pharmaceutics

MDPI

Article

Development of a Hierarchical Support Vector Regression-Based In Silico Model for Caco-2 Permeability

Giang Huong Ta [1], Cin-Syong Jhang [1], Ching-Feng Weng [2] and Max K. Leong [1,*]

[1] Department of Chemistry, National Dong Hwa University, Shoufeng, Hualien 974301, Taiwan; 810812203@gms.ndhu.edu.tw (G.H.T.); 610512002@gms.ndhu.edu.tw (C.-S.J.)

[2] Department of Physiology, School of Basic Medical Science, Xiamen Medical College, Xiamen 361023, China; cfweng@gms.ndhu.edu.tw

* Correspondence: leong@gms.ndhu.edu.tw; Tel.: +886-3-890-3609

Abstract: Drug absorption is one of the critical factors that should be taken into account in the process of drug discovery and development. The human colon carcinoma cell layer (Caco-2) model has been frequently used as a surrogate to preliminarily investigate the intestinal absorption. In this study, a quantitative structure–activity relationship (QSAR) model was generated using the innovative machine learning-based hierarchical support vector regression (HSVR) scheme to depict the exceedingly confounding passive diffusion and transporter-mediated active transport. The HSVR model displayed good agreement with the experimental values of the training samples, test samples, and outlier samples. The predictivity of HSVR was further validated by a mock test and verified by various stringent statistical criteria. Consequently, this HSVR model can be employed to forecast the Caco-2 permeability to assist drug discovery and development.

Keywords: intestinal absorption; intestinal permeability; human colon carcinoma cell layer (Caco-2); hierarchical support vector regression (HSVR)

check for updates

Citation: Ta, G.H.; Jhang, C.-S.; Weng, C.-F.; Leong, M.K Development of a Hierarchical Support Vector Regression-Based In Silico Model for Caco-2 Permeability. *Pharmaceutics* **2021**, *13*, 174. https://doi.org/10.3390/pharmaceutics13020174

Academic Editor: Maria Isabel Gonzalez-Alvarez

Received: 27 December 2020

Accepted: 21 January 2021

Published: 28 January 2021

1. Introduction

Clinically, the majority of drugs are orally administered [1]. Prior to reaching the blood circulation system, the administered pharmaceutical agents have to pass through the intestinal barrier via passive diffusion, active uptake, and/or efflux transport processes [2–4], as illustrated by Figure 10.2 of Proctor et al. [2]. In passive diffusion, drug molecules can permeate the epithelial cell layers through the transcellular pathway, in which they penetrate through the cell membrane, or the paracellular pathway, in which they can cross the epithelial cell layer through the tight junction between cells [5]. The significance of active transporters on intestinal absorption has been detailed elsewhere [6]. Principally, active transport can be modulated by the efflux transporters of the ATP-binding cassette (ABC) family as well the influx transporters of the solute carrier (SLC) family [6], of which the efflux transporters can pump the administrated drugs out of enterocytes, leading to the reduction of the accumulated concentration, whereas the influx can enhance the intestinal uptake, resulting in the increased drug accumulation [7]. Of various active influx and efflux transporters, P-glycoprotein (P-gp), also termed multidrug resistance 1 protein (MDR1/encoded by *ABCB1* gene), breast cancer resistance protein (BCRP/*ABCG2*), organic anion transporting polypeptide 2B1 (OATP2B1/*SLCO2B1*), and peptide transporter 1 (PEPT1/*SLC15A1*) play predominant roles in intestinal absorption [8].

Passive diffusion depends on a number of physicochemical properties, whereas active transport relies on the characteristics of specific binding sites on the transport proteins [9]. The uncharged and modest hydrophobic drugs such as testosterone [10] can permeate through the membrane. Conversely, it is very difficult for highly hydrophobic molecules to get across cells, since they can be adhered to the membrane [5]. On the other hand, hydrophilic drugs such as mannitol predominantly pass through the paracellular pathway [10].

Of various drug absorption, distribution, metabolism, elimination, and toxicity (ADME/Tox) properties, drug absorption plays a pivotal role in drug discovery, since they substantially contribute to the earlier preclinical go/no-go decisions for the drug candidates [10,11] to achieve the "fail fast, fail early" paradigm [12]. As such, numerous in vivo and in situ assays have been developed to evaluate the intestinal absorption [13,14]. For instance, the in situ single-pass intestinal perfusion (SPIP) model measures the appearance of the drug in plasma after intravenous and intraintestinal drug administration [13,15]. The drug is orally administrated or directly given into the intestine or stomach in some animal species in in vivo assay [13,14,16].

In addition to in vivo and in situ assays, various in vitro assays have been devised, since they have more advantages such as low cost and time efficiency as compared with their in situ and in vivo counterparts [15]. Of various in vitro assays to evaluate intestinal absorption, human colon carcinoma monolayer cells (Caco-2) [3], parallel artificial membrane permeability (PAMPA) [17,18], and Madin–Darby canine kidney cells (MDCK) [19] are most frequently used. In fact, a comprehensive drug absorption profile should include the Caco-2, MDCK, and PAMPA permeability data to explore drug solubility and bioavailability [20]. Moreover, Caco-2, which can be adopted to evaluate the drug permeability through the cytoplasm (transcellular uptake) or between cells (paracellular uptake) and active transport [6], has become the golden standard for predicting intestinal drug permeability and absorption because of its similarity in morphology and function with human enterocytes [21–23]. The Caco-2 protocol has been clearly described in detail by Hubatsch et al. As compared with the biological membrane, the Caco-2 system still suffers from a range of disadvantages such as high technical complexity, the limitations related to the differences between cell monolayers and intestinal membrane structurally and functionally [24], in addition to its long culture periods (21-24 days) with the significantly extensive costs, contributing to the major concerns in practical applications [21,25].

The Caco-2 permeability is normally expressed by the apparent permeability coefficient (P_{app}), in which the drug solution is added to the apical side, viz. the donor compartment, and the P_{app} value in the basolateral side, viz. the receiver compartment, is measured [23]

$$P_{\text{app}} = \frac{dQ}{dt} \times \frac{1}{(A \times C_0)} \quad (1)$$

where dQ/dt is the linear appearance rate of mass in the receiver solution transported during sink conditions, A is the membrane surface area, and C_0 is the initial concentration at the donor compartment [26]. However, it is not uncommon to observe in vitro permeability variations among different from research groups, because the cultured cells can vary based on culture conditions, passage number, monolayer age, seeding density, and stage of differentiation [27,28], as exemplified by those compounds listed in Table 3 of Lee et al. [29]. Furthermore, Yamashita et al. have found that the different pH values of apical medium and the different solvents can produce different drug absorption values [30]. For instance, the P_{app} values of alprenolol are $(6.06 \pm 0.18) \times 10^{-6}$ cm/s and $(30.0 \pm 1.8) \times 10^{-6}$ cm/s at pH 6.0 and pH 7.4, respectively. More examples of P_{app} variations at different pH values can be found in Table 1 of Yamashita et al. [30].

In silico technologies have become an essential component in drug discovery and development according to the fact that they can provide guidance in the early stages in the drug discovery process such as the activity classification (high/moderate/poor) or quantitative predictions [31,32]. As such, a great number of in silico models have been established to predict the ADME/Tox properties [33]. The relationship between biological activity and chemical characteristics can be established by quantitative structure–activity or structure–property relationships (QSAR and QSPR) [34]. Numerous QSAR models have been generated to predict Caco-2 permeability based on a variety of physicochemical and physiological descriptors [35–51]. Nevertheless, the difficulties in developing sound in silico models to predict the intestinal permeability still remain unanswered mainly due to the fact that Caco-2 permeability is a dramatically perplexing process that can take place through numerous non-linear routes (vide supra).

More specifically, the ABC transporters, which are efflux transporters, can reduce the drug absorption, whereas the SLC transporters, which are influx transporters, can enhance the drug uptake, leading to the decrease and/or increase of drug absorption, respectively. In fact, such controversy can establish a paramount barrier in model development. For instance, the number of aromatic rings (n_{Ar}) can enhance the compound hydrophobicity [52] and facilitate the passive diffusion consequently. Conversely, n_{Ar} is also an important feature for P-gp substrate recognition and modulates the compound efflux correspondingly [53]. Thus, n_{Ar} can simultaneously affect the active efflux and passive diffusion.

It is exceedingly difficult, if not nearly impossible, to derive a robust in silico model, which can properly render the complex relationships between the selected descriptors and Caco-2 permeability. However, the hierarchical support vector regression (HSVR) scheme, which is an innovative machine learning-based scheme initially developed by Leong et al. [54], can properly address the complicated and varied dependencies of descriptors that, in turn, can be greatly contributed to its advantageous features of both a local model and a global model, namely wider coverage of applicability domain (AD) and a higher capability of prediction, respectively. When comparing with most theoretical models, which are vulnerable to the outliers that represent mathematic extrapolations, HSVR can still show consistent performance, as demonstrated elsewhere [1,54–57]. Herein, the objective of this study was to develop an in silico model based on the HSVR scheme to predict Caco-2 permeability in conjunction with previously published PAMPA permeability, intestinal absorption, and MDCK efflux in silico models [1,55,57] to facilitate drug discovery and development, since medicinal chemists can employ these models to predict the drug absorption of (virtual) hit compounds as well as drug metabolism and pharmacokinetics (DM/PK) scientists can adopt these models to prioritize the lead compounds.

2. Materials and Methods

2.1. Data Collection

The P_{app} values were collected from the various sources after a comprehensive literature search [22,23,58–66]. Assay systems were carefully scrutinized to ensure data consistency, since various assay conditions such as pH value and solvent system, for example, can affect the Caco-2 permeability [30]. Only P_{app} values, which were measured in the Hank's balanced salt solution (HBSS) buffer and 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) including ca. 1% dimethylsulfoxide (DMSO) at pH 7.4 were chosen in this study. The average P_{app} value was selected to warrant better consistency in case there was more than one P_{app} value for a given compound within a near range. Finally, 144 compounds were chosen in this study and their corresponding logarithm P_{app} values, simplified molecular input line entry system (SMILES) strings, Chemical Abstracts Service (CAS) registry numbers, and references to the literature are listed in Table S1.

2.2. Molecular Descriptors

The density functional theory (DFT), Becke 3-parameter Lee–Yang–Parr (B3LYP) method was employed to do full geometry optimization by the Gaussian package (Gaussian, Wallingford, CT, USA) for all recruited samples with the selection of basis set 6-31G (*d*,*p*). The solvent system was taken into consideration by the polarizable continuum model (PCM) [67,68]. The atomic charges, upon which the dipole moments depend, were calculated by the molecular electrostatic potential (MEP) [69]. The frontier orbitals energies, namely the highest occupied molecular orbital energy (E_{HOMO}) and the lowest unoccupied molecular orbital energy (E_{LUMO}), molecular dipole (μ), as well as the maximum absolute component of μ ($|\mu|_{\mathrm{max}}$) were also recovered from the optimization calculations.

In total, more than 100 descriptors, which feature one-, two-, and three-dimensional ones and can be categorized into a variety of classes consisting of topological descriptors, electronic descriptors, thermodynamic descriptors, structure descriptors, spatial descriptors, and *E*-state indices, were enumerated by Discovery Studio (BIOVIA, San Diego, CA, USA) and E-Dragon (available at the website http://www.vcclab.org/lab/edragon/). The logarithm of the *n*-octanol–water partition coefficient at pH 7.4 (log *P*) was calculated by *XLOGP3* of SwissADME (available at the website http://www.swissadme.ch/index.php). Furthermore, the cross-sectional area (CSA), which has been implicated in membrane permeability [70,71], was calculated using the method modified by Muehlbacher et al. [72]. The collected compounds were divided into 4 ion classes [73], namely zwitterion, base, acid, and neutral ions according to their pK_a values. The neutral ions only have one pK_a value, the zwitterion ions are those whose strongest acidic pK_a values are larger than 7 and the strongest basic ones are smaller than 7, the acidic ions have all their pK_a values smaller than 7, whereas the basic ions have all their pK_a values larger than 7.

2.3. Descriptor Selection

Descriptor selection was initially executed by removing those descriptors missing more than one molecule or displaying little or no distinction among all molecules. Furthermore, the Spearman's matrix between calculated descriptors was constructed to minimize the chance of spurious correlations, and those descriptors with intercorrelation values of $r^2 > 0.80$ were discarded, since the threshold was proposed by Topliss and Edwards [74]. In this study, a more conservative value of $r^2 \geq 0.64$ was taken to further ensure the quality of derived models.

Descriptor values can span a wide range due to their diverse nature (vide supra). It is of necessity to transfer descriptors into a more consistent range to decrease the chance of descriptors with broader ranges overriding those with narrower ranges [75]. Accordingly, descriptors were subjected to normalization by centering and scaling

$$\hat{x}_{ij} = \frac{x_{ij} - \langle x_j \rangle}{\sqrt{\sum_{i=1}^{n} \left(x_{ij} - \langle x_j \rangle \right)^2 / (n-1)}} \tag{2}$$

where x_{ij} and $\hat{x}_{ij}$ symbolize the *j*th original and normalized descriptors of the *i*th molecule, respectively; $\langle x_j \rangle$ is the average value of the original *j*th descriptor; and n is the number of molecules.

The descriptor selection is of pivotal importance in the performance of QSAR models [76]. Thus, genetic function approximation (GFA) bundled in the QSAR module of Discovery Studio was used for the initial descriptor because of its effectiveness and efficiency [77]. The recursive feature elimination (RFE) scheme was adopted for additional selection, in which the model was repeatedly generated by all but one descriptor. The descriptor, which had the less contribution in predictive performance, was removed after ranking their contributions [78].

2.4. Dataset Selection

It is not uncommon to identify the outliers and remove them from data collection for model development [79]. As such, outliers were recognized by inspecting molecular distribution in the chemical space [80], which was created by principal components (PCs) using the Diverse Molecules/Principal Component Analysis embedded in Discovery Studio, followed by discovering the outliers.

The remaining molecules were arbitrarily allocated into the training set and test set with an about 4:1 portion as recommended [81] to generate and verify the built model, respectively, using the Diverse Molecules/Library Analysis function within Discovery Studio. Golbraikh et al. have postulated that a sound model can be resulted only when both samples in the training set and test set can show high levels of chemical and biological similarity [82]. Thus, the data distributions in the training and test set were carefully checked to ensure the high similarity degrees biologically and chemically in both datasets.

