Drug Discovery Today • Volume xxx, Number xx • xxxx 2021 REVIEWS

ELSEVIER

A decade of machine learning-based predictive models for human pharmacokinetics: Advances and challenges

INFORMATICS

Danishuddin [a], Vikas Kumar [a], Mohammad Faheem [b], Keun Woo Lee [a,*]

[a] Department of Bio & Medical Big Data (BK4), Division of Life Sciences, Research Institute of Natural Sciences (RINS), Gyeongsang National University (GNU), 501 Jinju-daero, Jinju 52828, Republic of Korea
[b] Department of Biotechnology, Indian Institute of Technology, Roorkee, Uttarakhand 247667, India

Traditionally, *in vitro* and *in vivo* methods are useful for estimating human pharmacokinetics (PK) parameters; however, it is impractical to perform these complex and expensive experiments on a large number of compounds. The integration of publicly available chemical, or medical Big Data and artificial intelligence (AI)-based approaches led to qualitative and quantitative prediction of human PK of a candidate drug. However, predicting drug response with these approaches is challenging, partially because of the adaptation of algorithmic and limitations related to experimental data. In this report, we provide an overview of machine learning (ML)-based quantitative structure–activity relationship (QSAR) models used in the assessment or prediction of PK values as well as databases available for obtaining such data.

Keywords: Pharmacokinetics; QSAR; Chemical Big Data; Drug development

Introduction

The discovery and optimization of any therapeutic agent with desirable PKs and pharmacodynamics (PDs) properties is the key focus of the drug development process. This is especially important in any drug discovery process, because, for any new drug candidate, a poorly designed study with incorrect dosing threshold can cause misleading results, which might be costly to the pharmaceutical company.[1,2] Druggability is mainly dependent on the metabolism and PK properties of the drug (the DMPK), which are the main hurdles in pharmaceutical R&D. Human PK evaluation of preclinical drugs are vital for reducing the failure rate in clinical trials.[3,4] Although issues related to PK have improved over the past few decades, nearly half of all the therapeutic candidates in drug development are lost because of poor absorption, distribution, metabolism, excretion (ADME) and toxicities.[5–7] Minimizing the toxicity and optimizing PK is crucial for any clinical trial. Therefore, the estimation of PK properties in humans before the first-in-human clinical trial is one of the main purposes of nonclinical studies during the drug development process. Traditionally, *in vitro–in vivo* extrapolation (IVIVE) and allometric scaling (AS) methods have been widely used for estimation of human PK values.[8,9] Although these methods are useful tools, it is challenging and impractical to apply these complex and expensive experiments to a large number of compounds. As a cost-effective and high-throughput alternative to experimental methods, computational approaches to predict PK parameters have proven increasingly appealing.[10–13] Over the past few decades, PK/PD modeling approaches have made many advances from the basic concept of the dose–response relationship to extended mechanism-based methods. Csajka and Verotta discussed the historical progression of the modeling of the concentration–response relationship models to some of the more sophisticated approaches.[14] Although such mechanistic methods require a priori knowledge, as well as high-quality data for reliable prediction, ML-guided predictions are based on multivariate experimental data for training. A basic introduction to

* *Corresponding author.Woo Lee, K. (kwlee@gnu.ac.kr)*

https://doi.org/10.1016/j.drudis.2021.09.013

Please cite this article in press as: Danishuddin et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.013

these ML methods in the context of drug discovery is provided elsewhere.[15] *In silico*-based based approaches are being used extensively for modeling ADMET or PK properties (Figs. 1 and 2). Application of such prediction models to assess human PK not only avoids preclinical to clinical extrapolation, but is also significantly less resource-intensive compared with traditional approaches. Validating and interpretating these approaches for predicting PK in humans are beneficial to the design of *in vitro and in vivo* experiments. Additionally, *in silico*-based PK prediction can also reduce the direct use of animal data, which are hard to extrapolate to humans. Recent advances in publicly available chemical Big Data open a new chapter in the search for more optimized and accurate computational prediction models.[16] However, regardless of their potential, it is important to scrutinize them and compare their performance to distinguish progress made with them. Such careful and rigorous analytical is

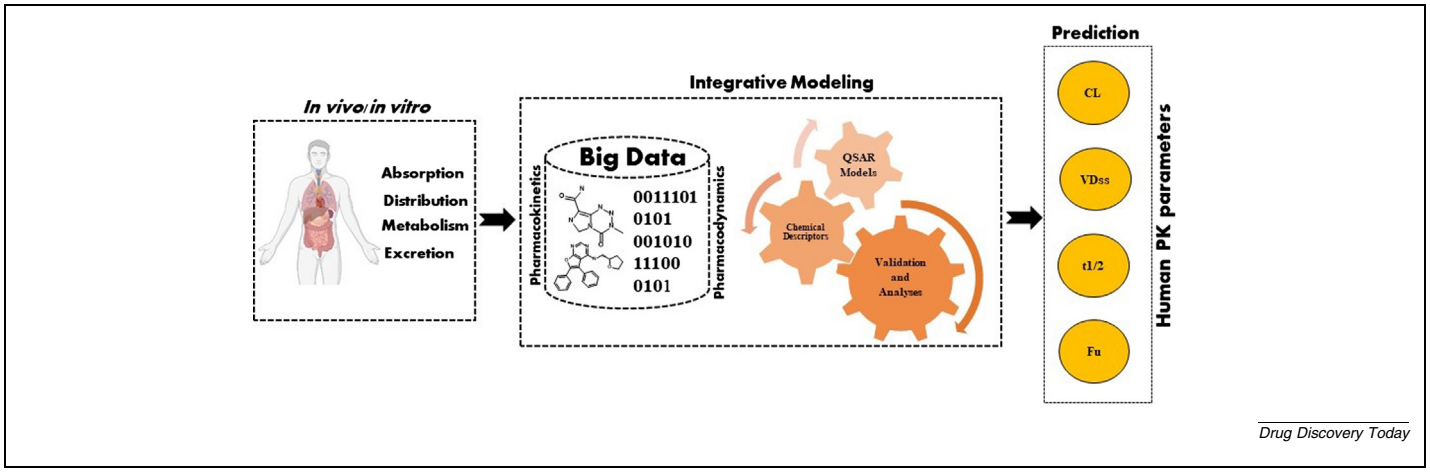

FIGURE 1

Modeling process for human pharmacokinetic (PK) parameter prediction. Abbreviations: CL, clearance; F_{up}, fraction of unbound protein; QSAR, quantitative structure–activity relationship; $t_{½}$, half-life; VDss, volume of distribution at steady state.

