Applying Linear and Non-Linear Methods for Parallel Prediction of Volume of Distribution and Fraction of Unbound Drug

Eva M. del Amo¹,², Leo Ghemtio¹, Henri Xhaard¹, Marjo Yliperttula², Arto Urtti¹, Heidi Kidron¹*²

¹Centre for Drug Research, Faculty of Pharmacy, University of Helsinki, Helsinki, Finland; ²Division of Biopharmaceutics and Pharmacokinetics, Faculty of Pharmacy, University of Helsinki, Helsinki, Finland

Abstract

Volume of distribution and fraction unbound are two key parameters in pharmacokinetics. The fraction unbound describes the portion of free drug in plasma that may extravasate, while volume of distribution describes the tissue access and binding of a drug. Reliable in silico predictions of these pharmacokinetic parameters would benefit the early stages of drug discovery, as experimental measuring is not feasible for screening purposes. We have applied linear and nonlinear multivariate approaches to predict these parameters: linear partial least square regression and non-linear recursive partitioning classification. The volume of distribution and fraction of unbound drug in plasma are predicted in parallel within the model, since the two are expected to be affected by similar physicochemical drug properties. Predictive models for both parameters were built and the performance of the linear models compared to models included in the commercial software Volsurf+. Our models performed better in predicting the unbound fraction (Q² 0.54 for test set compared to 0.38 with Volsurf+ model), but prediction accuracy of the volume of distribution was comparable to the Volsurf+ model (Q² 0.70 for test set compared to 0.71 with Volsurf+ model). The nonlinear classification models were able to identify compounds with a high or low volume of distribution (sensitivity 0.81 and 0.71, respectively, for test set), while classification of fraction unbound was less successful. The interrelationship between the volume of distribution and fraction unbound is investigated and described in terms of physicochemical descriptors. Lipophilicity and solubility descriptors were found to have a high influence on both volume of distribution and fraction unbound, but with an inverse relationship.

Introduction

The extent of drug distribution determines the access of a drug to its sites of action and to other tissues, which might give rise to adverse effects. A primary parameter for drug distribution is the volume of distribution (Vd) that is defined as:

\[ V_d = \frac{A}{C} \]

where A is the amount of drug in the body, and C is the drug concentration in plasma (both free drug and protein-bound drug). Volume of distribution is an apparent volume that increases with concentration in plasma (both free drug and protein-bound drug). However, only the fraction of free drug in plasma (fu) is described by the ratio Cu/C. Consequently, extensive drug binding outside the blood vessels leads to increasing values of A/C ratio. As tissue binding of drugs varies considerably, volume of distribution displays a wide range of values. For example, erythropoietin is confined to the vascular space presenting a Vd of 4 L (approximately the anatomical volume of vascular space) [1], while hydroxychloroquine with a Vd of 49 000 L strongly accumulates into the cells and tissues [2]. Volume of distribution at steady state (Vss) is measured at equilibrium, therefore, it describes the molecular tissue binding more reliably than other volume of distribution parameters that are dependent on the time after measurement. Vss depends on the access of the drug to the cells and tissues, its affinity to plasma proteins and tissue components, and number of binding sites in plasma and tissues.

Drug concentration in plasma (C) includes both unbound (Cu) and protein-bound drug in plasma. However, only the fraction of free drug in plasma permeates across the cellular membranes and vascular walls in most tissues. The free fraction of drug in plasma (fu) is described by the ratio Cu/C. Likewise the drug in the tissues also includes both free (Ct) or tissue bound parts. The unbound fraction of drug in tissues is: fuT = CuT/CT, where CuT is the total drug concentration in the tissue. Drug binding to plasma proteins and tissue components influences drug partitioning between the tissues and plasma. Thus, Vss can be presented using the following equation:

\[ V_{ss} = V_p + V_{T1} \left( \frac{f_u}{f_uT1} \right) + V_{T2} \left( \frac{f_u}{f_uT2} \right) \ldots \]
where \( V_p \) is the anatomical volume of plasma and \( V_T \) is the true anatomical volume of each tissue. \( V_a \) depends on the anatomical volumes of the tissues, and the relative extent of drug binding in the plasma and tissues described as \( f_a/f_T \) ratios.

As volume of distribution describes the extent of drug distribution, it is important to predict its value early in drug development before experimental measuring in humans. \( V_a \) in humans may be extrapolated from the in vitro animal data that is obtained during the drug discovery process, but computational approaches are useful at early stages before animal data has been collected. The volume of distribution used for computational modeling should be collected from intravenous and not from oral pharmacokinetics studies as in some cases [3], [4]. The benefit of intravenous administration is the defined quantity of the drug that is subject to distribution, which avoids the uncertainty associated with incomplete bioavailability after extravascular administration.

Even though quantitative structure-property relationship (QSPR) has been widely used for prediction of \( V_a \) [3–16], it remains a challenging problem that has not been adequately solved. The early attempts for predicting volume of distribution were based on small data sets and did not specify the type of volume of distribution that was used as the endpoint or in some cases used several types of volume of distribution for the model building [3-8], [11], [13], [14], [17]. In 2008, a major advance was the publication of a clean, manually curated dataset of \( V_a \) [18] that subsequently has been used successfully to build predictive models for \( V_a \) [12], [16].

