 doi:10.20944/preprints201807.0092.v1

Theoretical Prediction of the Complex P-Glycoprotein Substrate Efflux Based on the Novel Hierarchical Support Vector Regression Scheme

Chun Chen[1], Ming-Han Lee[1], Ching-Feng Weng[2], and Max K. Leong[1,2,*]

[1]Department of Chemistry, National Dong Hwa University, Shoufeng, Hualien 97401, Taiwan

[2]Department of Life Science and Institute of Biotechnology, National Dong Hwa University, Shoufeng, Hualien 97401, Taiwan

[*]Correspondence: leong@mail.ndhu.edu.tw

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

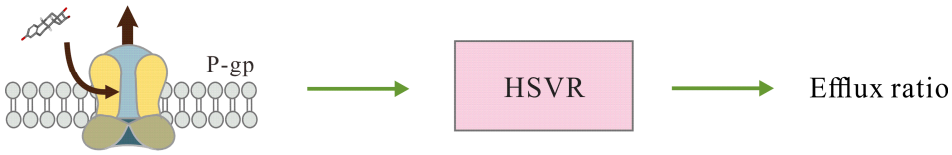
Abstract

P-glycoprotein (P-gp), a membrane-bound transporter, can eliminate xenobiotics by transporting them out of the cells or blood-brain barrier (BBB) at the expense of ATP hydrolysis. Thus, P-gp mediated efflux plays a pivotal role in altering the absorption and disposition of a wide range of substrates. Nevertheless, the mechanism of P-gp substrate efflux is rather complex since it can take place through active transport and passive permeability in addition to multiple P-gp substrate binding sites. A nonlinear quantitative structure-activity relationship (QSAR) model was developed in this study using the novel machine learning-based hierarchical support vector regression (HSVR) scheme to explore the perplexing relationships between descriptors and efflux ratio. The predictions by HSVR were found to be in good agreement with the observed values for the molecules in the training set (n = 50, r^2 = 0.96, q^2_{CV} = 0.94, RMSE = 0.10, s = 0.10) and test set (n = 13, q^2 = 0.80–0.87, RMSE = 0.21, s = 0.22). When subjected to a variety of statistical validations, the developed HSVR model consistently met the most stringent criteria. A mock test also asserted the predictivity of HSVR. Consequently, this HSVR model can be adopted to facilitate drug discovery and development.

Keywords: P-glycoprotein; efflux ratio; *in silico*; machine learning; hierarchical support vector regression; absorption, distribution, metabolism, excretion, and toxicity

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

Graphical abstract



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

1. Introduction

Permeability glycoprotein also known as P-glycoprotein (P-gp), which belongs to the ATP-binding cassette (ABC) superfamily of transporters, can actively transport a wide range of structurally and mechanistically diverse endogenous and xenobiotic chemical agents across the cell membrane at the energy expense of ATP hydrolysis [1]. P-gp, a 170-kDa plasma membrane protein encoded by the multidrug resistance gene (*MDR1*/*ABCB1*), is expressed at high levels in various tissues such as blood-brain-barriers (BBB), gastrointestinal tract (GIT), liver, kidney, and placenta [2-6]. In addition, P-gp plays significant roles in cell and tissue detoxification and elimination of harmful substances *per se* [1]. For example, the accumulation of neurotoxic amyloid-β (Aβ) peptides in the brain represents a pathogenic hallmark of Alzheimer's disease (AD), which is the most common form of dementia in aging populations [7]. It has been found that the decreased clearance rather than production of Aβ is the primary formation of the deleterious Aβ plaques in the brain [8]. The decreased elimination of Aβ from the brain into the blood can be partially attributed to the dysfunction of P-gp function, leading to the progression of AD [9-11]. Furthermore, it has been shown that Aβ can downregulate the P-gp expression along with other transporters and consequently lead to further accelerated neurodegeneration [12]. Hence, it has been suggested to increase Aβ clearance from the brain by restoring P-gp function of BBB to reduce Aβ brain accumulation as a new strategy in the medical treatment of the early stages of AD [13,14].

Additionally, P-gp efflux can profoundly implicate the role of drug absorption, distribution, metabolism, excretion, and toxicity (ADME/Tox) [15] that can clinically alter the administrated drug efficacy or even lead to various adverse side-effects due to drug-drug interaction (DDI) in case of polypharmacy [16]. For instance, rifampin can interact with the P-gp substrate digoxin,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

leading to a lower accumulation of digoxin as demonstrated by a clinical study [17]. Moreover, it is of particular interest to observe the subtle role played by P-gp in the central nervous system (CNS) since P-gp can affect the BBB penetration and pharmacological activities of administrated drugs [18]. The CNS-related side-effects of non-CNS drugs can be eliminated by P-gp because of their limited BBB penetration [19,20]. For instance, the P-gp substrate loperamide, which is a long-acting anti-diarrheal agent by agonizing the μ-opioid receptor, does not cause any CNS side-effects when administrated alone due to the blockage of the BBB penetration by P-gp [21]. When co-administrated with the P-gp inhibitor quinidine, loperamide produces adverse respiratory depression without significant alteration of the plasma accumulation due to its central opioid effect [22]. Conversely, P-gp can restrict or even eliminate the entry of CNS-targeted drugs into the brain, resulting in the reduction of the clinical efficacy [23].

In addition to normal tissues and organs, various types of tumor can over-express P-gp, producing multidrug resistance (MDR) [24], in which a single drug causes a non-drug resistant cell or cell line to become cross-resistant to other pharmacologically unrelated drugs due to the increase of administrated drug efflux and the decrease of intracellular drug accumulation [25]. As a result, P-gp efflux remains a major obstacle in the success of various kinds of cancer treatment [26] as well as infectious diseases [3,27]. For instance, brain tumor is one of the leading forms of malignancy and one of highest causes of cancer-related mortality among young adults aged less than 40 years and children [28] and glioma is the most common type of primary brain cancer with limited survival time and rate [29]. The CNS penetration of cediranib, which is a tyrosine kinase inhibitor for the treatment of glioma, is severely limited by the P-gp active efflux [30]. Co-administration of P-gp inhibitors is conceptually plausible and yet infeasible to circumvent MDR because of ineffective P-gp inhibitors in practical clinical applications [31,32]. Alternatively, P-gp

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

can be considered as an anti-target in pharmaceutical research [33] especially in the field of CNS-targeted therapeutics [34,35]. Nevertheless, not all of marketed drugs have to be P-gp non-substrates provided that their therapeutic index is large with respect to the P-gp efflux ratio (ER) [36,37]. For instance, risperidone and 9-hydroxyl risperidone are clinically approved therapeutic agents for the treatment of schizophrenia despite that they are P-gp substrates [38]. Accordingly, it is conceivable to expect that quantitative measure, *viz*. P-gp substrate efflux ratio, is more clinically relevant than qualitative classification, *viz*. substrate/non-substrate classification.

Of various *in vitro* assays to measure the efflux ratio [39-42], the monolayer efflux assay is the most relevant to drug distribution and the most commonly used in practice [20], in which the polarized epithelial cells, such as Madin-Darby canine kidney (MDCK) cells, are transfected with the *MDR1* gene, followed by measuring the ratios between basolateral-to-apical (B→A) apparent permeability (P_{app}) and apical-to-basolateral (A→B) P_{app} [43]

$$\mathrm{ER} = \frac{P_{app}\left(\mathrm{B} \rightarrow \mathrm{A}\right)}{P_{app}\left(\mathrm{A} \rightarrow \mathrm{B}\right)} \tag{1}$$

where P_{app} is evaluated by

$$P_{app} = \frac{1}{AC_0} \cdot \frac{dQ}{dt} \tag{2}$$

using the membrane surface area (A), initial dosing concentration of the test molecule (C_o) in the donor compartment, and the amount of molecule transported per time (dQ/dt) in the receiver compartment [44]. Normally, molecules with ER > 2 are classified as P-gp substrates [39].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

In contrast to *in vitro* and *in vivo* assays, *in silico* approaches are usually swift, inexpensive, labor less intensive, and less time-consuming for drug discovery and ADME/Tox profiling [45,46]. In fact, numerous P-gp classification structure–activity relationship (CSAR) models have been published in elsewhere [47-69], whereas *in silico* quantitative studies of efflux ratio are scanty [70-72]. Nevertheless, it is highly challenging to accurately model P-gp-substrate interactions [73] since P-gp is highly promiscuous *per se* as the result of the fact that P-gp can undergo substantial conformational changes upon binding with various ligands as illustrated by Figure 1 of Leong *et al*. [74]. In addition, P-gp has multiple substrate binding sites as reported [73,75-78]. The mechanism of P-gp substrate efflux is even far more complicated than P-gp-substrate interactions since P-gp substrate efflux can take place through various routes in that substrates can be actively transported by P-gp from the cytoplasm into the extracellular environment in an energy-dependent manner or through a protein channel positioned between the inner and outer leaflets of the lipid membrane as illustrated by Figure 2 of Edwards [79]. In addition to active transport, P-gp substrates also can passively diffuse from the cytoplasm into the extracellular environment through transcellular diffusion and/or paracellular route as illustrated by Figure 1 of Balimane *et al*. [80]. Notably, the P-gp substrate vinblastine, for instance, can be both passively diffused and actively transported [81]. As such, those modeling schemes employed by previously published investigations can only render the direct protein-ligand interactions and they are not suitable to model the efflux ratio. Conversely, any quantitative structure-activity relationship (QSAR) schemes, which are a mathematic means to establish the relationship between biological activity and chemical characteristics, provide the better approaches to model the efflux ratio since they can take into account any mechanisms that can occur through complex routes [82].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

The complexity of P-gp mediated efflux can be even problematic once the delicate roles played by those associated chemical features, *viz*. descriptors in QSAR models, are taken into considerations. For instance, inhibitors, modulators, and substrates can interact with P-gp using the hydrophobicity, hydrogen-bond acceptor (HBA), and hydrogen-bond donor (HBD) features [47,74,83]. Accordingly, hydrophobicity, HBA, and HBD can simultaneously enhance and reduce the P-gp efflux, and it is plausible to expect extremely nonlinear relationships between those chemical features and efflux ratio, suggesting that those linear models can yield significant prediction errors once applied to the test samples that are very different from their training patterns.

Thus, it seems extremely difficult, if not completely impossible, to develop a sound *in silico* model to predict the P-gp substrate efflux ratio to compressively take into account those critical factors mentioned above. A solution to such challenge, however, can be obtained by the novel hierarchical support vector regression (HSVR) scheme proposed by Leong *et al*. [84] because HSVR can render the complex and varied dependencies of descriptors. As such, HSVR can simultaneously possess the advantageous characteristics of a local model and a global model, *viz*. broader coverage of applicability domain and higher level of predictivity, respectively. Furthermore, HSVR is designated to circumvent the "mesa effect" [85] in that the performance of a developed model deteriorates dramatically when applied to extrapolated predictions as demonstrated elsewhere [86,87]. In another words, HSVR is insensitive to outliers as compared with the other predictive models that is of critical importance to a predictive model [88]. Herein, the objective of this investigation was to develop an accurate, fast, and predictive *in silico* model based on the HSVR scheme to predict the P-gp substrate efflux ratio to facilitate drug discovery to design molecules with a more preferable ADME/Tox profile.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

2. Materials and Methods

2.1 Data compilation

A sound predictive model can only be built based on good quality of sample data [89]. To compile quality data for this study, a comprehensive literature search was conducted to retrieve efflux ratio values from various sources to maximize the structural diversity. If there were two or more available efflux ratio data for a given compound and in close range, the average values were then taken in order to warrant better consistency. Further data curation was carried out by cautiously inspecting molecular structures to remove those molecules without definite stereochemistry.

2.2 Molecular descriptors

All of the molecules enlisted in this study were subjected to full geometry optimization using the density functional theory (DFT) B3LYP method with the basis set 6-31G(d,p) by the *Gaussian 09* package (Gaussian, Wallingford, CT) in the dimethyl sulfoxide (DMSO) solvent system using the polarizable continuum model (PCM) [90,91] to mimic the experimental conditions. These geometries were confirmed to be real minima on the potential energy surface by force calculations when no imaginary frequency was obtained. Additionally, atomic charges were also calculated by the molecular electrostatic potential-based method of Merz and Kollman [92] and the highest occupied molecular orbital energy (E_{HOMO}), lowest unoccupied molecular orbital energy (E_{LUMO}), free energy (ΔG), and dipole (μ) were also retrieved from the optimization calculations since those quantum mechanics descriptors have been adopted previously. As such, it is of necessity to employ

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

a more sophisticated quantum mechanics method to optimize those selected molecules and to calculate their associated descriptors.

The *Discovery Studio* package (BIOVIA, San Diego, CA) and *E-Dragon* (available at the web site http://www.vcclab.org/lab/edragon/) were also utilized to calculate more than 200 one-, two-, and three-dimensional molecular descriptors of those optimized molecules. These descriptors can be classified as electronic descriptors, spatial descriptors, structural descriptors, thermodynamic descriptors, topological descriptors, and E-state indices.

Data filtering was initially performed by removing those descriptors missing for at least one sample or showing little or no discrimination against all samples. Furthermore, only one descriptor should be kept among those descriptors with intercorrelation values of $r^2 > 0.8$ to reduce the probability of spurious correlations as postulated by Topliss and Edwards [93]. It is not uncommon to observe that certain descriptors with broader ranges outweigh those with narrower ranges because of substantial variations in magnitudes. Nevertheless, such problem can be resolved when the non-descriptive descriptors, *viz*. real variable descriptors, are normalized with the following equation [94]

$$\chi_{ij} = \left(x_{ij} - \left\langle x_j \right\rangle\right) \Big/ \left[\sum_{i=1}^{n} \left(x_{ij} - \left\langle x_j \right\rangle\right)^2 \Big/ (n-1)\right]^{1/2} \tag{3}$$

where x_{ij} and χ_{ij} represent the original and normalized *j*th descriptors of the *i*th compound, respectively; $\left\langle x_j \right\rangle$ stands for the mean value of the original *j*th descriptor; and *n* is the number of samples.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

Descriptor selection plays a pivotal role in determining the performance of predictive models [95]. More descriptors will be needed once there are more training samples with more diverse structures [89]. Conversely, it is highly possible to yield an over-trained model when there are too many selected descriptors [96]. The descriptor selection was initially executed by genetic function approximation (GFA) using the QSAR module of *Discovery Studio* due to its effectiveness and efficiency [97]. Further descriptor selection was carried out by the recursive feature elimination (RFE) method, in which the predictive model was repeatedly generated by all but one of descriptors. The descriptors were then ranked according to their contributions to the predictive performance; and the descriptor with least contribution was discarded [98].

