Predicting PAMPA permeability using the 3D-RISM-KH theory: are we there yet?

Roy, Dipankar; Dutta, Devjyoti; Wishart, David S.; Kovalenko, Andriy

Publisher's version / Version de l'éditeur:
https://doi.org/10.1007/s10822-020-00364-4
Journal of Computer-Aided Molecular Design, 35, 2, pp. 261-269, 2021-01-04

Predicting PAMPA permeability using the 3D-RISM-KH

theory: are we there yet?

Dipankar Roy, [1]

Devjyoti Dutta, [2]

David S. Wishart, [2]

Andriy Kovalenko, [1✉,3]

Email andriy.kovalenko@ualberta.ca

[1] Department of Mechanical Engineering, University of Alberta, 10-203 Donadeo Innovation Centre for Engineering, 9211-116 Street NW, Edmonton, AB, T6G 1H9 Canada

[2] Departments of Biological Sciences and Computing Science, University of Alberta, Edmonton, AB, T6G 2E8 Canada

[3] Nanotechnology Research Centre, National Research Council of Canada, 11421 Saskatchewan Drive, Edmonton, AB, T6G 2M9 Canada

Received: 10 September 2020 / Accepted: 26 November 2020

Abstract

The parallel artificial membrane permeability assay (PAMPA), a non-cellular lab-based assay, is extensively used to measure the permeability of pharmaceutical compounds. PAMPA experiments provide a working mimic of a molecule passing through cells and PAMPA values are widely used to estimate drug absorption parameters. There is an increased interest in developing computational methods to predict PAMPA permeability values. We developed an in silico model to predict the permeability of compounds based on the PAMPA assay. We used the

three-dimensional reference interaction site model (3D-RISM) theory with the Kovalenko–Hirata (KH) closure to calculate the excess chemical potentials of a large set of compounds and predicted their apparent permeability with good accuracy (mean absolute error or MAE = 0.69 units) when compared to a published experimental data set. Furthermore, our in silico PAMPA protocol performed very well in the binary prediction of 288 compounds as being permeable or impermeable (precision = 94%, accuracy = 93%). This suggests that our in silico protocol can mimic the PAMPA assay and could aid in the rapid discovery or screening of potentially therapeutic drug leads that can be delivered to a desired tissue.

AQ1

Keywords

PAMPA
3D-RISM-KH
Molecular solvation theory
Machine learning
Classification

Electronic supplementary material

The online version of this article (https://doi.org/10.1007/s10822-020-00364-4) contains supplementary material, which is available to authorized users.

Introduction

Drug development is often hindered by the failure to obtain correct absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiles of candidate drug compounds. For a compound to be effective as a drug, it must travel to the target organ/tissue, via a series of solvation–desolvation processes, while avoiding breakdown or decomposition via metabolism on the way. As most of the drugs are carried to the target site via the blood stream, poor solubility and/or a poor distribution profile between plasma and target tissues will render a drug ineffective. An ideal way to select or screen for candidate molecules with good distribution profiles would be to quantify the