2.5. Hierarchical Support Vector Regression

Leong et al. originally invented HSVR [54] which was evolved from support vector machine (SVM) proposed by Vapnik et al. [83]. Initially, SVM was designed for classification only and the regression function, termed as support vector regression (SVR), was introduced later [84]. HSVR has a higher level of predictivity and broader applicability domain (AD) as compared with SVR, since it can seamlessly combine the advantages of the local model and global model [56]. More significantly, the superiority of HSVR has been revealed by some studies [1,54–57].

The theory and fulfillment of HSVR have been delineated in detail elsewhere, and the schematic presentation of HSVR can be depicted by Figure 1 of Leong et al. [54]. Basically, an SVR ensemble (SVRE) is used to build an HSVR model, and SVR models in the ensemble are generated from different descriptor combinations and function as local models with their own ADs. Briefly, the svm-train module in *LIBSVM* (software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was employed to build various SVR models using those samples in the training set with different descriptor combinations and SVR run conditions. The module svm-predict in *LIBSVM* was adopted to validate the produced SVR models using the samples in the test set. Radial basis function (RBF) was the designated kernel function due to its simplicity and better functionality [85]. Both ε-SVR and ν-SVR regression functions were tested. The SVR runtime conditions including ε-SVR and ν-SVR, their associated ε and ν, the kernel width γ, and cost C were tuned by the grid-search technique.

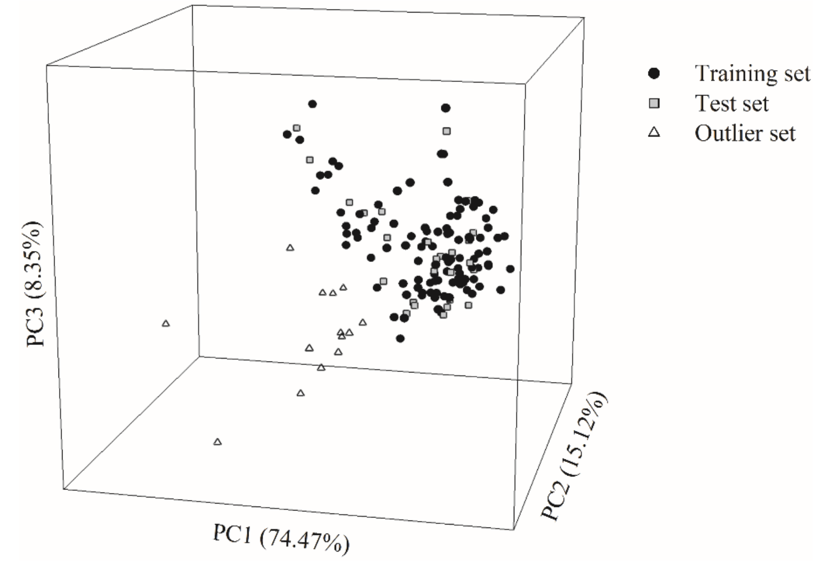

Figure 1. The chemical space spanned by three principle components (PCs) displays the distribution of the data samples in the training set (solid circle), test set (gray square), and outlier set (open triangle).

According to the principle of Occam's razor, i.e., the principle of parsimony, the number of descriptors selected to build SVR models should be minimized as much as possible. This principle was also applied to the construction of SVRE, which demanded the minimum number of ensemble members [86]. Initially, the combinations of two SVR models were adopted to generate the HSVR model; this process was repeated until the production of a predictive HSVR. Otherwise, the combinations of three- or even four-member SVRE were used to develop the HSVR models if the two-SVR ensembles failed to perform well.

2.6. Predictive Evaluation

The residual yielded by the difference between the observed value (y_i) and the predicted value $(\hat{y}_i)$ for the ith molecule was computed based on the following equation:

$$\Delta_i = y_i - \hat{y}_i \tag{3}$$

In addition, standard deviation (s), maximum residual (Δ_{Max}), root mean square error (RMSE), and mean absolute error (MAE) in a dataset with n samples were evaluated.

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \triangle_i^2 / n} \tag{4}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| \Delta_i \right| \tag{5}$$

Various statistic metrics were adopted to evaluate the produced models. The squared correlation coefficients including r^2 and q^2 in the training set and external set, respectively, were computed by the following equation.

$$r^2, q^2 = 1 - \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 / \sum_{i=1}^{n} (y_i - \langle \hat{y} \rangle)^2 \tag{6}$$

where $\langle \hat{y}_i \rangle$ represents the average predicted value, and n is the number of samples in the dataset. The derived models were subjected to the 10-fold cross-validation using the function embedded in *LIBSVM* to give rise to the squared correlation coefficient of 10-fold cross-validation q^2_{CV}. Another internal validation was carried out by the Y-scrambling test [87], in which the log P_{app} values were randomly permuted and then reapplied to the previous developed model without altering the descriptors. This process was repeated 25 times as suggested [87] to generate the average squared correlation coefficient $\langle r_s^2 \rangle$.

The external dataset was evaluated predictivity by the squared correlation coefficients q^2_{F1}, q^2_{F2}, and q^2_{F3}, and the concordance correlation coefficient (*CCC*) [88–93] using *QSARINS* [94,95].

$$q^2_{\text{F1}} = 1 - \sum_{i=1}^{n_{\text{EXT}}} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{n_{\text{EXT}}} (y_i - \langle y_{\text{TR}} \rangle)^2 \tag{7}$$

$$q^2_{\text{F2}} = 1 - \sum_{i=1}^{n_{\text{EXT}}} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{n_{\text{EXT}}} (y_i - \langle y_{\text{EXT}} \rangle)^2 \tag{8}$$

$$q^2_{\text{F3}} = 1 - \left[\sum_{i=1}^{n_{\text{EXT}}} (y_i - \hat{y}_i)^2 / n_{\text{EXT}} \right] / \left[\sum_{i=1}^{n_{\text{EXT}}} (y_i - \langle y_{\text{TR}} \rangle)^2 / n_{\text{TR}} \right] \tag{9}$$

$$\text{CCC} = \frac{2 \sum_{i=1}^{n_{\text{EXT}}} (y_i - \langle y_{\text{EXT}} \rangle)(\hat{y}_i - \langle \hat{y}_{\text{EXT}} \rangle)}{\sum_{i=1}^{n_{\text{EXT}}} (y_i - \langle y_{\text{EXT}} \rangle)^2 + \sum_{i=1}^{n_{\text{EXT}}} (\hat{y}_i - \langle \hat{y}_{\text{EXT}} \rangle)^2 + n_{\text{EXT}} (\langle y_{\text{EXT}} \rangle - \langle \hat{y}_{\text{EXT}} \rangle)^2} \tag{10}$$

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

In addition, some modified squared correlation coefficients r^2 were estimated [96,97]

$$r_m^2 = r^2 \left(1 - \sqrt{|r^2 - r_o^2|} \right) \tag{11}$$

$$r'^2_m = r^2 \left(1 - \sqrt{\left| r^2 - r'^2_o \right|} \right) \tag{12}$$

$$\left\langle r_m^2 \right\rangle = \left(r_m^2 + {r'}_m^2 \right)/2 \tag{13}$$

$$\Delta r_m^2 = \left| r_m^2 - {r'}_m^2 \right| \tag{14}$$

$$\left(r^2 - r_o^2 \right)/r^2 < 0.10 \text{ and } 0.85 \leq k \leq 1.15. \tag{15}$$

To externally evaluate the predictivity of the generated models, the most stringent criteria validation values jointly proposed by Golbraikh et al. [82], Ojha et al. [96], Roy et al. [98], and Chirico and Gramatica [89] were adopted

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

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

$$\left| r_0^2 - {r'}_0^2 \right| < 0.30 \tag{18}$$

$$r_m^2 \geq 0.65 \tag{19}$$

$$\left\langle r_m^2 \right\rangle \geq 0.65 \text{ and } \Delta r_m^2 < 0.20 \tag{20}$$

$$\mathrm{CCC} \geq 0.85 \tag{21}$$

where r^2 in Equations (15) and (18)-(20) symbolize r^2 and q^2 in the training set and external set, respectively. The q_{Fn}^2 in Equation (16) stands for q_{F1}^2, q_{F2}^2, and q_{F3}^2.

3. Results

3.1. Dataset Selection

Of all the molecules enrolled in this study, 104 and 26 molecules were randomly selected as the training set and test set, respectively, giving rise to a ca. 4:1 ratio as suggested [81]. The chemical space with the projection of all molecules is displayed in Figure 1. Three principle components (PCs), which accounted for 97.94% of the variance in the original data, were used to create the chemical space. This figure shows that samples in the training set and test set had similar distribution in the chemical space. The high levels of the biological and chemical similarity between both datasets can be illustrated by the histograms of log P_{app}, molecular weight (MW), surface area (SA), polar surface area (PSA), number of hydrogen bond acceptor (HBA), number of hydrogen bond donor (HBD), and *n*-octanol-water partition coefficient (log P) in the density form (Figure S1). Thus, it is plausible to assert that the substantial bias did not appear in the data partition.

It is of great significance to characterize the AD of the predictive model and to exclude the outliers from data collection [94]. Various methods to detect outliers have been proposed [99]. The scheme based on the chemical similarity/dissimilarity using principle component analysis (PCA) was adopted in this study [94]. Accordingly, 14 molecules were specified as outliers, which are substantially dissimilar to those ones in both the training and test sets, as shown in the chemical space (Figure 1), from which it can be observed that they are located far from the others. The distinction between the outliers and the others can be actually recognized by the fact that they contain more than nine rings or more than 12 HBAs as compared with the other molecules.

3.2. SVR Models

Numerous SVR models were generated using different descriptor combinations and runtime conditions. Three SVR models, coined as SVR A, SVR B, and SVR C, were assembled to establish the SVR ensemble, which was successively utilized to generate the HSVR model by another SVR. The optimal runtime conditions of SVR A, SVR B, SVR C, and HSVR are listed in Table S2.

SVR A, SVR B, and SVR C adopted five, five, and seven descriptors, respectively, with different combinations (Table 1). These SVR models in the ensemble were assembled according to their performances on the molecules and statistical assessments in the training set and test set. Their runtime conditions and their predicted log P_{app} values are listed in Tables S1 and S2, respectively. Tables 2 and 3 record their associated statistical evaluations in the training set and test set, respectively.

Table 1. The list of ensemble support vector regression (SVR) models and their descriptors, the correlation coefficient (r) with P_{app}, and their descriptions.

Descriptor	SVR A	SVR B	SVR C	r	Description
log P	X †	X		0.15	Logarithm of the n-octanol-water partition coefficient
n_{Ar}			X	−0.07	Number of aromatic rings
PSA	X	X		−0.56	Polar surface area
μ		X		−0.27	Dipole moment
$\mid\mu\mid_{\mathrm{max}}$	X		X	−0.08	The maximum dipole component
α	X			−0.34	Sum of atomic polarizabilities over all the molecule atoms
n_{Ring}			X	−0.31	Number of rings
V_{m}			X	−0.35	Molecular volume
n_{Rot}			X	−0.21	Number of rotatable bonds in a molecule
HBD		X	X	−0.40	Number of hydrogen-bond donors
p$K_{\mathrm{a(Max)}}$	X		X	−0.13	The maximum pK_{a} for a molecule
ion class		X		N/A ‡	Four classes are separated by the pK_{a} of molecules

† Selected. ‡ Not applicable.

Table 2. Statistic metrics including r^2, Δ_{Max}, mean absolute error (MEA), s, root mean square error (RMSE), q^2_{CV}, and $\langle r^2_{\mathrm{s}}\rangle$ assessed by support vector regression (SVR) A, SVR B, SVR C, and hierarchical support vector regression (HSVR) in the training set.

Statistic Metrics	SVR A	SVR B	SVR C	HSVR
r^2	0.69	0.77	0.76	0.91
Δ_{Max}	1.31	1.19	1.66	0.98
MAE	0.28	0.17	0.17	0.1
s	0.25	0.28	0.29	0.18
RMSE	0.38	0.32	0.33	0.2
q^2_{CV}	0.16	0.19	0.21	0.81
$\langle r^2_{\mathrm{s}}\rangle$	0.05	0.03	0.03	0.03

Table 3. Statistic metrics including q^2, q^2_{F1}, q^2_{F2}, q^2_{F3} *CCC*, Δ_{Max}, MAE, s, and RMSE assessed by SVR A, SVR B, SVR C, and HSVE in the test set.

Statistic Metrics	SVR A	SVR B	SVR C	HSVR
q^2	0.50	0.58	0.60	0.75
q^2_{F1}	0.42	0.58	0.59	0.71
q^2_{F2}	0.41	0.57	0.59	0.71
q^2_{F3}	0.30	0.50	0.50	0.70
CCC	0.62	0.74	0.77	0.85
Δ_{Max}	1.27	1.06	0.88	0.72
MAE	0.42	0.35	0.39	0.33
s	0.35	0.31	0.23	0.20
RMSE	0.54	0.46	0.45	0.38

The observed versus the predicted log P_{app} values by SVR A, SVR B, SVR C, and HSVR are displayed by the scatter plot in Figure 2, from which it can be observed that SVR A, SVR B, and SVR C predicted the observed values well for the majority of the molecules in the training set, producing small MAE and s values consequently (Table 2). Moreover, it can be found from Figure 2 that the points predicted by SVR B are generally closer to the regression line than SVR A and SVR C. SVR B, consequently, gave rise to the lowest

Δ_{Max} (1.19), MAE (0.17), and RMSE (0.32), and the largest r^2 (0.77), suggesting that SVR B performed marginally better than SVR A and SVR C in the training set.