The main difference in the work presented here compared to the previously published models of \( V_a \) is that we have included another pharmacokinetic parameter, \( f_u \), to the modeled responses. \( f_u \) in plasma depends on the binding affinity and capacity of plasma proteins, which also affect the volume of distribution. The fraction of unbound drug in plasma can be estimated relatively easily in vitro, but computational models for predicting \( f_u \) are also available [19-21]. The VolSurf+ software includes prediction tools for both volume of distribution and plasma protein binding, however, there is limited information of the methodology behind the models and their prediction capacity have not been evaluated in an unbiased manner in the literature. The two parameters, \( V_a \) and \( f_u \), are expected to be affected by similar physicochemical drug properties and our hypothesis was that modeling them in parallel would benefit their prediction. We have applied both linear and nonlinear multivariate approaches: linear partial least square (PLS) regression combined with principal component analysis (PCA) and non-linear recursive partitioning (RP) classification. RP has been shown to perform well when dealing with complex endpoints associated with multiple mechanisms, while PLS allows many responses (in our case \( V_a \) and \( f_u \)) to be incorporated in one regression model, but to our knowledge, this approach has not been used previously in pharmacokinetic QSPR modeling.

Materials and Methods

1. Data Set

The initial dataset collated by Obach and co-workers [18] contains 670 compounds with \( V_a \) and \( f_u \) values determined after intravenous administration to healthy people. The collection steps, the quality and the diversity of the data have been meticulously detailed in the publication.

The 2D structures of the compounds were obtained from the ACD/Dictionary version 11 [22] or the PubMed compound database (http://www.ncbi.nlm.nih.gov/pcmcompound Accessed 2010 October). If the compounds were represented as salts in the 2D structure, the counter ion was discarded. The 3D structures were generated using Concuor within SYBYL 8.0 [23]. A set of 648 drugs with both 2D and 3D structures were obtained. For the remaining 22 compounds in Obach’s data set either a 2D structure or minimized 3D structure was not obtained or it was not possible to calculate descriptors from the structures. The \( V_a \) of artesunate was corrected to 1.5 L/kg based on the work of White [24]. Furthermore, we excluded ibandronic, pamodronic, risedronic and zolendronic bisphosphonates from the set, since these compounds are sequestered to the bones, preventing their detection in the plasma, and leading to underestimated values of \( V_a \) [25]. The antimalarial drugs hydroxychloroquine and chloroquine have \( V_a \) values of 700 L/kg and 140 L/kg, respectively. These values are far beyond the range of other \( V_a \) values (0.035–60 L/kg) and they were excluded to avoid biasing the model.

The final data set of 642 drugs (Figure 1) displays \( V_a \) values of 0.035–60 L/kg and \( f_u \) values (541 drugs) of 0.0002–1.

2. Calculation of Molecular Descriptors

In this study, molecular descriptors were calculated using ACDlabs [26], VolSurf+ [17] and MOE [27]. Input molecular structures were two-dimensional for ACDlabs and three-dimensional for VolSurf+ and MOE, for the later Gasteiger-Huckel charges were added. Identical descriptors (i.e. molecular weight, molecular volume) were excluded before combining descriptor sets for modeling. The descriptors that were used for model building are listed in Table 1 and the calculated descriptor values for the data set are available in File S1.

3. PCA and PLS Regression Models

QSPR models were built using linear multivariate analysis tools PCA and PLS (Simca plus Version 10.5) [28]. All descriptors were transformed with unit variance scaling and mean centering before PCA and PLS analysis. Moreover, the descriptors with a broad range or unequal distribution across the range were logarithmically transformed to obtain better distributions. Three sets of molecular descriptors were assembled for the regression modeling: (1) ACDlabs descriptors and MOE logS descriptor; (2) VolSurf+ descriptors; (3) the combination of ACDlabs, MOE and VolSurf+ descriptors.

A workflow of the modeling process is presented in Figure 2. Before modeling, a foreign set of 101 drugs was randomly excluded from the final 642 compound set. The descriptor matrix of the remaining 541 drugs was analysed with PCA to identify the drugs that fall outside the general chemical space of the compound set and descriptors that should be excluded from the model (model calibration). Drugs that were outliers based on their distribution in the PCA plot and whose descriptor values fell outside the boundaries outlined in Table 2 were excluded. Based on the scatter plot of the final PCA plot, an external test set (Figure 3) of 101 compounds representative of the chemical space was selected. The external set comprises molecules within the chemical space of the model, while the foreign set, which was selected before the PCA and model calibration, also includes compounds outside the chemical space used for model building. The remaining compounds constitute the training set for the PLS model building (365 drugs for model 1; 357 drugs for model 2; 361 drugs for model 3). The training sets were used to build PLS models that relate the descriptors to the two simultaneously modelled responses, log \( V_a \) and \( f_u \). During initial stages of the analytical process, the number of highly correlated variables observed in the PLS weight plot was gradually reduced in order to equilibrate the influence of the overall set of descriptors on the responses. Subsequent models with improved statistic parameters were
obtained and variables deemed least influential to the modelled pharmacokinetic parameters were excluded. The decisions were based on the PLS weight plot and confirmed by the variable importance plot results. Moreover, the distribution of the drugs was followed up by the PLS score and Dmod plots, in order to detect outliers.

4. Recursive Partitioning Classification Models

A RP analysis was carried out using Discovery Studio version 3.5 (Accelrys Inc.) to develop decision trees that categorize the compounds into classes that are based on the $V_{ss}$ values or both $V_{ss}$ and $f_u$ values (Table 3 and 4). Volume of distribution is defined by drug interactions with the main volumes in the body: extracellular space and cellular tissue space. We used these anatomical volumes as rough guidance to classify the volumes into three classes. Class 1 represents the volume of the extracellular fluid (0–0.3 L/Kg), class 2 represents $V_{ss}$ values that take into consideration distribution to the tissues (0.3–1 L/Kg), and class 3 values of $V_{ss}$ represent significant binding to the cellular components (>1 L/Kg). However, it should be noted that $V_{ss}$ is an apparent volume that does not strictly obey anatomical volumes, therefore the anatomical distribution of the compounds cannot be concluded from the $V_{ss}$. Distribution of compounds into the three classes is shown in Figure 1A. When both $V_{ss}$ and $f_u$ values were predicted, each class