2.3 Data partition

The collected molecules were divided into two data sets, namely the training set and test set, to develop and to verify the predictive models using the Kennard-Stone (KS) algorithm [99] implemented in *MATLAB* (The Mathworks, Natick, MA) with an approximate 4:1 ratio as suggested [100]. It has been suggested that a sound model can be derived only based on chemically and biologically similar training samples and test samples [101]. As such, the data distribution was carefully examined to ensure the high levels of biological and chemical similarity in both data sets.

2.4 Hierarchical support vector regression

Support vector machine (SVM) proposed by Vapnik *et al.* [102] was initially designated for use in classification and consequently modified for regression problems by nonlinearly mapping the input data into a higher-dimension space, in which a linear regression is performed [103]. SVM regression takes into account both the training error and the model complexity as compared with the traditional regression algorithms, which develop predictive models by minimizing the training

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

error. As such, SVM performs better than traditional regression methods because of its advantageous characteristics, namely dimensional independence, limited number of freedom, excellent generalization capability, global optimum, and easy to implement [104].

Like any other linear or machine learning (ML)-based QSAR techniques, SVM has to tradeoff between the characteristics of a global model, *viz*. broader coverage of applicability domain (AD), and a local model, *viz*. higher level of predictivity [105]. This seeming dilemma, nevertheless, can be plausibly resolved using the hierarchical support vector regression (HSVR) scheme, which was initially proposed by Leong *et al*. and was derived from SVM [84], because HSVR can simultaneously take into consideration both seemingly mutually exclusive characteristics. Practically speaking, it has been demonstrated that HSVR outperformed a number of ML-based models, namely artificial neural network (ANN), genetic algorithm (GA), and SVM [86].

The detail of HSVR has been mentioned elsewhere [84]. Briefly, a panel of SVR models was built by the *LIBSVM* package (software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm) based on various descriptor combinations, and each SVR model represented a local model. The model generation and verification were executed using the modules *svm-train* and *svm-predict*, respectively, implemented in the *LIBSVM* package. The regression modes, namely, ε-SVR and γ-SVR, were adopted, and radial basis function (RBF) was employed as the kernel due to its simplicity and better performance when compared with the others [106]. The runtime parameters, namely regression modes ε-SVR and v-SVR, the associated ε and v, cost C, and the kernel width γ, were scanned by the systemic grid search algorithm using an in-house Perl script [107], in which all parameters were changed independently in a parallel fashion.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

Two SVR models were initially adopted to develop an SVR ensemble (SVRE), which, in turn, was further subjected to regression by another SVR to yield the final HSVR model. The two-member SVREs were continuously assembled until the HSVR model performed well. Otherwise, the three- or even four-member ensembles were built by adding one or more SVR models, respectively, if all two-member ensembles failed to perform well. The descriptor selection and ensemble assembly were predominantly governed by the principle of Occam's razor [108] by adopting the least numbers of descriptors and SVR models.

2.5 Predictive evaluation

The predictivity of a generated model was evaluated by several statistic metrics. The coefficients r^2 and q^2 in the training set and external set, respectively, for the linear least square regression were computed by the following equation

$$r^2, q^2 = 1 - \sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2 \Big/ \sum_{i=1}^{n} \left(y_i - \langle \hat{y} \rangle\right)^2 \tag{4}$$

where $\hat{y}_i$ and y_i are the predicted and observed values, respectively; and $\langle \hat{y} \rangle$ and n stand for the average predicted value and the number of samples in the data set, respectively.

Furthermore, the residual Δ_i, which is the difference between y_i and $\hat{y}_i$, was calculated

$$\Delta_i = y_i - \hat{y}_i \tag{5}$$

The root mean square error (RMSE) and the mean absolute error (MAE) for n samples in the data set were computed

$$\mathrm{RMSE} = \left[\sum_{i=1}^{n} \Delta_i^2 \Big/ n \right]^{1/2} \tag{6}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\Delta_i| \tag{7}$$

The produced model was further subjected to 10-fold cross-validation instead of the widely used leave-one-out due to its better performance [109], giving rise to the correlation coefficient of 10-fold cross validation q^2_{CV}. In addition to cross-validation, the developed models were also internally validated by the *Y*-scrambling test [89], which was carried out by randomly permuting the log ER values, *viz*. *Y* values, to refit the previously developed models while the descriptors were remained unaltered, giving rise to the correlation coefficient r_s^2. The observed log ER values were scrambled 25 times as suggested [110] to produce the average correlation coefficient $\langle r_s^2 \rangle$. Furthermore, various modified versions of r^2 proposed by Ojha *et al*. [111] were also computed

$$r_m^2 = r^2\left(1 - \sqrt{\left|r^2 - r_o^2\right|}\right) \tag{8}$$

$$r_m'^2 = r^2\left(1 - \sqrt{\left|r^2 - r_o'^2\right|}\right) \tag{9}$$

$$\langle r_m^2 \rangle = \left(r_m^2 + r_m'^2\right)/2 \tag{10}$$

$$\Delta r_m^2 = \left|r_m^2 - r_m'^2\right| \tag{11}$$

where the correlation coefficient r_o^2 and the slope of the regression line *k* were calculated from the regression line (predicted *vs*. observed values) through the origin, whereas $r_o'^2$ was calculated from the regression line (observed *vs*. predicted values) through the origin.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

Moreover, the correlation coefficients q^2_{F1}, q^2_{F2}, and q^2_{F3} and concordance correlation coefficient (*CCC*) proposed by Shi *et al.* [112], Schüürmann *et al.* [113], Consonni *et al.* [114], and Chirico and Gramatica [115] were also computed by *QSARINS* [116,117] to evaluate the model performance in the external data set

$$q^2_{F1} = 1 - \sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2 \Big/ \sum_{i=1}^{n_{EXT}} (y_i - \langle y_{TR} \rangle)^2 \tag{12}$$

$$q^2_{F2} = 1 - \sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2 \Big/ \sum_{i=1}^{n_{EXT}} (y_i - \langle y_{EXT} \rangle)^2 \tag{13}$$

$$q^2_{F3} = 1 - \left[\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2 / n_{EXT} \right] \Big/ \left[\sum_{i=1}^{n_{TR}} (y_i - \langle y_{TR} \rangle)^2 / n_{TR} \right] \tag{14}$$

$$CCC = \frac{2 \sum_{i=1}^{n_{EXT}} (y_i - \langle y_{EXT} \rangle)(\hat{y}_i - \langle \hat{y}_{EXT} \rangle)}{\sum_{i=1}^{n_{EXT}} (y_i - \langle y_{EXT} \rangle)^2 + (\hat{y}_i - \langle \hat{y}_{EXT} \rangle)^2 + n_{EXT} (\langle y_{EXT} \rangle - \langle \hat{y}_{EXT} \rangle)^2} \tag{15}$$

where n_{TR} and n_{EXT} are the numbers of samples in the training set and external set, respectively; $\langle \hat{y}_{TR} \rangle$ is the average predicted value in the training set; and $\langle y_{EXT} \rangle$ and $\langle \hat{y}_{EXT} \rangle$ are the average observed and predicted values in the external set, respectively.

Various criteria for those statistical parameters have been proposed to gauge the model predictivity [118]. For instance, Chirico and Gramatica considered that both q^2_{F3} and *CCC* are the best validation parameters to measure the predictivity [115], whereas Roy *et al.* suggested that $\langle r^2_m \rangle$ and Δr^2_m are the most stringent metrics [119]. Recently, Todeschini *et al.* have demonstrated that q^2_{F3} is the most reliable metric [120]. The parameter q^2_{F2} has been adopted by Organization for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

Economic Co-operation and Development (OECD) to assess the performance of QSAR models [113].

More importantly, a model can be considered as predictive if it can meet the most stringent criteria collectively proposed by Golbraikh *et al*. [121], Ojha *et al*. [111], Roy *et al*. [119], and Chirico and Gramatica [120].

$$r^2, q^2_{\mathrm{CV}}, q^2, q^2_{\mathrm{Fn}} \geq 0.70 \tag{16}$$

$$\left|r^2 - q^2_{\mathrm{CV}}\right| < 0.10 \tag{17}$$

$$\left(r^2 - r_o^2\right)/r^2 < 0.10 \text{ and } 0.85 \leq k \leq 1.15 \tag{18}$$

$$\left|r_o^2 - r_o'^2\right| < 0.30 \tag{19}$$

$$r_m^2 \geq 0.65 \tag{20}$$

$$\left\langle r_m^2 \right\rangle \geq 0.65 \text{ and } \Delta r_m^2 < 0.20 \tag{21}$$

$$CCC \geq 0.85 \tag{22}$$

where r in Eqs. (18)–(21) represents the parameters r and q in the training set and external set, respectively; and q_{Fn} in Eq. (16) stands for q_{F1}, q_{F2}, and q_{F3}.

3. Results

3.1 *Data Compilation*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

More than 550 compounds were collected after comprehensive literature search. Data curation was carefully carried out by eliminating those compounds i) with only qualitative array results (*i.e.* substrate or non-substrate), ii) without specific ER values, or iii) chemical structures. In addition, cells used to express P-gp protein also play a significant role in determining ER values. For instance, the measured ER values of astemizole were 2.16 and 0.6 assayed in MDCK and human colon adenocarcinoma (Caco-2) cells, respectively [51,122]. Of various assayed cells, 63 molecules tested in MDCK cells were selected from various sources [39,122-136] since it constituted the largest amount of data. The data size is seemingly small since a number of CSAR models have been derived based on rather large amounts of data. For instance, Li *et al*. [68] built various predictive models based on 423 P-gp substrates and 399 non-substrates compiled from numerous sources. Nevertheless, their data were generated from different assay conditions (*e.g.* different cell lines), leading to high levels of data heterogeneity. QSAR models, conversely, are vulnerable to data inhomogeneity [89]. Additionally, some molecules such as those selenium-containing ones [137] were excluded due to the fact that their topological descriptors, for instance, cannot be enumerated. Those ER values were discarded when they were not consistent with their measured P_{app} (B→A) and P_{app} (A→B) values [138]. Recently, the efflux ratios of more than 4,000 Amgen in-house compounds were measured [139]. It is plausible to expect that the great sample amount and data consistency can furnish a good ER pool. Unfortunately, those chemical structures are proprietary, leading to the fact that there are only limited quantitative data with chemical structures available in the public domain to date. Those factors partially contribute to the fact that there is no genuine QSAR model has been published.

As such, only very limited data samples with available chemical structures and consistent assay conditions were recruited in this study to maximize the structural diversity and to maintain data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

homogeneity after purging inappropriate data based on above-mentioned criteria. Table S1 lists the SMILES strings, CAS registry numbers, efflux ratio values, and literature references of all molecules collected in this study.

3.2 *Data Partition*

Of all molecules adopted in this study, 50 and 13 molecules were randomly assigned to the training set and test set, respectively, with a *ca*. 4:1 ratio as suggested [100]. Figure S1 displays the projection of all molecules enrolled in this investigation in chemical space, spanned by the first three principal components (PCs), explaining 94.6% of the variance in the original data. As illustrated, both data sets exhibited high levels of similarity in the chemical space. Furthermore, the high levels of biological and chemical similarity between both data sets can also be validated by Figure S2, which shows the histograms of log ER, molecular weight (MW), polar surface area (PSA), number of HBA, and number of HBD in density form for all molecules in the training set and test set. Thus, it can be asserted that there was no substantial bias in data sets.

3.3 SVRE

Of all generated SVR models using various combinations of descriptors and runtime parameters, three SVR models, denoted by SVR A, SVR B, and SVR C, were assembled to construct the SVR ensemble, which was further subjected to regression by another SVR to generate the HSVR model. Table S2 summarizes the optimal runtime parameters of SVR A, SVR B, and SVR C. These three SVR models, which adopted 4, 6, and 3 descriptors (Table 1), respectively, were selected based on their individual performances on all molecules and statistical analyses in the training set and test set. Table S1 lists the predicted log ER values. Tables 2 and 3 summarize the associated statistical analyses of these three SVR models in the training set and test set, respectively. Figures

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

1 and 2 display the scatter plots of observed versus the predicted log ER values by SVR A, SVR B, and SVR C for the molecules in the training set and test set, respectively.

Table 1. Descriptor selected as the input of SVR models in the ensemble and their description.

Descriptor	SVR A	SVR B	SVR C	Description
SA			x	Total surface area
n_{N+O}	x	x		Number of nitrogen and oxygen atoms
V_m	x	x	x	Molecule volume
PSA	x	x		Polar surface area
HBD	x	x		Number of hydrogen bond donating groups
n_{Rot}		x	x	Number of rotatable bonds
n_{Ar}		x		Number of aromatic rings

Figure 1 shows that the predictions by SVR A, SVR B, and SVR C are in good agreement with the observed values for most of the molecules in the training set as further manifested by their small RMSDs, average deviations, standard deviations (s), and larger r^2 parameters (Table 2). Of 50 training samples, SVR A, SVR B, and SVR C gave rise to 28, 3, and 2 predictions, which deviated from the experimental values by more than 0.10, respectively. It can be further observed from Figure 1 that most of the points predicted by SVR C generally lie on or are closer to the regression line when compared with SVR A and SVR B. As a result, SVR C produced the lowest MAE (0.02), s (0.06), and RMSE (0.06) and the highest r^2 parameter (0.98), suggesting that SVR C performed better than SVR A and SVR B for the molecules in the training set. Nevertheless, the predictions of quinidine (**48**) by SVR A, SVR B and SVR C unanimously yielded the maximum

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

residuals of 0.32, 0.51 and 0.40, respectively, denoting that SVR A executed better than SVR B and SVR C.