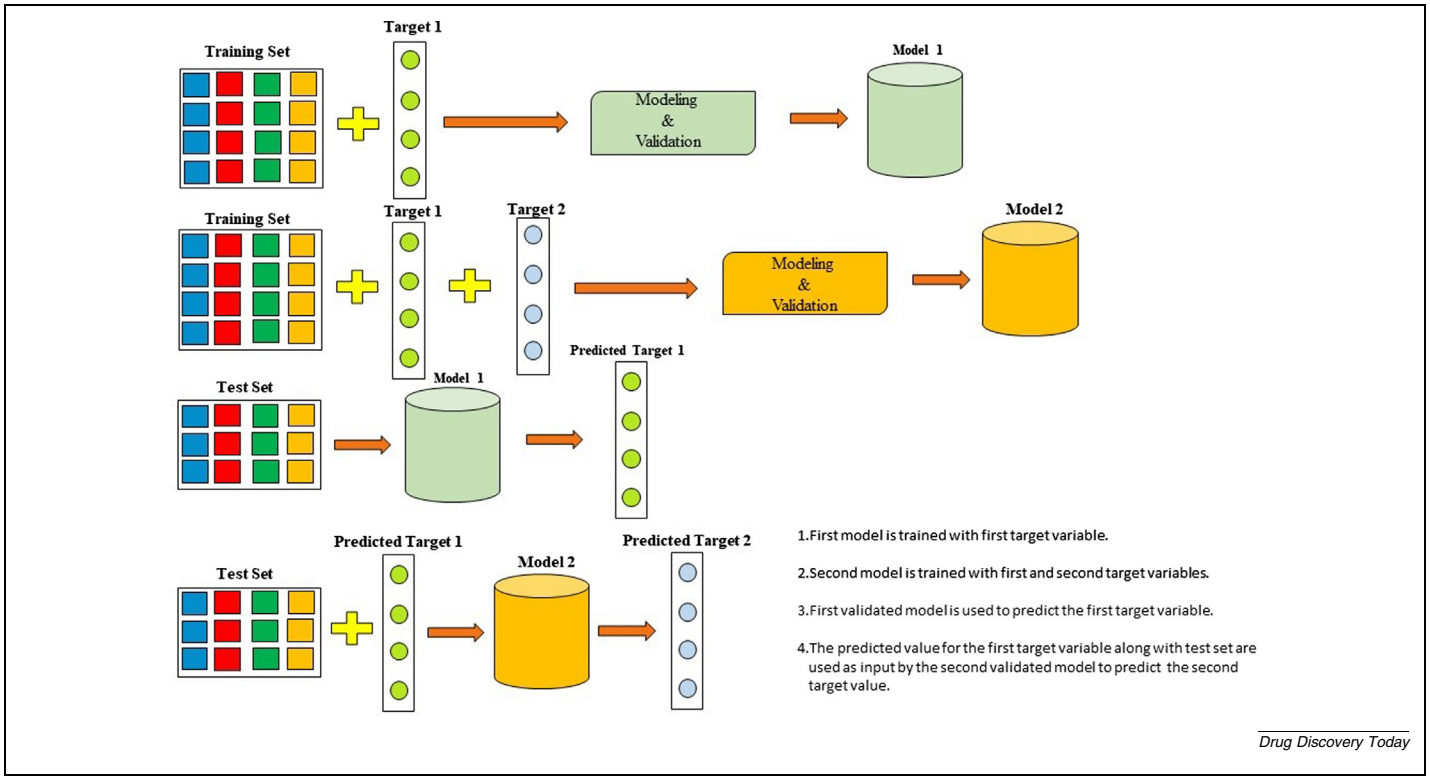

FIGURE 2

Proposed architecture for parallel modeling of pharmacokinetics (PK) parameters.

Please cite this article in press as: Danishuddin et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.013

necessary to understand the data and its integration into ML approaches. In this review, we focus on ML-based models for PK parameter prediction, because they are directly available to researchers, more transparent, and enable the prediction of the efficacy PK values of drugs. We explore recent advances in human PK prediction and discuss ongoing work to address challenges in evaluating, interpreting, and implementing artificial techniques for molecular design.

Volume of distribution steady-state

The volume of distribution at steady state (VDss) is a vital PK parameter that is considered a reliable indicator of drug distribution in the body; it measures the relative affinity of a drug for tissues and plasma.[17] Drugs that bind equally to plasma proteins and are extensively taken up by tissues have a high volume of distribution, whereas drugs that bind extensively to plasma proteins have a lower volume of distribution. Generally, acidic drugs have low (<1 l/kg), neutrals and zwitterions have moderate (1–3 l/kg), and basic drugs have high (>3 l/kg) VDss values.[18,19] In addition, VDss also has an important role in calculating the mean residence time (MRT) and half-life ($t_{½}$) of a drug, thus establishing a suitable dosing regimen.[20,21] Early assessment of VDss is important for developing medication decision and dose measurement plans. Methods used to predict VDss include: (i) extrapolation of *in vivo* animal data; (ii) physiologically based PK (PBPK) modeling; and (3) computational approaches. Various computational-based QSAR models have been reported for VDss prediction. These models have been instrumental in the application of the prediction of VDss values that were either not feasible, costly, or too time consuming previously (Table 1).

