partition coefficient; logS, aqueous solubility; TPSA, topological polar surface area.

The vNN web server (https://vnnadmet.bhsai.org/) is a freely accessible resource that includes 15 models for the prediction of ADMET endpoints (Fig. 4c) [68]. As a singular feature, the site allows the user to customize the inbuilt models by adding their own experimental data. It is also possible to generate entirely new regression and classification analyses solely based on the user's data. vNN models are based on a *k*-NN algorithm that is adapted to apply a structural similarity function to define the model's applicability domain (AD) [69]. The user can choose between results based on a restricted AD only or results based on an unrestricted AD. The platform provides predictions for safety-related properties, such as cytotoxicity to HepG2 cells, mitochondrial toxicity, mutagenicity (AMES test), cardiotoxicity (hERG blocking), P-gp binding, liver injury and the maximum recommended therapeutic dose. PK endpoints, such as CYP450 inhibition for five isoforms, BBB permeability and human liver microsomal stability, are also provided.

Stork *et al.* recently reported Hit Dexter 2.0, a ML method to predict frequent hitters [70]. Addressing primary screening and confirmatory dose–response tests, the development of Hit Dexter relied on ML tools such as ETC, RF, AdaBoost and Bagging Classifier, with the best results being obtained with ETC (MCC values

from 0.56 to 0.58). In the external validation step, the best models correctly classified promiscuous and nonpromiscuous compounds with MCC values of up to 0.64 and AUROC values of up to 0.96. A noteworthy resource of Hit Dexter 2.0 is the ability of predicting promiscuous compounds among marketed drugs. Hit Dexter 2.0 predicted as promiscuous compounds ~13% of the marketed drugs; 6% were predicted as highly promiscuous. In addition, the webserver addresses a wide range of chemical spaces, for example dark chemical matter, aggregators, HTS compounds, drug-like molecules, PAINS and natural products. Hit Dexter 2.0 is available online (http://hitdexter2.zbh.uni-hamburg.de) where the users can access the inbuilt ML models and rules for identifying frequent hitters and substructures that are unsuitable for drug discovery.

Concluding remarks

In silico ADMET prediction has played an increasingly important part in drug R&D by providing an effective way to assess multiple PK properties in hit-to-lead and lead-optimization campaigns. The importance of such a strategy has grown alongside the evolution of chemoinformatics, which, from the landmark developments in the 1960s and through the widely applied drug-likeness concepts of the 1990s, has entered the big data age. In this progressively challenging scenario and given the need to handle increasingly larger and more-heterogeneous datasets, advanced ML approaches have stood out over traditional chemometrics techniques. In the field of ADMET modeling, the importance of ML approaches is corroborated by studies published during 2017 and 2018 that predominantly relied on RF, SVM and tree-based methods. It is worth mentioning that, in most of the cases discussed herein, RF and SVM outperformed other approaches that were used in parallel. Another important aspect to note is the translation of the resulting models into well-structured and user-friendly online platforms that can be freely accessed by the worldwide drug discovery community. Out of the 17 selected studies, seven have been converted into open-web servers for ADMET prediction and two have originated software that is available free of charge. Some of these tools were built on comprehensive and curated datasets and provide tens of well-validated PK models. By including other relevant properties, such as safety endpoints, structural alerts, drug- and lead-likeness, and physicochemical features, these resources have effectively contributed to the advancement of drug R&D.

Conflicts of interest

The authors declare no competing financial interests.

Acknowledgments

We gratefully acknowledge the financial support from the State of Sao Paulo Research Foundation (FAPESP, Fundação de Amparo à Pesquisa do Estado de São Paulo) and the National Council for Scientific and Technological Development (CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil.