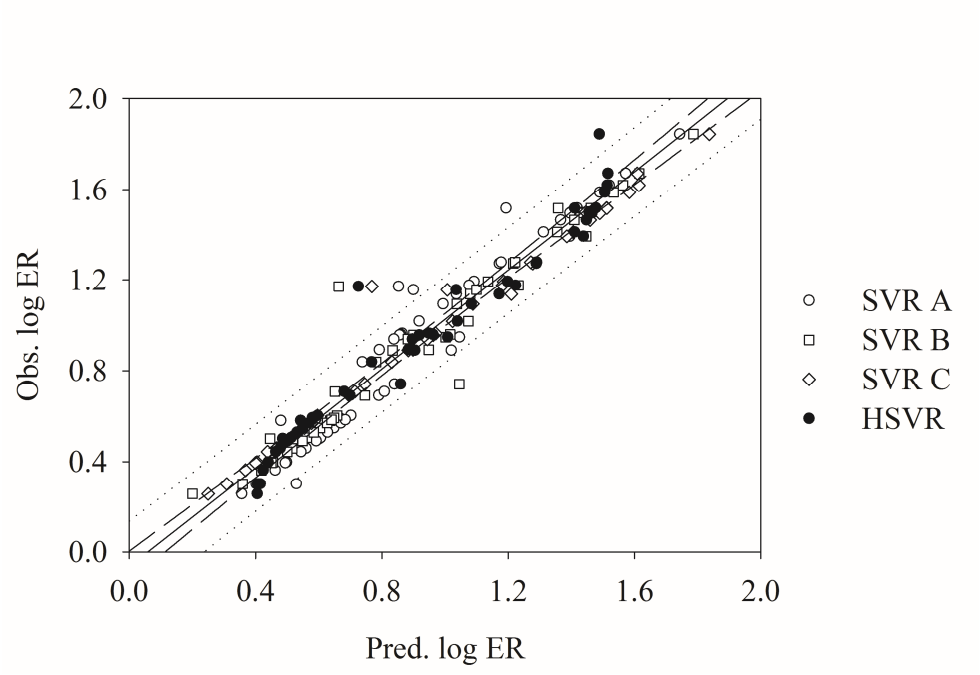


Figure 1. Observed log ER *vs*. the log ER predicted by SVR A (open circle), SVR B (open square), SVR C (open diamond), and HSVR (solid circle) for the molecules in the training set. The solid line, dashed line, and dotted lines correspond to the HSVR regression of the data, 95% confidence.

Table 2. Statistic evaluations, namely correlation coefficient (r^2), maximum residual (Δ_{Max}), mean absolute error (MAE), standard deviation (s), RMSE, and 10-fold cross-validation correlation coefficient (q^2_{CV}) evaluated by SVR A, SVR B, SVR C, and HSVR in the training set.

	SVR A	SVR B	SVR C	HSVR
r^2	0.95	0.95	0.98	0.96
Δ_{Max}	0.32	0.51	0.40	0.45
MAE	0.11	0.07	0.02	0.06

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

s	0.12	0.10	0.06	0.10
RMSE	0.12	0.10	0.06	0.10
q^2_{CV}	0.01	0.01	0.07	0.94

The predictions by SVR A, SVR B, and SVR C in the test set are also in good agreement with the experimental values (Figure 2). Nevertheless, most of the residuals obtained by the three SVR models in the test set are more than 0.15 (11, 11, and 8, respectively). It can be further observed from Table 3 that the mean absolute errors computed by SVR A, SVR B, and SVR C unequivocally increase from 0.11, 0.07, and 0.02 in the training set to 0.29, 0.22, and 0.24 in test set, respectively. The other statistical parameters also suggest that the performances of these three models in the SVRE slightly decline from the training set to the test set (Tables 2 and 3). The maximum residual computed by SVR C in the test set was yielded from the prediction of cimetidine (**13**) with an absolute residual of 0.55, which were only 0.34 and 0.10 by SVR A and SVR B, respectively. Similarly, vinblastine (**58**) was best predicted by SVR C with an absolute residual of 0.01, whereas SVR A and SVR B gave rise to absolute errors of 0.60 and 0.41, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

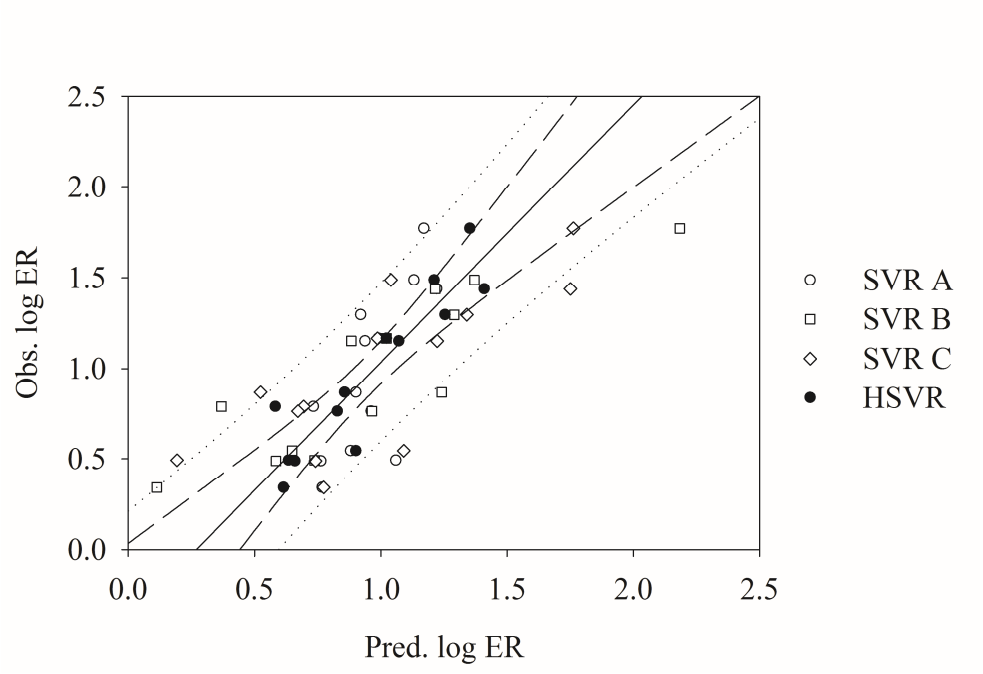


Figure 2. Observed log ER *vs*. the log ER predicted by SVR A (open circle), SVR B (open square), SVR C (open diamond), and HSVR (solid circle) for the molecules in the test set. The solid line, dashed line, and dotted lines correspond to the HSVR regression of the data, 95% confidence interval for the HSVR regression, and 95% confidence interval for the prediction, respectively.

Table 3. Statistic evaluations, correlation coefficients q^2, q^2_{F1}, q^2_{F2}, and q^2_{F3}, concordance correlation coefficient (*CCC*), maximal absolute residual (Δ_{Max}), mean absolute error (MAE), standard deviation (*s*), and RMSE evaluated by SVR A, SVR B, SVR C, and HSVR in the test set.

	SVR A	SVR B	SVR C	HSVR
q^2	0.54	0.75	0.60	0.83
q^2_{F1}	0.39	0.67	0.55	0.80
q^2_{F2}	0.39	0.67	0.54	0.80
q^2_{F3}	0.38	0.66	0.54	0.80

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

CCC	0.45	0.86	0.78	0.87
Δ_{Max}	0.60	0.42	0.55	0.42
MAE	0.29	0.22	0.24	0.17
s	0.35	0.26	0.30	0.22
RMSE	0.34	0.25	0.29	0.21

Furthermore, SVR A, SVR B, and SVR C yielded the q^2 values of 0.54, 0.75, and 0.60 in the test and the cross-validation correlation coefficients q^2_{CV} of 0.01, 0.01, and 0.07 in the training set, respectively (Tables 2 and 3). When subjected to the *Y*-scrambling test, SVR A, SVR B, and SVR C gave rise to the $\langle r_s^2 \rangle$ values of 0.02, 0.03, and 0.03, respectively (Table 1). The almost zero values of $\langle r_s^2 \rangle$ as well as substantial differences between corresponding r^2 and $\langle r_s^2 \rangle$ signify that those three SVR models in the ensemble are not the result of chance correlation [110]. Conversely, the substantial differences between r^2 and q^2 and between r^2 and q^2_{CV} imply the over-fitting characteristics of these three models that actually can be further manifested by their small q^2_{F1}, q^2_{F2}, q^2_{F3}, and *CCC* values (Table 3). As a result, it is plausible to expect that these models will have limited coverage of applicability domain.

3.4 HSVR

The HSVR model was produced by the regression of the SVR ensemble based on the predictions of all molecules and statistical evaluations in the training set (Table S1 and Table 2). Table S2 lists the optimal runtime conditions for the final SVR model. It can be observed from Figure 1 that the HSVR model showed better prediction accuracy than SVR A, SVR B, and SVR C for the molecules in the training set because the distances between the predictions by HSVR and regression line are generally between the largest ones and smallest ones produced by its SVR

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

counterparts in the ensemble. However, HSVR executed better than any of SVR models in the ensemble in some cases. The predictions of desloratadine (**19**) by SVR A, SVR B, SVR C, and HSVR, for instance, yielded absolute residuals of 0.10, 0.06, 0.01, and 0.00, respectively. Statistically, HSVR performed better than SVR A and SVR B, whereas SVR C, in turn, functioned negligibly better than HSVR as manifested by those parameters listed in Table 2. For example, SVR A, SVR B, SVR C, and HSVR yielded the r^2 values of 0.95, 0.95, 0.98, and 0.96, respectively.

When applied to the test samples, HSVR only showed insignificant performance decreases from the training set to the test set. For instance, RMSE increased from 0.10 in the training set to 0.21 in the test set (Tables 2 and 3). However, the maximum residual declined from 0.45 in the training set to 0.42 in the test set. Figure 2 displays that HSVR showed better performance than SVR A, SVR B, and SVR C in the test set. The performance predominance of HSVR can be further manifested by those statistical parameters listed in Table 3. For instance, SVR A, SVR B, SVR C, and HSVR gave rise to MAE values of 0.29, 0.22, 0.24, and 0.17, respectively. Similar observation that HSVR generated smaller absolute residuals than its counterparts in the ensemble can also be found in the test set. The absolute prediction error of paliperidone (**41**), for instance, was 0.14 given rise by HSVR, whereas SVR A, SVR B, and SVR C produced residuals of 0.57, 0.25, and 0.30, respectively. Additionally, HSVR yielded the smallest differences between r^2 and q^2_{CV} (0.02) and between r^2 and q^2 (0.13), indicating that HSVR was a well-trained model or no over-fitting effect was observed because it will otherwise produce at least one significant difference among those parameters. Similarly, the possibility of chance correlation of HSVR can be eliminated by *Y*-scrambling since it also produced an almost zero of $\langle r_s^2 \rangle$ (0.03) and marked difference between r^2 and $\langle r_s^2 \rangle$ (Table 2) [110].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

3.5 Predictive evaluations

Figure 3 displays the scatter plots of the residual *vs*. the log ER values predicted by HSVR for the molecules in the training set and test set. It can be observed that the residuals are approximately evenly distributed on both sides of *x*-axis along the range of predicted values in both data sets, suggesting that there is no systematic error associated with the HSVR model [89]. The unbiased predictions can be further exhibited by its almost negligible average residuals that were −0.02 and −0.02 in the training set and test set, respectively (Table S1).

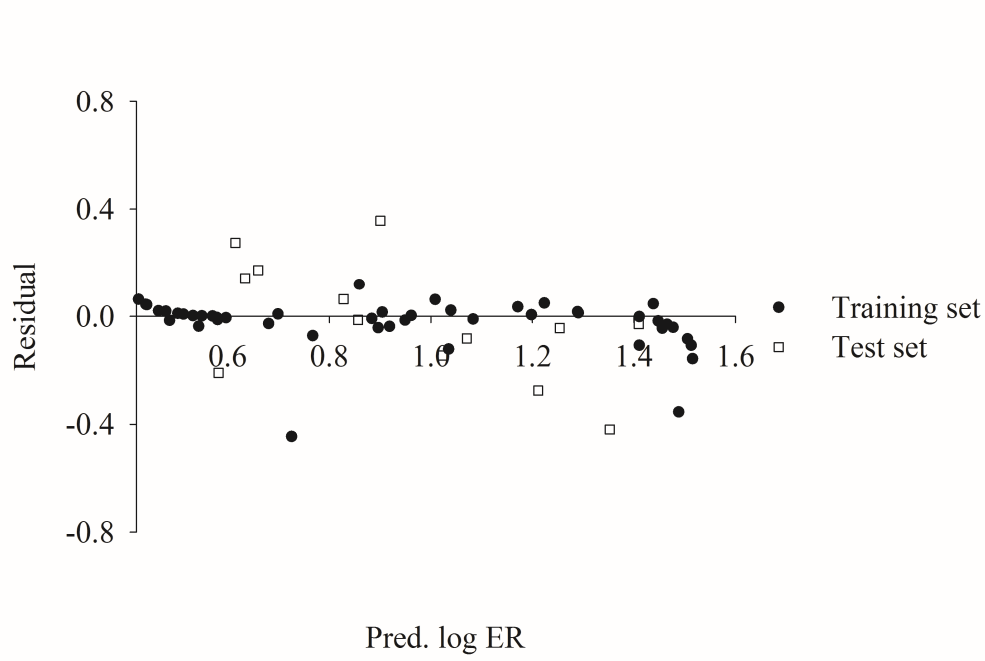


Figure 3. Residual *vs*. the log ER predicted by HSVR in the training set (solid circle) and test set (open square).