Predictive models for VD_{SS}

Zhivkova and Doytchinova[22] developed QSAR models using 132 acidic drugs. Models were built using a total of 178 molecular descriptors and a stepwise linear regression approach. The developed models outperformed XXXX with $Q^2_{\text{Loo-CV}}$ of 0.581 and 0.625. The presence of sulfur and sulfuric groups was found to significantly impact the VDss value predictions for acidic drugs. Despite the smaller descriptors set and drug molecules, results from these predictions were consistent with experimental studies. In addition, Zhivkova *et al.* also developed models for predicting VDss values for 216 basic drugs.[23] The models were developed by applying genetic algorithms, stepwise regressions, multiple linear regressions, and 179 structural and physicochemical descriptors. The best model performed with R^2 of 0.663. Descriptors such as lipophilicity, fraction ionized, number of

TABLE 1

ML-based human PK predictive models.[a,b]

Data set size	Descriptors	Method(s)	Performance/Comments	Refs
VDss				
132 (acidic drugs)	Molecular descriptors (total descriptors = 178)	GA, SR	Mean fold error values from 1.58 (cross-validation) to 2.25 (external validation)	22
216 (basic drugs): training set, 180; test set, 36	Structural, physicochemical	GA, SR, MLR	$R^2 = 0.663$, $Q^2_{\text{LOO-CV}} = 0.606$, $R^2_{\text{pred}} = 0.593$	23
1236	Mordred, PaDEL, MOE	RF, SVM, XGB, GBM	Best model: SVM ($R^2_{\text{test}} = 0.87$, $\text{RMSE}_{\text{test}} = 0.2$, $Q^2 = 0.77$)	24
1303: training set, 1090; test set, 213	logD7.4, MW, max basic pKa, min acidic pKa, Rule 13, aromatic atoms	RF	59% and 78% within a twofold and threefold error, respectively	26
604	ACDlabs, MOE, Kt:p	Decision tree	Mean fold error of best model = 2.29	32
412	ECFP, PubChem, molecular descriptors	Multitask DL	~63.33% accuracy	34
670	Volsurf+, ACDlabs, MOE	RP, PLS	$Q^2 = 0.70$	43
569	E-State	SVR, MLR	$R^2_{\text{test}} = 0.782$ (MLR Model), $Q^2 = 0.55$(SVR model)	47
F_{up}				
878	Mordred, PaDEL, MOE	RF, SVM, XGB, GBM	Best model: RF ($R^2_{\text{test}} = 0.81$, $\text{RMSE}_{\text{test}} = 0.29$; $Q^2 = 0.84$)	24
2738	Mordred, PaDEL, ADMET	RF, SVM, PLS, k-NN, ANN, AdaBoost		40
5471	PubChem fingerprint	PyCaret, TPOT, AutoKeras, Auto-sklearn	Best model: PyCaret (RMSE = 8.44)	41
220	176 molecular descriptors	GA, MLR, SR	59% drugs with twofold error, $Q^2 = 0.532$	42
$t_{½}$				
1253	Mordred, PaDEL, MOE	RF, SVM, XGB, GBM	Best model: RF ($R^2_{\text{test}} = 0.83$, $\text{RMSE}_{\text{test}} = 0.15$, $Q^2 = 0.73$)	24
1105	Codessa	GBM	$R^2_{\text{test}} = 0.82$, $\text{RMSE}_{\text{test}} = 0.55$	50
CL				
1268	Mordred, PaDEL, MOE	RF, SVM, XGB, GBM	Best model: RF ($R^2_{\text{test}} = 0.87$, $\text{RMSE}_{\text{test}} = 0.10$, $Q^2 = 0.76$)	24
89	ALOGP, Dragon	ANN	$R^2_{\text{test}} = 0.64$, $\text{RMSE}_{\text{test}} = 0.54$	46
525	E State	MLR	$R^2_{\text{test}} = 0.70$	47

[a] Search was carried out from January 2010 to February 2021.

[b] Abbreviations: ANN, artificial neural network; GA, genetic algorithm; k-NN, k-nearest neighbors; PLS, partial least squares; RF, random forest; SR, stepwise regression; SVM, support vector machine; XGB, XGBoost.

Please cite this article in press as: Danishuddin et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.013

cycles and fused aromatic rings, and presence of Cl and F atoms had positive effects, whereas polarity and the presence of strong electrophiles had a negative effect. Wang *et al.* used various ML methods to develop regression models for four human PK parameters, including VDss.[24] Out of four methods used, support vector machines (SVMs) showed the best performance with R^2_{test} of 0.870 and $RMSE_{test}$ of 0.208. LogP and LogS were among the top-ten descriptors that were highly correlated with VDss. Similar observations were also observed in early studies evaluating the effect of various chemical descriptors in predicting VDss with either *in silico* or experimental approaches.[25]

Although the models reported in these studies are applicable to marketed drugs, the extent to which they are applicable to other types of molecule is unclear. Recently, Lombardo *et al.*[26] reported regression models for predicting VDss using Random Forest (RF) and six molecular descriptors, including logD 7.4, molecular weight (MW), Max basic pKa, Min acidic pKa, and aromatic atoms. The model included 1090 compounds in the training set and 213 compounds in the test set. Prediction accuracies of 59% and 78% within twofold and threefold errors, respectively, was observed.

Predicting VDss accurately remains to be fully solved.[27] PBPK modeling, also known as a mechanism-based approach, provides an opportunity to understand the underlying distribution processes and is superior to empirical methods for predicting PK with high levels of accuracy.[28] In addition, VDss can also be extrapolated from tissue-to-plasma partition coefficients (Kp) from preclinical species. These approaches are resource intensive and depend on the synthesis of compounds, which further hinder the ability to predict human VDss for early drug development. Considerable efforts have been made to develop *in silico*-based models for predicting tissue partitioning based on physicochemical properties, such as pKa and logP, plasma protein binding, and the blood-to-plasma partition ratio (BPR).[29] Poulin and Theil proposed a mechanistic-based Kp prediction method. The use of such tissue composition-based equations to predict VDss has the advantage of providing a model with a clear interpretation of where the drug is distributed.[30,31] Freitas *et al.* addressed the problem of estimating the VDss of drugs using decision tree-based regression.[32] Regression models were built using either molecular descriptors only or both the molecular descriptors and tissue:plasma partition coefficients (Kt:p). Models were developed in two phases: in Phase I, the authors extracted experimentally derived Kt:p values for 110 compounds from the available literature and built the models. These models were then used in the second phase to estimate Kt:p values for 604 compounds; these predicted values were then combined with molecular descriptors to estimate VDss. Models built with only molecular descriptors had a mean fold error of 2.33, whereas models using predicted Kt:p values in addition to molecular descriptors had a mean fold error of 2.29. The authors stated that predicted Kt:p values could be beneficial for accurate prediction of VDss using decision trees if prior feature selection is applied. Recently, Murad *et al.* developed *in silico* models for the prediction of VDss directly from larger data of structurally diverse clinical compounds ($N = 956$).[33] They also investigated the utility of adipocyte and myocyte partitioning in predicting the VDss. The predicted models outperformed other computational methods with R^2 of 0.5. Measured fraction of unbound protein (F_{up}) and BPR parameters improved the performance of VDss predictions with R^2 of 0.6. Various combinations of both adipocyte and myocyte partitions with different strategies did not lead to any significant improvement in VDss predictions. However, these were found to have slight improvements in predicting low VDss values.