![Figure 1. Distribution of compounds in the data set.](image-url)

**Table 1. The descriptors included in modeling.**

| ACDlabs descriptors | Volsurf- descriptors | MOE descriptors |
|---------------------|---------------------|----------------|
| ALogD5              | V                   | LgS5           |
| ALogD55             | S                   | LgS6           |
| ALogD7              | R                   | LgS7           |
| ALogD7.4            | G                   | LgS7.5         |
| APSA                | W1                  | LgS8           |
| HDonors             | W2                  | LgS9           |
| HAcceptors          | W3                  | LgS10          |
| FR8                 | W4                  | Lxlgs5         |
| Rule Of 5           | W5                  | DRDRO          |
| Molar Volume        | W6                  | L11gLs5        |
| MW                  | W7                  | L2lgs5         |
| Surface Tension     | W8                  | L3lgS5         |
| Polarizability      | D1                  | L4lgS5         |
| C ratio             | D2                  | L4lgS5         |
| N ratio             | D3                  | L4lgS5         |
| NO ratio            | D4                  | L4lgS5         |
| Num Rings           | D5                  | L4lgS5         |
| Num Ar Rings        | D6                  | D4             |
| D7                  | CW3                 | LgS4           |
| D8                  | CW4                 | D7             |
|                     | CP                  | D8             |

doi:10.1371/journal.pone.0074758.g001

doi:10.1371/journal.pone.0074758.t001
Table 2. Statistical parameters of the PCA models and the chemical boundaries chosen during the PCA modelling.

| A  | R²X | Q²X | Criteria of model calibration |
|----|-----|-----|--------------------------------|
| Model 1 | 7  | 0.90 | 0.58 | MW≤940 | PSA≤205 | POL≤71 | HBD≤15 | HBA≤-7.71 | LogS≤0.38 |
| Model 2 | 7  | 0.79 | 0.73 | MW≤940 | WO4≤100 | WO6≤2 | PSA≤205 | SOLY≤9.35 | V≤1353 | POL≤71 | LogS9≤5.3 | W4≤483 |
| Model 3 | 7  | 0.79 | 0.73 | MW≤940 | WO4≤100 | PSA≤205 | SOLY≤0.93 | V≤1353 | POL≤71 | LogS9≤5.3 | W4≤483 |

*MW: molecular weight; ^PSA: polar surface area; ^POL: polarizability; ^HBD: hydrogen bond donors; ^HBA: hydrogen bond acceptors; ^LogS: log of solubility; ^WO4 and ^WO6: hydrogen bond donor volume at different energy levels; ^SOLY: intrinsic solubility; ^V: molecular volume; ^LogS9: log of solubility at pH 9; ^W4: hydrophilic volume; ^MV: molar volume.

doi:10.1371/journal.pone.0074758.t002

Figure 2. Flowchart of the work process to obtain regression and classification models for Vss and fu. MFE = mean fold error, SI = Sensitivity, SPEC = specificity.
doi:10.1371/journal.pone.0074758.g002
was further divided into compounds with low to intermediate (<0.7) or high (>0.7) \( f_u \). Compounds with missing \( f_u \) values were addressed by assigning them the mean value of all \( f_u \) values and distributing them equally in the training and external test set, which is a standard approach to handle missing values in RP analysis. In our study, balanced forest of RP was used, since it is the appropriate method for imbalanced data [29]. This type of RP contains a relatively small number of trees (in average 10) using a separate bootstrap sample of the original data for each tree. For each tree, the number of members in all classes is equal to the number of members in the smallest class. The number of descriptors that was used as split criterion within each tree was set to the square root of total descriptors. The weighing method was set to “uniform” and the equalize class sizes to true. All others parameters were set to default.

A training set was used to build the decision trees and an external test set was utilized to evaluate the predictive power of the models. To generate the training and external test set for RP analyses, all compounds were first clustered by similarity based on root mean square deviation and each cluster was divided into training and test sets to ensure that both sets included compounds from each cluster. The data set used to train the model consisted of 382 compounds, while 260 compounds were used as an external test set (Figure 2).

5. Model validation

The prediction accuracy of the PLS models was determined by internal and external validation. The internal validation is based on the cross-validation value \( Q^2 \) that is calculated by leaving out 1/7 of the data, and predicting these compounds based on a model trained by the remaining data. The external validation is conducted with the external test set. The model was used to predict the log \( V_m \) and \( f_u \) of the external test set. The predicted responses were plotted against the observed responses (i.e. experimental \( V_m \) and \( f_u \)). The \( R^2 \) value of the regression line for the plot was considered as the \( Q_e^2 \) (goodness of prediction of the external test set).

We estimated the predictive ability of the RP classification models using out-of-bag statistics. The external test set was used to estimate the fitting ability of the model on a new dataset that was not used in the model construction. The performance of the RP models is based on three metrics: true positive rate (recall or sensitivity), specificity, and the area under the curve (AUC) of the receiver operating characteristics (ROC) plot [30]. AUC represents the probability that a classifier will be estimated correctly, with values >0.5 indicating better than random prediction and 1 signifying perfect prediction. In the case of more than two classes (multiclassification), a confusion matrix is a square of NxN, where N is the number of classes. AUC is computed as defined by Hand and Till (2001) as an average over components generated from several ROC plots for a Y property and cannot be plotted [30]. For instance, when N (A, B, C) is 3, the classifier’s performance is computed per class as follows for class A:

Table 3. Division of training and test set compounds into three classes according to observed \( V_{ss} \).