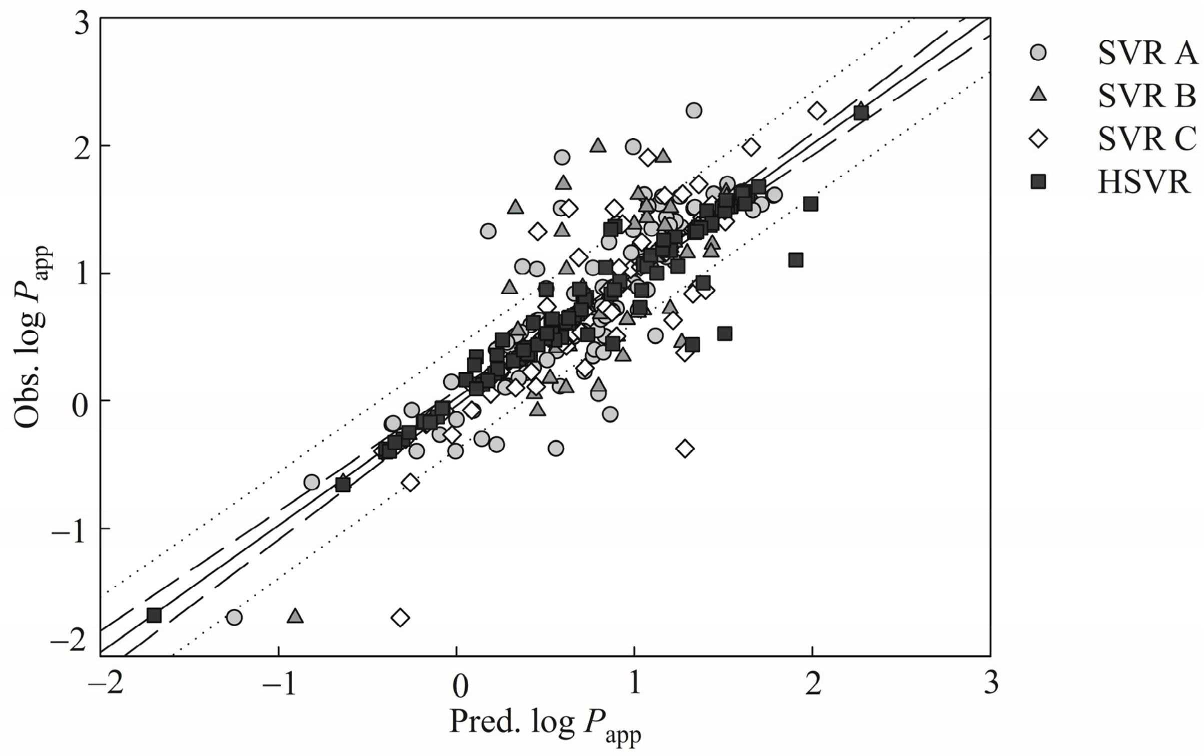

Figure 2. Observed log P_{app} versus the log P_{app} predicted by SVR A (gray circle), SVR B (gray triangle), SVR C (open diamond), and HSVR (solid square) for the training samples. The solid, dashed, and dotted lines represent to the HSVR regression of the data, 95% confidence intervals for the HSVR regression, and 95% confidence intervals for the prediction, respectively.

Furthermore, the difference between r^2 and q^2_{CV} evaluated by SVR B was 0.58 when subjected to the leave-one-out cross-validation, indicating that SVR B was over-trained which, in turn, can severely limit its application. Over-training was also associated with SVR A and SVR C as manifested by their extremely low q^2_{CV} values. The $\langle r_s^2 \rangle$ values produced by SVR A, SVR B, and SVR C were 0.05, 0.03, and 0.03 (Table 2), respectively, when subjected in Y-scrambling. These near zero values suggest that there is an almost zero chance correlation associated with those SVR models [87].

The predicted values by SVR A, SVR B, and SVR C are in moderate agreement with the observed values for those test molecules depicted by Figure 3, which shows the scatter plot of observed versus the log P_{app} predictions by SVR A, SVR B, SVR C, and HSVR for those samples in the test set. The MAE values generated by SVR A, SVR B, and SVR C increase from 0.28, 0.17, and 0.17 in the training set to 0.42, 0.35, and 0.39 in the test set, respectively (Table 3). RMSE along with the other statistic values also reveal deteriorating performances of these models in SVRE from the training set to the test set (Tables 2 and 3). Moreover, the q^2 values produced by SVR A, SVR B, and SVR C were 0.50, 0.58, and 0.60 in the test set, respectively, which are much less than their r^2 counterparts in the training set.

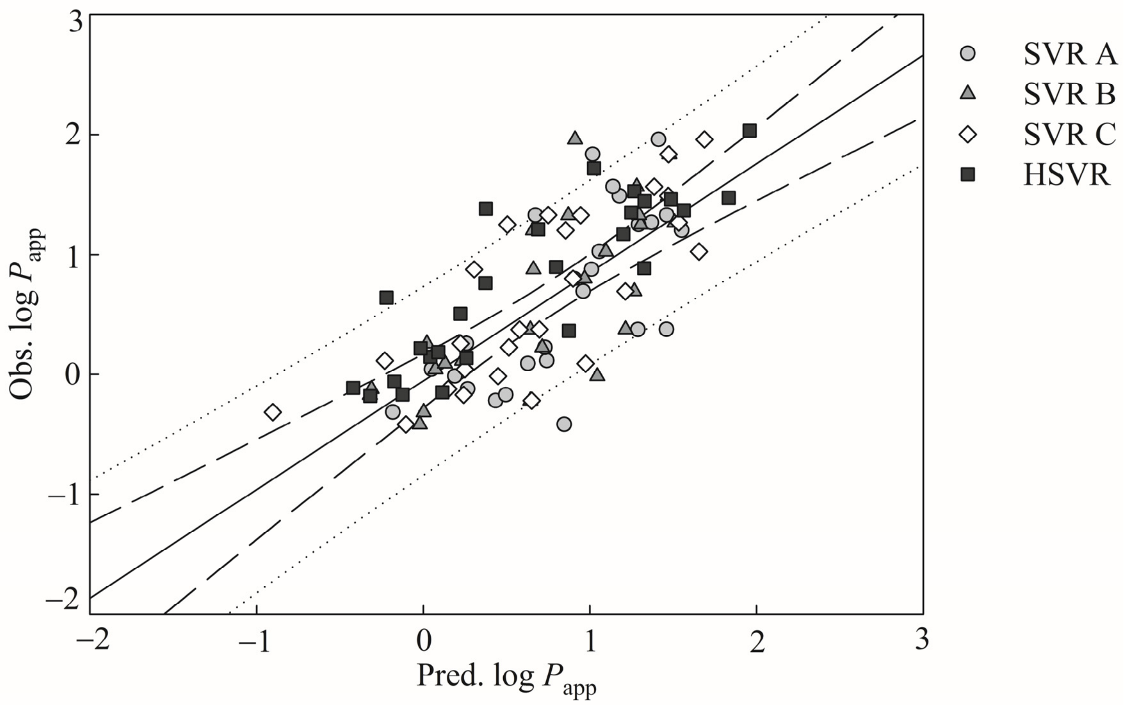

Figure 3. Observed log P_{app} versus the log P_{app} predicted by SVR A (gray circle), SVR B (gray triangle), SVR C (open diamond), and HSVR (solid square) for the test samples. The solid, dashed, and dotted lines represent the HSVR regression of the data, 95% confidence intervals for the HSVR regression, and 95% confidence intervals for the prediction, respectively.

The prediction performances of those SVR models in the SVRE were significantly decreased when applied to those samples in the outlier set as suggested by the statistical metrics listed in Table 4. For example, SVR A, SVR B, and SVR C yielded the q^2_{F2} values of −0.18, −0.41, and 0.16, respectively, which are substantially smaller than the r^2 values in the training set and the q^2_{F2} values in the test set (Tables 2 and 3). Furthermore, the distances between the points and the regression line in the outlier set were much greater than those in the training set shown in Figure 4. As such, it can be asserted that those three models in the SVRE are vulnerable to the outliers that, actually, are not uncommon for most predictive models [100].

Table 4. Statistic metrics including q^2, q^2_{F1}, q^2_{F2}, q^2_{F3}, *CCC*, Δ_{Max}, MAE, *s*, and RMSE assessed by SVR A, SVR B, SVR C, and HSVE in the outlier set.

Statistic Metrics	SVR A	SVR B	SVR C	HSVR
q^2	0.45	0.36	0.40	0.76
q^2_{F1}	0.75	0.70	0.82	0.95
q^2_{F2}	−0.18	−0.41	0.16	0.76
q^2_{F3}	0.39	0.27	0.56	0.87
CCC	0.49	0.56	0.58	0.87
Δ_{Max}	1.58	0.91	0.82	0.49
MAE	0.35	0.47	0.16	0.17
s	0.41	0.34	0.56	0.17
RMSE	0.52	0.57	0.58	0.24

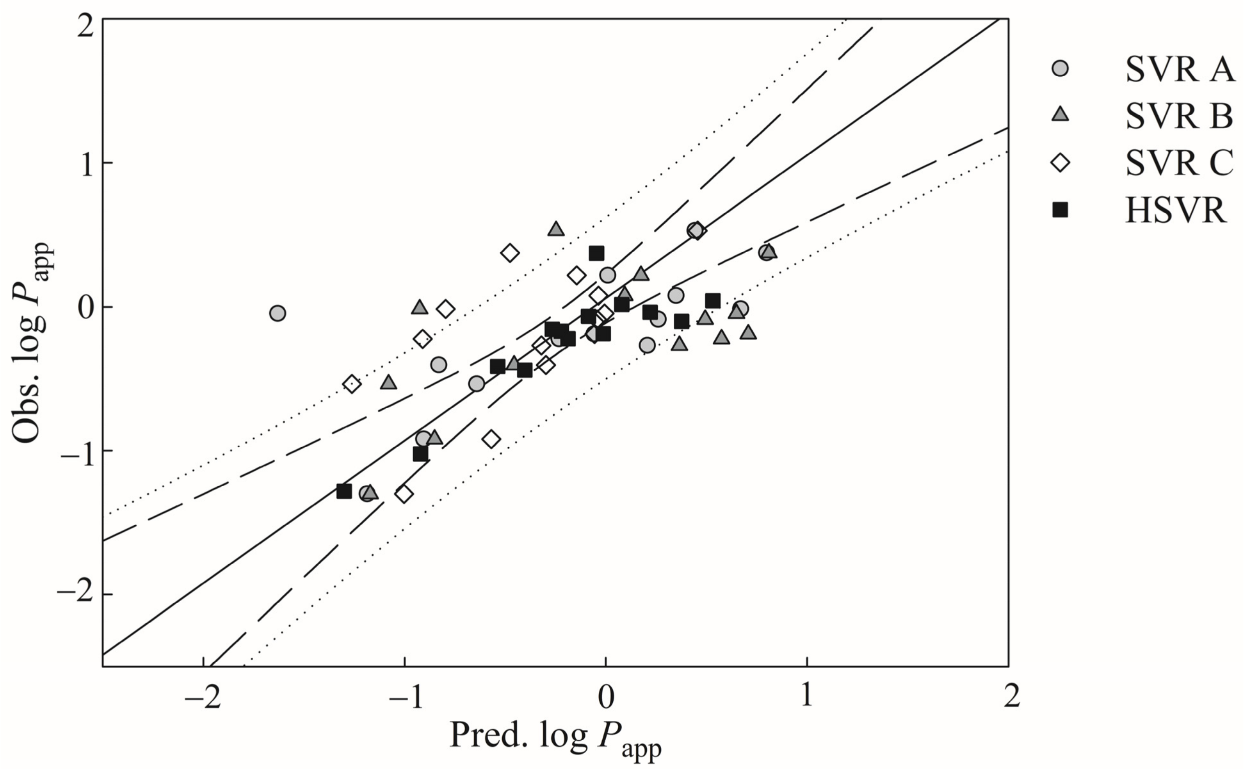

Figure 4. Observed log P_{app} versus the log P_{app} predicted by SVR A (gray circle), SVR B (gray triangle), SVR C (open diamond), and HSVR (solid square) for the outlier samples. The solid, dashed, and dotted lines represent the HSVR regression of the data, 95% confidence intervals for the HSVR regression, and 95% confidence intervals for the prediction, respectively.

3.3. HSVR Model

The HSVR model was generated by the regression of SVRE according to the predictions of all molecules and statistical assessments in the training set (Table S1 and Table 2), and its runtime parameters are recorded in Table S2. HSVR commonly predicted better than SVR A, SVR B, and SVR C for the samples in the training set, as demonstrated by Figure 2, from which it can be noticed that most of predictions by HSVR lie in the range between the largest and the smallest ones predicted by those models in the SVRE. HSVR can improve the predictions in some cases. For instance, the prediction of compound **101** (omeprazole) by HSVR yielded an absolute residual of 0.02, whereas SVR A, SVR B, and SVR C produced the absolute errors of 0.34, 1.10, and 0.18, respectively (Table S1). In addition, HSVR produced the highest r^2 (0.91) and q^2_{CV}(0.81) and the lowest Δ_{Max} (0.98), MAE (0.10), s (MAE), and RMSE (0.20) values when compared with those models in the SVRE, suggesting that HSVR statistically performed better SVR A, SVR B, and SVR C in the training set. Furthermore, HSVR gave rise to a $\langle r_s^2 \rangle$ value of 0.03, indicating that it is least possible that HSVR was created by chance correlation [87].

When applied to the test molecules, marginal performance deteriorations can be found for HSVR. For example, s increased from 0.18 in the training set to 0.20 in the test set (Tables 2 and 3). However, Δ_{Max} dropped from 0.98 in the training set to 0.72 in the test set. HSVR still executed better than SVR A, SVR B, and SVR C in the test set as shown in Figure 3. The other statistical parameters listed in Table 3 also assert the performance dominance of HSVR. For instance, the q^2 values were 0.50, 0.58, 0.60, and 0.75 generated by SVR A, SVR B, SVR C, and HSVR, respectively. Similarly, HSVR also produced smaller absolute deviations than its counterparts in the SVRE in the test set. For example, the absolute residuals of compound 36 (clozapine) were 0.35, 0.54, 0.35, and 0.03 yielded by SVR A, SVR B, SVR C, and HSVR, respectively (Table S1). HSVR generally produced consistent and small deviations in both training and test sets as asserted by those parameters listed in Tables 2 and 3 in comparison with its counterparts in the SVRE. More importantly, the HSVR model generated the largest q^2 (0.75) in the test set and the smallest difference between r^2 and q^2_{CV} (0.10), suggesting that it is less likely that HSVR model was over-trained or over-fitted.