Overall, these studies present an interestingly approach to the prediction of VDss. These models could be used if implemented in the form of free available web applications, which would be useful to nonspecialist researchers. Although mechanistic rationalization of the modeling approach was provided in both studies, a disadvantage is that such approaches require computational and human metabolism expertise. Models based on traditional ML algorithms also rely on feature engineering, which is as time-consuming and challenging, and can result in poor model performance. Nevertheless, deep learning (DL) can automatically extract the crucial features or molecular descriptors from raw data without the need for feature engineering. Ouyang *et al.* used an integrated transfer learning and multitask learning approach to develop QSAR models to predict VDss.[34] The developed model was found to be more generalized and have better accuracies compared with conventional approaches.

Fraction of unbound protein

F_{up} is a significant parameter in PK studies of a drug. It reflects the concentration of drug capable of interacting with pharmacological targets and able to diffuse between plasma and tissues.[35] The value of F_{up} affects the volume of distribution (VD) and the total clearance (CL_{tot}) of a drug.[36] VDss is directly proportional to F_{up} and inversely proportional to the fraction of unbound in tissues (F_{ut}).[37] Any sudden increase in free concentration of a drug could cause toxicity.[38] F_{up} can also be helpful to estimate the effective concentration of drugs that can precipitate metabolism (or transporter)-related drug–drug interactions. The value range also has an important role because small differences in protein binding can have large effects on F_{up} and, thus the drug efficacy can also change dramatically.[39] Therefore, it is necessary to have an optimal value of F_{up} during drug development. Thus, unbound drug concentration is significant factor in establishing safety margins and efficacious doses for humans.

Predictive models for F_{up}

Watanabe *et al.* developed F_{up} prediction models using several ML techniques with PaDEL, MODRED, and other descriptors, and focused specifically on the low value ranges of F_{up}.[40] Prediction models were trained on a large data set of 2738 experimental values. The training and test sets comprised 2192 and 546 compounds, respectively. The classification model showed a high true positive rate of 0.826 for the low fraction unbound class in the test set. Lipophilicity and conjugated double bond-related features were the key components for predicting F_{up} values. The strongly biased distribution of F_{up} was mitigated by a logarithmic transformation in the regression model, leading to improved accuracy at lower values. These models showed better performance than those of previously published methods, including some commercial software. Wang and coworkers used

Please cite this article in press as: Danishuddin et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.013

878 drugs and 121 descriptors in their research. Among the four methods used for model construction, RF performed better with R^2_{test} of 0.818 and $RMSE_{test}$ of 0.291.[24] Recently, Mulpuru and Mishra developed F_{up} prediction models using a chemical PubChem fingerprint and four different AutoML frameworks (Auto-sklearn, AutoKeras, PyCaret, and TPOT).[41] These models were trained on data sets of 5471 experimental values from the ChEMBL database. PyCaret was found to be best framework, with $R^2 = 0.85$ and RMSE = 8.44. Despite the strength of these models, one of their weaknesses is the use of a limited descriptor set. Zhivkova reported regression models using a data set of 220 basic drugs for the prediction the F_{up} values.[42] The models were developed by applying genetic algorithms, stepwise regression, and multiple linear regression. These models predicted ~ 59% of the drugs from an external validation set within the twofold error of the experimental values with squared correlation coefficient of prediction of 0.532, and mean absolute error (MAE) of 0.17. The author defined the criteria and an empirical rule for distinguishing between low, high, and very high plasma protein binders. Based on this criteria, lipophilicity (logP > 3), presence of aromatic nonsubstituted C-atoms and molar volume (>300 cm^3) increased F_{up} (plasma protein binding percentage; PPB), whereas the presence of quaternary C-atom reduced the binding. Despite the well-studied relationship between the drug features and F_{up} values, the limitations of this study were the size of the drug molecules and descriptors, which might be not sufficient for drawing any firm conclusions.

Del Amo *et al.* used linear and nonlinear methods to predict the VDss and F_{up}.[43] They used both measurements in parallel for model building. Partial least square (PLS) and recursive partition methods were used for model development, along with Volsurf+, MOE, and ACDlab descriptors. Models built with Volsurf + descriptors performed better in predicting F_{up}, with Q^2_{test} of 0.54, whereas models of VDss performed with Q^2 of 0.70. Nonlinear classification models were able to identify compounds with high or low VDss with sensitivities of 0.81 and 0.71, respectively for the test set, whereas the classification accuracy of F_{up} was less accurate. Lipophilicity and solubility descriptors were found to significantly impact both VDss and F_{up}. The parallel modeling approach developed in this study is interesting, because VDss and F_{up} are expected to be affected by similar physicochemical properties; such models can be helpful to establish the mechanism for accurately predicting PK values.

Clearance

The concept of clearance (CL) was introduced during the 1970s, and it has remained a powerful tool for explaining the PK and predicting changes in the blood concentration time course of drugs.[44] Drug CL has a significant role in determining dosing regimens. Drugs can be cleared through renal, hepatic, and biliary CL. The total systemic CL of a drug is the sum of all CLs by various organs, and can be defined as the volume of blood from which all drug is removed per minute (ml/min).[45].