|               | Class 1 \( V_{ss} = 0-0.3 \) L/kg | Class 2 \( V_{ss} = 0.3-1 \) L/kg | Class 3 \( V_{ss} > 1 \) L/kg | Total |
|---------------|-----------------------------------|----------------------------------|-------------------------------|-------|
| Training      | 105                               | 96                               | 181                           | 382   |
| Test          | 62                                | 71                               | 127                           | 260   |
| Total         | 167                               | 167                              | 308                           | 642   |

doi:10.1371/journal.pone.0074758.t003

Figure 3. PCA score plot. The final PCA score plot obtained after model 3 calibration where the two principal components explain 27% and 20%, respectively, of the variance in the data set. The open squares represent the drugs in the external test set and the filled triangles the drugs in the training set. The ellipse depicts the 95% confidence region of the model (Hotelling T^2).

doi:10.1371/journal.pone.0074758.g003
of compounds within the applicability domain. Increasing $Z$ will
these distances were calculated. $Z$ is an arbitrary parameter to
3. Y-randomization test
In addition to the internal and external validation, the Y-
randomization test (response permutation test) was performed,
which estimates the robustness of models [31]. The X data are left
intact, whereas the Y data are permuted to appear in a different
order (random shuffling). A model is then fitted to the permuted Y-
data and the model statistics are computed for the derived model.
It is expected that the models from randomized activities would have
significantly lower accuracy values.

6. The applicability domain of models
An applicability domain (AD) of the model is needed to avoid
making predictions for compounds, which differ substantially from
the training set molecules. The AD is used to estimate which
compounds are suitable for model predictions and avoid
unjustified extrapolation of predictions. We used a method
introduced by Zhang et al. (2006) for defining the AD based on
the distribution of similarities between each compound and its
nearest neighbours in the training sets [32]. The AD was
calculated as follows:

$$AD = <d> + Z\sigma$$

The average of Euclidean distances between all points of the
training set were calculated from Similarity and Clustering Canvas
of Schrödinger modeling package [33], with 32 bit linear Daylight
fingerprint. Data for estimation of the Euclidean distance and
application of the AD on new compounds are available in
Files S2-S5. Then, using the distances lower than the average, a
new average distance $<d>$ and standard deviation $\sigma$ between
these distances were calculated. $Z$ is an arbitrary parameter to
control the significance level and considerably affects the number
of compounds within the applicability domain. Increasing $Z$ will
include compounds that are more dissimilar in the AD. We set the
value of $Z$ to 0.7 to calculate the compounds within the AD of the
models in the foreign test set.

Results

1. PLS Regression Models
The linear regression model of log $V_{ss}$ and $f_u$ was attempted
with three descriptor sets: (1) 19 descriptors from ACDlabs and
MOE, (2) 121 descriptors from VolSurf+ and (3) 140 descriptors
from the combination of the two previous sets. The three sets were
first analyzed with PCA. In Table 2, the final PCA model statistics
for the three strategies are presented as well as the criteria of
selection chosen in each case. In Figure 3, the score plot of the
final PCA model of data set 3 is shown as an example. Similar
plots were obtained for the other data sets.

The statistical values of the final models are present in Figure 2.
Model 1 resulted in a non-predictive model, yielding a Q^2Y
value of $Z$ to 0.7 to calculate the compounds within the AD of the
models in the foreign test set.

$$V_{ss} = 0.1521 - 0.1173L1LgS + 0.2858L3LgS$$
$$- 0.0123SOLY + 0.0122LOGPn - Oct + 0.0463LgD9$$
$$- 0.0083WN5 - 0.0002W1 + 0.2811ID3 + 0.0026A$$

Model 3.

$$V_{ss} = 0.2464 + 0.0909LgS3 - 0.269LgS10 - 0.0099 \log S$$
$$+ 0.3894L3LgS + 0.0465LgD10 + 0.0514ALogD5.5$$
$$+ 0.0010%FU10 + 0.0004MV - 0.0005W1 + 0.0023D4$$
$$+ 0.0174HD$$
Where L1LgS and L3LgS are solubility profiling coefficients, logS is the logarithm of solubility, LgS3 and LgS10 are the logarithms of solubility at pH 3 and pH 10, respectively, SOLY is intrinsic solubility, LOGP n-Oct is the partitioning coefficient in octanol/water, LgD9, LgD10 and ALogD5.5 are distribution coefficients at pH 9, pH 10 and pH 5.5, respectively, WN5 is hydrogen bond acceptor volume, W1 is hydrophilic volume, ID3 is hydrophobic interglycogen moment, A is amphiphillic moment, %FU10 is % of fraction unionized at pH 10 (not to be confused with fu), MV is molar volume, D4 is hydrophobic volume and HD is hydrogen bond donor.

Model 2 and model 3 were internally validated by cross-validation, gaining Q^2 values of 0.58 and 0.55, respectively. In external validation of the models we determined their accuracy in predicting log Vss and fu with the external test sets. In log Vss prediction by model 2, two outliers were excluded (ribavirin and bilobalide), while in fu prediction by model 2, four outliers (acetylcysteine, amiodarone, aripiprazole, repaglinide) were excluded and in fu prediction by model 3, five outliers were excluded (ethambutol, atovaquone, beclomethasone dipropionate, drotaquine, irbesartan). The statistical results of the predictions are presented in Figure 2. The Y-randomization test after 50 permutations provided R^2Y- and Q^2Y-intercepts smaller than the recommended limits of 0.3 and 0.05 for both log Vss and fu, respectively (data not shown).

The AD was estimated from the compounds belonging to the training set as:

\[
A = 24.22 + Z2.03
\]

With Z = 0.7, AD is 25.641 that represent the maximum distance between compounds in the training set and new compound to be predicted. The compounds in the foreign test set that fell inside this AD were selected, yielding a set of 35 drugs for model 2, and 30 drugs for model 3. The statistical parameters of log Vss and fu predictions for the foreign set are presented in Table 5 and plots of the observed and predicted responses of
model 3 and VolSurf+ ADME models are presented in Figure 5. A comparison of the predicted and the observed values is found Table S1. Increasing Z increases the number of compounds in the foreign test set that are considered to be within the applicability domain but decreases the accuracy of prediction due to inclusion of dissimilar nearest neighbors (Figure 6).