References

1 Segall, M. (2014) Advances in multiparameter optimization methods for *de novo* drug design. *Expert Opin. Drug Discov.* 9, 803–817
2 Cheng, F. *et al.* (2013) *In silico* ADMET prediction: recent advances, current challenges and future trends. *Curr. Top. Med. Chem.* 13, 1273–1289
3 González-Medina, M. *et al.* (2017) Open chemoinformatic resources to explore the structure, properties and chemical space of molecules. *RSC Adv.* 7, 54153–54163
4 Laoui, A. and Polyakov, V.R. (2011) Web services as applications' integration tool: QikProp case study. *J. Comput. Chem.* 32, 1944–1951
5 Sander, T. *et al.* (2015) DataWarrior: an open-source program for chemistry aware data visualization and analysis. *J. Chem. Inf. Model.* 55, 460–473
6 Rudik, A.V. *et al.* (2017) MetaTox: web application for predicting structure and toxicity of xenobiotics' metabolites. *J. Chem. Inf. Model.* 57, 638–642
7 Trunzer, M. *et al.* (2009) Metabolic soft spot identification and compound optimization in early discovery phases using MetaSite and LC-MS/MS validation. *J. Med. Chem.* 52, 329–335
8 Waring, M.J. *et al.* (2015) An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discov.* 14, 475–486
9 Dahlin, J.L. *et al.* (2015) PAINS in the assay: chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging HTS. *J. Med. Chem.* 58, 2091–2113
10 Fleming, N. (2018) How artificial intelligence is changing drug discovery. *Nature* 557, S55–S57
11 Ray, L.C. and Kirsch, R.A. (1957) Finding chemical records by digital computers. *Science* 126, 814–819
12 Hansch, C. *et al.* (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 194, 178–180
13 Hansch, C. and Fujita, T. (1964) Rho sigma pi analysis: a method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* 86, 1616–1626
14 Topliss, J.G. and Costello, R.J. (1972) Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* 15, 1066–1068
15 Free, S.M. and Wilson, J.W. (1964) A mathematical contribution to structure-activity studies. *J. Med. Chem.* 7, 395–399
16 Cramer, R.D. *et al.* (1974) Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem.* 17, 533–535
17 Timmermans, P.B. *et al.* (1977) Lipophilicity and brain disposition of clonidine and structurally related imidazolidines. *Naunyn. Schmiedebergs. Arch. Pharmacol.* 300, 217–226
18 Hinderling, P.H. *et al.* (1984) Quantitative relationships between structure and pharmacokinetics of beta-adrenoceptor blocking agents in man. *J. Pharmacokinet. Biopharm.* 12, 263–287
19 Lipinski, C.A. *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
20 Ghose, A.K. *et al.* (1999) A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* 1, 55–68
21 Veber, D.F. *et al.* (2002) Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45, 2615–2623
22 Egan, W.J. *et al.* (2000) Prediction of drug absorption using multivariate statistics. *J. Med. Chem.* 43, 3867–3877
23 Muegge, I. *et al.* (2001) Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* 44, 1841–1846
24 Lavecchia, A. (2015) Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* 20, 318–331
25 Tao, L. *et al.* (2015) Recent progresses in the exploration of machine learning methods as in-silico ADME prediction tools. *Adv. Drug Deliv. Rev.* 86, 83–100
26 Sakiyama, Y. (2009) The use of machine learning and nonlinear statistical tools for ADME prediction. *Expert Opin. Drug Metab. Toxicol.* 5, 149–169
27 Shen, M. *et al.* (2003) Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J. Med. Chem.* 46, 3013–3020
28 Heikamp, K. and Bajorath, J. (2014) Support vector machines for drug discovery. *Expert Opin. Drug Discov.* 9, 93–104
29 Doniger, S. *et al.* (2002) Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms. *J. Comput. Biol.* 9, 849–864
30 Cano, G. *et al.* (2017) Automatic selection of molecular descriptors using random forest: application to drug discovery. *Expert Syst. Appl.* 72, 151–159
31 Breiman, L. (2001) Random Forests. *Mach. Learn.* 45, 5–32
32 Svetnik, V. *et al.* (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43, 1947–1958
33 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
34 Gobburu, J.V. and Shelver, W.H. (1995) Quantitative structure–pharmacokinetic relationships (QSPR) of beta blockers derived using neural networks. *J. Pharm. Sci.* 84, 862–865