Predictive models for CL

Several QSAR models are available for CL prediction that could provide insights into the mechanism of CL. Paulo *et al.* used artificial neural networks for modeling human hepatocyte intrinsic CL.[46] The model was trained with 71 drugs, whereas the performance of model was set with 18 drugs. The resulting performance of the models was 0.953 and 0.804 for the train and test sets, respectively. Gombar and Hall used support vector regression (SVR) and multiple linear regression (MLR) to develop and evaluate QSAR models for CL.[47] These models were developed from 525 compounds and 77 fingerprints. For SVR models, the best average Q^2 of 0.32 was achieved from 50 cross-validation models, each created from randomly selected training (60%) and test (40%) compounds. For the MLR model, the training set of 525 compounds led to a unique set of 653 two-atom fragments, the E-state values of which were used as structure descriptors. A preliminary model was developed with 358 descriptors. After refinement, an 89-descriptor model from 510 compounds was built, with R^2 of 0.70. However, these models focused on a small group of drugs, which limits their applicability. Wang *et al.* developed various models using 1268 drugs.[24] These models were built using four different statistical modeling methods [SVM, FR, gradient-boosting machine (GBM) and XGBoost] and 134 variables from Mordred, PaDEL, and MOE. Polar surface area (PSA), water solubility, and hydrogen bonds of the molecules were among the top-ten descriptors influencing CL prediction. Among the four methods, RF showed the better performance, with R^2_{test} of 0.875 and $RMSE_{test}$ of 0.103. These models are generally associated with the applicability domain, which is defined based on inclusion or exclusion rules regarding the compound class in the training set. Moreover, these models are transparent and have the advantage of being based on molecular descriptors computed with freely available software.

Half life

The half-life ($t_{1/2}$) of a drug is defined as the amount of time required for the drug concentration measured in plasma to be reduced to exactly half of its starting concentration.[48] While its value can be extracted from PK concentration versus time graphs, it is closely related to CL and VDss. Elimination $t_{1/2}$ is increased by an increase in volume of distribution or a decrease in CL, and vice versa.[49] Understanding the concept of $t_{1/2}$ is useful for determining excretion rates as well as steady-state concentrations for any specific drug. The clinical significance of $t_{1/2}$ tends to arise in situations involving drug toxicity, which results from dosing more or less frequently than the $t_{1/2}$ of the drug. Such a clinical scenario illustrates why medical professionals rely on $t_{1/2}$ of drug so often in practice, and why it is still an emphasis in medical education.

Predictive models for half-life

Lu *et al.* used several ML approaches to build high-quality models for predicting $t_{1/2}$.[50] A data set of 1105 chemicals was used for model building. Among the seven models, GBM showed the best performance with $R^2 = 0.820$ and RMSE = 0.555 for the test set. LogP, logD, and logS are usually identified as important parameters of for modeling $t_{1/2}$. LogP and logD correlated positively with $R^2 = 0.666$ and 0.687, respectively and logS correlated negatively with $R^2 = -0.632$, indicating that highly lipophilic chemicals were more liable to show long $t_{1/2}$. Wang and coworkers proposed various models using Mordred, PaDEL, and MOE descrip-

Please cite this article in press as: Danishuddin et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.013

tors and four different ML methods (SVM, RF, GBM, and XGBoost) to predict the $t_{1/2}$ of 1253 drugs.[24] Compared with statistical results, RF performed better with R^2_{test} value of 0.832. This study highlighted the importance of descriptors, such as autocorrelations, physical properties, or lipophilicity. Partial charge descriptors also exhibited high correlation with the log$t_{1/2}$ values. Autocorrelation descriptors were found to have a positive correlation with log $t_{1/2}$. Although these are simple modeling studies, as a result of using reasonable and different sets of descriptors, the multidimensional space defined by each model provides an opportunity to discuss and interpret the influence of descriptors in predicting these values more broadly.

Databases and software for PK

The performance of any QSAR or predictive model depends on the quality of data and modeling methodology. The collection of large amounts of structurally derived molecules and associated experimental values for modeling is facilitated by an increasing number of public databases for human PK. Such curated databases provide a link between indexed 2D chemical structures and biological targets, from single proteins, to protein complexes, subcellular components, cell lines, to tissue and whole-organism *in vivo* data. Although the number of entries in these databases is important, there is an increased awareness concerning the quality of data rather the number of entries. There are numerous well-defined human PK data available, making ML-based prediction models easier to design (Table 2). Here, we investigate the resources available for human PK parameters and briefly describe recent advances.

Obach *et al*. developed a database that contains human PK information for 1352 drugs.[51] These data were obtained from the literature and regulatory agencies. Although the database is not available online, the authors provided a link to an Excel spreadsheet that can be downloaded. This spreadsheet contains information, such as the name of the drug, VDss, CL, MRT, F_{up}, $t_{1/2}$, physiochemical characteristics, and SMILE of each drug. PK/DB is a web-based and easy-to-access database containing PK information for structurally diverse drug-like and lead-like molecules extracted from the literature.[52] The online database also allows researchers to predict the ADMET properties of a molecule. PKKB provides PK information for 1685 drug and drug-like molecules and offers structure property data sets including: Caco-2, LogBB, P-gp inhibition, human intestinal absorption, and oral bioavailability, which might be complementary to the PK data set.[53] PK-DB, provides high-quality PK data enriched from experimental and clinical studies, important for computational modeling.[54] ChEMBL (www.ebi.ac.uk/chembl/) is an open-access database containing functional and ADMET-related information for numerous drug-like molecules. DrugBank contains PK information for US Food and Drug Administration (FDA)-approved drugs.[55] E-Drug3D is a manually curated database of PK and PD properties for 1852 FDA-approved drugs.[56] This database holds the manually curated experimental PK properties from drug labels, including the volume of distribution (VD), CL, PPB, terminal $t_{1/2}$, and bioavailability (F).

In addition to these publicly available databases, some pharmaceutical companies maintain their own PK data originating from in-house drug discovery projects. There are several other

TABLE 2

Examples of web resources for PK data or their prediction.