The predictivity of generated HSVR model was further evaluated by those validation requirements proposed by Golbraikh *et al*. [121], Ojha *et al*. [111], Roy *et al*. [119], and Chirico and Gramatica [120] (Eqs. (18)–(21)) in the training set and test set. Table 4 summarizes the results,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

from which it can be observed that HSVR maintained similar high levels of performance in the training set and test set. Additionally, HSVR fulfilled all validation requirements, indicating that this predictive model is highly accurate and predictive.

Table 4. Validation verification of HSVR based on prediction performance of those molecules in the training set and test set.

	Training set	Test set
n	50	13
r_o^2	0.95	0.77
k	1.03	1.05
$r_o'^2$	0.94	0.52
r_m^2	0.90	0.72
$r_m'^2$	0.85	0.60
$\langle r_m^2 \rangle$	0.88	0.66
Δr_m^2	0.05	0.12
$r^2, q_{\mathrm{CV}}^2, q^2, q_{\mathrm{Fn}}^2 \geq 0.70$	x	x
$\left\|r^2 - q_{\mathrm{CV}}^2\right\| < 0.10$	x	N/A
$\left(r^2 - r_o^2\right)/r^2 < 0.10$ and $0.85 \leq k \leq 1.15$	x	x
$\left\|r_o^2 - r_o'^2\right\| < 0.30$	x	x
$r_m^2 \geq 0.65$	x	x
$\langle r_m^2 \rangle \geq 0.65$ and $\Delta r_m^2 < 0.20$	x	x
$CCC \geq 0.85$	N/A[†]	x

3.6 Mock test

To mimic real world challenges, the developed HSVR model was further tested by those P-gp substrates assayed by Crivori *et al.* [51]. Of all marketed drugs measured by Crivori *et al.*, 12 were

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

also enrolled in this study, yielding a good way to calibrate the testing system. However, these molecules were measured in Caco-2 cells, whereas all of the molecules adopted in this study were tested in MDCK cells, suggesting that those compounds assayed by Crivori *et al*. are not qualified as the second external or test set since those validation criteria (*vide supra*) are not applicable to these compounds. To eliminate the discrepancy between both assay systems, the linear correlation between both assay systems for those common molecules was first inspected and the obtained scatter plot is illustrated in Figure 4. It can be observed that the experimental values in both systems were modestly correlated with each other well with an *r* value of 0.78. Thus, it is plausible to examine the HSVR model with those novel P-gp substrates assayed in Caco-2 cells.

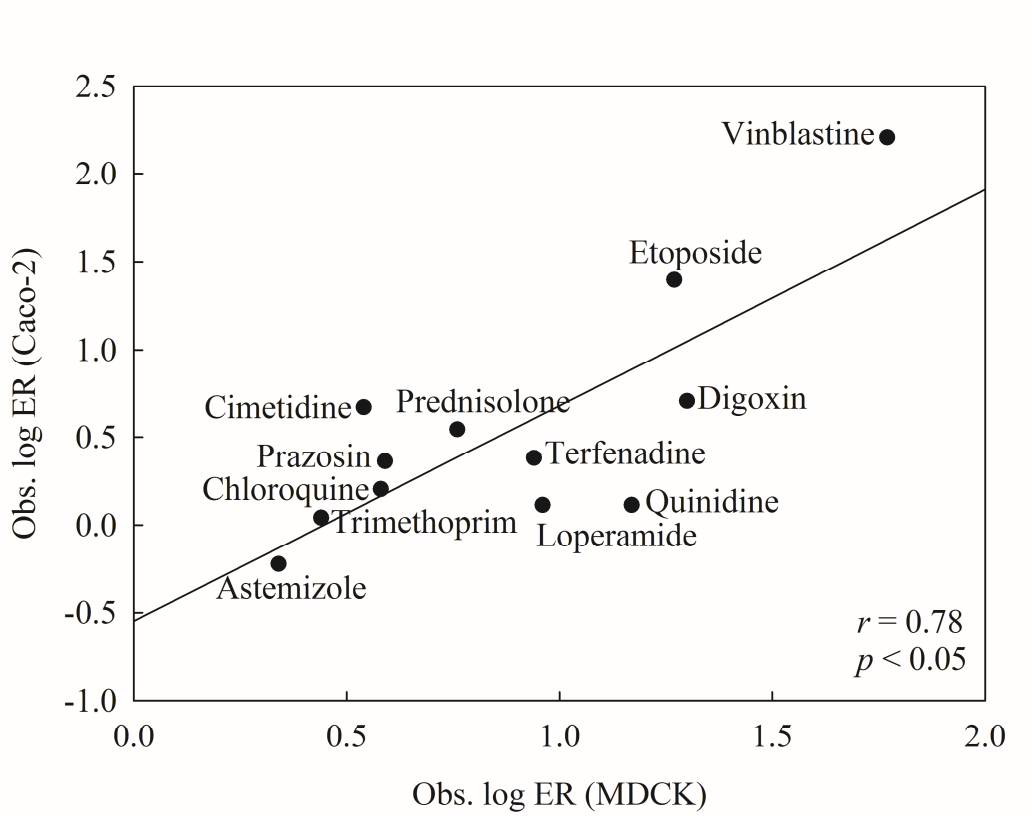


Figure 4. The observed log ER values (Caco-2) *vs*. the observed log ER values (MDCK).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

Figure 5 displays the tested results of the 9 novel drugs. It can be observed that the *r* value between experimental log ER obtained in the Caco-2 cells and predicted log ER in the MDCK cells was 0.77. The negligible difference between both numbers (0.78 *vs*. 0.77) suggests that the predictions by the HSVR model can almost reproduce the experimental observations and this mock test unequivocally assured the predictive capability of HSVR.

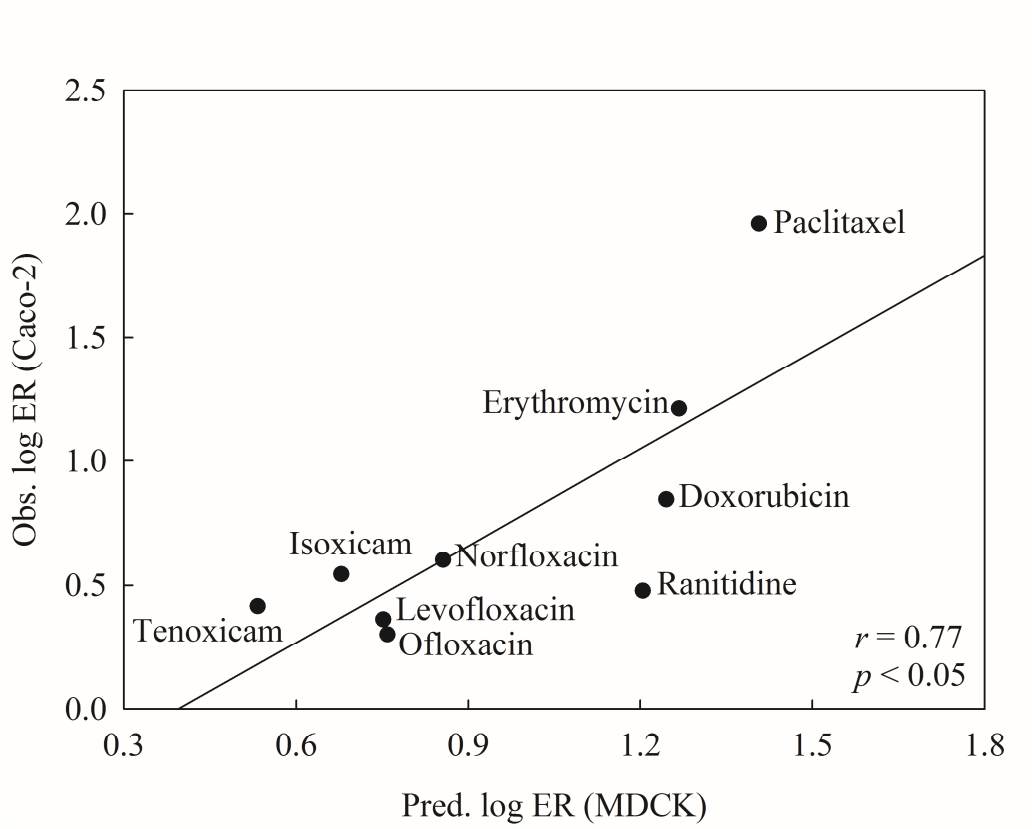


Figure 5. The observed log ER values (Caco-2) *vs*. the predicted log ER values (MDCK).

4. Discussion

Collectively, 7 descriptors were adopted in this study. Intrinsically, the sample-to-descriptor ratio was *ca*. 7:1, which is significantly larger than 5, *viz*. the minimal requirement to lessen the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

probability of chance correlations in a predictive model [93]. However, the process of P-gp substrate efflux is complex since it can take place thought various routes (*vide supra*). As such, different descriptors were adopted by different classification models. Of various descriptors selected by qualitative predictive models, hydrophobic, HBA, and HBD are the most frequently selected chemical features as illustrated by the model proposed by Penzotti *et al*. [47]. However, the analysis of Amgen in-house compounds can reveal that HBD and topological PSA (tPSA) are the predominant factors associated with ER [139].

Figure 6 displays the average log ER for each histogram bin of HBD for all molecules selected in this study. It can be observed that the average log ER value initially increased with HBD when HBD was no more than 6 and then subsequently decreased when HBD was more than 6. Such positive dependence of log ER on HBD is, in fact, consistent with the analysis made by Hitchcock *et al.* [139]. However, those Amgen in-house compounds had HBD of no more than 5, leading to an only positive relationship between log ER and HBD. Such discrepancy in both systems can be conceivably attributed to the fact that the initial P-gp substrate binding can be enhanced by HBD as illustrated by the pharmacophore models of Penzotti *et al*. [47], whereas the consequent transport of the substrates into the extracellular environment can be hampered by too many HBDs, plausibly because of the increase in water desolvation energy [140] and the decrease in membrane fluidity [141]. As such, a nonlinear relationship between HBD and log ER was yielded consequently.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

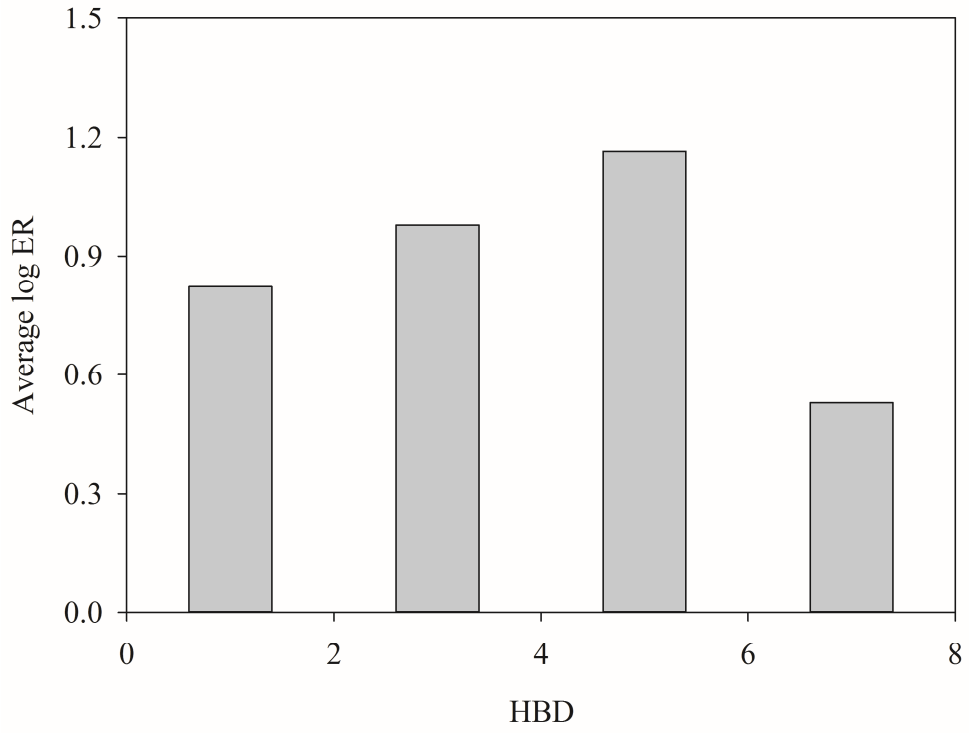


Figure 6. Average log ER *vs*. the distribution of HBD.

It has been observed that hydrophobicity, which normally can be represented by log *P*, plays an important role in P-gp-substrate interaction due to the hydrophobic nature of the substrate binding pocket, resulting in stronger P-gp substrate binding for those more hydrophobic substrates [142]. Nevertheless, the interaction between substrates and lipid bilayer as well as the release of substrates into the extracellular environment also depend on the hydrophobicity of substrates (*vide supra*), leading to a nonlinear relationship between log *P* and log ER. Figure 7 displays the average log ER for histogram bin of log *P* for all molecules enlisted in this study. It can be observed that the average log ER initially increased with log *P* when log *P* was smaller and decreased with log

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

P when log *P* became higher. Such observation is qualitatively similar to the trend of P_{app} (A→B) found by Hitchcock *et al.* [139].