permeability of the compound from plasma to the tissues of interest through an in vitro or model-organism-based assay. However, such in vitro cell- or tissue-based permeability experiments are complicated, time consuming, and difficult to reproduce. As an alternative, the parallel artificial membrane permeability assay (PAMPA) was developed by Kansy et al. as a non-cellular permeability assay to mimic passive trans-cellular permeability of small molecules through cells or tissues [1, 2]. The PAMPA assay consists of a hydrophobic membrane coated with lecithin in an organic solvent. Over the years, the type of membrane used in PAMPA assays changed based on the nature of experiments, therefore the permeability values are relative rather than absolute [3]. The PAMPA assay has been widely used to study intestinal absorption and the permeability of drugs or drug candidates across the blood–brain barrier [4, 5]. The underlying principle of PAMPA assay is based on the assumption of passive diffusion of the solutes. While this may be true for the majority of solutes, the presence of transporters and modulators in real biological membranes is well known. The Caco-2 assay is another permeability assay that is extensively used to study oral drug absorption. A strong correlation has been shown between the Caco-2 and PAMPA permeability profile by several research groups [6, 7, 8]. A common term for both Caco-2 and PAMPA assays for recording the permeability of a compound is P_e (effective permeability), defined as the number of molecules diffusing across a unit cross-section of the membrane in a unit of time under a unit concentration gradient. P_e is expressed as cm s^{-1} and varies largely with the lipid content of the membrane. The log of the effective permeability ($\log P_e$), calculated based on the formulations by Faller et al. and Sugano et al. [9, 10], is commonly used as the measure of compound permeability. As the PAMPA permeability assay has had broad success and applicability across many diverse tissue types, it is an interesting benchmark for developing in silico permeability models. Furthermore, as PAMPA assays can be expensive and time consuming to perform for high throughput screening, these in silico models could be used as a first step in screening early stage drug candidates.

Over the years several quantitative structure activity relationship (QSAR) models for predicting PAMPA parameters have been developed and validated on different types of molecules. For instance, Kansy et al. proposed a QSAR model for PAMPA flux for a large set of drug-like compounds and reported that their test set of ~ 100 compounds depended on the maximum flux on lipophilicity. However, these authors noted that logD is not an ideal descriptor in expressing PAMPA flux under steady state approximations [2]. QSAR models for PAMPA permeability developed by Akamatsu and coworkers for predicting Caco-2 permeability used octanol–water partition coefficients ($\log P_{OW}$), pH, pK_a, and surface area for hydrogen bond donors and acceptors as descriptors for modeling [11, 12]. They reported a

correlation coefficient of 0.78 for 57 compounds with a standard deviation of 0.37 [11]. Introducing apparent hydrophobicity to their QSAR model did not lead to significant improvements in its prediction quality (correlation coefficient 0.72 and standard deviation 0.36 for 97 compounds) [12]. Verma et al. have also developed QSAR models of PAMPA permeability of drugs with reference to the Caco-2 oral permeability and human intestinal permeability and reported correlation coefficients in the range of 0.76–0.89 for data sets of ~ 20 compounds at different pH values [13]. While these reports are encouraging, the modest size of the testing and training datasets and the limited diversity in the compound types, constrains the extension of these models to other scenarios. Recently, Sun and coworkers reported an excellent area under the receiver operating characteristic curve (AUC-ROC) of 0.88 using an in silico model of PAMPA permeability for a large number of compounds. For the training set, a correlation coefficient of 0.90 was reported. The prediction accuracy of this model had a significant decrease in the effective permeability range of 2.0–2.5 units (AUC 0.61). However, their dataset is not publicly available [14]. A brief collection of published in silico modeling or prediction of PAMPA permeability is provided in Table 1. The list is not comprehensive but provides an overview of the methods used to predict PAMPA permeability and the different databases or datasets used in training and testing these methods.

Table 1

QSAR modeling of PAMPA reported in literature

Database	Dependent variable	Statistical method	Performance
57 Compounds (35 drugs, 22 peptides) [11]	Apparent permeability ($logP_{app}$)	Multiple linear regression (MLR) and partial least square (PLS)	$r^2 = 0.65$, $SE_{pred} = 0.43$
Database of 97 compounds [12]	$logP_{app}$	MLR, NLR	$r^2 = 0.54–0.76$ (pH 7.3)
A maximum of 94 compounds depending on dataset [13]	$logP_{app}$, and Flux (F)	MLR and non-linear regression (NLR)	$r^2 = 0.76–0.89$; depending on dataset for different pH
Databases of 4079 and 2346 compounds [14]	$logP_e$	Support vector regression (SVR) models and support vector classifier (SVC)	$r^2 = 0.9$ (training set), AUC-ROC = 0.88–0.90
Databases of 74 compounds (28 acidic, 46 basic) and 58 compounds (12 neutral, 46	$logP_e$	MLR	$r^2 > 0.8$ (acidic compounds), $r^2 > 0.7$