HSVR even displayed better performance than the SVR models in the ensemble in the outlier set as depicted by those statistical assessments listed in Table 4. The HSVR model generated the largest q^2 value (0.76) and yet SVR A, SVR B, and SVR C yielded 0.45, 0.36, and 0.40, respectively. The superiority of HSVR in the outlier set can also be assured by the other statistical parameters, which is mainly due to the broader application domain of HSVR when compared with its counterparts in the ensemble. That robust HSVR feature makes it more utilizable in practical applications [101].

3.4. Predictive Evaluations

The scatter plot of residual versus the log P_{app} prediction by HSVR for the training, test, and outlier samples is shown in Figure 5, from which it can be found that the residuals are commonly situated on both sides of x-axis along with the prediction range in those three datasets, suggesting that it is least likely that systematic error is associated with HSVR. Additionally, the training set, test set, and outlier set had the average residuals of 0.02, −0.13, and 0.06, respectively (Table S1), denoting that there is no biased prediction by HSVR.

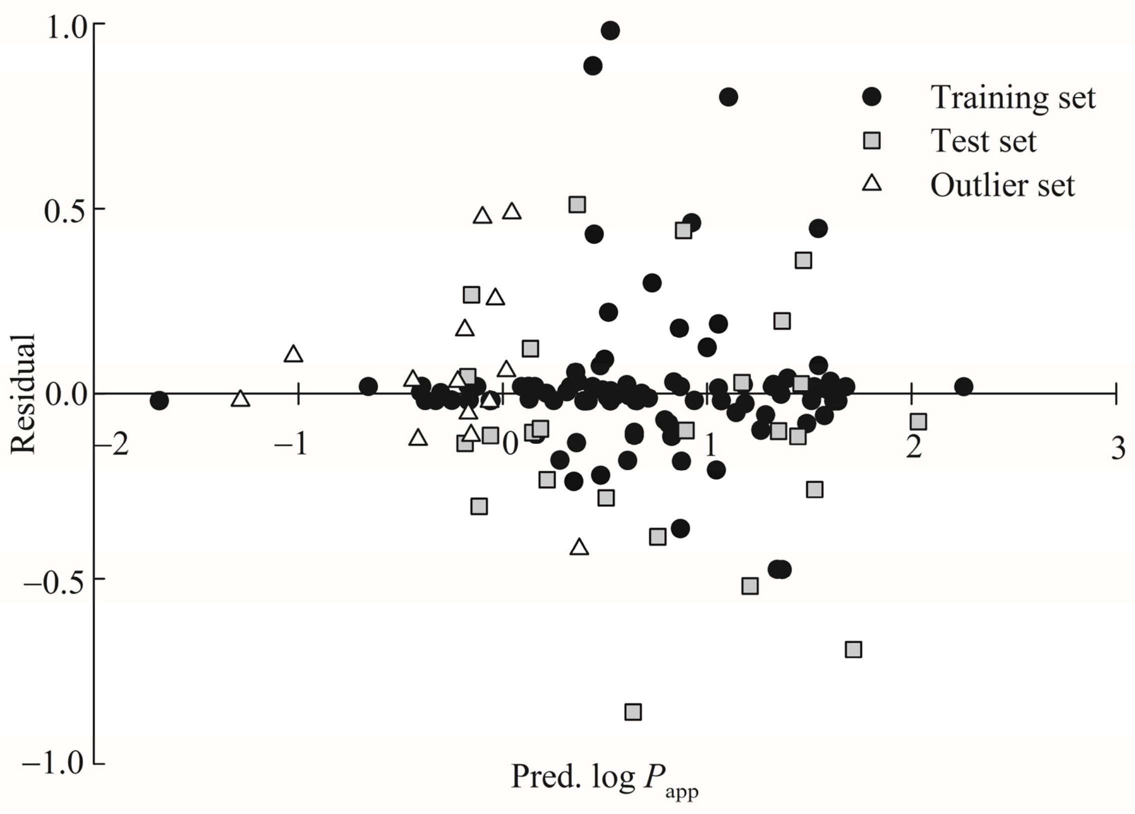

Figure 5. Residual versus the log P_{app} prediction by HSVR in the training set (solid circle), test set (gray square), and outlier set (open triangle).

Table 5 lists the results when the developed HSVR model was further subjected to the most stringent validation criteria collectively recommended by Golbraikh et al. [82], Ojha et al. [96], Roy et al. [98], and Chirico and Gramatica [89] in the three datasets (Equations (15)–(21)). It can be observed that HSVR completely met those proposed validation requirements in addition to the fact that HSVR exhibited a similar degrees of performance in the training set, test set, and outlier set. As such, it can be asserted that HSVR is an extremely accurate and predictive theoretical model.

Table 5. Validation verification of HSVR based on prediction performance of the training, test, and outlier samples.

Validation Verification	Training Set	Test Set	Outlier Set
r_0^2	0.91	0.75	0.75
k	1.01	0.86	0.93
r'^2_0	0.91	0.68	0.71
r_m^2	0.84	0.71	0.68
$r_{\prime m}^2$	0.91	0.75	0.76
$\langle r_m^2 \rangle$	0.87	0.73	0.72
Δr_m^2	0.06	0.04	0.08
$r^2 \geq 0.70$	X [†]	X	X
Equation (15)	X	X	X
Equation (16)	X	N/A	N/A
Equation (17)	X	X	X
Equation (18)	X	X	X
Equation (19)	X	X	X
Equation (20)	X	X	X
Equation (21)	N/A [‡]	X	X

[†] Fulfilled; [‡] Not applicable.

3.5. Mock test

To verify the practical applicability of the generated HSVR model, this model was applied to those drugs measured by Yamashita et al. [30]. There were eight compounds commonly adopted by this study and Yamashita et al., furnishing a sound way to calibrate the challenging system. However, Yamashita et al. assayed the P_{app} values at pH 6.0, instead of pH 7.4 used by those compounds collected in this study, suggesting that some P_{app} variations can be resulted from both systems (vide supra). These discrepancies make those drugs assayed by Yamashita et al. not appropriate as the second external dataset or the test set because those validation criteria listed in Table 5 cannot be applied to those drugs. The relationship between both different experimental conditions was initially constructed for those eight common compounds, and the resulting scatter plot is exhibited in Figure 6, from which it can be found that both assay systems were reasonably correlated with each other with an r value of 0.86), suggesting that this HSVR can be adopted to predict those novel compounds measured by Yamashita et al.

Figure 7 shows the predicted results of seven novel drugs in the mock test. The correlation coefficient r value between the predicted log P_{app} (pH 7.4) and observed log P_{app} (pH 6.0) was 0.86, suggesting that the HSVR model can nearly reproduce the experimental results. In addition, the produced p-value was <0.05. This mock test ensured the predictive ability of generated HSVR when applied to the novel compounds with different experimental conditions.

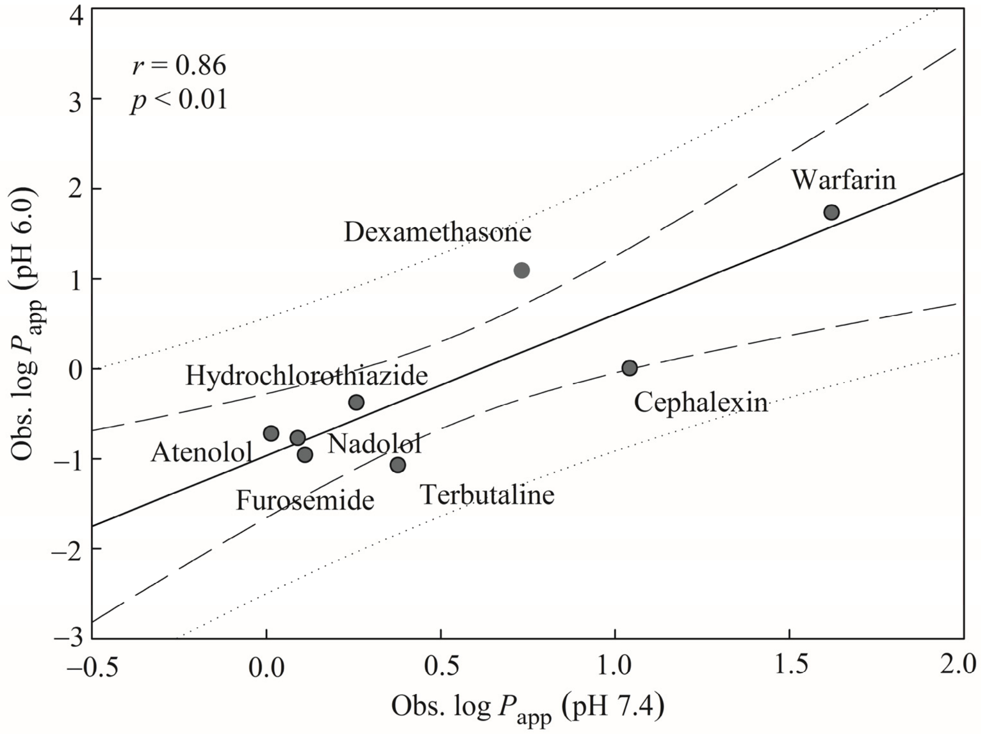

Figure 6. Observed log P_{app} at pH 7.4 versus observed log P_{app} at pH 6.0 for the common drugs in the mock test. The solid, dashed, and dotted lines represent the mock test regression of the observed data, 95% confidence interval for the mock test regression, and 95% confidence interval for the observation, respectively.

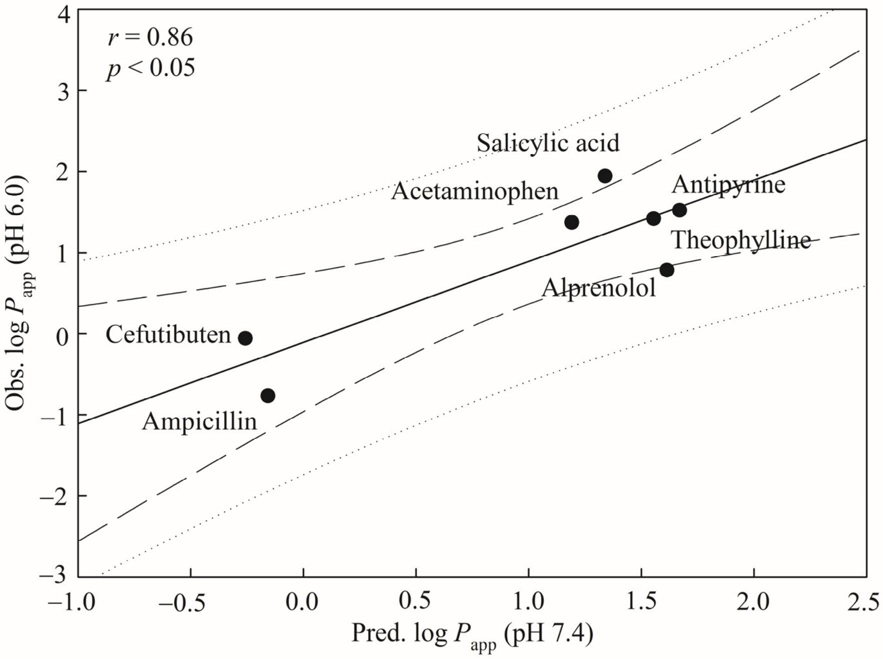

Figure 7. Predicted log P_{app} at pH 7.4 versus observed log P_{app} at pH 6.0 by the HSVR model for the drugs in the mock test. The solid, dashed, and dotted lines represent the HSVR regression data, 95% confidence interval for the HSVR regression, and 95% confidence interval for prediction, respectively.

3.6. Classification

It is of interest to verify the qualitative predictivity of HSVR, since a number of qualitative models have been published [25,102]. Accordingly, compounds enlisted in this study were classified as Caco-2 permeable (Caco-2^{+}) and Caco-2 impermeable (Caco-2^{-}) based on the threshold value of P_{app} (8 × 10 $^{-6}$ cm/s) as suggested [25,102]. Initially, the confusion matrix was constructed (Table S3), and the Cooper statistics and Kubat's G-mean [103] (Table S4) were employed to qualitatively evaluate the predictivity of HSVR. The results were also compared with predictions made by *admetSAR* [104] (available at website: http://lmmd.ecust.edu.cn/admetsar2/), since *admetSAR* has been adopted by DrugBank (available at: https://go.drugbank.com/) to qualitatively predict Caco-2 permeability. The results are listed in Table 6, from which it can be asserted that HSVR outperformed *admetSAR* in every aspect. For instance, the parameter accuracy was 93.1% produced by HSVR, which is substantially higher than that generated by *admetSAR* (50.7%). The metric MCC is the most distinction between HSVR and *admetSAR* (85.0% vs. −8.0%). Thus, it can be asserted that HSVR is also an accurate and predictive qualitative predictive model.

Table 6. Statistical parameters of qualitative predictions by HSVR and *admetSAR*.

Statistical Parameters	HSVR	*admetSAR*
Se	90.0%	32.0%
Sp	94.7%	60.6%
Acc	93.1%	50.7%
PP	90.0%	30.2%
NP	94.7%	62.6%
MCC	85.0%	−8.0%
G-mean	92.3%	44.1%
F-measure	90.0%	31.1%
κ	85.0%	−8.0%

4. Discussion

Caco-2 has been commonly adopted to predict the intestinal permeability in the process of drug discovery because of its morphological and functional similarity with human enterocytes [105]. The mechanism of Caco-2 permeation is rather complex, since it can take place through passive diffusion, which can go through the paracellular and transcellular routes and active transport. The passive diffusion is predominately governed by the concentration gradient, and most hydrophilic drugs prefer to penetrate between cells in a paracellular fashion, whereas hydrophobic drugs are inclined to get across the cells via the transcellular route. Drugs that can permeate the Caco-2 cells by the active transport can interact with the influx and/or efflux transporters expressed on the cell surface [106]. As such, Caco-2 permeability is affected by some physicochemical and physiological properties [106].