Name	Description	Website
PK/DB	Robust databases for PK studies and *in silico* ADME prediction; contains 1203 compounds representing 2973 PK measurements, including five models for *in silico* ADME prediction	www.pkdb.ifsc.usp.br
PK-DB	Open-access database of PK data from experimental and clinical studies. Information of high-quality PK data significant for computational modeling and data integration	https://pk-db.com/
Obach *et al.'s* data set	Human intravenous PK data in a data set of 1352 drugs. *In vivo* data obtained or derived from original references, either through literature or regulatory agency reports, exclusively from studies suing intravenous administration	[49]
Pharmacokinetic Knowledge Base (PKKB)	Provides PK information for drugs and drug-like molecules and offers structure–property data sets containing Caco-2, LogBB, P-gp inhibitory, human intestinal absorption, and oral bioavailability data	http://cadd.suda.edu.cn/admet
e-Drug3D	Database of 1852 FDA-approved drugs with information on eight additional pharmacokinetic parameters (solubility, VD, CL, $t_{½}$, PPB, F, C_{max}, and T_{max})	http://chemoinfo.ipmc.cnrs.fr/edrug3d
ChEMBL	Open-access database containing ADMET information for numerous drug-like compounds	www.ebi.ac.uk/chembl/
ADMETlab	Freely available and user-friendly web interface for systematic ADMET evaluation of compounds; comprehensive database with 288 967 entries	http://admet.scbdd.com/
pkCSM	Uses graph-based signatures encoding distance patterns between atoms. Predictive regression and classification models have been developed for five different PK properties	http://biosig.unimelb.edu.au/pkcsm/
ChemPK	Commercially available software that uses chemical structures to predict human PK values	www.cyprotex.com/insilico/physiological_modelling/chempk
ADMET Predictor	Commercially available software that provides strategic pharmacology and pharmacometric consulting services, PK/PD model development with NONMEM, etc.	www.simulations-plus.com
BIOVIA Discovery Studio	Commercially available software use for predicting ADMET properties of drug	www.discngine.com/discovery-studio

Please cite this article in press as: Danishuddin et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.013

available web resources that can predict PK values for any drug molecule. Dong *et al*. developed a webserver, called ADMETlab, for systematic ADMET evaluation of molecules.[57] Predictions are based on comprehensively collected information for 288 967 compounds. Another webserver, pkCSM, was developed using graph theory for predicting PK properties. The predictions are based on distance patterns between atoms and are used to represent the small molecule and to train predictive models.[58] It includes 14 quantitative regression and 16 classification models for predicting an array of ADMET properties. Commercially available software packages such as ChemPK (www.cyprotex.com/insilico), ADMET Predictor (www.simulations-plus.com/) and BIOVIA Discovery Studio (https://discover.3ds.com/discovery-studio-visualizer-download) are also used for PK predictions. VLS3D.COM is a curated directory of tools and databases collected over the past two decades and available from www.vls3d.com/index.php.

Challenges and future directions

ML-based QSAR models for predicting drug PK have undergone significant advances.[59] However, most of the models lack sufficient interpretability and offer poor prediction for novel drugs. Therefore, it is necessary to identify and address these in the development of QSAR. Dearden *et al*. discussed 21 types of error that are common in QSAR-based approaches and each is allocated to the relevant OECD principle.[60] Avoiding such errors could improve and enhance the model predictability. Another challenge posed for QSAR approaches in medicine is the lack of quality experiment data. A fundamental assumption of any QSAR modeling study is the use of correct experimental data. Such data sets have more of an influence on the prediction performance of QSAR models compared with model optimization techniques. Thus, any erroneous structures represented by erroneous descriptors can directly affect the model performance. Small changes in chemical structure can lead to major changes in the target activity and can be one of the reasons for poor predictions. Fourches *et al*. discussed the cleaning or optimizing of chemical records in a database that cannot be appropriately handled by conventional cheminformatics approaches.[61] Addressing and identifying such 'activity cliffs' is statistically important to avoid factors limiting predictions.[62] The collection of a reasonably diverse data set in terms of structures and a range of data is not a trivial task, especially when aimed at robust model development. In terms of PK data, it is important to access consistent clinical data for a range of structures along with highly correlated properties of interest. Given that experiments in this area are expensive and time consuming, most research focuses only on limited available data sets. PK parameters do relate to each other and are assumed to be affected by similar physiochemical and other drug properties. Parallel modeling mechanism can be used to exploit this effect to increase the accuracy of the model (Fig. 2). Therefore, further research and development in this area is encouraged. The transferability and complexity of the models are additional issues that cannot be avoided. Transparent and less complex models can be helpful for identifying the data pattern that can relate to experiments results. Addressing these problems is daunting, and it might not be possible to address some of them in any substantive way. Thus, any QSAR model developed might be erroneous to some degree because of limitations derived from one or more of the problems described herein. Nevertheless, advances in AI, such as DL approaches, could mitigate some of these difficulties, thus helping to develop better predictive models. Despite their limitations, these models might be useful when combined with expert systems.

Concluding remarks

Given that QSAR-based predictive models of human PK are associated with different methods and descriptor sets for prediction, it is unlikely that any single model will be capable of making reliable predictions for all drugs of interest. Thus, the predictive abilities of these models should be compared and refinement of literature models explored. The public availability of chemical Big Data and advancement in artificial intelligence (AI)-based techniques boosted targeted drug discovery at an unprecedented speed. With AI assistance, the ADMET profile of any drug candidate can be predicted at low cost and in a short period of time. The evaluation of human PK prediction is significant for improvements in any prediction method. Here, we also discussed the challenges and advances in the assessment or reliable prediction of human PK. We believe that ongoing and emerging developments in these approaches and chemical Big Data for human PK will enable to the development of more effective and robust prediction models.

Acknowledgments

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) & funded by Korean government (MSIT) (No. NRF-2018M3A9A70-57263).

References

1 P.M. Glassman, V.R. Muzykantov, Pharmacokinetic and pharmacodynamic properties of drug delivery systems, J. Pharmacol Exp Ther 370 (2019) 570–580.
2 J.M. Gallo, Pharmacokinetic/pharmacodynamic-driven drug development, Mt Sinai J Med 77 (4) (2010) 381–388.
3 A. Jorda, M. Zeitlinger, Preclinical pharmacokinetic/pharmacodynamic studies and clinical trials in the drug development process of EMA-approved antibacterial agents: a review, Clin Pharmacokinet 59 (9) (2020) 1071–1084.
4 P.JH. Webborn, The role of pharmacokinetic studies in drug discovery: where are we now, how did we get here and where are we going?, Future Med Chem 6 (11) (2014) 1233–1235
5 L.L.G. Ferreira, A.D. Andricopulo, ADMET modeling approaches in drug discovery, Drug Discov. Today. 24 (2019) 1157–1165.
6 E.A.G. Blomme, Y. Will, Toxicology strategies for drug discovery: present and future, Chem. Res Toxicol 29 (4) (2016) 473–504.
7 P.A. Walker, S. Ryder, A. Lavado, C. Dilworth, R.J. Riley, The evolution of strategies to minimise the risk of human drug-induced liver injury (DILI) in drug discovery and development, Arch Toxicol 94 (8) (2020) 2559–2585.
8 J. Houston, A. Galetin, Methods for predicting *in vivo* pharmacokinetics using data from *in vitro* assays, Curr. Drug Metab 9 (2008) 940–951.
9 M. Davies, R.D.O. Jones, K. Grime, R. Jansson-Löfmark, A.J. Fretland, S. Winiwarter, P. Morgan, D.F. McGinnity, Improving the accuracy of predicted human pharmacokinetics: lessons learned from the AstraZeneca drug pipeline over two decades, Trends Pharmacol Sci 41 (6) (2020) 390–408.