2. RP Classification Models
The AUC for the in-bag training data for all trees in the forest model is 0.96 and 0.92, and the out-of-bag AUC is 0.81 and 0.79 for the Vss and Vss & fu models, respectively. The in-bag results use predictions for the records used to train the tree, while the out-of-bag results use predictions for the left-out records. The statistics for the training set data presented in Figure 2 are derived from the in-bag results. The external test set including 260 compounds (described in Methods section) was used to evaluate the predictive ability of the two models. All compounds were classified according to their Vss or Vss & fu values without applying AD. The overall prediction accuracy, calculated as ROC curve, was 0.78 and 0.82, respectively, and the sensitivity and specificity values are presented

| Table 5. Statistical parameters for log Vss and fu predictions of the foreign set compounds inside the applicability domain of the models, calculated with Z = 0.7. |
|-----------------|-----------------|-----------------|
| Log Vss prediction of foreign set | fu prediction of foreign set |
| Q² | MFE | %<2-fold | Q² | MFE | %<2-fold |
| Model 2 | 0.62 | 2.85 | 60 | 0.59 | 5.58 | 54 |
| Model 3 | 0.70 | 2.41 | 67 | 0.54 | 7.04 | 52 |

doi:10.1371/journal.pone.0074758.t005

Figure 5. Log Vss and fu prediction plots of model 3 versus VolSurf+ ADME models (Vd and protein binding). Dot lines represent 2-fold error, dashed lines represent 3-fold error and long dash lines represent 5-fold error. MFE: mean fold error.
doi:10.1371/journal.pone.0074758.g005
in Figure 2. The confusion matrices are presented in Tables S2-S7.

In general, the sensitivity of the models is high for compounds with a very low or high volume of distribution, while compounds belonging to class 2, with \( V_{ss} \) values between 0.3 and 1 L/kg are more difficult to classify correctly. The \( V_{ss} \) model performed well on the training set, with sensitivity 0.79 in class 2, but less than half of the class 2 compounds in the training set (42 of 93 compounds, leading to a sensitivity of 0.45) were predicted to the correct class in the out-of-bag results (Table S3). Similarly, the model was able to identify class 1 and class 3 compounds form the external test set (sensitivity 0.71 and 0.81, respectively), while recognition of class 2 test set compounds was not as successful (sensitivity 0.32, Figure 2, Table S6). Interestingly, in the \( V_{ss} \) & \( f_u \) model, compounds with high \( f_u \) were predicted more accurately, with 10 of 17 compounds of the test set compounds correctly classified (sensitivity 0.59), but only 7 of 51 compounds with low \( f_u \) (sensitivity 0.14) (Figure 2, Table S7). The \( Y \)-randomization test was performed four times, and the AUC values for the model using the data set with experimental \( V_{ss} \) and \( V_{ss} \) & \( f_u \) values were significantly higher than those obtained from the dataset with randomized values (data not shown), indicating that our models are statistically robust. The AD was applied to the test set and its effect analyzed on the \( V_{ss} \) & \( f_u \) model (Figure 6). The prediction accuracy was highest with low \( Z \) cutoff, as expected, and slowly decreasing as the cutoff was increased to 1. However, increasing the cutoff from 1 to 20 did not markedly affect the prediction accuracy, while increasing the coverage of the test set from 39% to 100%. The small decrease in prediction accuracy is probably due to the cluster-based approach used to select the training and test set (described in Methods) that make the chemical space covered by two set similar.

One aid for interpretation of forest models is a set of descriptor importance measures, which indicate the relative importance of the descriptors in distinguishing among the different classes in the data. The percent selection frequency empirically appears to best distinguish truly important descriptors from others. It represents the percent of the time that the descriptor was selected for a split when a split was possible. A summary of descriptors ranked as top 10 based on their frequency of occurrences in the models are given in Table 6. It should be noted that size, polarity and lipophilicity are predominant in all models. The simple importance measures reported here are known to have bias in some cases [34]. However, if all descriptors have the same character as in our cases (e.g. they are all continuous numerical properties), then bias is generally not an issue.

**Discussion**

We have predicted \( V_{ss} \) and \( f_u \) with linear PLS models and nonlinear RP classification models, aiming for models that rely on in silico descriptors only and therefore are suitable for screening. \( V_{ss} \) is affected by the \( f_u \) in plasma, and we wanted to explore if predicting both parameters in parallel would help to find relevant physicochemical descriptors affecting these parameters. PLS can easily be used to correlate descriptors with several related responses, but to our knowledge, this approach has not been used in pharmacokinetic QSAR modeling.

The RP classification model was reasonably successful in classifying compounds with high (≥1 L/kg) or low (0–0.3 L/kg) \( V_{ss} \), while it had difficulties to identify the compounds with moderate (0.3–1 L/kg) \( V_{ss} \). Interestingly, the level of binding to plasma proteins had an influence on the prediction accuracy, which was seen most clearly in the moderate \( V_{ss} \) class, where compounds with high \( f_u \) were correctly predicted in 59% of the test set, but only 14% of those with low \( f_u \) (Figure 2, Table S7). The attempt to create a PLS model for \( V_{ss} \) and \( f_u \) (model 1) starting with only 19 descriptors from ACDlabs and MOE was not successful, but using a wider range of descriptors from Volsurf resulted in a predictive model (model 2) (Table 1 and Figure 2). The combination of all descriptors to model 3 did not significantly improve the prediction of the external set (\( V_{ss} \) \( Q_e^2 \) = 0.50, \( f_u \) \( Q_e^2 \) = 0.54) compared to model 2 (\( V_{ss} \) \( Q_e^2 \) = 0.52, \( f_u \) \( Q_e^2 \) = 0.51) (Figure 2). However, model 3 had better success in predicting the \( V_{ss} \) of the compounds in the foreign set (model 3 \( Q_e^2 \) = 0.70, model 2 \( Q_e^2 \) = 0.62) (Table 5). Notably, the prediction of the compounds in the foreign set within the AD was better than for the