Hydrophobicity or lipophilicity plays an important role in passive diffusion through membranes as well as the drug–receptor interactions [17,107,108]. In addition, hydrophobicity, which can represent by the *n*-octanol-water partition coefficient, *viz*. log *P*, is also an important factor affecting the interaction between the molecules and the target protein, since more lipophilic molecules tend to have stronger interactions with both target protein and biological membrane. Therefore, the very lipophilic molecules have poor oral absorption from the stomach [107,109]. Polar and hydrophobic drug must penetrate through the Caco-2 cell membrane [17,110]. In addition, it has been observed that log *P*, hydrogen bond propensity, weight, and volume are closely related with P_{app} [43]. As such, log *P* was adopted in this study (Table 1), which is consistent with the fact that numerous published in silico models to predict intestinal absorption, PAMPA permeability [1,111], and Caco-2 permeability also have employed this descriptor [40,112–114]. It can be observed from Figure 8, which displays the average log P_{app} for each histogram bin of log *P* for all molecules included in this investigation, that log P_{app} increased with log *P* value initially and then decreased afterward, leading to a seemingly bilinear relationship between log P_{app} and log *P*. This perplexing dependency can be realized by the fact that the more hydrophobic solutes can easier approach the lipid bilayer to penetrate the membrane. The opposite relationship between hydrophobicity and permeability will be resulted when the solutes are too hydrophobic due to stronger attractions between solutes and the membrane as well as stronger repulsive forces from the solvent molecules upon the entrance to the solvent environment that can be illustrated by the PAMPA permeability [1,115,116]. Complexity can be even profound when taking into account the fact that P-gp and BCRP, which are efflux transporters in Caco-2 (vide supra), can interact with substrates by hydrophobicity [117], subsequently leading to a low correlation between log P_{app} and log P (r = 0.15).

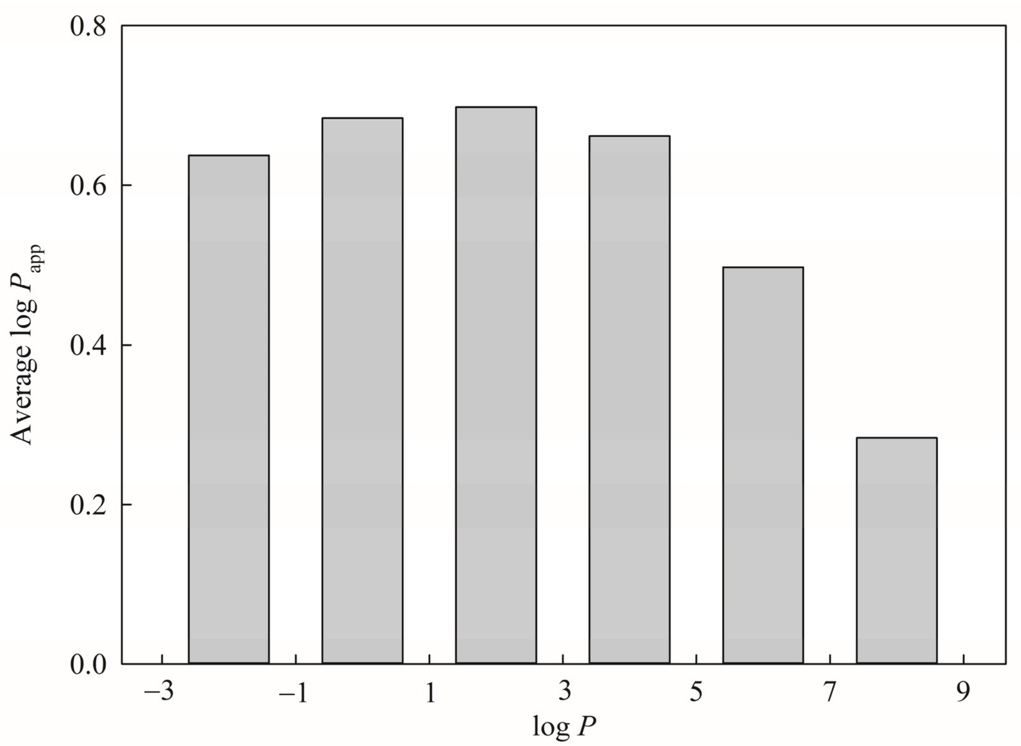

Figure 8. Histogram of average log P_{app} versus the distribution of log P.

It has been observed that the number of aromatic rings (n_{Ar}) has a positive correlation with log P with an r value of 0.67 [118], suggesting that a predictive model can be overtrained once both log P and n_{Ar} are adopted simultaneously. However, this issue was not concerned in this study, since only SVR C adopted this descriptor, whereas SVR A and SVR B included log P (Table 1). In addition, the aromatic ring is a non-polar group, which can enhance the hydrophobicity [52] and increase the passive diffusion [119,120]. In addition, aromatic ring moieties have been implicated in P-gp substrate recognition and efflux modulation [53], leading to the fact that n_{Ar} can be an important factor in P-gp modulation action [121] and BCRP-substrate interactions [122]. As such, n_{Ar} plays a complex role in both passive diffusion and active transport in Caco-2 permeability.

It has been recognized that both PSA and μ are associated with passive diffusion [37,123–125]. In addition, these descriptors have been adopted by published in silico Caco-2 permeability models [37,45–49,126–128]. It has been reported in the PAMPA permeability study that larger PSA, μ, and polarity can enhance the solute-solute and solute-solvent interactions, which, in turn, require more desolvation energy when the solutes penetrate through the lipophilic membrane to the donor compartment [123,129–132], and conversely decrease the passive diffusion [1], consequently, making permeability less favorable. Therefore, it has been shown that PSA has a negative impact in the permeation rate [133,134]. In addition, Joung et al. have indicated that PSA shows an important role in distinguishing the P-gp substrate from the non-substrates [135]. Accordingly, PSA and μ were adopted in this study due to their pivotal roles in Caco-2 permeability.

It is seemingly unusual to include the descriptor $|\mu|_{\mathrm{max}}$, which is the absolute maximum component of the molecular dipole, in this study, since it has never been employed by any published model before. This inconsistency actually can be manifested by Figure 9, which displays the average $|\mu|_{\mathrm{max}}$ for each histogram bin of μ, that the larger μ, the larger $|\mu|_{\mathrm{max}}$, suggesting that they were positively correlated with each other. In addition, μ was recruited by SVR A and SVR C, whereas $|\mu|_{\mathrm{max}}$ was enlisted by SVR B only, suggesting that it is less likely to produce an over-trained HSVR, since no single model adopted both two descriptors simultaneously. More importantly, the empirical observation has revealed that HSVR including these selections executed better than the others (data not shown) plausibly because of the descriptor-descriptor interaction [1]. Any other traditional linear or machine learning-based QSAR schemes, conversely, cannot properly render such contradictory descriptor selections.

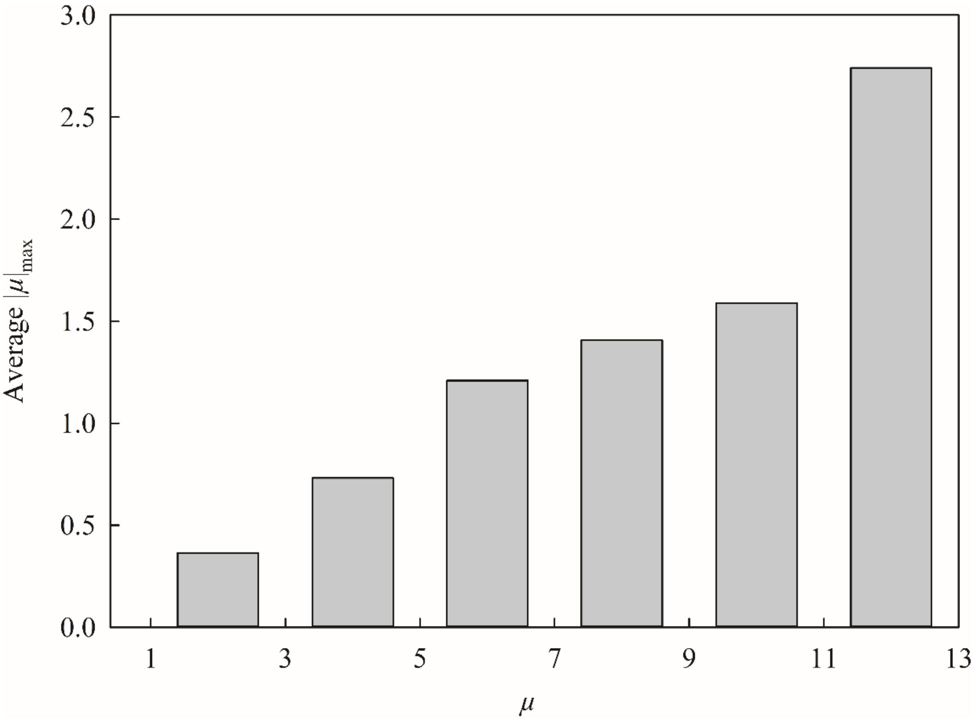

Figure 9. Histogram of average $|\mu|_{\mathrm{max}}$ versus the distribution of μ.

It has been reported that the molecular size of the solute molecule is of critical importance in the diffusivity of the biological membrane [37,125,136], and the intestinal absorption can decrease with the increase of molecular size [137]. Furthermore, the molecular size also affects passive diffusion through membranes [138,139] and active transport through the P-gp-substrate interactions [121,138]. Molecular size can be represented by a number of descriptors such α, n_{Ring}, V_{m}, and n_{rot} [140–142], which were adopted in the investigation and negatively associated with log P_{app} (Table 1). Conversely, Fujiwara et al. adopted the descriptor molecule weight (MW) to develop a theoretical Caco-2 permeability model [37], whereas MW was not included in this study. This discrepancy can be realized by the fact that α was highly correlated with MW with an r value of 0.98 for all molecules enlisted in this study, suggesting that it is plausible to replace MW by α in order not to produce an over-trained model. In addition, it has been observed that α is positively correlated to log P [143] and is highly associated with absorption [50].

The descriptor n_{Ring}, which is reportedly related to molecular size [136,141], has never been adopted by any published Caco-2 permeability predictive model and yet was selected by SVR C (Table 1). This disagreement can be recognized by the fact that n_{Ring} was greatly correlated with α with an r value of 0.78 for all molecules recruited in this study. As such, it is plausible to expect that both n_{Ring} and α play similar roles in Caco-2 permeability. The relationship among log P_{app}, n_{Ring}, and log P can be further perplexing as illustrated by Figure 10, which shows the 3D plot of log P_{app}, n_{Ring}, and log P. The relationship between n_{Ring} and log P has been detailed by Pham-The et al. [125].

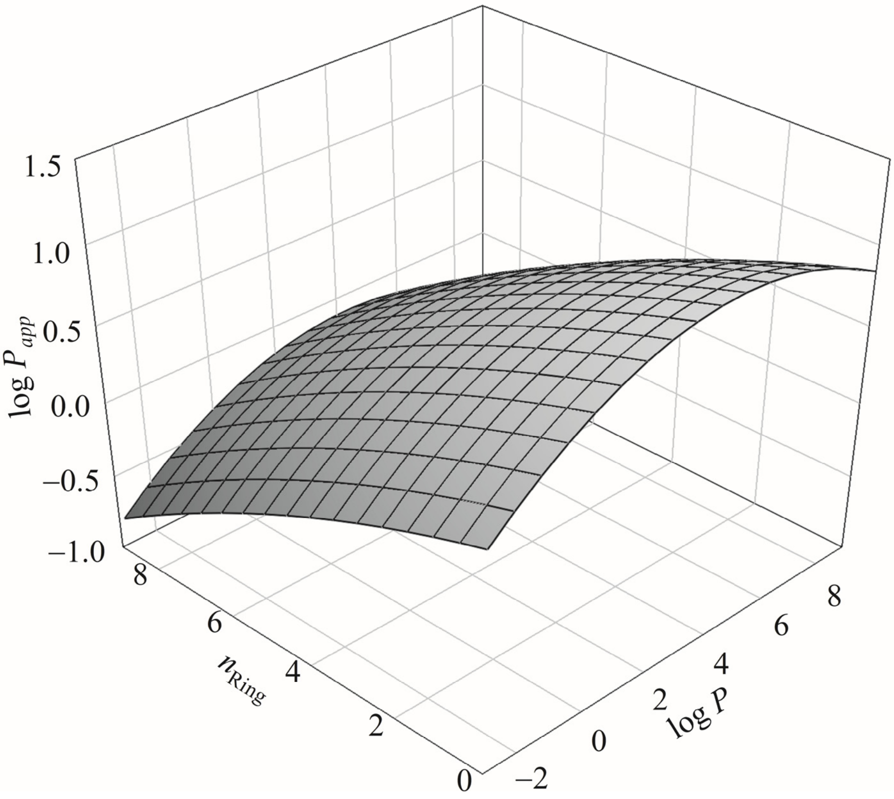

Figure 10. The relationship among log P_{app}, n_{Ring}, and log P in 3D presentation.

It has been observed that V_{m} plays an important role in passive absorption [9,144,145] and it is adopted by a published Caco-2 permeability model [146] as well as in this study. It has been observed in the rat that fewer rotatable bonds, viz. smaller n_{rot}, can lead to better oral bioavailability, and n_{rot} can also exert a positive effect on the permeation rate [133,143], since more rigid molecules will have smaller n_{rot} values that, in turn, can enhance permeability [125]. Furthermore, n_{rot} is of importance in intestinal absorption [147], since increased n_{rot} can reduce the permeability [133]. Furthermore, a number of published membrane permeability models have also employed the descriptor CSA, which is another feature associated with molecular size and also plays a pivotal role in membrane permeability [70,71]. However, n_{rot} was greatly associated with CSA with an r value of 0.80 for all molecules enrolled in this investigation, suggesting that using n_{rot} in lieu of CSA without producing the over-trained model is plausible. Li et al. also have found that n_{rot} is another feature to discriminate P-gp substrates from non-substrates [148]. As such, it is of necessity to recruit n_{rot} in model development to properly render Caco-2 permeability as suggested [71,72].

Hydrogen bonding potential, which can be expressed by HBD and HBA, is another important factor in determining the solute–solvent interactions [37], and it is the main contributor for the passive diffusion [143]. It has been observed that Caco-2 permeability is a function of HBD and/or HBA, since more permeable solutes tend to have smaller HBD and/or HBA [130,131,149]. Between HBD and HBA, HBD seemingly shows a more profound effect on Caco-2 permeability as compared with HBA [150] as manifested by the fact that several published in silico models have selected HBD to predict Caco-2 permeability instead of HBA [35,42]. Mechanistically, HBD is one of the features associated with P-gp-substrate interactions [148,151]. In addition to efflux transport, HDB is one of the features linked to substrate binding with OATP2B1 [7] as well as PepT1 [152]. Thus, it is of necessity to include in Caco-2 predictive models to take into consideration the passive diffusion as well as the active influx/efflux transport.