Please cite this article in press as: Danishuddin et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.013

10 F. Wu, Y. Zhou, L. Li, X. Shen, G. Chen, X. Wang, et al., Computational approaches in preclinical studies on drug discovery and development, Front. Chem 8 (2020) 726.
11 E.N. Muratov, J. Bajorath, R.P. Sheridan, I.V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I.I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D.A. Winkler, D. Agrafiotis, A. Cherkasov, A. Tropsha, QSAR without borders, Chem. Soc Rev 49 (11) (2020) 3525–3564.
12 D.S. Wishart, Improving early drug discovery through ADME modelling, Drugs R D 8 (6) (2007) 349–362.
13 X. Wang, M. Liu, L. Zhang, Y. Wang, Y.u. Li, T. Lu, Optimizing pharmacokinetic property prediction based on integrated datasets and a deep learning approach, J. Chem Inf Model 60 (10) (2020) 4603–4613.
14 C. Csajka, D. Verotta, Pharmacokinetic-pharmacodynamic modelling: history and perspectives, J. Pharmacokinet Pharmacodyn 33 (3) (2006) 227–279.
15 A. Lavecchia, Machine-learning approaches in drug discovery: methods and applications, Drug Discov Today 20 (3) (2015) 318–331.
16 H. Zhu, Big Data and artificial intelligence modeling for drug discovery, Annu. Rev Pharmacol Toxicol 60 (2020) 573–589.
17 G. Berellini, F. Lombardo, An accurate *in vitro* prediction of human VD ss based on the Äie-Tozer equation and primary physicochemical descriptors. 3. Analysis and assessment of predictivity on a large dataset, Drug Metab Dispos 47 (12) (2019) 1380–1387.
18 T. Rodgers, M. Rowland, Mechanistic approaches to volume of distribution predictions: understanding the processes, Pharm. Res 24 (2007) 918–933.
19 D.F. McGinnity, J. Collington, R.P. Austin, R.J. Riley, Evaluation of human pharmacokinetics, therapeutic dose and exposure predictions using marketed oral drugs, Curr. Drug Metab 8 (2007) 463–479.
20 Mahmood I. Interspecies scaling: predicting volumes, mean residence time and elimination half-life. some suggestions. J. Pharm Pharmacol, 2011; 50: 493–499.
21 L.M. Berezhkovskiy, On the accuracy of calculation of the mean residence time of drug in the body and its volumes of distribution based on the assumption of central elimination, Xenobiotica 46 (6) (2016) 477–482.
22 Z. Zhivkova, I. Doytchinova, Prediction of steady-state volume of distribution of acidic drugs by quantitative structure–pharmacokinetics relationships, J. Pharm Sci 101 (3) (2012) 1253–1266.
23 Z.D. Zhivkova, T. Mandova, I. Doytchinova, Quantitative structure-pharmacokinetics relationships analysis of basic drug volume of distribution, J. Pharm. Pharm Sci 18 (2015) 515–527.
24 Y. Wang, H. Liu, Y. Fan, X. Chen, Y. Yang, L.u. Zhu, J. Zhao, Y. Chen, Y. Zhang, In silico prediction of human intravenous pharmacokinetic parameters with improved accuracy, J. Chem Inf Model 59 (9) (2019) 3968–3980.
25 D.A. Smith, K. Beaumont, T.S. Maurer, L.i. Di, Volume of distribution in drug design, J. Med Chem 58 (15) (2015) 5691–5698.
26 F. Lombardo, J. Bentzien, G. Berellini, I. Muegge, In silico models of human PK parameters. prediction of volume of distribution using an extensive data set and a reduced number of parameters, J. Pharm Sci 110 (1) (2021) 500–509.
27 N.A. Miller, M.B. Reddy, A.T. Heikkinen, V. Lukacova, N. Parrott, Physiologically based pharmacokinetic modelling for first-in-human predictions: an updated model building strategy illustrated with challenging industry case studies, Clin. Pharmacokinet 58 (6) (2019) 727–746.
28 Graham H, Walker M, Jones O, Yates J, Galetin A, Aarons L. Comparison of in-vivo and in-silico methods used for prediction of tissue: plasma partition coefficients in rat. J. Pharm Pharmacol, 2012; 64: 383–396.
29 P. Poulin, F.-P. Theil, Prediction of pharmacokinetics prior to *in vivo* studies. 1. Mechanism-based prediction of volume of distribution, J. Pharm Sci 91 (1) (2002) 129–156.
30 J.H. Lin, Y. Sugiyama, S. Awazu, M. Hanano, *In vitro* and *in vivo* evaluation of the tissue-to-blood partition coefficient for physiological pharmacokinetic models, J. Pharmacokinet Biopharm 10 (6) (1982) 637–647.
31 K. Holt, M. Ye, S. Nagar, K. Korzekwa, Prediction of tissue-plasma partition coefficients using microsomal partitioning: incorporation into physiologically based pharmacokinetic models and steady-state volume of distribution predictions, Drug. Metab Dispos 47 (10) (2019) 1050–1060.
32 A.A. Freitas, K. Limbu, T. Ghafourian, Predicting volume of distribution with decision tree-based regression methods using predicted tissue: plasma partition coefficients, J. Cheminform 7 (2015) 6.
33 N. Murad, K.K. Pasikanti, B.D. Madej, A. Minnich, J.M. McComas, S. Crouch, J.W. Polli, A.D. Weber, Predicting volume of distribution in humans: performance of in silico methods for a large set of structurally diverse clinical compounds, Drug Metab Dispos 49 (2) (2021) 169–178.
34 Z. Ye, Y. Yang, X. Li, D. Cao, D. Ouyang, An integrated transfer learning and multitask learning approach for pharmacokinetic parameter prediction, Mol Pharm 16 (2) (2019) 533–541.
35 T. Bohnert, L.-S. Gan, Plasma protein binding: from discovery to development, J. Pharm Sci 102 (9) (2013) 2953–2994.