Figure 6. The effect of the AD on the prediction accuracy and chemical space coverage. Dashed black line: \( Q_e^2 \) of the \( V_{ss} \) foreign set predicted with PLS model 3. Dashed purple line: percentage of compounds from the \( V_{ss} \) foreign set predicted with PLS model 3. Black line: \( Q_e^2 \) of the \( f_u \) foreign set predicted with PLS model 3. Purple line: percentage of compounds from the \( f_u \) foreign set predicted with PLS model 3. Dotted black line: \( Q_e^2 \) of the test set predicted with RP classification. Dotted purple line: percentage of compounds from the test set predicted with RP classification.

doi:10.1371/journal.pone.0074758.g006
external set for both model 2 and 3. The AD was not used to filter compounds for prediction in the external set, which might be one reason for the improved performance on the foreign set. The use of an AD prevents extrapolation beyond the limits of chemical space that was used to build the model and can be used to identify the compounds for which predictions are reliable.

The impact of the descriptors on the responses can be observed graphically in the PLS weight plot (model 3 in Figure 4, model 2 in Figure S1). In model 3, the descriptors L3LgS, %FU10 and LgD10 have the highest positively correlated impact to Vss (L3LgS, LogP n-oct and LgD9 in model 2, Figure S1), while LgS3, D4, Molar Volume and ALOGD5.5 have a more moderate positive influence on Vss (A in model 2). LgS10 has the largest negative correlation to Vss (L1LgS in model 2), while W1, HDonors and LogS have smaller negative correlation in model 3 (Wn5, SOLY, W1 and ID3 in model2). On the other hand, LogS, LgS3 and LgS10 have the highest positive correlation with fu (SOLY in model 2), while LogD10 and ALOGD5.5 have the highest negative correlation (LogD9 and LOGP n-Oct in model 2). All in all, this suggests that charge and lipophilicity of the drug affect drug distribution, albeit with an inverse relationship. Thus, the lipophilicity descriptors have high correlation with the two responses, positive with Vss and negative with fu, while reversely, the charge and solubility descriptors have negative correlation with Vss and positive with fu. There is a complex relationship between fu and Vss and increasing the fu of a compound does not inevitably lead to a higher volume of drug distribution, as is stated in many pharmacokinetic textbooks [35], [36]. This is easy to understand, since structural changes influencing the drugs ability to bind to plasma proteins may also affect the tissue binding of the drug.

Similar descriptors were found to be important in both the RP classification models (Table 6) and the PLS models. These include solubility descriptors, LogD at pH 9 or 5, as well as hydrophilic and hydrophobic area and volume descriptors. Due to the complexity of Vss and fu, many descriptors were always required to yield good prediction capability. Previously, trends have been observed between Vss and LogP, polar surface area and hydrogen bond descriptors for the data set we have used [18]. Using the same data set, Berellini et al. (2009) found hydrogen bonding, LogD at pH 5–10, flexibility of the molecule and the VolSurf descriptors DRDRDO, DRDRAc to be important in their Vss model [12]. DRDRDO and DRDRAc are pharmacophoric descriptors of the maximum area of the triangles derived from Dry (DR), H-bond acceptor (AC) and H-bond donor (DO) points in a molecule. DRDRDO and flexibility were among the ten most influential descriptors in the RP models, but in the PLS models they did not have equally high importance. However, when comparing our descriptor selection to previous models of Vss it must be kept in mind that we have modeled both Vss and fu parameters. Therefore a comparison is not directly applicable as descriptors having high influence on one parameter, but no correlation with the other parameter, are likely to be removed in our models.

Outliers are usually interesting, and the analysis of outliers can sometimes give a deeper understanding of the mechanisms under investigation. However, it is difficult to analyse the outliers in this study, because we do not know the reason for their exceptional behavior. Deviations in Vss may be due to the active transport (influx or efflux) or compound specific binding to the tissues. As an example, let’s consider the outliers in the prediction of Vss by the PLS models (ribavirin, bilobalide, tamsulosin, decitabine). Ribavirin and decitabine are substrates of widely expressed nucleoside transporters, and extensive active transport might lead to outlier profiles of ribavirin and decitabine [37]. Tamsulosin is a substrate of alpha1 adrenergic receptors and bilobalide binds to GABA, glycine, and 5-HT3 receptors [38]. We cannot be sure, however, if these transport and binding phenomena take place substantially enough to cause exceptional Vss values. Clearly, Vss and fu are complex phenomena that are affected by numerous factors. Therefore, explanations for the outlier behavior are not on firm ground and the reasons can be identified only by extensive experimental work.