Database	Dependent variable	Statistical method	Performance
amphoteric) [15, 16, 17]			(basic compounds) $r^2 = 0.95$ (neutral compounds); at different pH values
73 commercial drugs and organic acids [18]	Intrinsic permeability ($logP_0$)	MLR	$r^2 = 0.72–0.89$
248 drugs and drug like compounds in the training set [19]	$logP_0$	MLR	$r^2 = 0.74$
Databases of 47–61 compounds [20]	$logP_{app}$	MLR, Artificial neural network	$r^2 = 0.74$ at pH 5.5 $r^2 = 0.791$ at pH 7.4

Central to the accurate prediction of molecular partitioning via (bio)-chemical processes is the calculation of solvation free energy (SFE). Significant developments have been made in calculating SFEs with accuracy using different solvation models [21, 22]. The physics-based reference interaction site model (RISM) molecular solvation theory is built on the first principle statistical mechanics and provides direct correlation functions (DCFs) for all species in a solution [23, 24, 25]. The principle components of the RISM formalism are in representing a molecule of arbitrary shape by a six-dimensional vector consisting of three positional {**r**} and orientational degrees of freedom {**Θ**}, each in the molecular Ornstein–Zernike equation (MOZ) via the pair correlation functions (PCF) of r and Θ of liquids, in three dimensions (3D). Briefly, a solvent is represented with a finite number of sites (γ) around a solute with the 3D correlation function ($h_\gamma(r)$):

$$h_\gamma(\boldsymbol{r}) = \sum_\alpha \int d\boldsymbol{r}' c_\alpha\left(\boldsymbol{r} - \boldsymbol{r}'\right) \chi_{\alpha\gamma}\left(r'\right) \quad (1)$$

The 3D-site distribution function (g_γ) is calculated as $g_\gamma(\mathbf{r}) = h_\gamma(\mathbf{r}) + 1$ and consists of all sorts of interactions for all solvent sites, irrespective of the number of chemically distinct species [26, 27, 28]. The construct is dependent on the physical characteristics of a given solvent/solution, e.g. density, dielectric constant. This information is used as input

along with atomic charges, molecular size, and interatomic interaction potentials. The computational speed and accuracy of a RISM calculation depends on the choice of a closure relation for solving Eq. 1. A closure relation is a mathematical handle to integrate the infinite chain of diagrams produced through Eq. 1. A handful of closure relations have been developed for applications with the RISM formalism. The Kovalenko–Hirata closure (KH) has proven to be numerically stable and provides a solvation structure with reasonable accuracy at a modest computational cost [29]. The KH closure approximation accounts for both electrostatic and non-polar features of the liquid. The KH closure has the functional form

$$g_\gamma(r) = \begin{cases} \exp\left(-u_\gamma(r)/(k_B T) + h_\gamma(r) - c_\gamma(r)\right) & \text{for} \quad g_\gamma(r) \le 1 \\ 1 - u_\gamma(r)/(k_B T) + h_\gamma(r) - c_\gamma(r) & \text{for} \quad g_\gamma(r) > 1 \end{cases} \qquad 2$$