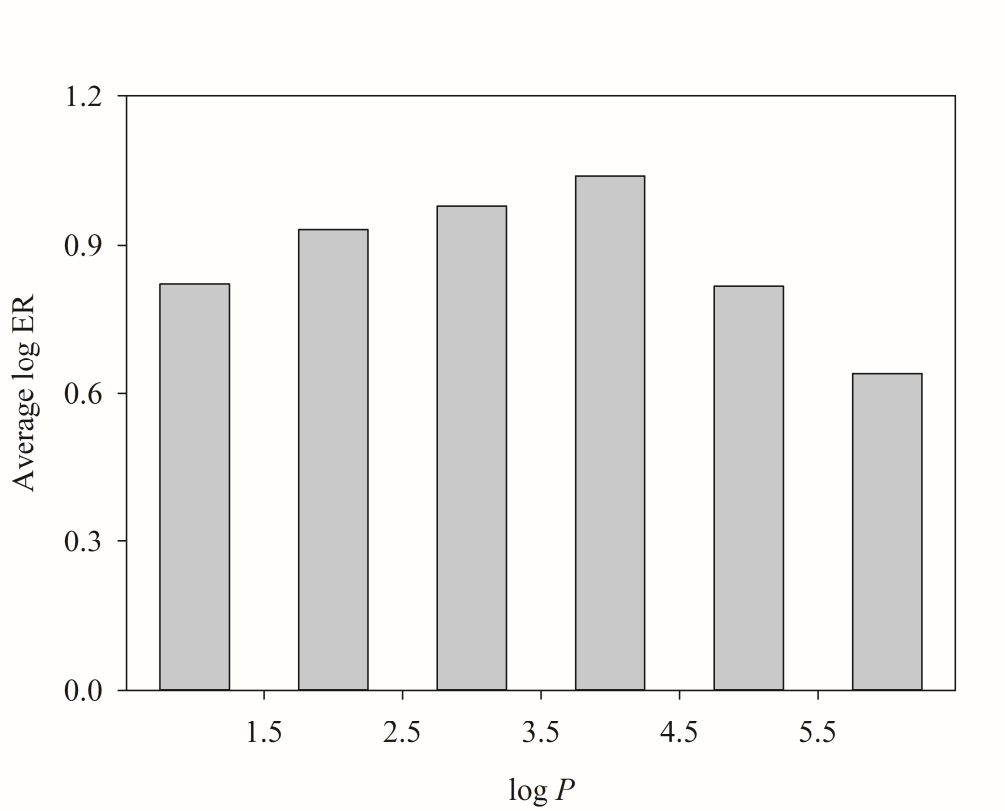


Figure 7. Average log ER *vs*. the distribution of log *P*.

Nevertheless, it is unusual to observe that log *P* was not included in this study, whereas the number of aromatic rings (n_{Ar}) was enlisted in this study. Such inconsistency can be realized by the fact that the average log *P* values increased with n_{Ar} for all of molecules included in this study as illustrated in Figure 8, which displays the average log *P* versus the distribution of n_{Ar}. As such, it is plausible to replace log *P* by n_{Ar}. Furthermore, it has been found that aromatic ring moieties are important in substrate recognition and efflux modulation [143,144]. More importantly, the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

empirical observation has indicated that models with the selection of n_{Ar} unanimously showed better performance than those with the selection of log *P* (data not shown).

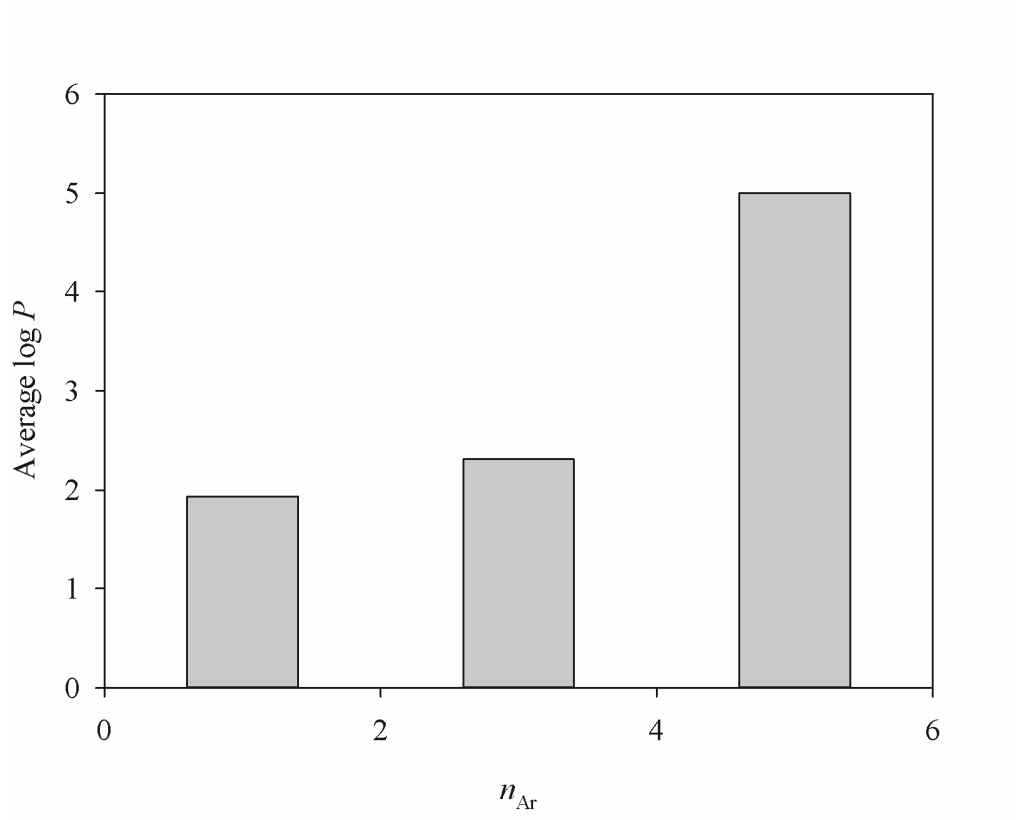


Figure 8. Average log *P* *vs*. the distribution of number of aromatic ring (n_{Ar}).

The significant role of HBA in the P-gp-substrate interaction has been manifested by molecular docking simulations [72] as well as numerous qualitative models. Additionally, it has been suggested that HBA can enhance P-gp-mediated efflux [57]. Nevertheless, it is unusual to observe that none of SVR models in the ensemble has adopted HBA, plausibly because the descriptor number of nitrogen and oxygen atoms (n_{N+O}) correlated well with HBA as demonstrated by Figure 9, which displays n_{N+O} versus HBA. In fact, Desai *et al*. [57] adopted n_{N+O} instead of HBA as the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

substrate classification criterion. Furthermore, empirical model development has shown that models with the selection of n_{N+O} executed better than those with the selection of HBA (data not shown). As a result, the descriptor n_{N+O} was selected in lieu of HBA.

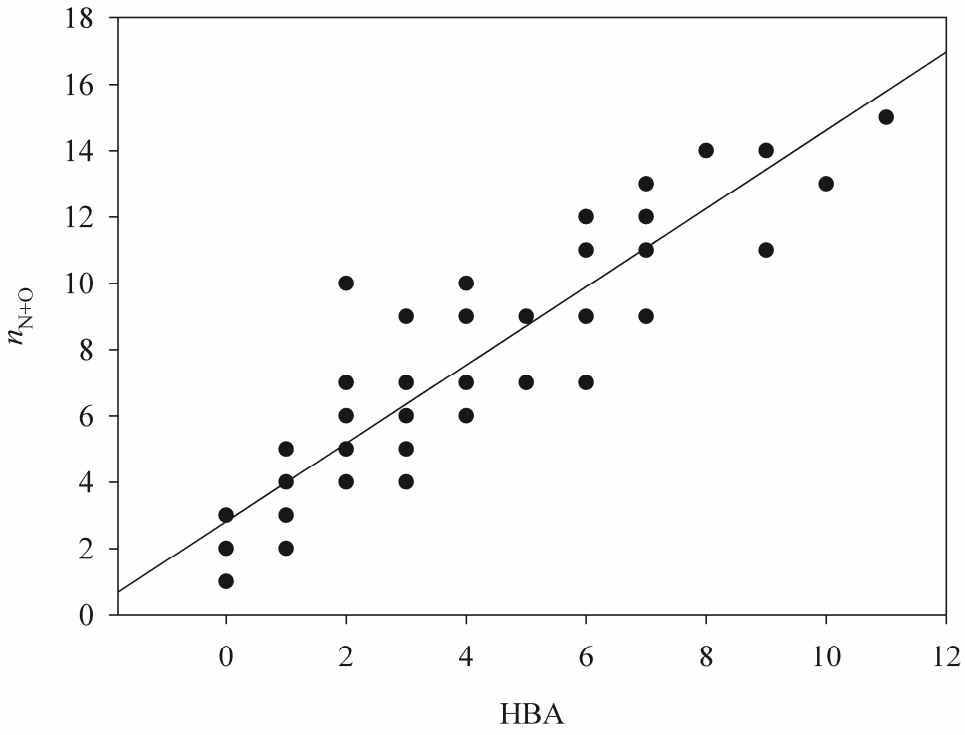


Figure 9. The number of nitrogen and oxygen (n_{N+O}) *vs*. HBA.

The descriptor tPSA is a modified version to swiftly calculate the polar surface area only based on the additive polar surface areas [145]. The recursive partitioning (RP) model of Joung *et al.* [56] indicated the significant role of PSA in classifying molecules as P-gp substrates/non-substrates. Moreover, Hitchcock *et al.* also found the profound contribution of tPSA to P-gp

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

mediated efflux (*vide supra*). Accordingly, the more sophisticated version of PSA was adopted in this study since it can function as polarity and hydrogen-bonding features [68].

It has been observed that the substrate size, which can be characterized by molecular weight (MW), molecular volume (V_m), and total surface area (SA), can have a large impact on P-gp-substrate interaction as well as passive permeability [146]. Nevertheless, it has been suggested that both V_m and SA can be better metrics to estimate the actual molecular size [147], and MW, conversely, was closely associated with V_m with an r^2 values of 0.98 for the molecules enlisted in this study. In fact, it has been postulated that V_m rather than MW is a better metric to associate with ER [148]. Accordingly, V_m and SA were adopted to render the size effects, whereas MW was discarded to reduce the probability of spurious correlations.

It has been found that P-gp substrates generally have more rotatable bonds than non-substrates since more flexible molecules can be more easily to adopt favorable orientation to interact with P-gp [49,68,149]. In fact, non-CNS drugs are more flexible than their CNS counterparts [23] since molecules with more conformational flexibility can favor the internal H-bond formation, which, in turn, can enhance the passive membrane permeability [150]. As such, substrate conformational flexibility, which can be characterized by the number of rotatable bond (n_{Rot}), can facilitate not only the active transport but passive permeability of P-gp substrates, and n_{Rot} was adopted in this investigation.

Gunaydin *et al.* [70] only took into account the contribution of the differences between free energy in water (G_{H_2O}) and that in chloroform (G_{CHCl_3}), *viz.* $\Delta G_{H_2O-CHCl_3}$, since it was hypothesized that P-gp undergoes a conformation change from the intercellular-facing state to extracellular-facing state upon binding with substrates. As such, the transported substrates experience from a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

lipophilic environment into a hydrophilic one. In addition to $\Delta G_{H_2O-CHCl_3}$, the contribution of $\Delta G_{DMSO-CHCl_3}$ was also computed in this study to mimic the assay conditions. Nevertheless, neither of solvation free energy differences was selected in this study due to their insignificant contribution to ERs (data not shown), plausibly because the P-gp conformation change can only account for a small part of the whole complicated efflux process and, additionally, passive permeability is not resulted from the P-gp conformation change. The predictive model of Gunaydin *et al.* [89], nevertheless, was derived only based on 12 marketed drugs that cannot comprehensively render the complex efflux. As such, more descriptors will be required in case of more diverse samples.

Didziapetris *et al.* [64] has proposed the "rule-of-fours," which states that molecules with i) $n_{N+O} \geq 8$, ii) MW > 400, and iii) acid pK_a > 4 are likely to be P-gp substrates. Of all molecule with ER > 2 selected in this study, *viz*. substrates, approximately 32%, 52%, and 100% can meet the criteria $n_{N+O} \geq 8$, MW > 400, and acid pK_a > 4, respectively, and only 29% can completely fulfill those 3 criteria. Actually, Li *et al*. [68] also found that only *ca*. 34% of samples can simultaneously meet those 3 criteria. Furthermore, it is not unusual to observe that different rules have been proposed to classify molecules into P-gp substrates/non-substrates. Desai *et al*. [57], for instance, have proposed the molecules with TPSA >100 Å^2 and most basic pK_a > 8 have higher probability to be substrates. The inconsistency in various proposed rules can be plausibly due to the fact that those rules were derived only based on linear analyses of those P-gp substrates/non-substrates. However, such bisection is not always true as manifested by the naïve Bayesian classifiers built by Li *et al*. [68]. In addition, the size and hydrophobicity of substrates can affect the substrate-membrane interactions nonlinearly [151]. Further complexity can be raised once the P-gp substrate efflux is considered instead of P-gp substrate/non-substrate classification since the P-gp substrate efflux can take place through various routes (*vide supra*), leading to nonlinear relationships

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

between some descriptors and log ER, such as HBD and log *P* (Figures 6 and 7). Numerous attempts have been made in this study to develop various partial least square (PLS) models in order to accommodate the novel 2-QSAR scheme [87] and no satisfactory models were produced (data not shown). Conversely, the accurate and predictive HSVR can comprehensively describe such nonlinear dependence of log ER on descriptors.