36 Z.Q. Yuan, C. Qiao, Z.C. Yang, L. Yu, L.N. Sun, Y. Qian, et al., The impact of plasma protein binding characteristics and unbound concentration of voriconazole on its adverse drug reactions, Front. Pharmacol 11 (2020) 505.
37 L.J. Francis, J.B. Houston, D. Hallifax, Impact of plasma protein binding in drug clearance prediction: a data base analysis of published studies and implications for *in vitro-in vivo* extrapolation, Drug Metab Dispos 49 (3) (2021) 188–201.
38 P. Poulin, Drug distribution to human tissues: prediction and examination of the basic assumption in *in vivo* pharmacokinetics-pharmacodynamics (PK/PD) research, J Pharm Sci 104 (6) (2015) 2110–2118.
39 J.A. Roberts, F. Pea, J. Lipman, The clinical relevance of plasma protein binding changes, Clin Pharmacokinet 52 (1) (2013) 1–8.
40 R. Watanabe, T. Esaki, H. Kawashima, Y. Natsume-Kitatani, C. Nagao, R. Ohashi, et al., Predicting fraction unbound in human plasma from chemical structure: improved accuracy in the low value ranges, Mol. Pharm 15 (2018) 5302–5311.
41 V. Mulpuru, N. Mishra, In silico prediction of fraction unbound in human plasma from chemical fingerprint using automated machine learning, ACS Omega 6 (10) (2021) 6791–6797.
42 Z.D. Zhivkova, Quantitative structure–pharmacokinetics relationships for plasma protein binding of basic drugs, J. Pharm Pharm Sci 20 (1) (2017) 349, https://doi.org/10.18433/J33633.
43 E.M. del Amo, L. Ghemtio, H. Xhaard, M. Yliperttula, A. Urtti, H. Kidron, P. Taylor, Applying linear and non-linear methods for parallel prediction of volume of distribution and fraction of unbound drug, PLoS ONE. 8 (10) (2013) e74758.
44 R. Mehvar, Clearance concepts: fundamentals and application to pharmacokinetic behavior of drugs, J. *Pharm Pharm Sci* 21 (2018) 88–102.
45 D.-S. Yim, Tutorial for beginners: the concept of clearance explained using the example of a vacuum cleaner, Transl. Clin Pharmacol 25 (1) (2017) 1, https://doi.org/10.12793/tcp.2017.25.1.1.
46 P. Paixão, L.F. Gouveia, J.A. Morais, Prediction of the *in vitro* intrinsic clearance determined in suspensions of human hepatocytes by using artificial neural networks, *Eur. J. Pharm. Sci.* 39 (2010) 310–321.
47 V.K. Gombar, S.D. Hall, Quantitative structure–activity relationship models of clinical pharmacokinetics: clearance and volume of distribution, J. Chem Inf Model 53 (4) (2013) 948–957.
48 A. Grover, L.Z. Benet, Intermittent drug dosing intervals guided by the operational multiple dosing half-lives for predictable plasma accumulation and fluctuation, J. Pharmacokinet Pharmacodyn 38 (3) (2011) 369–383.
49 P. Vermeir, D. Vandijck, S. Degroote, R. Peleman, R. Verhaeghe, E. Mortier, G. Hallaert, S. Van Daele, W. Buylaert, D. Vogelaers, Communication in healthcare: a narrative review of the literature and practical recommendations, Int. J Clin Pract. 69 (11) (2015) 1257–1267.
50 J. Lu, D. Lu, X. Zhang, Y. Bi, K. Cheng, M. Zheng, et al., Estimation of elimination half-lives of organic chemicals in humans using gradient boosting machine, Biochim. Biophys. Acta Gen. 1860 (2016) 2664–2671.
51 F. Lombardo, G. Berellini, R.S. Obach, Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 1352 drug compounds, Drug. Metab Dispos. 46 (11) (2018) 1466–1477.
52 T.L. Moda, L.G. Torres, A.E. Carrara, A.D. Andricopulo, PK/DB: database for pharmacokinetic properties and predictive in silico ADME models, Bioinformatics. 24 (19) (2008) 2270–2271.
53 D. Cao, J. Wang, R. Zhou, Y. Li, H. Yu, T. Hou, ADMET evaluation in drug discovery. 11. PharmacoKinetics Knowledge Base (PKKB): a comprehensive database of pharmacokinetic and toxic properties for drugs, J. Chem Inf Model. 52 (5) (2012) 1132–1137.
54 Grzegorzewski J, Brandhorst J, Green K, Eleftheriadou D, Duport Y, Barthorscht F, et al. PK-DB: pharmacokinetics database for individualized and stratified computational modeling. Nucleic Acids Res. 2021;49: D1358–D1364.
55 Wishart DS. et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008; 36: D901–D906.
56 D. Douguet, Data sets representative of the structures and experimental properties of fda-approved drugs, ACS Med Chem Lett 9 (3) (2018) 204–209.

Please cite this article in press as: Danishuddin et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.013

57 J. Dong, N.-N. Wang, Z.-J. Yao, L. Zhang, Y. Cheng, D. Ouyang, A.-P. Lu, D.-S. Cao, ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database, J. Cheminform 10 (1) (2018), https://doi.org/10.1186/s13321-018-0283-x.

58 D.E. Pires, T.L. Blundell, D.B. Ascher, pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures, J. Med Chem 58 (2015) 4066–4072.

59 A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Mol Inform. 29 (6-7) (2010) 476–488.

60 J.C. Dearden, M.T.D. Cronin, K.L.E. Kaiser, How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR), SAR QSAR Environ Res. 20 (3-4) (2009) 241–266.

61 D. Fourches, E. Muratov, A. Tropsha, Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research, J Chem Inf Model 50 (7) (2010) 1189–1204.

62 D. Young, T. Martin, R. Venkatapathy, P. Harten, Are the chemical structures in your QSAR correct?, QSAR Comb Sci 27 (11-12) (2008) 1337–1345

Please cite this article in press as: Danishuddin et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.09.013