The descriptor p$K_{a(Max)}$ was selected in this study due to the fact that higher p$K_{a(Max)}$ can lead to the lower ionized form of drugs in the donor compartment, which, in turn, can increase the penetration through hydrophobic membrane [153]. Furthermore, it has been recognized that neutral compounds can have higher membrane permeability than the other ion classes [154]. Accordingly, all molecules included in this investigation were categorized into different ion classes based on their pK_a values. In addition, ABC and/or SLC substrates were also identified based on the drug information retrieved from DrugBank to understand if the dependence of ion class can be varied by their ion classes. It can be found from Figure 11, which displays the histograms of median log P_{app} versus all molecules, ABC substrates, SLC substrates, as well as ABC and SLC substrates for four different ion classes, that the median log P_{app} values of neutral compounds are substantially larger than the others, suggesting that neutral compounds exhibit higher Caco-2 permeability regardless of active transporter substrate classes, viz. influx transporter or efflux transporter. This observation actually is very similar to the PAMPA permeability, since the ionized compounds will demand larger desolvation energies, which, in turn, can hinder their penetration [134].

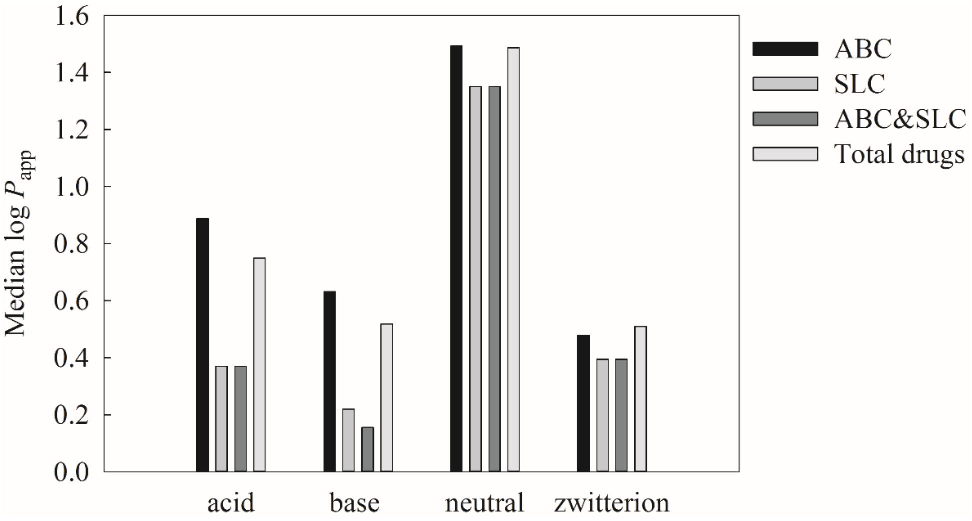

Figure 11. The histogram represents the log P_{app} versus the molecules belong to ATP-binding cassette (ABC) substrate, solute carrier (SLC) substrate, both ABC and SLC substrate, and total drugs in the acid class, base class, neutral class, and zwitterion class, respectively.

Initially, numerous efforts were made in attempting to build assorted 2-QSAR models by employing the partial least square (PLS) scheme, and yet no productive models were produced (data not shown) [1]. This challenge can be realized by the fact that the correlations between the designated descriptors and log P_{app} for all molecules included in this investigation were small, and the largest absolute maximum r was only 0.56 between PSA and log P_{app} (Table 1), signifying the high non-linearity between them. More significantly, the substantial difference in 2-QSAR development between the passive diffusion, viz. the PAMPA system, and Caco-2 permeability can be greatly attributed to the complex active (influx and efflux) transport. Thus, it is extremely difficult, if not absolutely impossible, to derive a linear Cacao-2 permeability QSAR model. Conversely, the accurate and predictive HSVR model can properly render such non-linear dependence of log P_{app} on descriptors.

5. Conclusions

Intestinal permeability is one of the important ADME/Tox metrics that should be addressed in the process of drug discovery and development. The Caco-2 system has been frequently used as a surrogate to preliminarily investigate the intestinal absorption. An in silico model can be a useful approach to predict Caco-2 permeability in assisting drug discovery and development. However, Caco-2 permeability can occur through passive diffusion and active transport, leading to a complex process. Therefore, it is of necessity to include different descriptor combinations and diverse relationships to address these variations in distinct mechanisms. The innovative machine learning-based HSVR scheme, which possesses the superior features of a local model (greater predictivity) and a global model (larger coverage of the application domain), was employed in this study to construct a theoretical model to predict the Caco-2 permeability. The generated HSVR models unveiled great prediction accuracy for the training, test, and outlier samples. When challenged by a group of drugs assayed at different experimental conditions, the developed HSVR model also executed equivalently well. In addition, HSVR showed excellent qualitative performance in recognizing Caco-2 permeable and impermeable compounds, and the selected descriptors can completely justify the diverse mechanisms related to the passive diffusion and active transport. Thus, it can be assured that this HSVR model can be useful to accurately and swiftly predict the Caco-2 permeability of novel compounds in order to assist drug discovery and development.

Supplementary Materials: The following are available online at https://www.mdpi.com/1999-4923/13/2/174/s1, Table S1. Selected compounds for this study; their names, SMILES strings, CAS numbers, and observed log P_{app} values; their predicted values by SVR A, SVR B, SVR C, and HSVR; data partitions; and references; Table S2. Optimal runtime parameters for the SVR models; Table S3. Confusion matrix for the qualitative predictive model; Table S4. The Cooper statistics and Kubat's G-mean calculated from the confusion matrix. Optimal runtime parameters for the SVR models; Figure S1. Histograms of: (A) log *P*app, (B) molecular weight (MW), (C) surface area (SA), (D) polar surface area (PSA), (E) number of hydrogen bond acceptor (HBA), (F) number of hydrogen bond donor (HBD), (G) and the *n*-octanol-water partition coefficient (log *P*) in the training set, test set, and outlier set.

Author Contributions: G.H.T., C.-F.W., and M.K.L. conceived and designed the study; G.H.T., C.-S.J., and M.K.L. performed the experiments and analyzed the data; and G.H.T., C.-F.W., and M.K.L. wrote the paper. The final version of manuscript is reviewed and approved by all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of Science and Technology, Taiwan.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Parts of calculations were performed at the National Center for High-Performance Computing, Taiwan. The authors are grateful to Paola Gramatica for providing the free license of *QSARINS*.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chi, C.-T.; Lee, M.-H.; Weng, C.-F.; Leong, M.K. *In Silico* Prediction of PAMPA Effective Permeability Using a Two-QSAR Approach. *Int. J. Mol. Sci.* **2019**, *20*, 3170. [CrossRef] [PubMed]
2. Proctor, W.R.; Ming, X.; Thakker, D.R. In Vitro Techniques to Study Drug–Drug Interactions Involving Transport: Caco-2 Model for Study of P-Glycoprotein and Other Transporters. In *Enzyme- and Transporter-Based Drug-Drug Interactions: Progress and Future Challenges*; Pang, S.K., Rodrigues, D.A., Peter, M.R., Eds.; Springer: New York, NY, USA, 2010; pp. 257–282.
3. Volpe, D.A. Advances in cell-based permeability assays to screen drugs for intestinal absorption. *Expert. Opin. Drug Discov.* **2020**, *15*, 539–549. [CrossRef] [PubMed]

4. Dobson, P.D.; Kell, D.B. Carrier-mediated cellular uptake of pharmaceutical drugs: An exception or the rule? *Nat. Rev. Drug Discov.* **2008**, *7*, 205–220. [CrossRef]
5. Artursson, P.; Bergström, C.A.S. Intestinal Absorption: The Role of Polar Surface Area. In *Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability*; van de Waterbeemd, H., Lennernäs, H., Artursson, P., Eds.; WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2004; Volume 18, pp. 339–357.
6. Müller, J.; Keiser, M.; Drozdzik, M.; Oswald, S. Expression, regulation and function of intestinal drug transporters: An update. *Biol. Chem.* **2017**, *398*, 175–192. [CrossRef] [PubMed]
7. Varma, M.V.; Ambler, C.M.; Ullah, M.; Rotter, C.J.; Sun, H.; Litchfield, J.; Fenner, K.S.; El-Kattan, A.F. Targeting Intestinal Transporters for Optimizing Oral Drug Absorption. *Curr. Drug Metab.* **2010**, *11*, 730–742. [CrossRef]
8. Seithel, A.; Karlsson, J.; Hilgendorf, C.; Björquist, A.; Ungell, A.L. Variability in mRNA expression of ABC- and SLC-transporters in human intestinal cells: Comparison between human segments and Caco-2 cells. *Eur. J. Pharm. Sci.* **2006**, *28*, 291–299. [CrossRef]
9. Agatonovic-Kustrin, S.; Beresford, R.; Yusof, A.P.M. Theoretically-derived molecular descriptors important in human intestinal absorption. *J. Pharm. Biomed. Anal.* **2001**, *25*, 227–237. [CrossRef]
10. Li, A.P. Screening for human ADME/Tox drug properties in drug discovery. *Drug Discov. Today* **2001**, *6*, 357–366. [CrossRef]
11. Caldwell, G.W.; Yan, Z.; Tang, W.; Dasgupta, M.; Hasting, B. ADME optimization and toxicity assessment in early- and late-phase drug discovery. *Curr. Top. Med. Chem.* **2009**, *9*, 965–980. [CrossRef]
12. Wishart, D.S. Improving early drug discovery through ADME modelling: An overview. *Drugs R D* **2007**, *8*, 349–362. [CrossRef]
13. Dahlgren, D.; Lennernäs, H. Intestinal Permeability and Drug Absorption: Predictive Experimental, Computational and In Vivo Approaches. *Pharmaceutics* **2019**, *11*, 411. [CrossRef] [PubMed]
14. Petri, N.; Lennernäs, H. In Vivo Permeability Studies in the Gastrointestinal Tract of Humans. In *Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability*; van de Waterbeemd, H., Lennernäs, H., Artursson, P., Eds.; WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2004; Volume 18, pp. 155–188.
15. Dezani, T.M.; Dezani, A.B.; Junior, J.B.; Serra, C.H. Single-Pass Intestinal Perfusion (SPIP) and prediction of fraction absorbed and permeability in humans: A study with antiretroviral drugs. *Eur. J. Pharm. Biopharm.* **2016**, *104*, 131–139. [CrossRef]
16. Lennernäs, H. Human in Vivo Regional Intestinal Permeability: Importance for Pharmaceutical Drug Development. *Mol. Pharm.* **2014**, *11*, 12–23. [CrossRef] [PubMed]
17. Chmiel, T.; Mieszkowska, A.; Kempińska-Kupczyk, D.; Kot-Wasik, A.; Namieśnik, J.; Mazerska, Z. The impact of lipophilicity on environmental processes, drug delivery and bioavailability of food components. *Microchem. J.* **2019**, *146*, 393–406. [CrossRef]
18. Kansy, M.; Fischer, H.; Kratzat, K.; Senner, F.; Wagner, B.; Parrilla, I. High-Throughput Artificial Membrane Permeability Studies in Early Lead Discovery and Development. In *Pharmacokinetic Optimization in Drug Research*; Testa, B., Van de Waterbeend, H., Folkers, G., Guy, R., Eds.; Verlag Helvetica Chimica Acta/Wiley/VCH: Zurich, Switzerland; Weinheim, Germany, 2001; pp. 447–464.
19. Irvine, J.D.; Takahashi, L.; Lockhart, K.; Cheong, J.; Tolan, J.W.; Selick, H.E.; Grove, J.R. MDCK (Madin–Darby canine kidney) cells: A tool for membrane permeability screening. *J. Pharm. Sci.* **1999**, *88*, 28–33. [CrossRef] [PubMed]
20. Avdeef, A.; Nielsen, P.E.; Tsinman, O. PAMPA—a drug absorption in vitro model: 11. Matching the in vivo unstirred water layer thickness by individual-well stirring in microtitre plates. *Eur. J. Pharm. Sci.* **2004**, *22*, 365–374. [CrossRef]
21. Wang, N.-N.; Dong, J.; Deng, Y.-H.; Zhu, M.-F.; Wen, M.; Yao, Z.-J.; Lu, A.-P.; Wang, J.-B.; Cao, D.-S. ADME Properties Evaluation in Drug Discovery: Prediction of Caco-2 Cell Permeability Using a Combination of NSGA-II and Boosting. *J. Chem. Inf. Model.* **2016**, *56*, 763–773. [CrossRef]
22. Uchida, R.; Okamoto, H.; Ikuta, N.; Terao, K.; Hirota, T. Investigation of Enantioselective Membrane Permeability of α-Lipoic Acid in Caco-2 and MDCKII Cell. *Int. J. Mol. Sci.* **2016**, *17*, 155. [CrossRef]
23. Zeng, Z.; Shen, Z.L.; Zhai, S.; Xu, J.L.; Liang, H.; Shen, Q.; Li, Q.Y. Transport of curcumin derivatives in Caco-2 cell monolayers. *Eur. J. Pharm. Biopharm.* **2017**, *117*, 123–131. [CrossRef]
24. Sánchez, A.B.; Calpena, A.C.; Mallandrich, M.; Clares, B. Validation of an Ex Vivo Permeation Method for the Intestinal Permeability of Different BCS Drugs and Its Correlation with Caco-2 In Vitro Experiments. *Pharmaceutics* **2019**, *11*, 638. [CrossRef]
25. Ponce, Y.M.; Pérez, M.A.C.; Zaldivar, V.R.; Sanz, M.B.; Mota, D.S.; Torrens, F. Prediction of Intestinal Epithelial Transport of Drug in (Caco-2) Cell Culture from Molecular Structure using *in silico* Approaches During Early Drug Discovery. *Internet Electron. J. Mol. Des.* **2005**, *4*, 124–150.
26. Petri, N.; Tannergren, C.; Rungstad, D.; Lennernäs, H. Transport Characteristics of Fexofenadine in the Caco-2 Cell Model. *Pharm. Res.* **2004**, *21*, 1398–1404. [CrossRef] [PubMed]
27. Volpe, D.A. Variability in Caco-2 and MDCK cell-based intestinal permeability assays. *J. Pharm. Sci.* **2008**, *97*, 712–725. [CrossRef] [PubMed]
28. Hosey, C.M.; Benet, L.Z. Predicting the Extent of Metabolism Using *in Vitro* Permeability Rate Measurements and *in Silico* Permeability Rate Predictions. *Mol. Pharm.* **2015**, *12*, 1456–1466. [CrossRef] [PubMed]
29. Lee, J.B.; Zgair, A.; Taha, D.A.; Zang, X.; Kagan, L.; Kim, T.H.; Kim, M.G.; Yun, H.-y.; Fischer, P.M.; Gershkovich, P. Quantitative analysis of lab-to-lab variability in Caco-2 permeability assays. *Eur. J. Pharm. Biopharm.* **2017**, *114*, 38–42. [CrossRef]
30. Yamashita, S.; Furubayashi, T.; Kataoka, M.; Sakane, T.; Sezaki, H.; Tokuda, H. Optimized conditions for prediction of intestinal drug permeability using Caco-2 cells. *Eur. J. Pharm. Sci.* **2000**, *10*, 195–204. [CrossRef]