We compared the performance of our model 3 with the volume of distribution and plasma protein binding models available in the VolSurf+ package (Figure 5). As no AD is reported for the VolSurf+ models, we have applied our AD with the Z cutoff value of 0.7 to select the compounds from the foreign test set for both models. For the practical use of AD in Vss and fu prediction, see File S2. It should be noted that we are not aware of which compounds have been used to train the VolSurf+ model, and it is possible that some, or all, of the compounds used in our test set have been used to for that purpose. The same considerations apply for the VolSurf+ plasma protein binding model. Our model achieved higher accuracy than the VolSurf+ model in predicting fu (Q2 = 0.54 and

### Table 6. Most influential descriptors in the classification models.

| Descriptor | Type | fu | Vss | Comment |
|------------|------|----|-----|---------|
| Rule Of 5  | Drug like | 12 | 8.3 | |
| CD3        | Hydrophobic area | 31 | 6.4 | |
| FLEX_RB    | Size/Shape | 69 | 5.8 | |
| L1LgS      | Solubility | 53 | 5.7 | |
| CD6        | Hydrophobic area | 37 | 5.4 | DRDRDO Pharmacoforic |
| CW8        | Hydrophilic area | 20 | 5.0 | |
| LgS11      | Solubility | 66 | 4.5 | |
| C ratio    | Topology | 45 | 4.4 | |
| NO ratio   | Topology | 48 | 4.2 | |
| LgS5       | Solubility | 72 | 4.2 | |

DOI: 10.1371/journal.pone.0074758.t006
Parallel Prediction of $V_{ss}$ and $f_u$ for Drugs

$Q^2 = 0.38$, respectively) (Figure 5), while the prediction of $V_{ss}$ was comparable to the Volsurf+ model ($V_{ss}$, $Q^2 = 0.70$ and $Q^2 = 0.71$ for model 3 and Volsurf+ models, respectively). The best predictions with our model were obtained at $f_u$ values above 0.05. Predictions of the compounds with $f_u$ values above 0.03 in the data set had a MFE of only 2.2 for model 3, compared to 7.04 for the whole foreign set (Figure 5, table 5). The predictions at $f_u$ values below 0.05 give high FE values (>5-fold), whereas % error in this region is low. However, FE is pharmacologically a more relevant parameter, because the free drug concentration in plasma, $C_p$, is defined as $f_u \times C$. Therefore, 3-fold change in $f_u$ is expected to result in 3 fold change in $C_p$. Unfortunately, we do not have an explanation for the poor results for the compounds that have very low $f_u$ values, however, the compounds that were badly predicted by our models were also badly predicted by the Volsurf+ model (Table S1), suggesting that the exceptional behavior is drug dependent and not due to the model.

The physical complexity of the $V_{ss}$ and $f_u$ parameters makes their prediction very challenging, and we were not able to reach models with optimal predictability. One way to improve prediction accuracy is to build the model using a narrower range of more similar compounds. We divided the data set of 642 compounds based on structural features or chemical properties and used these data sets to build several sub-models. However, the models were not able to achieve much higher accuracy than the more global models presented here (data not shown), but presented a much narrower AD and therefore more limited use.

Conclusions

The PLS models of $V_{ss}$ showed similar performance to the commercial Volsurf+ model, while the $f_u$ prediction accuracy was slightly better. The RP classification models were able to distinguish between compounds with high or low $V_{ss}$ values, but accurate classification of moderate $V_{ss}$ or low $f_u$ values were not as successful. Due to the complex nature of $V_{ss}$ and $f_u$ parameters, a fairly large number of descriptors were needed for meaningful models. The advantages of the models compared to previous models is that they are based on a large set of structurally unrelated compounds, they are open, and they have a defined AD, which aids in identifying compounds for which reliable predictions can be made.

Supporting Information

Figure S1 PLS model 2 weight plot. (TIF)

Table S1 Table of predicted vs. observed values for foreign set with PLS models. (XLSX)

References

1. Lim VS (1991) Recombinant human erythropoietin in predialysis patients. Am J Kidney Dis 18: 34–37.
2. Tett SE, Catler DJ, Dey RO, Brown KF (1998) A dose-ranging study of the pharmacokinetics of hydroxy-chloroquine following intravenous administration to healthy volunteers. Br J Clin Pharmacol 26: 303–313.
3. Paul Y, Dhake AS, Parle M, Singh B (2010) In Silico Quantitative Structure Pharmacokinetic Relationship Modeling for Quinolone Drugs: Biological Half-Life. 22: 4880–4890.
4. Gleson MP, Waters NJ, Paine SW, Davis AM (2006) In silico human and rat Vss quantitative structure-activity relationship models. J Med Chem 49: 1953–1963.
5. Hirota S, Nakagome I, Hirota H, Yoshii F, Moriguchi I (1994) Non-congeneric structure-pharmacokinetic property correlation studies using fuzzy adaptive least-squares: volume of distribution. Biol Pharm Bull 17: 686–690.
6. Karalis V, Tsantili-Kakoulidou A, Macheras P (2002) Multivariate statistics of disposition pharmacokinetic parameters for structurally unrelated drugs used in therapeutics. Pharm Res 19: 1027–1034.
7. Karalis V, Tsantili-Kakoulidou A, Macheras P (2003) Quantitative structure-pharmacokinetic relationships for disposition parameters of cephalosporins. 20: 115–123.
8. Lombrano MD, Obach RS, Dicapua FM, Baker GA, Lu J, et al. (2006) A hybrid mixture discriminant analysis-random forest computational model for the prediction of volume of distribution of drugs in humans. J Med Chem 49: 2262–2267.
9. Ng C, Xiao Y, Putnam W, Lum B, Tropsha A (2004) Quantitative structure-pharmacokinetic parameters relationships (QSPPR) analysis of antimicrobial agents in humans using simulated annealing k-nearest-neighbor and partial least-square analysis methods. J Pharm Sci 93: 2535–2544.

Author Contributions

Conceived and designed the experiments: EMA LG HX MY AU HK. Performed the experiments: EMA LG. Analyzed the data: EMA LG HX MY AU HK. Contributed reagents/materials/analysis tools: HX MY AU. Wrote the paper: EMA LG HX MY AU HK.