Moreover, it has been observed that P-gp and other ABC members, namely breast cancer resistant protein (BCRP/ABCG2) and multidrug resistance-associated protein 4 (ABCC4/MRP4), play a critical role in BBB permeability [152], which can take place via various routes [153] in addition to the already complicated P-gp mediated efflux. As such, it is plausible to expect that it is extremely difficult to develop a sound *in silico* model to predict BBB permeability if not entirely impossible [154]. The development of an accurate *in silico* model in this study to predict the P-gp substrate efflux can pave the way to establish a sound theoretical model to predict the BBB permeability in the future. Most of molecules adopted in this study are marketed drugs for treating various illnesses, such as HIV infection, allergy symptoms, rheumatoid arthritis, hypertension, diarrhea, and different types of cancer in addition to assorted CNS-related disorders (Table S1). The broad spectrum of therapeutic agents unequivocally indicates that the data samples are structurally diverse that can be further manifested by the fact that the average minimum distance between two molecules, *viz*. the distance between two nearest neighbors, in the chemical space was 2.06 with an standard deviation of 1.39 and the maximum distance between two collected samples was 29.57 (Figure S1), giving rise to an ratio of *ca*. 1:14. As such, it is plausible to expect that developed HSVR should have a larger coverage of applicability domain accordingly, which is an important characteristic for a predictive model in practical application. More importantly, the derived HSVR model and published P-gp substrate/non-substrate classification models can work

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

in a synergistic fashion, in which the latter can be used to identify those P-gp substrates and the former can be deployed to predict their efflux ratios.

5. Conclusions

P-gp substrate efflux can be a major obstacle in the success of CNS-targeted therapeutic delivery as well as a critical pharmacokinetic factor for causing DDIs. On the other hand, the CNS-related side-effects of non-CNS drugs can be reduced by P-gp mediated efflux. As such, P-gp substrate efflux is of critical importance to drug discovery and development regardless of CNS drugs or non-CNS drugs. An *in silico* model to predict the P-gp substrate efflux can be valuable to drug discovery and development. Nevertheless, P-gp substrate efflux is a complex process that can take place through various routes, namely active transport and passive permeability, leading to different descriptor combinations as well as different relationships to render these variations in different mechanisms. In this study, a QSAR predictive model derived from the novel hierarchical support vector regression (HSVR) scheme, which can simultaneously possess the advantageous characteristics of a local model and a global model, *viz*. broader coverage of applicability domain and higher level of predictivity, respectively, was developed to envisage the P-gp substrate efflux ratio. The developed HSVR showed great prediction accuracy for the 50 and 13 molecules in the training set and test set, respectively, with excellent predictivity and statistical significance. When mock tested by a group of molecules to mimic real challenges, the derived HSVR model also executed accordantly well. Furthermore, the HSVR model can elucidate the discrepancies among all published P-gp substrate classifiers, indicating its superiority. Hence, it can be affirmed that this HSVR model can be adopted as an accurate and reliable predictive tool, even in the high

 doi:10.20944/preprints201807.0092.v1

throughput fashion, to facilitate drug discovery and development by designing drug candidates with a more desirable pharmacokinetic profile.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0092.v1

Supplementary Materials

Table S1. Selected compounds for this study, their names, SMILES strings, CAS numbers, observed log ER values and predicted values by SVR A, SVR B, and HSVR, data partitions, and references.

Table S2. Optimal runtime parameters for the SVR models.

Figure S1. Molecular distribution for the samples in the training set (solid circle) and test set (open square) in the chemical space spanned by three principal components.

Figure S2. Histograms of (A) observed log ER, (B) molecular weight (MW), (C) polar surface area (PSA), (D) number of hydrogen bond acceptor (HBA), and (E) number of hydrogen bond donor (HBD) in density form for all molecules in the training set and test set.

Acknowledgments: This work was supported by the Ministry of Science and Technology, Taiwan. Parts of calculations were performed at the National Center for High-Performance Computing, Taiwan. The authors are grateful to Prof. Paola Gramatica for providing free license of *QSARINS* and Yi-Lung Ding for helping data analysis.

Author Contributions: C.C., C.F.W., and M.K.L. conceived and designed the study; C.C., M.H.L., and M.K.L. performed the experiments and analyzed the data; C.C., C.F.W., and M.K.L. wrote the paper.

Conflict of Interest: The authors declare that they have no conflict of interest.

 doi:10.20944/preprints201807.0092.v1

References

1. Schinkel, A.H.; Jonker, J.W. Mammalian drug efflux transporters of the atp binding cassette (abc) family: An overview. *Adv. Drug Deliv. Rev.* **2003**, *55*, 3-29.
2. Thiebaut, F.; Tsuruo, T.; Hamada, H.; Gottesman, M.M.; Pastan, I.; Willingham, M.C. Cellular localization of the multidrug-resistance gene product p-glycoprotein in normal human tissues. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 7735-7738.
3. Kim, R.B.; Fromm, M.F.; Wandel, C.; Leake, B.; Wood, A.J.; Roden, D.M.; Wilkinson, G.R. The drug transporter p-glycoprotein limits oral absorption and brain entry of hiv-1 protease inhibitors. *J. Clin. Invest.* **1998**, *101*, 289-294.
4. Cordon-Cardo, C.; O'Brien, J.P.; Casals, D.; Rittman-Grauer, L.; Biedler, J.L.; Melamed, M.R.; Bertino, J.R. Multidrug-resistance gene (p-glycoprotein) is expressed by endothelial cells at blood-brain barrier sites. *Proc. Natl. Acad. Sci. U. S. A.* **1989**, *86*, 695-698.
5. Schinkel, A.H. P-glycoprotein, a gatekeeper in the blood-brain barrier. *Adv. Drug Deliv. Rev.* **1999**, *36*, 179-194.
6. Vähäkangas, K.; Myllynen, P. Drug transporters in the human blood-placental barrier. *Br. J. Pharmacol.* **2009**, *158*, 665-678.
7. Gosselet, F.; Saint-Pol, J.; Candela, P.; Fenart, L. Amyloid-β peptides, alzheimer's disease and the blood-brain barrier. *Curr. Alzheimer Res.* **2013**, *10*, 1015-1033.
8. Mawuenyega, K.G.; Sigurdson, W.; Ovod, V.; Munsell, L.; Kasten, T.; Morris, J.C.; Yarasheski, K.E.; Bateman, R.J. Decreased clearance of cns β-amyloid in alzheimer's disease. *Science* **2010**, *330*, 1774.
9. van Assema, D.M.E.; Lubberink, M.; Bauer, M.; van der Flier, W.M.; Schuit, R.C.; Windhorst, A.D.; Comans, E.F.I.; Hoetjes, N.J.; Tolboom, N.; Langer, O.*, et al.* Blood–brain barrier p-glycoprotein function in alzheimer's disease. *Brain* **2012**, *135*, 181-189.
10. Jedlitschky, G.; Vogelgesang, S.; Kroemer, H.K. Mdr1-p-glycoprotein (abcb1)-mediated disposition of amyloid-β peptides: Implications for the pathogenesis and therapy of alzheimer's disease. *Clin. Pharmacol. Ther.* **2010**, *88*, 441-443.
11. Cascorbi, I.; Flüh, C.; Remmler, C.; Haenisch, S.; Faltraco, F.; Grumbt, M.; Peters, M.; Brenn, A.; Thal, D.R.; Warzok, R.W.*, et al.* Association of atp-binding cassette transporter variants with the risk of alzheimer's disease. *Pharmacogenomics* **2013**, *14*, 485-494.
12. Brenn, A.; Grube, M.; Peters, M.; Fischer, A.; Jedlitschky, G.; Kroemer, H.K.; Warzok, R.W.; Vogelgesang, S. Beta-amyloid downregulates mdr1-p-glycoprotein (abcb1) expression at the blood-brain barrier in mice. *Int. J. Alzheimers Dis.* **2011**, *2011*, Article ID 690121.
13. Neuwelt, E.A.; Bauer, B.; Fahlke, C.; Fricker, G.; Iadecola, C.; Janigro, D.; Leybaert, L.; Molnár, Z.; O'Donnell, M.E.; Povlishock, J.T.*, et al.* Engaging neuroscience to advance translational research in brain barrier biology. *Nat. Rev. Neurosci.* **2011**, *12*, 169-182.
14. Wolf, A.; Bauer, B.; Hartz, A. Abc transporters and the alzheimer's disease enigma. *Front. Psychiatry* **2012**, *3*.
15. Selick, H.E.; Beresford, A.P.; Tarbit, M.H. The emerging importance of predictive adme simulation in drug discovery. *Drug Discov. Today* **2002**, *7*, 109-116.
16. Montanari, F.; Ecker, G.F. Prediction of drug–abc-transporter interaction — recent advances and future challenges. *Adv. Drug Deliv. Rev.* **2015**, *86*, 17-26.

17. Greiner, B.; Eichelbaum, M.; Fritz, P.; Kreichgauer, H.P.; von Richter, O.; Zundler, J.; Kroemer, H.K. The role of intestinal p-glycoprotein in the interaction of digoxin and rifampin. *J. Clin. invest.* **1999**, *104*, 147-153.
18. Padowski, J.M.; Pollack, G.M. Influence of time to achieve substrate distribution equilibrium between brain tissue and blood on quantitation of the blood–brain barrier p-glycoprotein effect. *Brain Res.* **2011**, *1426*, 1-17.
19. Bagal, S.; Bungay, P. Restricting cns penetration of drugs to minimise adverse events: Role of drug transporters. *Drug Discov. Today Technol.* **2014**, *12*, e79-e85.
20. Hochman, J.H.; Ha, S.N.; Sheridan, R.P. Establishment of p-glycoprotein structure–transport relationships to optimize cns exposure in drug discovery. In *Blood-brain barrier in drug discovery: Optimizing brain exposure of cns drugs and minimizing brain side effects for peripheral drugs*, Di, L.; Kerns, E.H., Eds. John Wiley & Sons, Inc: Hoboken, NJ, 2015; pp 113-124.
21. Schinkel, A.H.; Wagenaar, E.; Mol, C.A.; van Deemter, L. P-glycoprotein in the blood-brain barrier of mice influences the brain penetration and pharmacological activity of many drugs. *J. Clin. Invest.* **1996**, *97*, 2517-2524.
22. Aszalos, A. Drug–drug interactions affected by the transporter protein, p-glycoprotein (abcb1, mdr1): Ii. Clinical aspects. *Drug Discov. Today* **2007**, *12*, 838-843.
23. Doan, K.M.M.; Humphreys, J.E.; Webster, L.O.; Wring, S.A.; Shampine, L.J.; Serabjit-Singh, C.J.; Adkison, K.K.; Polli, J.W. Passive permeability and p-glycoprotein-mediated efflux differentiate central nervous system (cns) and non-cns marketed drugs. *J. Pharmacol. Exp. Ther.* **2002**, *303*, 1029-1037.
24. Hennessy, M.; Spiers, J.P. A primer on the mechanics of p-glycoprotein the multidrug transporter. *Pharmacol. Res.* **2007**, *55*, 1-15.
25. Gottesman, M.M.; Pastan, I. Biochemistry of multidrug resistance mediated by the multidrug transporter. *Ann. Rev. Biochem.* **2003**, *62*, 385-427.
26. Breier, A.; Gibalova, L.; Seres, M.; Barancik, M.; Sulova, Z. New insight into p-glycoprotein as a drug target. *Anticancer Agents Med. Chem.* **2013**, *13*, 159-170.
27. Ambudkar, S.V.; Dey, S.; Hrycyna, C.A.; Ramachandra, M.; Pastan, I.; Gottesman, M.M. Biochemical, cellular, and pharmacological aspects of the multidrug transporter. *Annu. Rev. Pharmacol. Toxicol.* **1999**, *39*, 361-398.
28. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2017. *CA Cancer J. Clin.* **2017**, *67*, 7-30.
29. Clarke, J.; Penas, C.; Pastori, C.; Komotar, R.J.; Bregy, A.; Shah, A.H.; Wahlestedt, C.; Ayad, N.G. Epigenetic pathways and glioblastoma treatment. *Epigenetics* **2013**, *8*, 785-795.
30. Wang, T.; Agarwal, S.; Elmquist, W.F. Brain distribution of cediranib is limited by active efflux at the blood-brain barrier. *J. Pharmacol. Exp. Ther.* **2012**, *341*, 386-395.
31. Palmeira, A.; Sousa, E.; H. Vasconcelos, M.; M. Pinto, M. Three decades of p-gp inhibitors: Skimming through several generations and scaffolds. *Curr. Med. Chem.* **2012**, *19*, 1946-2025.
32. van Hoppe, S.; Schinkel, A.H. What next? Preferably development of drugs that are no longer transported by the abcb1 and abcg2 efflux transporters. *Pharmacol. Res.* **2017**.
33. Crivori, P. Computational models for p-glycoprotein substrates and inhibitors. In *Antitargets: Prediction and prevention of drug side effects*, Vaz, R.J.; Klabunde, T., Eds.

Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2008; Vol. 38, pp 367-397.