35 Haixiang, G. *et al.* (2017) Learning from class-imbalanced data: review of methods and applications. *Expert Syst. Appl.* 73, 220–239
36 Norinder, U. *et al.* (2014) Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. *J. Chem. Inf. Model.* 54, 1596–1603
37 Svensson, F. *et al.* (2016) Modelling compound cytotoxicity using conformal prediction and PubChem HTS data. *Toxicol. Res.* 6, 73–80
38 Yang, M. *et al.* (2018) A novel adaptive ensemble classification framework for ADME prediction. *RSC Adv.* 8, 11661–11683
39 Wishart, D.S. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082
40 Stratton, T.P. *et al.* (2017) Addressing the metabolic stability of antituberculars through machine learning. *ACS Med. Chem. Lett.* 8, 1099–1104
41 Zhu, L. *et al.* (2018) ADME properties evaluation in drug discovery: *in silico* prediction of blood–brain partitioning. *Mol. Divers.* 22, 979–990
42 Norinder, U. *et al.* (1998) Theoretical calculation and prediction of brain–blood partitioning of organic solutes using MolSurf parametrization and PLS statistics. *J. Pharm. Sci.* 87, 952–959
43 Ooms, F. *et al.* (2002) A simple model to predict blood–brain barrier permeation from 3D molecular fields. *Biochim. Biophys. Acta* 1587, 118–125
44 Wang, N.N. *et al.* (2017) ADME properties evaluation in drug discovery: prediction of plasma protein binding using NSGA-II combining PLS and consensus modeling. *Chemometr. Intell. Lab. Syst.* 170, 84–95
45 Kumar, R. *et al.* (2017) Prediction of human intestinal absorption of compounds using artificial intelligence techniques. *Curr. Drug Discov. Technol.* 14, 244–254
46 Kumar, R. *et al.* (2018) Prediction of drug–plasma protein binding using artificial intelligence based algorithms. *Comb. Chem. High Throughput Screen.* 21, 57–64
47 Finkelmann, A.R. *et al.* (2017) Site of metabolism prediction based on *ab initio* derived atom representations. *ChemMedChem* 12, 606–612
48 Kirchmair, J. *et al.* (2013) FAst MEtabolizer (FAME): a rapid and accurate predictor of sites of metabolism in multiple species by endogenous enzymes. *J. Chem. Inf. Model.* 53, 2896–2907
49 Geurts, P. *et al.* (2006) Extremely randomized trees. *Mach. Learn.* 63, 3–42
50 ícho, M. *et al.* (2017) FAME 2: simple and effective machine learning model of cytochrome P450 regioselectivity. *J. Chem. Inf. Model.* 57, 1832–1846
51 Bocci, G. *et al.* (2017) ADME–Space: a new tool for medicinal chemists to explore ADME properties. *Sci. Rep.* 7, 6359
52 Kohonen, T. (1990) The self-organizing map. *Proc. IEEE* 78, 1464–1480
53 Dong, J. *et al.* (2018) ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. *J. Cheminform.* 10, 29
54 Varma, M.V.S. *et al.* (2010) Physicochemical space for optimum oral bioavailability: contribution of human intestinal absorption and first-pass elimination. *J. Med. Chem.* 53, 1098–1108
55 Oprea, T.I. (2000) Property distribution of drug-related chemical databases. *J. Comput. Aid. Mol. Des.* 14, 251–264
56 Tian, S. *et al.* (2018) CypReact: a software tool for *in silico* reactant prediction for human cytochrome P450 enzymes. *J. Chem. Inf. Model.* 58, 1282–1291
57 Bolton, J.L. and Dunlap, T. (2017) Formation and biological targets of quinones: cytotoxic versus cytoprotective effects. *Chem. Res. Toxicol.* 30, 13–37
58 Hughes, T.B. and Swamidass, S.J. (2017) Deep learning to predict the formation of quinone species in drug metabolism. *Chem. Res. Toxicol.* 30, 642–656
59 Miteva, M.A. *et al.* (2006) FAF-Drugs: free ADME/tox filtering of compound collections. *Nucleic Acids Res.* 34, W738–W744
60 Lagorce, D. *et al.* (2017) FAF-Drugs4: free ADME-tox filtering computations for chemical biology and early stages drug discovery. *Bioinformatics* 33, 3658–3660
61 Podlewska, S. and Kafel, R. (2018) MetStabOn – online platform for metabolic stability predictions. *Int. J. Mol. Sci.* 19, 1040
62 Shevade, S.K. *et al.* (2000) Improvements to the SMO algorithm for SVM regression. *IEEE Trans. Neural. Netw.* 11, 1188–1193
63 Daina, A. *et al.* (2017) SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep* 7, 42717
64 Brenk, R. *et al.* (2008) Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem* 3, 435–444
65 Teague, S.J. *et al.* (1999) The design of leadlike combinatorial libraries. *Angew. Chem. Int. Ed.* 38, 3743–3748
66 Potts, R.O. and Guy, R.H. (1992) Predicting skin permeability. *Pharm. Res.* 9, 663–669
67 Daina, A. and Zoete, V. (2016) A BOILED-Egg to predict gastrointestinal absorption and brain penetration of small molecules. *ChemMedChem* 11, 1117–1121
68 Schyman, P. *et al.* (2017) vNN web server for ADMET predictions. *Front. Pharmacol.* 8, 889
69 Liu, R. *et al.* (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
70 Stork, C. *et al.* (2019) Hit Dexter 2.0: machine-learning models for the prediction of frequent hitters. *J. Chem. Inf. Model.* . http://dx.doi.org/10.1021/acs.jcim.8b00677