Table S2 Confusion matrix in-bag training results for the $V_{ss}$ classification model. (DOCX)

Table S3 Confusion matrix out-of-bag training results for the $V_{ss}$ classification model. (DOCX)

Table S4 Confusion matrix in-bag training results for the $V_{ss}$ & $f_u$ classification model. (DOCX)

Table S5 Confusion matrix out-of-bag training results for the $V_{ss}$ & $f_u$ classification model. (DOCX)

Table S6 Confusion matrix external test results for the $V_{ss}$ classification model. (DOCX)

Table S7 Confusion matrix external test results for the $V_{ss}$ and $f_u$ classification model. (DOCX)

File S1 Final data set used for models. (SDF)

File S2 Instructions for use of applicability domain. (DOCX)

File S3 Training set for RP models. (SDF)

File S4 Training set for PLS model 2. (SDF)

File S5 Training set for PLS model 3. (SDF)

Acknowledgments

The authors would like to thank Dr. Ossi Korhonen and Dr. Jari Pajander (School of Pharmacy, University of Eastern Finland) for their valuable help in the initial stages of the project, Dr. Veli-Pekka Kanta for his readiness to discuss pharmacokinetic questions (School of Pharmacy, University of Eastern Finland) and Dr. Mikko Kolehmainen (Dpt of Environmental Science, University of Eastern Finland) for his experience in the use of multivariate analyses tools.

Author Contributions

Conceived and designed the experiments: EMA LG HX MY AU HK. Performed the experiments: EMA LG. Analyzed the data: EMA LG HX MY AU HK. Contributed reagents/materials/analysis tools: HX MY AU. Wrote the paper: EMA LG HX MY AU HK.
10. Turner JV, Maddalena DJ, Cutler DJ (2004) Pharmacokinetic parameter prediction from drug structure using artificial neural networks. Int J Pharm 270: 209–219.
11. Balakin KV, Ivanenkov YA, Srechuk NP, Zeluchenchkou AA, Ekims S (2005) Comprehensive computational assessment of ADME properties using mapping techniques. Curr Drug Discov Technol 2: 99–113.
12. Berellini G, Springer C, Waters NJ, Lombardo F (2009) In silico prediction of volume of distribution in human using linear and nonlinear models on a 669 compound data set. J Med Chem 52: 4489–4495.
13. Fateni MH, Ghobaramezad Z (2011) Estimation of the volume of distribution of some pharmacologically important compounds from their structural descriptors. 76: 1003–1014.
14. Demir-Kavuk O, Benzien J, Muegge I, Knapp E (2011) DemQSAR: predicting human volume of distribution and clearance of drugs. J Comput Aided Mol Des 25: 1121–1133.
15. Zhivkova Z, Doytchinova I (2012) Prediction of steady-state volume of distribution of acidic drugs by quantitative structure-pharmacokinetics relationships. J Pharm Sci 101: 1253–1266.
16. Gombar VK, Hall SD (2013) Quantitative Structure-Activity Relationship Models of Clinical Pharmacokinetics: Clearance and Volume of Distribution. J Chem Inf Model.
17. Cruciani G, Crivori P, Currant P, Testa B (2000) Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. 503: 17–30.
18. Obach RS, Lombardo F, Waters NJ (2008) Trend Analysis of a Database of Intravenous Pharmacokinetic Parameters in Humans for 670 Drug Compounds. Drug Metab Dispos 36: 1385–1405.
19. Gleeson MP (2007) Plasma protein binding affinity and its relationship to molecular structure: an in silico analysis. J Med Chem 50: 101–112.
20. Moda TL, Montanari CA, Andricopulo AD (2007) Hologram QSAR model for the prediction of human oral bioavailability. Bioorg Med Chem 15: 7738–7745.
21. Chen C, Liaw A, Breiman L (2004) Using Random Forest to Learn Imbalanced Data. Technical report 666, Department of Statistics, University of Berkeley.
22. ACD/Dictionary (2013) version 9.3, Advanced Chemistry Development, Inc., Toronto, ON, Canada. Available: www.acdlabs.com.
23. SI tính 8.8, Tripos International, St. Louis, Missouri, USA.
24. White NJ (1994) Clinical pharmacokinetics and pharmacodynamics of artemisinin and derivatives. Trans R Soc Trop Med Hyg 88 Suppl 1: S41–S.
25. Watts NB, Diab DL (2010) Long-term use of bisphosphonates in osteoporosis. J Clin Endocrinol Metab 95: 1555–1565.
26. ACDlabs software, Advanced Chemistry Development, Inc., Toronto, On, Canada. Available: www.acdlabs.com.
27. MOE, Molecular Operating Environment, Chemical Computing Group: Montreal, Canada, 2008.
28. ACD/abs software, Advanced Chemistry Development, Inc., Toronto, On, Canada. Available: www.acdlabs.com.
29. Chen C, Liaw A, Breiman L (2004) Using Random Forest to Learn Imbalanced Data. Technical report 666, Department of Statistics, University of Berkeley.
30. Hand DJ, Till RJ (2001) A Simple Generalisation of the Area Under the ROC Curve for Multiple-Class Classification Problems. Machine Learning 45: 171–186.
31. Trophys A, Gramatica P, Gombar VK (2003) The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSAR Models. 22: 69–77.
32. Zhang S, Goldsbrough A, Oldfield, H, Trophys A (2006) A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models. 46: 1984–1995.
33. Stute 2012, Matoss (2012) version 9.3, Schrödinger, LLC, New York, NY.
34. Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8: 25.
35. Lüllmann H, Mohr K, Hein L, Bieger D (2005) Color Atlas of Pharmacology. p. 262.
36. Rowland M, Tozer TN (2011) Clinical Pharmacokinetics and Pharmacodynamics: Concepts and Applications. p. 864.
37. King AE, Ackley MA, Cass CE, Young JD, Baldwin SA (2006) Nucleoside transporters: from scavengers to novel therapeutic targets. Trends Pharmacol Sci 27: 416–425.
38. Thompson AJ, Duke RK, Lunnis SC (2011) Binding sites for bilobalide, diluzem, ginkgoside, and picrorrhizin at the 5-HT3 receptor. Mol Pharmacol. 88: 103–190.