34. Terasaki, T.; Hosoya, K.-i. The blood-brain barrier efflux transporters as a detoxifying system for the brain. *Adv. Drug Deliv. Rev.* **1999**, *36*, 195-209.
35. Garg, P.; Verma, J. In silico prediction of blood brain barrier permeability: An artificial neural network model. *J. Chem. Inf. Model.* **2006**, *46*, 289-297.
36. Kalvass, J.C.; Maurer, T.S.; Pollack, G.M. Use of plasma and brain unbound fractions to assess the extent of brain distribution of 34 drugs: Comparison of unbound concentration ratios to in vivo p-glycoprotein efflux ratios. *Drug Metab. Dispos.* **2007**, *35*, 660-666.
37. Di, L.; Rong, H.; Feng, B. Demystifying brain penetration in central nervous system drug discovery. *J. Med. Chem.* **2013**, *56*, 2-12.
38. Inoue, T.; Osada, K.; Tagawa, M.; Ogawa, Y.; Haga, T.; Sogame, Y.; Hashizume, T.; Watanabe, T.; Taguchi, A.; Katsumata, T., *et al.* Blonanserin, a novel atypical antipsychotic agent not actively transported as substrate by p-glycoprotein. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **2012**, *39*, 156-162.
39. Polli, J.W.; Wring, S.A.; Humphreys, J.E.; Huang, L.; Morgan, J.B.; Webster, L.O.; Serabjit-Singh, C.S. Rational use of in vitro p-glycoprotein assays in drug discovery. *J. Pharmacol. Exp. Ther.* **2001**, *299*, 620-628.
40. Hochman, J.H.; Yamazaki, M.; Ohe, T.; Lin, J.H. Evaluation of drug interactions with p-glycoprotein in drug discovery: In vitro assessment of the potential for drug-drug interactions with p-glycoprotein. *Curr. Drug Metab.* **2002**, *3*, 257-273.
41. Schwab, D.; Fischer, H.; Tabatabaei, A.; Poli, S.; Huwyler, J. Comparison of in vitro p-glycoprotein screening assays: Recommendations for their use in drug discovery. *J. Med. Chem.* **2003**, *46*, 1716-1725.
42. Zhang, Y.; Bachmeier, C.; Miller, D.W. In vitro and in vivo models for assessing drug efflux transporter activity. *Adv. Drug Deliv. Rev.* **2003**, *55*, 31-51.
43. Sugano, K.; Shirasaka, Y.; Yamashita, S. Estimation of michaelis–menten constant of efflux transporter considering asymmetric permeability. *Int. J. Pharm.* **2011**, *418*, 161-167.
44. Storch, C.H.; Nikendei, C.; Schild, S.; Haefeli, W.E.; Weiss, J.; Herzog, W. Expression and activity of p-glycoprotein (mdr1/abcb1) in peripheral blood mononuclear cells from patients with anorexia nervosa compared with healthy controls. *Int. J. Eating Disord.* **2008**, *41*, 432-438.
45. Egan, W.J. Computational models for adme. In *Annual reports in medicinal chemistry*, John, E.M., Ed. Academic Press: San Diego, CA, 2007; Vol. 42, pp 449-467.
46. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E.W. Computational methods in drug discovery. *Pharmacol. Rev.* **2014**, *66*, 334-395.
47. Penzotti, J.E.; Lamb, M.L.; Evensen, E.; Grootenhuis, P.D.J. A computational ensemble pharmacophore model for identifying substrates of p-glycoprotein. *J. Med. Chem.* **2002**, *45*, 1737-1740.
48. Gombar, V.K.; Polli, J.W.; Humphreys, J.E.; Wring, S.A.; Serabjit-Singh, C.S. Predicting p-glycoprotein substrates by a quantitative structure-activity relationship model. *J. Pharm. Sci.* **2004**, *93*, 957-968.
49. Xue, Y.; Yap, C.W.; Sun, L.Z.; Cao, Z.W.; Wang, J.F.; Chen, Y.Z. Prediction of p-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1497-1505.

50. Wang, Y.-H.; Li, Y.; Yang, S.-L.; Yang, L. Classification of substrates and inhibitors of p-glycoprotein using unsupervised machine learning approach. *J. Chem. Inf. Model.* **2005**, *45*, 750-757.
51. Crivori, P.; Reinach, B.; Pezzetta, D.; Poggesi, I. Computational models for identifying potential p-glycoprotein substrates and inhibitors. *Mol. Pharmaceutics* **2006**, *3*, 33-44.
52. de Cerqueira Lima, P.; Golbraikh, A.; Oloff, S.; Xiao, Y.; Tropsha, A. Combinatorial qsar modeling of p-glycoprotein substrates. *J. Chem. Inf. Model.* **2006**, *46*, 1245-1254.
53. Huang, J.; Ma, G.; Muhammad, I.; Cheng, Y. Identifying p-glycoprotein substrates using a support vector machine optimized by a particle swarm. *J. Chem. Inf. Model.* **2007**, *47*, 1638-1647.
54. Li, W.-X.; Li, L.; Eksterowicz, J.; Ling, X.B.; Cardozo, M. Significance analysis and multiple pharmacophore models for differentiating p-glycoprotein substrates. *J. Chem Inf. Model.* **2007**, *47*, 2429-2438.
55. Wang, Z.; Chen, Y.; Liang, H.; Bender, A.; Glen, R.C.; Yan, A. P-glycoprotein substrate models using support vector machines based on a comprehensive data set. *J. Chem. Inf. Model.* **2011**, *51*, 1447-1456.
56. Joung, J.Y.; Kim, H.; Kim, H.M.; Ahn, S.K.; Nam, K.-Y.; No*, K.T. Prediction models of p-glycoprotein substrates using simple 2d and 3d descriptors by a recursive partitioning approach. *Bull. Korean Chem. Soc.* **2012**, *33*, 1123-1127.
57. Desai, P.V.; Sawada, G.A.; Watson, I.A.; Raub, T.J. Integration of in silico and in vitro tools for scaffold optimization during drug discovery: Predicting p-glycoprotein efflux. *Mol. Pharmaceutics* **2013**, *10*, 1249-1261.
58. Ecker, G.F.; Stockner, T.; Chiba, P. Computational models for prediction of interactions with abc-transporters. *Drug Discov. Today* **2008**, *13*, 311-317.
59. Adenot, M. A practical approach to computational models of the blood–brain barrier. In *Handbook of neurochemistry and molecular neurobiology: Neural membranes and transport*, Lajtha, A.; Reith, M.E.A., Eds. Springer: New York, 2007; pp 109-150.
60. Ivanciuc, O. Artificial immune systems in drug design: Recognition of p-glycoprotein substrates with airs (artificial immune recognition system). *Internet Electron. J. Mol. Des.* **2006**, *5*, 542-554.
61. Bikadi, Z.; Hazai, I.; Malik, D.; Jemnitz, K.; Veres, Z.; Hari, P.; Ni, Z.; Loo, T.W.; Clarke, D.M.; Hazai, E., *et al.* Predicting p-glycoprotein-mediated drug transport based on support vector machine and three-dimensional crystal structure of p-glycoprotein. *PLoS ONE* **2011**, *6*, e25815.
62. Erić, S.; Kalinić, M.; Ilić, K.; Zloh, M. Computational classification models for predicting the interaction of drugs with p-glycoprotein and breast cancer resistance protein. *SAR QSAR Environ. Res.* **2014**, *25*, 939-966.
63. Pan, X.; Mei, H.; Qu, S.; Huang, S.; Sun, J.; Yang, L.; Chen, H. Prediction and characterization of p-glycoprotein substrates potentially bound to different sites by emerging chemical pattern and hierarchical cluster analysis. *Int. J. Pharm.* **2016**, *502*, 61-69.
64. Didziapetris, R.; Japertas, P.; Avdeef, A.; Petrauskas, A. Classification analysis of p-glycoprotein substrate specificity. *J. Drug Target.* **2003**, *11*, 391-406.
65. Joung, J.-Y.; Kim, H.-J.; Kim, H.-M.; Ahn, S.-K.; Nam, K.-Y.; No, K.-T. Prediction models of p-glycoprotein substrates using simple 2d and 3d descriptors by a recursive partitioning approach. *Bull. Korean Chem. Soc.* **2012**, *33*, 1123-1127.

66. Broccatelli, F. Qsar models for p-glycoprotein transport based on a highly consistent data set. *J. Chem. Inf. Model.* **2012**, *2*, 2462-2470.
67. Poongavanam, V.; Haider, N.; Ecker, G.F. Fingerprint-based in silico models for the prediction of p-glycoprotein substrates and inhibitors. *Bioorg. Med. Chem.* **2012**, *20*, 5388-5395.
68. Li, D.; Chen, L.; Li, Y.; Tian, S.; Sun, H.; Hou, T. Admet evaluation in drug discovery. 13. Development of *in silico* prediction models for p-glycoprotein substrates. *Mol. Pharmaceutics* **2014**, *11*, 716-726.
69. Estrada, E.; Molina, E.; Nodarse, D.; Uriarte, E. Structural contributions of substrates to their binding to p-glycoprotein. A topsmode approach. *Curr. Pharm. Design* **2010**, *16*, 2676-2709.
70. Gunaydin, H.; Weiss, M.M.; Sun, Y. De novo prediction of p-glycoprotein-mediated efflux liability for druglike compounds. *ACS Med. Chem. Lett.* **2013**, *4*, 108-112.
71. Dolghih, E.; Jacobson, M.P. Predicting efflux ratios and blood-brain barrier penetration from chemical structure: Combining passive permeability with active efflux by p-glycoprotein. *ACS Chem. Neurosci.* **2012**, *4*, 361-367.
72. Dolghih, E.; Bryant, C.; Renslo, A.R.; Jacobson, M.P. Predicting binding to p-glycoprotein by flexible receptor docking. *PLoS Comput. Biol.* **2011**, *7*, e1002083.
73. Subramanian, N.; Condic-Jurkic, K.; O'Mara, M.L. Structural and dynamic perspectives on the promiscuous transport activity of p-glycoprotein. *Neurochem. Int.* **2016**, *98*, 146-152.
74. Leong, M.K.; Chen, H.-B.; Shih, Y.-H. Prediction of promiscuous p-glycoprotein inhibition using a novel machine learning scheme. *PLoS One* **2012**, *7*, e33829.
75. Garrigues, A.; Loiseau, N.; Delaforge, M.; Ferté, J.; Garrigos, M.; André, F.; Orlowski, S. Characterization of two pharmacophores on the multidrug transporter p-glycoprotein. *Mol. Pharmacol.* **2002**, *62*, 1288-1298.
76. Chufan, E.E.; Sim, H.-M.; Ambudkar, S.V. Molecular basis of the polyspecificity of p-glycoprotein (abcb1): Recent biochemical and structural studies. In *Advances in cancer research: Abc transporters and cancer*, John, D.S.; Toshihisa, I., Eds. Academic Press: 2015; Vol. 125, pp 71-96.
77. Ferreira, R.J.; Ferreira, M.-J.U.; dos Santos, D.J.V.A. Molecular docking characterizes substrate-binding sites and efflux modulation mechanisms within p-glycoprotein. *J. Chem. Inf. Model.* **2013**, *53*, 1747-1760.
78. Aller, S.G.; Yu, J.; Ward, A.; Weng, Y.; Chittaboina, S.; Zhuo, R.; Harrell, P.M.; Trinh, Y.T.; Zhang, Q.; Urbatsch, I.L.*, et al.* Structure of p-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science* **2009**, *323*, 1718-1722.
79. Edwards, G. Ivermectin: Does p-glycoprotein play a role in neurotoxicity? *Filaria J.* **2003**, *24*.
80. Balimane, P.V.; Han, Y.-H.; Chong, S. Current industrial practices of assessing permeability and p-glycoprotein interaction. *AAPS J.* **2006**, *8*, E1-E13.
81. Roger, P.; Sahla, M.E.; Mäkelä, S.; Gustafsson, J.Å.; Baldet, P.; Rochefort, H. Decreased expression of estrogen receptor β protein in proliferative preinvasive mammary tumors. *Cancer Res.* **2001**, *61*, 2537-2541.
82. Speck-Planche, A.; Cordeiro, M.N.D.S. Multi-target qsar approaches for modeling protein inhibitors. Simultaneous prediction of activities against biomacromolecules present in gram-negative bacteria. *Curr. Top. Med. Chem.* **2015**, *15*, 1801-1813.

83. Ferreira, R.J.; dos Santos, D.J.V.A.; Ferreira, M.-J.U.; Guedes, R.C. Toward a better pharmacophore description of p-glycoprotein modulators, based on macrocyclic diterpenes from euphorbia species. *J. Chem. Inf. Model.* **2011**, *51*, 1315-1324.
84. Leong, M.K.; Chen, Y.-M.; Chen, T.-H. Prediction of human cytochrome p450 2b6-substrate interactions using hierarchical support vector regression approach. *J. Comput. Chem.* **2009**, *30*, 1899-1909.
85. Caudill, M. Using neural networks: Hybrid expert networks. *AI Expert* **1990**, *5*, 49-54.
86. Leong, M.K.; Lin, S.-W.; Chen, H.-B.; Tsai, F.-Y. Predicting mutagenicity of aromatic amines by various machine learning approaches. *Toxicol. Sci.* **2010**, *116*, 498-513.
87. Ding, Y.-L.; Lyu, Y.-C.; Leong, M.K. *In silico* prediction of the mutagenicity of nitroaromatic compounds using a novel two-qsar approach. *Toxicol. Vitro* **2017**, *40*, 102-114.
88. Gnanadesikan, R.; Kettenring, J.R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **1972**, *28*, 81-124.
89. Cherkasov, A.; Muratov, E.N.; Fourches, D.; Varnek, A.; Baskin, I.I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y.C.; Todeschini, R., *et al.* Qsar modeling: Where have you been? Where are you going to? *J. Med. Chem.* **2014**, *57*, 4977-5010.
90. Cammi, R.; Tomasi, J. Remarks on the use of the apparent surface charges (asc) methods in solvation problems: Iterative versus matrix-inversion procedures and the renormalization of the apparent charges. *J. Comput. Chem.* **1995**, *16*, 1449-1458.
91. Miertuš, S.; Scrocco, E.; Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilizaion of ab initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* **1981**, *55*, 117-129.
92. Besler, B.H.; Merz, K.M.J.; Kollman, P.A. Atomic charges derived from semiempirical methods. *J. Comput. Chem.* **1990**, *11*, 431-439.
93. Topliss, J.G.; Edwards, R.P. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **1979**, *22*, 1238-1244.
94. Kettaneh, N.; Berglund, A.; Wold, S. Pca and pls with very large data sets. *Comput. Stat. Data Anal.* **2005**, *48*, 69-85.
95. Tseng, Y.J.; Hopfinger, A.J.; Esposito, E.X. The great descriptor melting pot: Mixing descriptors for the common good of qsar models. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 39-43.
96. Burden, F.R.; Ford, M.G.; Whitley, D.C.; Winkler, D.A. Use of automatic relevance determination in qsar studies using bayesian neural networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1423-1430.
97. Rogers, D.; Hopfinger, A.J. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854-866.
98. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389-422.
99. Kennard, R.W.; Stone, L.A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137-148.
100. Tropsha, A.; Gramatica, P.; Gombar, Vijay K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of qspr models. *QSAR Comb. Sci.* **2003**, *22*, 69-77.