31. Bergström, C.A. In silico predictions of drug solubility and permeability: Two rate-limiting barriers to oral drug absorption. *Basic Clin. Pharmacol. Toxicol.* **2005**, *96*, 156–161. [CrossRef]
32. Parrott, N.; Lavé, T. Prediction of intestinal absorption: Comparative assessment of gastroplus™ and idea™. *Eur. J. Pharm. Sci.* **2002**, *17*, 51–61. [CrossRef]
33. Pelkonen, O.; Turpeinen, M.; Raunio, H. In vivo-in vitro-in silico pharmacokinetic modelling in drug development: Current status and future directions. *Clin. Pharm.* **2011**, *50*, 483–491. [CrossRef]
34. 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. [CrossRef]
35. Subramanian, G.; Kitchen, D.B. Computational approaches for modeling human intestinal absorption and permeability. *J. Mol. Model.* **2006**, *12*, 577–589. [CrossRef] [PubMed]
36. Karelson, M.; Karelson, G.; Tamm, T.; Indrek, T.; Jänes, J.; Tämm, K.; Lomaka, A.; Savchenko, D.; Dobcheva, D. QSAR study of pharmacological permeabilities. *Arkivoc* **2009**, *2*, 218–238. [CrossRef]
37. Fujiwara, S.-I.; Yamashita, F.; Hashida, M. Prediction of Caco-2 cell permeability using a combination of MO-calculation and neural network. *Int. J. Pharm.* **2002**, *237*, 95–105. [CrossRef]
38. Van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Raevsky, O.A. Estimation of Caco-2 Cell Permeability using Calculated Molecular Descriptors. *Quant. Struct. Act. Relat.* **1996**, *15*, 480–490. [CrossRef]
39. Refsgaard, H.H.F.; Jensen, B.F.; Brockhoff, P.B.; Guldbrandt, M.; Christensen, M.S. *In Silico* Prediction of Membrane Permeability from Calculated Molecular Parameters. *J. Med. Chem.* **2005**, *48*, 805–811. [CrossRef]
40. Paixão, P.; Gouveia, L.F.; Morais, J.A.G. Prediction of the *in vitro* permeability determined in Caco-2 cells by using artificial neural networks. *Eur. J. Pharm. Sci.* **2010**, *41*, 107–117. [CrossRef]
41. Nordqvist, A.; Nilsson, J.; Lindmark, T.; Eriksson, A.; Garberg, P.; Kihlén, M. A General Model for Prediction of Caco-2 Cell Permeability. *QSAR Comb. Sci.* **2004**, *23*, 303–310. [CrossRef]
42. Ma, G.; Cheng, Y. Predicting Caco-2 Permeability Using Support Vector Machine and Chemistry Development Kit. *J. Pharm. Pharm. Sci.* **2006**, *9*, 210–221.
43. Di Fenza, A.; Alagona, G.; Ghio, C.; Leonardi, R.; Giolitti, A.; Madami, A. Caco-2 cell permeability modelling: A neural network coupled genetic algorithm approach. *J. Comput. Aided Mol. Des.* **2007**, *21*, 207–221. [CrossRef]
44. Chan, E.C.Y.; Tan, W.L.; Ho, P.C.; Fang, L.J. Modeling Caco-2 permeability of drugs using immobilized artificial membrane chromatography and physicochemical descriptors. *J. Chromatogr. A* **2005**, *1072*, 159–168. [CrossRef]
45. Santos-Filho, O.A.; Hopfinger, A.J. Combined 4D-Fingerprint and Clustering Based Membrane-Interaction QSAR Analyses for Constructing Consensus Caco-2 Cell Permeation Virtual Screens. *J. Pharm. Sci.* **2008**, *97*, 566–583. [CrossRef] [PubMed]
46. Tantishaiyakul, V. Prediction of Caco-2 cell permeability using partial least squares multivariate analysis. *Pharmazie* **2001**, *56*, 407–411.
47. Welling, S.H.; Clemmensen, L.K.H.; Buckley, S.T.; Hovgaard, L.; Brockhoff, P.B.; Refsgaard, H.H.F. *In silico* modelling of permeation enhancement potency in Caco-2 monolayers based on molecular descriptors and random forest. *Eur. J. Pharm. Biopharm.* **2015**, *94*, 152–159. [CrossRef]
48. Yamashita, F.; Wanchana, S.; Hashida, M. Quantitative structure/property relationship analysis of Caco-2 permeability using a genetic algorithm-based partial least squares method. *J. Pharm. Sci.* **2002**, *91*, 2230–2239. [CrossRef] [PubMed]
49. Yamashita, F.; Fujiwara, S.-i.; Hashida, M. The "Latent Membrane Permeability" Concept: QSPR Analysis of Inter/Intralaboratorically Variable Caco-2 Permeability. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 408–413. [CrossRef]
50. Norinder, U.; Osterberg, T.; Artursson, P. Theoretical calculation and prediction of intestinal absorption of drugs using Molsurf parametrization and PLS statistics. *Eur. J. Pharm. Sci.* **1999**, *8*, 49–56. [CrossRef]
51. Pham The, H.; González-Álvarez, I.; Bermejo, M.; Mangas Sanjuan, V.; Centelles, I.; Garrigues, T.M.; Cabrera-Pérez, M. In Silico Prediction of Caco-2 Cell Permeability by a Classification QSAR Approach. *Mol. Inform.* **2011**, *30*, 376–385. [CrossRef] [PubMed]
52. Shinde, R.N.; Srikanth, K.; Sobhia, M.E. Insights into the permeability of drugs and drug-likemolecules from MI-QSAR and HQSAR studies. *J. Mol. Model.* **2012**, *18*, 947–962. [CrossRef] [PubMed]
53. Wang, B.; Ma, L.-Y.; Wang, J.-Q.; Lei, Z.-N.; Gupta, P.; Zhao, Y.-D.; Li, Z.-H.; Liu, Y.; Zhang, X.-H.; Li, Y.-N.; et al. Discovery of 5-Cyano-6-phenylpyrimidin Derivatives Containing an Acylurea Moiety as Orally Bioavailable Reversal Agents against P-Glycoprotein-Mediated Mutidrug Resistance. *J. Med. Chem.* **2018**, *61*, 5988–6001. [CrossRef]
54. 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. [CrossRef] [PubMed]
55. Lee, M.-H.; Ta, G.H.; Weng, C.-F.; Leong, M.K. *In Silico* Prediction of Intestinal Permeability by Hierarchical Support Vector Regression. *Int. J. Mol. Sci.* **2020**, *21*, 3582. [CrossRef] [PubMed]
56. 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. [CrossRef] [PubMed]
57. Chen, C.; Lee, M.H.; Weng, C.F.; Leong, M.K. Theoretical Prediction of the Complex P-Glycoprotein Substrate Efflux Based on the Novel Hierarchical Support Vector Regression Scheme. *Molecules* **2018**, *23*, 1820. [CrossRef] [PubMed]
58. Bergström, C.A.S.; Bolin, S.; Artursson, P.; Rönn, R.; Sandström, A. Hepatitis C virus NS3 protease inhibitors: Large, flexible molecules of peptide origin show satisfactory permeability across Caco-2 cells. *Eur. J. Pharm. Sci.* **2009**, *38*, 556–563. [CrossRef] [PubMed]

59. Skolnik, S.; Lin, X.; Wang, J.; Chen, X.-H.; He, T.; Zhang, B. Towards prediction of *in vivo* intestinal absorption using a 96-well Caco-2 assay. *J. Pharm. Sci.* **2010**, *99*, 3246–3265. [CrossRef] [PubMed]
60. Lazorova, L.; Hubatsch, I.; Ekegren, J.K.; Gising, J.; Nakai, D.; Zaki, N.M.; Bergström, C.A.S.; Norinder, U.; Larhed, M.; Artursson, P. Structural features determining the intestinal epithelial permeability and efflux of novel HIV-1 protease inhibitors. *J. Pharm. Sci.* **2011**, *100*, 3763–3772. [CrossRef]
61. Nti-Addae, K.W.; Guarino, V.R.; Dalwadi, G.; Stella, V.J. Determination of the permeability characteristics of two sulfenamide prodrugs of linezolid across Caco-2 cells. *J. Pharm. Sci.* **2012**, *101*, 3134–3141. [CrossRef]
62. Deng, X.; Zhang, G.; Shen, C.; Yin, J.; Meng, Q. Hollow fiber culture accelerates differentiation of Caco-2 cells. *Appl. Microbiol. Biotechnol.* **2013**, *97*, 6943–6955. [CrossRef]
63. Yang, Y.; Bai, L.; Li, X.; Xiong, J.; Xu, P.; Guo, C.; Xue, M. Transport of active flavonoids, based on cytotoxicity and lipophilicity: An evaluation using the blood–brain barrier cell and Caco-2 cell models. *Toxicol. Vitro* **2014**, *28*, 388–396. [CrossRef]
64. Wu, S.; Xu, W.; Wang, F.-R.; Yang, X.-W. Study of the Biotransformation of Tongmai Formula by Human Intestinal Flora and Its Intestinal Permeability across the Caco-2 Cell Monolayer. *Molecules* **2015**, *20*, 18704. [CrossRef]
65. Zhou, L.; Lee, K.; Thakker, D.R.; Boykin, D.W.; Tidwell, R.R.; Hall, J.E. Enhanced Permeability of the Antimicrobial Agent 2,5-Bis(4-Amidinophenyl)Furan Across Caco-2 Cell Monolayers Via Its Methylamidoxime Prodrug. *Pharm. Res.* **2002**, *19*, 1689–1695. [CrossRef]
66. Troutman, M.D.; Thakker, D.R. Efflux Ratio Cannot Assess P-Glycoprotein-Mediated Attenuation of Absorptive Transport: Asymmetric Effect of P-Glycoprotein on Absorptive and Secretory Transport Across Caco-2 Cell Monolayers. *Pharm. Res.* **2003**, *20*, 1200–1209. [CrossRef] [PubMed]
67. 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. [CrossRef]
68. 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. [CrossRef]
69. Besler, B.H.; Merz, K.M., Jr.; Kollman, P.A. Atomic charges derived from semiempirical methods. *J. Comput. Chem.* **1990**, *11*, 431–439. [CrossRef]
70. Gerebtzoff, G.; Seelig, A. *In Silico* Prediction of Blood−Brain Barrier Permeation Using the Calculated Molecular Cross-Sectional Area as Main Parameter. *J. Chem. Inf. Model.* **2006**, *46*, 2638–2650. [CrossRef]
71. Leung, S.S.F.; Sindhikara, D.; Jacobson, M.P. Simple Predictive Models of Passive Membrane Permeability Incorporating Size-Dependent Membrane-Water Partition. *J. Chem. Inf. Model.* **2016**, *56*, 924–929. [CrossRef]
72. Muehlbacher, M.; Spitzer, G.M.; Liedl, K.R.; Kornhuber, J. Qualitative prediction of blood–brain barrier permeability on a large and refined dataset. *J. Comput. Aided Mol. Des.* **2011**, *25*, 1095–1106. [CrossRef]
73. Fridén, M.; Winiwarter, S.; Jerndal, G.; Bengtsson, O.; Wan, H.; Bredberg, U.; Hammarlund-Udenaes, M.; Antonsson, M. Structure−Brain Exposure Relationships in Rat and Human Using a Novel Data Set of Unbound Drug Concentrations in Brain Interstitial and Cerebrospinal Fluids. *J. Med. Chem.* **2009**, *52*, 6233–6243. [CrossRef]
74. Topliss, J.G.; Edwards, R.P. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244. [CrossRef]
75. Kettaneh, N.; Berglund, A.; Wold, S. PCA and PLS with very large data sets. *Comput. Stat. Data Anal.* **2005**, *48*, 69–85. [CrossRef]
76. 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. [CrossRef] [PubMed]
77. Rogers, D. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866. [CrossRef]
78. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
79. Yimin, A.; Jun, X. A New Support Vector Machine Model for Outlier Detection. In Proceedings of the International Conference on Graphic and Image Processing (ICGIP 2012), Hong Kong, China, 26–27 October 2013; p. 87680E.
80. Hubert, M.; Engelen, S. Robust PCA and classification in biosciences. *Bioinformatics* **2004**, *20*, 1728–1736. [CrossRef]
81. Tropsha, A.; Gramatica, P.; Gombar, V.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. [CrossRef]
82. Golbraikh, A.; Shen, M.; Xiao, Z.; 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. [CrossRef]
83. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
84. Vapnik, V.; Golowich, S.E.; Smola, A. Support vector method for function approximation, regression estimation and signal processing. In Proceedings of the 9th International Conference on Neural Information Processing Systems; MIT Press: Denver, Colorado, 1996; pp. 281–287.
85. Kecman, V. *Learning and Soft Computing: Support. Vector Machines, Neural Networks, and Fuzzy Logic. Models*; MIT Press: Cambridge, MA, USA, 2001; p. 608.
86. Dearden, J.C.; Cronin, M.T.; Kaiser, K.L. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009**, *20*, 241–266. [CrossRef]

87. Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357. [CrossRef]
88. 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. [CrossRef] [PubMed]
89. 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. [CrossRef] [PubMed]
90. Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the Definition of the Q2 Parameter for QSAR Validation. *J. Chem. Inf. Model.* **2009**, *49*, 1669–1678. [CrossRef] [PubMed]
91. Lin, L.I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **1989**, *45*, 255–268. [CrossRef]
92. Lin, L.I.K. Assay Validation Using the Concordance Correlation Coefficient. *Biometrics* **1992**, *48*, 599–604. [CrossRef]
93. 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. [CrossRef]
94. 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. [CrossRef]
95. 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. [CrossRef]
96. Ojha, P.K.; Mitra, I.; Das, R.N.; Roy, K. Further exploring rm2 metrics for validation of QSPR models. *Chemom. Intell. Lab. Syst.* **2011**, *107*, 194–205. [CrossRef]
97. Roy, P.P.; Roy, K. On Some Aspects of Variable Selection for Partial Least Squares Regression Models. *QSAR Comb. Sci.* **2008**, *27*, 302–313. [CrossRef]
98. 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. [CrossRef] [PubMed]
99. Gajewicz, A. How to judge whether QSAR/read-across predictions can be trusted: A novel approach for establishing a model's applicability domain. *Environ. Sci. Nano* **2018**, *5*, 408–421. [CrossRef]
100. Caudill, M. Using neural networks: Hybrid expert networks. *AI Expert* **1990**, *5*, 49–54.
101. Gnanadesikan, R.; Kettenring, J.R. Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. *Biometrics* **1972**, *28*, 81–124. [CrossRef]
102. Shin, M.; Jang, D.; Nam, H.; Lee, K.H.; Lee, D. Predicting the Absorption Potential of Chemical Compounds Through a Deep Learning Approach. *IEEE/ACM Trans. Comput Biol. Bioinform.* **2018**, *15*, 432–440. [CrossRef]
103. Ma, L.; Fan, S. CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. *BMC Bioinform.* **2017**, *18*, 169. [CrossRef]
104. Cheng, F.; Li, W.; Zhou, Y.; Shen, J.; Wu, Z.; Liu, G.; Lee, P.W.; Tang, Y. admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties. *J. Chem. Inf. Model.* **2012**, *52*, 3099–3105. [CrossRef]
105. Grandvuinet, A.S.; Gustavsson, L.; Steffansen, B. New Insights into the Carrier-Mediated Transport of Estrone-3-sulfate in the Caco-2 Cell Model. *Mol. Pharm.* **2013**, *10*, 3285–3295. [CrossRef]
106. Volpe, D.A. Drug-permeability and transporter assays in Caco-2 and MDCK cell lines. *Future Med. Chem.* **2011**, *3*, 2063–2077. [CrossRef]
107. Valkó, K.L. Lipophilicity and biomimetic properties measured by HPLC to support drug discovery. *J. Pharm. Biomed. Anal.* **2016**, *130*, 35–54. [CrossRef]
108. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26. [CrossRef]
109. Liu, X.; Testa, B.; Fahr, A. Lipophilicity and its relationship with passive drug permeation. *Pharm. Res.* **2011**, *28*, 962–977. [CrossRef]
110. Volpe, D.A. Drug Permeability Studies in Regulatory Biowaiver Applications. In *Drug Absorption Studies: In Situ, In Vitro and In Silico Models*; Ehrhardt, C., Kim, K.-J., Eds.; Springer US: Boston, MA, USA, 2008; pp. 665–680.
111. Akamatsu, M.; Fujikawa, M.; Nakao, K.; Shimizu, R. *In silico* prediction of human oral absorption based on QSAR analyses of PAMPA permeability. *Chem. Biodivers* **2009**, *6*, 1845–1866. [CrossRef] [PubMed]
112. Pham-The, H.; Cabrera-Pérez, M.Á.; Nam, N.-H.; Castillo-Garit, J.A.; Rasulev, B.; Le-Thi-Thu, H.; Casañola-Martin, G.M. *In silico* assessment of ADME properties: Advances in Caco-2 cell monolayer permeability modeling. *Curr. Top. Med. Chem.* **2018**, *18*, 2209–2229. [CrossRef] [PubMed]
113. Faassen, F.; Vogel, G.; Spanings, H.; Vromans, H. Caco-2 permeability, P-glycoprotein transport ratios and brain penetration of heterocyclic drugs. *Int. J. Pharm.* **2003**, *263*, 113–122. [CrossRef]
114. Faassen, F.; Kelder, J.; Lenders, J.; Onderwater, R.; Vromans, H. Physicochemical Properties and Transport of Steroids Across Caco-2 Cells. *Pharm. Res.* **2003**, *20*, 177–186. [CrossRef] [PubMed]
115. Verma, R.P.; Hansch, C.; Selassie, C.D. Comparative QSAR studies on PAMPA/modified PAMPA for high throughput profiling of drug absorption potential with respect to Caco-2 cells and human intestinal absorption. *J. Comput. Aided Mol. Des.* **2007**, *21*, 3–22. [CrossRef]

116. Fujikawa, M.; Nakao, K.; Shimizu, R.; Akamatsu, M. QSAR study on permeability of hydrophobic compounds with artificial membranes. *Bioorg. Med. Chem.* **2007**, *15*, 3756–3767. [CrossRef]
117. Billat, P.-A.; Roger, E.; Faure, S.; Lagarce, F. Models for drug absorption from the small intestine: Where are we and where are we going? *Drug Discov. Today* **2017**, *22*, 761–775. [CrossRef]
118. Ward, S.E.; Beswick, P. What does the aromatic ring number mean for drug design? *Expert. Opin. Drug Discov.* **2014**, *9*, 995–1003. [CrossRef]
119. Camenisch, G.; Alsenz, J.; van de Waterbeemd, H.; Folkers, G. Estimation of permeability by passive diffusion through Caco-2 cell monolayers using the drugs' lipophilicity and molecular weight. *Eur. J. Pharm. Sci.* **1998**, *6*, 317–324. [CrossRef]
120. Yang, N.J.; Hinner, M.J. Getting across the cell membrane: An overview for small molecules, peptides, and proteins. *Methods Mol. Biol.* **2015**, *1266*, 29–53. [PubMed]
121. Beck, W.T.; Qian, X.-d. Photoaffinity substrates for P-glycoprotein. *Biochem. Pharmacol.* **1992**, *43*, 89–93. [CrossRef]
122. Ebert, B.; Seidel, A.; Lampen, A. Identification of BCRP as transporter of benzo[a]pyrene conjugates metabolically formed in Caco-2 cells and its induction by Ah-receptor agonists. *Carcinogenesis* **2005**, *26*, 1754–1763. [CrossRef]
123. Clark, D.E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J. Pharm. Sci.* **1999**, *88*, 807–814. [CrossRef]
124. Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Artursson, P. Correlation of drug absorption with molecular surface properties. *J. Pharm. Sci.* **1996**, *85*, 32–39. [CrossRef]
125. Pham-The, H.; González-Álvarez, I.; Bermejo, M.; Garrigues, T.; Le-Thi-Thu, H.; Cabrera-Pérez, M.Á. The Use of Rule-Based and QSPR Approaches in ADME Profiling: A Case Study on Caco-2 Permeability. *Mol. Inf.* **2013**, *32*, 459–479. [CrossRef]
126. Değim, Z. Prediction of Permeability Coefficients of Compounds Through Caco-2 Cell Monolayer Using Artificial Neural Network Analysis. *Drug Dev. Ind. Pharm.* **2005**, *31*, 935–942. [CrossRef]
127. Hou, T.J.; Zhang, W.; Xia, K.; Qiao, X.B.; Xu, X.J. ADME Evaluation in Drug Discovery. 5. Correlation of Caco-2 Permeation with Simple Molecular Properties. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1585–1600. [CrossRef]
128. Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Beigi, F.; Lundahl, P.; Artursson, P. Evaluation of Dynamic Polar Molecular Surface Area as Predictor of Drug Absorption: Comparison with Other Computational and Experimental Predictors. *J. Med. Chem.* **1998**, *41*, 5382–5392. [CrossRef]
129. Bergström, C.A.S.; Strafford, M.; Lazorova, L.; Avdeef, A.; Luthman, K.; Artursson, P. Absorption Classification of Oral Drugs Based on Molecular Surface Properties. *J. Med. Chem.* **2003**, *46*, 558–570. [CrossRef] [PubMed]
130. Burton, P.S.; Conradi, R.A.; Hilgers, A.R.; Ho, N.F.H.; Maggiora, L.L. The relationship between peptide structure and transport across epithelial cell monolayers. *J. Control. Release* **1992**, *19*, 87–97. [CrossRef]
131. Kulkarni, A.; Han, Y.; Hopfinger, A.J. Predicting Caco-2 Cell Permeation Coefficients of Organic Molecules Using Membrane-Interaction QSAR Analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 331–342. [CrossRef]
132. Rossi Sebastiano, M.; Doak, B.C.; Backlund, M.; Poongavanam, V.; Over, B.; Ermondi, G.; Caron, G.; Matsson, P.; Kihlberg, J. Impact of Dynamically Exposed Polarity on Permeability and Solubility of Chameleonic Drugs Beyond the Rule of 5. *J. Med. Chem.* **2018**, *61*, 4189–4202. [CrossRef] [PubMed]
133. Veber, D.F.; Johnson, S.R.; Cheng, H.-Y.; Smith, B.R.; Ward, K.W.; Kopple, K.D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623. [CrossRef]
134. Yang, Y.; Engkvist, O.; Llinàs, A.; Chen, H. Beyond Size, Ionization State, and Lipophilicity: Influence of Molecular Topology on Absorption, Distribution, Metabolism, Excretion, and Toxicity for Druglike Compounds. *J. Med. Chem.* **2012**, *55*, 3667–3677. [CrossRef]
135. 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. [CrossRef]
136. Newby, D.A. *Data Mining Methods for the Prediction of Intestinal Absorption Using QSAR*; University of Kent: Kent, UK, 2014.
137. Helen Chan, O.; Stewart, B.H. Physicochemical and drug-delivery considerations for oral drug bioavailability. *Drug Discov. Today* **1996**, *1*, 461–473. [CrossRef]
138. 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. [CrossRef]
139. Litman, T.; Zeuthen, T.; Skovsgaard, T.; Stein, W.D. Structure-activity relationships of P-glycoprotein interacting drugs: Kinetic characterization of their effects on ATPase activity. *Biochim. Biophys. Acta Mol. Basis Dis.* **1997**, *1361*, 159–168. [CrossRef]
140. Kaliszan, R. QSRR: Quantitative Structure-(Chromatographic) Retention Relationships. *Chem. Rev.* **2007**, *107*, 3212–3246. [CrossRef] [PubMed]
141. Kouskoura, M.G.; Kachrimanis, K.G.; Markopoulou, C.K. Modeling the drugs' passive transfer in the body based on their chromatographic behavior. *J. Pharm. Biomed. Anal.* **2014**, *100*, 94–102. [CrossRef] [PubMed]
142. Winiwarter, S.; Ridderström, M.; Ungell, A.L.; Andersson, T.B.; Zamora, I. Use of Molecular Descriptors for Absorption, Distribution, Metabolism, and Excretion Predictions. In *Comprehensive Medicinal Chemistry II*; John, B.T., David, J.T., Eds.; Elsevier: Oxford, UK, 2007; pp. 531–554.
143. Raevsky, O.A. Physicochemical Descriptors in Property-Based Drug Design. *Mini Rev. Med. Chem.* **2004**, *4*, 1041–1052. [CrossRef] [PubMed]

144. Wolohan, P.R.N.; Clark, R.D. Predicting drug pharmacokinetic properties using molecular interaction fields and SIMCA. *J. Comput. Aided Mol. Des.* **2003**, *17*, 65–76. [CrossRef] [PubMed]
145. Platts, J.A.; Abraham, M.H.; Hersey, A.; Butina, D. Estimation of molecular linear free energy relationship descriptors. 4. Correlation and prediction of cell permeation. *Pharm. Res.* **2000**, *17*, 1013–1018. [CrossRef]
146. Deconinck, E.; Verstraete, T.; Van Gyseghem, E.; Vander Heyden, Y.; Coomans, D. Orthogonal Chromatographic Descriptors for Modelling Caco-2 Drug Permeability. *J. Chromatogr. Sci.* **2012**, *50*, 175–183. [CrossRef]
147. Varma, M.V.S.; Obach, R.S.; Rotter, C.; Miller, H.R.; Chang, G.; Steyn, S.J.; El-Kattan, A.; Troutman, M.D. Physicochemical Space for Optimum Oral Bioavailability: Contribution of Human Intestinal Absorption and First-Pass Elimination. *J. Med. Chem.* **2010**, *53*, 1098–1108. [CrossRef]
148. 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. Pharm.* **2014**, *11*, 716–726. [CrossRef]
149. Kim, D.C.; Burton, P.S.; Borchardt, R.T. A correlation between the permeability characteristics of a series of peptides using an in vitro cell culture model (Caco-2) and those using an in situ perfused rat ileum model of the intestinal mucosa. *Pharm. Res.* **1993**, *10*, 1710–1714. [CrossRef]
150. Goetz, G.H.; Shalaeva, M.; Caron, G.; Ermondi, G.; Philippe, L. Relationship between Passive Permeability and Molecular Polarity Using Block Relevance Analysis. *Mol. Pharm.* **2017**, *14*, 386–393. [CrossRef]
151. 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. [CrossRef] [PubMed]
152. Foley, D.W.; Rajamanickam, J.; Bailey, P.D.; Meredith, D. Bioavailability Through PepT1: The Role of Computer Modelling in Intelligent Drug Design. *Curr. Comput. Aided Drug Des.* **2010**, *6*, 68–78. [CrossRef] [PubMed]
153. Meshali, M.M.; Abdel-Aleem, H.M.; Sakr, F.M.; Nazzal, S.; El-Malah, Y.; El-Malah, Y. In vitro phonophoresis: Effect of ultrasound intensity and mode at high frequency on NSAIDs transport across cellulose and rabbit skin membranes. *Pharmazie* **2008**, *63*, 49–53. [PubMed]
154. Avdeef, A. The rise of PAMPA. *Expert Opin. Drug Metab. Toxicol.* **2005**, *1*, 325–342. [CrossRef] [PubMed]