The KH closure combines the so-called mean spherical approximation (applied to the spatial part with solvent density enrichment, $g_\gamma(\mathbf{r}) > 1$) with the hypernetted chain (HNC), applied to the spatial part with solvent density depletion, $g_\gamma(\mathbf{r}) < 1$) in a non-trivial way, thus providing numerical stability and accuracy. The 3D-RISM integral equation has an exact differential of the solvation free energy for the KH closure. This enables the generation of an analytical expression of Kirkwood's thermodynamic integration by gradually switching on the solute–solvent interaction. While the KH closure is known to underestimate the height of strong associative peaks, these errors are mitigated by broadening of the peak and this often corrects the solvation thermodynamics and structure. Details of the 3D-RISM-KH theory are provided in references [30, 31, 32]. The 3D-RISM-KH theory is applicable to polar liquids, largely and also to a lesser extent to non-polar liquids, and has been carefully validated with experimental solvation energy datasets [33, 34, 35, 36, 37, 38, 39, 40]. While the 3D-RISM-KH theory has been widely applied to the fields of chemistry and biology, the predicted solvation free energy measurements are not absolute. The excess chemical potentials obtained from the 3D-RISM theory have a qualitative relation with the experimental solvation free energies and can be calibrated using a so called "universal correction" scheme to generate more quantitative results [40].

Our goal in this study was to calculate excess chemical potentials of drugs/solute molecules in specific solvents using the 3D-RISM-KH molecular solvation theory and to use these calculated chemical potentials to develop QSAR

models that accurately predict PAMPA permeability. Our database of chemicals (for training and testing) consists of datasets from published sources: one dataset consists of experimental logP_e acquired from 190 compounds and a second dataset consists of 288 compounds subdivided into permeable or impermeable by binary classification [15, 16, 17, 41]. We developed the PAMPA permeability prediction model using the combined 3D-RISM-KH calculated excess chemical potentials, 2D-molecular descriptors, and machine learning methods.

AQ2

Computational methods

Database preparation

The experimental logP_e data was collected from Oja et al. [15, 16, 17]. The data for the binary classification of PAMPA permeability was extracted from Avdeef et al. [41]. This data set had 50% of its molecules (144 data points) listed as permeable and 50% of its molecules (144 data points) listed as nonpermeable molecules (found in the Electronic Supplementary Material (ESM)). We chose these data sets to fulfill our requirement for a balanced compound collection containing different classes of chemicals. This chemical diversity was intended to help us to achieve more broadly applicable QSAR models which could possibly include ionic forms in the permeability estimation process.

RISM-KH calculations

The lowest energy conformation of all the solutes was generated using the OpenBabel toolkit with MMFF94 force field [42]. The lowest energy conformations of all the solute molecules were further optimized using the M06-2X dispersion corrected density functional in combination with the MIDI! basis set as implemented in the Gaussian16 software package [43, 44, 45]. All stationary points were confirmed as true minima at the respective potential energy surfaces by vibrational mode analysis. All the molecules were used in their neutral form. These optimized geometries were used for all further calculations. The 3D-RISM-KH based excess chemical potential and partial molar volume (used as descriptors in the prediction) were calculated for all the solutes using our in-house 3D-RISM-KH code. A public version of this code is implemented in the AMBERTOOLS suite of programs [46]. We used five solvents, chloroform, cyclohexane, *n*-hexadecane, dimethyl sulfoxide (DMSO), *n*-octanol, and water, for the 1D-RISM

susceptibility calculations of the pure liquids. The parameters for these solvents were validated against experimental solvation free energy datasets, as reported by us previously [33, 34, 35, 36, 37, 38]. We have employed the UFF parameters with AM1 charges for all the solutes [47, 48]. The atomic charge calculations and van der Waals parameter assignments were automated by integrating the MOPAC and Openbabel software in our workflow [42, 49]. The 3D-RISM-KH calculations for solute molecules were performed using a uniform cubic 3D-grid of 128 × 128 × 128 points in the box of size 64 × 64 × 64 $Å^3$ to represent a solute with a few solvation layers. The convergence accuracy was set to 10^{-5} in the modified direct inversion in the iterative subspace (MDIIS) solver. The excess chemical potential (exchem) and the partial molar volumes (PMV) of the solutes in each solvent were used as additional descriptors for the QSAR models.