101. Tropsha, A. Recent trends in statistical qsar modeling of environmental chemical toxicity. In *Molecular, clinical and environmental toxicology: Volume 3: Environmental toxicology*, Luch, A., Ed. Springer Basel: 2012; Vol. 101, pp 381-411.
102. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273-297.
103. Vapnik, V.; Golowich, S.; Smola, A. In *Support vector method for function approximation, regression estimation, and signal processing*, Advances in Neural Information Processing Systems 9, 1997; Mozer, M.; Jordan, M.I.; Petsche, T., Eds. MIT Press, Cambridge, MA, USA: pp 281-287.
104. Schölkopf, B.; Smola, A. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. 1st ed.; MIT Press: Cambridge, MA, 2002.
105. Netzeva, T.I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M.T.D.; Gramatica, P.; Jaworska, J.S.; Kahn, S.; Klopman, G.; Marchant, C.A., *et al.* Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships : The report and recommendations of ecvam workshop 52. *Altern. Lab. Anim.* **2005**, *33*, 1-19.
106. Kecman, V. *Learning and soft computing : Support vector machines, neural networks, and fuzzy logic models*. MIT Press: 2001; p 576p.
107. Leong, M.K.; Syu, R.-G.; Ding, Y.-L.; Weng, C.-F. Prediction of *n*-methyl-d-aspartate receptor glun1-ligand binding affinity by a novel svm-pose/svm-score combinatorial ensemble docking scheme. *Sci. Rep.* **2017**, *7*, 40053.
108. Dearden, J.C.; Cronin, M.T.D.; Kaiser, K.L.E. How not to develop a quantitative structure–activity or structure–property relationship (qsar/qspr). *SAR QSAR Environ. Res.* **2009**, *20*, 241-266.
109. Breiman, L.; Spector, P. Submodel selection and evaluation in regression. The x-random case. *Int. Stat. Rev.* **1992**, *60*, 291-319.
110. Rücker, C.; Rücker, G.; Meringer, M. Y-randomization and its variants in qspr/qsar. *J. Chem. Inf. Model.* **2007**, *47*, 2345-2357.
111. Ojha, P.K.; Mitra, I.; Das, R.N.; Roy, K. Further exploring r_m^2 metrics for validation of qspr models. *Chemometrics Intell. Lab. Syst.* **2011**, *107*, 194-205.
112. Shi, L.M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R.M.; Branham, W.S.; Dial, S.L.; Moland, C.L.; Sheehan, D.M. Qsar models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186-195.
113. Schüürmann, G.; Ebert, R.-U.; Chen, J.; Wang, B.; Kühne, R. External validation and prediction employing the predictive squared correlation coefficient-test set activity mean vs training set activity mean. *J. Chem. Inf. Model.* **2008**, *48*, 2140-2145.
114. Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the q^2 parameter for qsar validation. *J. Chem. Inf. Model.* **2009**, *49*, 1669-1678.
115. Chirico, N.; Gramatica, P. Real external predictivity of qsar models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J. Chem. Inf. Model.* **2011**, *51*, 2320-2335.
116. Gramatica, P.; Chirico, N.; Papa, E.; Cassani, S.; Kovarich, S. Qsarins: A new software for the development, analysis, and validation of qsar mlr models. *J. Comput. Chem.* **2013**, *34*, 2121-2132.
117. Gramatica, P.; Cassani, S.; Chirico, N. Qsarins-chem: Insubria datasets and new qsar/qspr models for environmental pollutants in qsarins. *J. Comput. Chem.* **2014**, *35*, 1036-1044.

118. Gramatica, P.; Sangion, A. A historical excursus on the statistical validation parameters for qsar models: A clarification concerning metrics and terminology. *J. Chem. Inf. Model.* **2016**, *56*, 1127-1131.
119. Roy, K.; Mitra, I.; Kar, S.; Ojha, P.K.; Das, R.N.; Kabir, H. Comparative studies on some metrics for external validation of qspr models. *J. Chem. Inf. Model.* **2012**, *52*, 396-408.
120. Chirico, N.; Gramatica, P. Real external predictivity of qsar models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *J. Chem. Inf. Model.* **2012**, *52*, 2044-2058.
121. Golbraikh, A.; Shen, M.; Xiao, Z.Y.; Xiao, Y.D.; Lee, K.H.; Tropsha, A. Rational selection of training and test sets for the development of validated qsar models. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241-253.
122. Carrara, S.; Reali, V.; Misiano, P.; Dondio, G.; Bigogno, C. Evaluation of *in vitro* brain penetration: Optimized pampa and mdckii-mdr1 assay comparison. *Int. J. Pharm.* **2007**, *345*, 125-133.
123. Chen, C.; Hanson, E.; Watson, J.W.; Lee, J.S. P-glycoprotein limits the brain penetration of nonsedating but not sedating h1-antagonists. *Drug Metab. Dispos.* **2003**, *31*, 312-318.
124. Eriksson, U.G.; Dorani, H.; Karlsson, J.; Fritsch, H.; Hoffmann, K.-J.; Olsson, L.; Sarich, T.C.; Wall, U.; Schützer, K.-M. Influence of erythromycin on the pharmacokinetics of ximelagatran may involve inhibition of p-glycoprotein-mediated excretion. *Drug Metab. Dispos.* **2006**, *34*, 775-782.
125. Feng, B.; Mills, J.B.; Davidson, R.E.; Mireles, R.J.; Janiszewski, J.S.; Troutman, M.D.; de Morais, S.M. In vitro p-glycoprotein assays to predict the in vivo interactions of p-glycoprotein with drugs in the central nervous system. *Drug Metab. Dispos.* **2008**, *36*, 268-275.
126. Gertz, M.; Harrison, A.; Houston, J.B.; Galetin, A. Prediction of human intestinal first-pass metabolism of 25 cyp3a substrates from in vitro clearance and permeability data. *Drug Metab. Dispos.* **2010**, *38*, 1147-1158.
127. Huang, L.; Wang, Y.; Grimm, S. Atp-dependent transport of rosuvastatin in membrane vesicles expressing breast cancer resistance protein. *Drug Metab. Dispos.* **2006**, *34*, 738-742.
128. Luo, S.; Pal, D.; Shah, S.J.; Kwatra, D.; Paturi, K.D.; Mitra, A.K. Effect of hepes buffer on the uptake and transport of p-glycoprotein substrates and large neutral amino acids. *Mol. Pharmaceutics* **2010**, *7*, 412-420.
129. Mahar, D.K.M.; Humphreys, J.E.; Webster, L.O.; Wring, S.A.; Shampine, L.J.; Serabjit-Singh, C.J.; Adkison, K.K.; Polli, J.W. Passive permeability and p-glycoprotein-mediated efflux differentiate central nervous system (cns) and non-cns marketed drugs. *J. Pharmacol. Exp. Ther.* **2002**, *303*, 1029-1037.
130. Taub, M.E.; Podila, L.; Ely, D.; Almeida, I. Functional assessment of multiple p-glycoprotein (p-gp) probe substrates: Influence of cell line and modulator concentration on p-gp activity. *Drug Metab. Dispos.* **2005**, *33*, 1679-1687.
131. Troutman, M.D.; Thakker, D.R. Novel experimental parameters to quantify the modulation of absorptive and secretory transport of compounds by p-glycoprotein in cell culture models of intestinal epithelium. *Pharm. Res.* **2003**, *20*, 1210-1224.
132. Wager, T.T.; Chandrasekaran, R.Y.; Hou, X.; Troutman, M.D.; Verhoest, P.R.; Villalobos, A.; Will, Y. Defining desirable central nervous system drug space through the alignment

of molecular properties, in vitro adme, and safety attributes. *ACS Chem. Neurosci.* **2010**, *1*, 420-434.

133. Callegari, E.; Malhotra, B.; Bungay, P.J.; Webster, R.; Fenner, K.S.; Kempshall, S.; LaPerle, J.L.; Michel, M.C.; Kay, G.G. A comprehensive non-clinical evaluation of the cns penetration potential of antimuscarinic agents for the treatment of overactive bladder. *Br. J. Clin. Pharmacol.* **2011**, *72*, 235-246.
134. Obradovic, T.; Dobson, G.; Shingaki, T.; Kungu, T.; Hidalgo, I. Assessment of the first and second generation antihistamines brain penetration and role of p-glycoprotein. *Pharm. Res.* **2007**, *24*, 318-327.
135. Liu, Q.; Wang, C.; Meng, Q.; Huo, X.; Sun, H.; Peng, J.; Ma, X.; Sun, P.; Liu, K. Mdr1 and oat1/oat3 mediate the drug-drug interaction between puerarin and methotrexate. *Pharm. Res.* **2014**, *31*, 1120-1132.
136. Kim, W.Y.; Benet, L.Z. P-glycoprotein (p-gp/*mdr1*)-mediated efflux of sex-steroid hormones and modulation of p-gp expression *in vitro*. *Pharm. Res.* **2004**, *21*, 1284-1293.
137. McIver, Z.A.; Kryman, M.W.; Choi, Y.; Coe, B.N.; Schamerhorn, G.A.; Linder, M.K.; Davies, K.S.; Hill, J.E.; Sawada, G.A.; Grayson, J.M., *et al.* Selective photodepletion of malignant t cells in extracorporeal photopheresis with selenorhodamine photosensitizers. *Bioorg. Med. Chem.* **2016**, *24*, 3918-3931.
138. Lee, W.; Crawford, J.J.; Aliagas, I.; Murray, L.J.; Tay, S.; Wang, W.; Heise, C.E.; Hoeflich, K.P.; La, H.; Mathieu, S., *et al.* Synthesis and evaluation of a series of 4-azaindole-containing p21-activated kinase-1 inhibitors. *Bioorg. Med. Chem. Lett.* **2016**, *26*, 3518-3524.
139. Hitchcock, S.A. Structural modifications that alter the p-glycoprotein efflux properties of compounds. *J. Med. Chem.* **2012**, *55*, 4877-4895.
140. Desai, P.V.; Raub, T.J.; Blanco, M.-J. How hydrogen bonds impact p-glycoprotein transport and permeability. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 6540-6548.
141. Teixeira, V.H.; Vila-Viçosa, D.; Baptista, A.M.; Machuqueiro, M. Protonation of dmpc in a bilayer environment using a linear response approximation. *J. Chem. Theory Comput.* **2014**, *10*, 2176-2184.
142. Clay, A.T.; Sharom, F.J. Lipid bilayer properties control membrane partitioning, binding, and transport of p-glycoprotein substrates. *Biochemistry* **2013**, *52*, 343-354.
143. Raub, T.J. P-glycoprotein recognition of substrates and circumvention through rational drug design. *Mol. Pharmaceutics* **2006**, *3*, 3-25.
144. Suzuki, T.; Fukazawa, N.; San-nohe, K.; Sato, W.; Yano, O.; Tsuruo, T. Structure-activity relationship of newly synthesized quinoline derivatives for reversal of multidrug resistance in cancer. *J. Med. Chem.* **1997**, *40*, 2047-2052.
145. Prasanna, S.; Doerksen, R.J. Topological polar surface area: A useful descriptor in 2d-qsar. *Curr. Med. Chem.* **2009**, *16*, 21-41.
146. Ferté, J. Analysis of the tangled relationships between p-glycoprotein-mediated multidrug resistance and the lipid phase of the cell membrane. *Eur. J. Biochem.* **2000**, *267*, 277-294.
147. Johnson, T.W.; Dress, K.R.; Edwards, M. Using the golden triangle to optimize clearance and oral absorption. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 5560-5564.
148. Pettersson, M.; Hou, X.; Kuhn, M.; Wager, T.T.; Kauffman, G.W.; Verhoest, P.R. Quantitative assessment of the impact of fluorine substitution on p-glycoprotein (p-gp) mediated efflux, permeability, lipophilicity, and metabolic stability. *J. Med. Chem.* **2016**, *59*, 5284-5296.

149. Jabeen, I.; Wetwitayaklung, P.; Klepsch, F.; Parveen, Z.; Chiba, P.; Ecker, G.F. Probing the stereoselectivity of p-glycoprotein-synthesis, biological activity and ligand docking studies of a set of enantiopure benzopyrano[3,4-*b*][1,4]oxazines. *Chem. Commun*. **2011**, *47*, 2586-2588.
150. Rezai, T.; Bock, J.E.; Zhou, M.V.; Kalyanaraman, C.; Lokey, R.S.; Jacobson, M.P. Conformational flexibility, internal hydrogen bonding, and passive membrane permeability: Successful in silico prediction of the relative permeabilities of cyclic peptides. *J. Am. Chem. Soc*. **2006**, *128*, 14073-14080.
151. Rauch, C.; Paine, S.W.; Littlewood, P. Can long range mechanical interaction between drugs and membrane proteins define the notion of molecular promiscuity? Application to p-glycoprotein-mediated multidrug resistance (mdr). *Biochim. Biophys. Acta-Gen. Subj.* **2013**, *1830*, 5112-5118.
152. Declèves, X.; Jacob, A.; Yousif, S.; Shawahna, R.; Potin, S.; Scherrmann, J.-M. Interplay of drug metabolizing cyp450 enzymes and abc transporters in the blood-brain barrier. *Curr. Drug Metab*. **2011**, *12*, 732-741.
153. Passeleu-Le Bourdonnec, C.; Carrupt, P.-A.; Scherrmann, J.; Martel, S. Methodologies to assess drug permeation through the blood–brain barrier for pharmaceutical research. *Pharm. Res*. **2013**, *30*, 2729-2756.
154. Leong, M.K. *In silico* prediction of the blood-brain barrier permeation: Are we there yet? *Med. Chem*. **2015**, *5*, 130.