AQ3

2D-Molecular descriptor generation

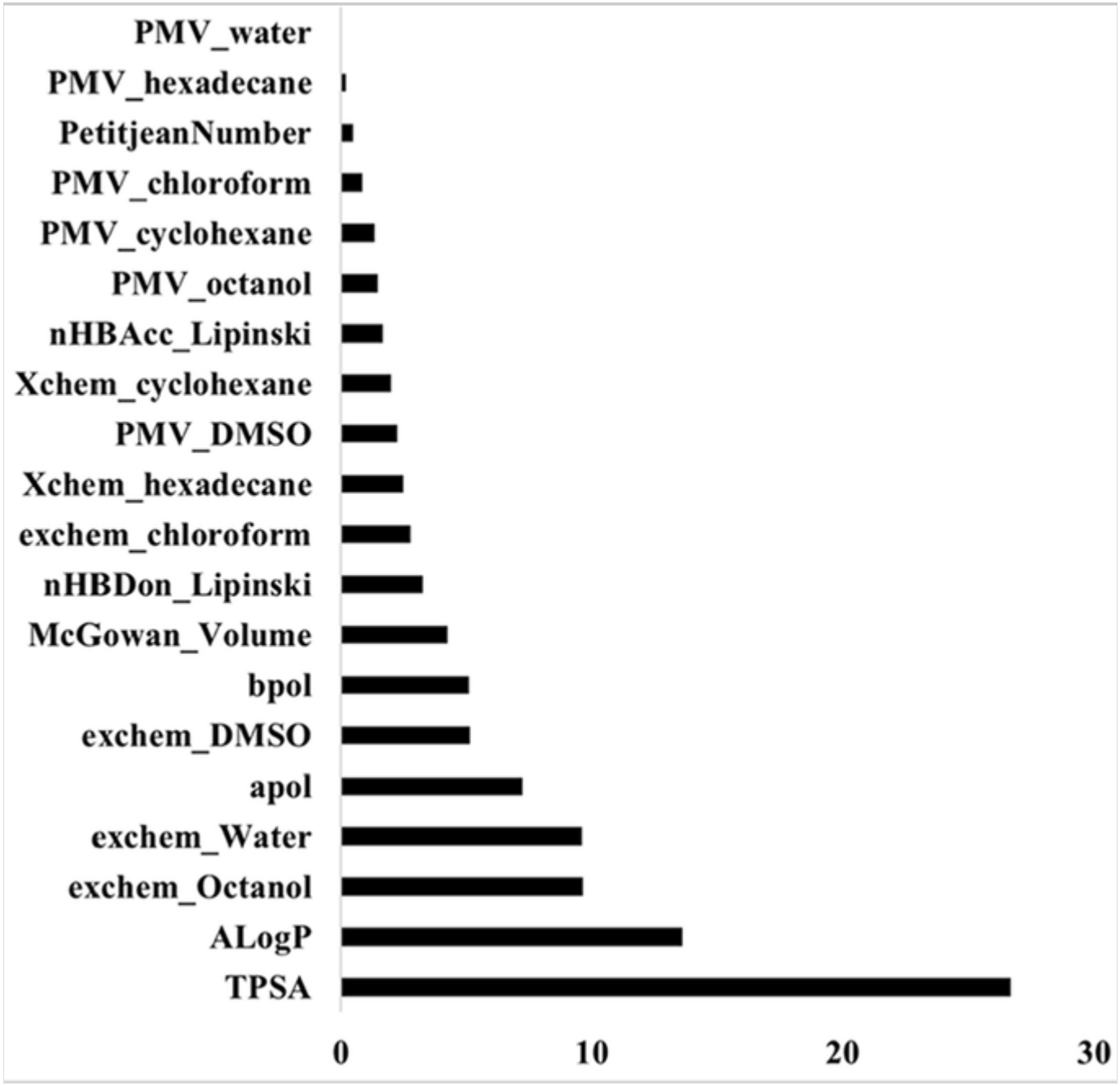
Molecular descriptors were generated from the corresponding SMILES strings of the solute molecules using the publicly available PaDEL-Descriptor software [50]. The 2D-descriptors used for the initial feature selection in the machine learning process are provided in the ESM. The logD values at *p*Hs 5.0 and 9.0 are calculated using the Calculator plugins in Marvin [51].

Machine learning and statistical modeling

The machine learning predictive models for PAMPA permeability was developed with the above-generated molecular descriptors. The statistical importance analysis of the descriptors, the machine learning calculations and the performance indices of models were calculated using the RStudio version 3.4.4 [52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63]. For predicting the logP_e values for the test drugs, a gradient boosting machine (GBM) was used with fivefold cross validation. The GBM technique minimizes the error in the model by building an ensemble of shallow and weak decision trees wherein each member tree learns and improves upon the previous tree. Additionally, the “Extreme Gradient Boosting” (XGBoost) technique was also used. Each method was used successively by optimizing the parameters reducing the relative mean square error (RMSE). The models containing 75% of the randomly chosen data from the apparent permeability dataset were trained using a fivefold cross validation with 2000 trees. Increasing the number of trees did not provide significant improvement in RMSE (Fig. 1).

Fig. 1

Statistical importance of each descriptor (in percentage) given on X-axis calculated for data at biological *p*H


PMV_water
PMV_hexadecane
PetitjeanNumber
PMV_chloroform
PMV_cyclohexane
PMV_octanol
nHBAcc_Lipinski
Xchem_cyclohexane
PMV_DMSO
Xchem_hexadecane
exchem_chloroform
nHBDon_Lipinski
McGowan_Volume
bpol
exchem_DMSO
apol
exchem_Water
exchem_Octanol
ALogP
TPSA
0
10
20
30

AQ4

To predict the permeability/non-permeability status, the binary database of the PAMPA permeable compounds was divided into a training set (75% of compounds) and a test set (25% of compounds) by randomly assigning molecules to each set. Several machine learning methods were assessed including GBM, GLM (Generalized linear models), SVM (support vector machines), and weighted-kNN (weighted K-nearest neighbor) to classify the compounds (i.e. PAMPA permeable = 1, PAMPA impermeable = 0). The performance indices (accuracy, precision, sensitivity, specificity, and F1-score) were calculated using R by generating a confusion matrix for each classification run. The confusion matrix consisted of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) counts for each of the machine learning protocols used for the test set. The performance indices were calculated as follows: Accuracy = (TP + TN)/(TP + TN + FP + FN); Precision = TP/(TP + FP); Sensitivity = TP/(TP + FN); Specificity = TN/(TN + FP); F1-Score = 2 × (Precision × Sensitivity)/(Precision + Sensitivity). The RMSE values were calculated as $\sqrt{\frac{\sum_i (x_i - x)^2}{n}}$, where x_i is the predicted data point for *i*-solute and *n* is the total number of data points. The mean absolute error (MAE) is calculated as $\frac{1}{n}\sum_1^n |x_i - x|$.

Results and discussion

In the following sections we have detailed our findings on the applicability of 3D-RISM-KH computed descriptors in predicting PAMPA permeability both quantitatively and qualitatively. The quantitative findings were based on a database collected from published literatures [15, 16, 17]. This is one of the largest publicly available databases for applications in QSAR modeling of PAMPA permeability. The dataset has a maximum molecular weight range of 230–390 Daltons and contains both polar (protic and aprotic) and non-polar systems. Our findings are divided into two sections: (1) QSAR based predictions of logP_e, and (2) binary predictions of PAMPA permeability.

QSAR based predictions of logP_e

The feature selection for QSAR modeling of the apparent permeability identified topological polar surface area (TPSA), *A*log*P*, number of H-bond acceptors and donors as important 2D-molecular descriptors. Among the excess chemical potentials and PMVs in six different solvents, those calculated in DMSO and water were found to be of significance. The other three solvents, i.e. chloroform, n-hexadecane, and cyclohexane were excluded from the important feature list. Both the GBM and XGBoost (XGB) methods yielded similar results (please see the ESM). The

GBM and XGB methods yielded a perfect prediction for the training set of 133 compounds ($r^2 = 0.99$) with a MAE of 0.03 units for apparent permeability calculations at *p*H 7.4. This indicates that the machine learning methods reported here are adequate for predicting PAMPA permeability. The final model (*Model2a*) yielded a good correlation (RMSE = 0.90, MAE = 0.69 units) with the experimental data and contained select 2D-descriptors as well as 3D-RISM-KH calculated solvent excess chemical potentials ($r^2 = 0.38$). Interestingly, the feature selection process did not choose the solvation energy in octanol as an informative feature. However, as the modelling strongly depends on the predicted log*P* values, (which is the octanol–water partition coefficient), the absence of the solvation energy in octanol in the final feature set can be easily rationalized. As a result, we removed the *A*log*P* descriptor from the third model (*model3a*), which resulted in a slight increase in the MAE and RMSE. The model built with only 2D-descriptors, *model4a*, based on feature selection, was best at predicting the apparent PAMPA permeability ($\log P_{e\text{-}pH7.4}$). The reliability of these models is summarized in Table 2. The comparative prediction of $\log P_{e\text{-}pH7.4}$ for the test dataset by *model2a* and *model4a* are shown in Fig. 2. Our findings justify the use of the 3D-RISM-KH based solvation energy descriptors as apparent PAMPA permeability predictors, as a proof of concept. The larger deviation in $\log P_{e\text{-}pH7.4}$ prediction compared to experimental results that were calculated with 3D-RISM-KH can be a direct result of the use of Gaussian-fluctuation (GF) excess chemical potentials for our prediction. Alternatively, one could apply a "universal-correction" scheme to all GF-excess chemical potentials to be calibrated against experimental solvation energy dataset and then apply correction factors to predict $\log P_{e\text{-}pH7.4}$. Here we did not employ such correction factors as we wanted to ensure that the method was more applicable and independent of additional regression analysis, depending on the choice of solvent. Additionally, the slightly poorer performance of the QSAR modeling can be attributed to the small size of the training and test dataset. While the diverse sets of molecules presented in the chosen dataset helped cover a large chemical space, the small number of datapoints was a handicap in better quantitative prediction. The machine learning methods performed consistently well for all the models. Incorporation of ionic species (i.e. replacing log*P* with log*D*) in the QSAR equation may improve quantitative predictions. To compare our machine learning based quantitative predictive models with those reported by Oja et al., we have calculated apparent permeability at *p*Hs 5.0 and 9.0. The excess chemical potentials are calculated for different ionic forms of compounds, based on *p*H whenever applicable. The modeling was done using the model2a to compare with the reported values. The results are summarized in Table 3. The incorporation of logD of compounds in acidic and basis *p*Hs improved the calculations of apparent permeabilities at these two different *p*H values. The performance of *model2a* is more or less similar to the previously reported correlations [15, 16, 17].

Table 2

Mean absolute error (MAE) and RMSE of different QSAR models (in log unit) in predicting log$P_{e\text{-}pH7.4}$

Performance	*model1a*		*model2a*		*model3a*		*model4a*	
	GBM	**XGBoost**	**GBM**	**XGBoost**	**GBM**	**XGBoost**	**GBM**	**XGBoost**
MAE	0.67	0.54	0.69	0.70	0.75	0.72	0.56	0.50
RMSE	0.90	0.68	0.90	0.91	0.94	0.96	0.72	0.65
r^2	0.37	0.61	0.37	0.38	0.37	0.33	0.55	0.64

Fig. 2

Apparent PAMPA permeability prediction performances for the test dataset using *model2a* (o) and *model4a* (•). The correlation (r^2) between the calculated and predicted log$P_{e\text{-}pH7.4}$ for model2a was 0.38, while the correlation between the calculated and predicted log$P_{e\text{-}pH7.4}$ for model4a was 0.62

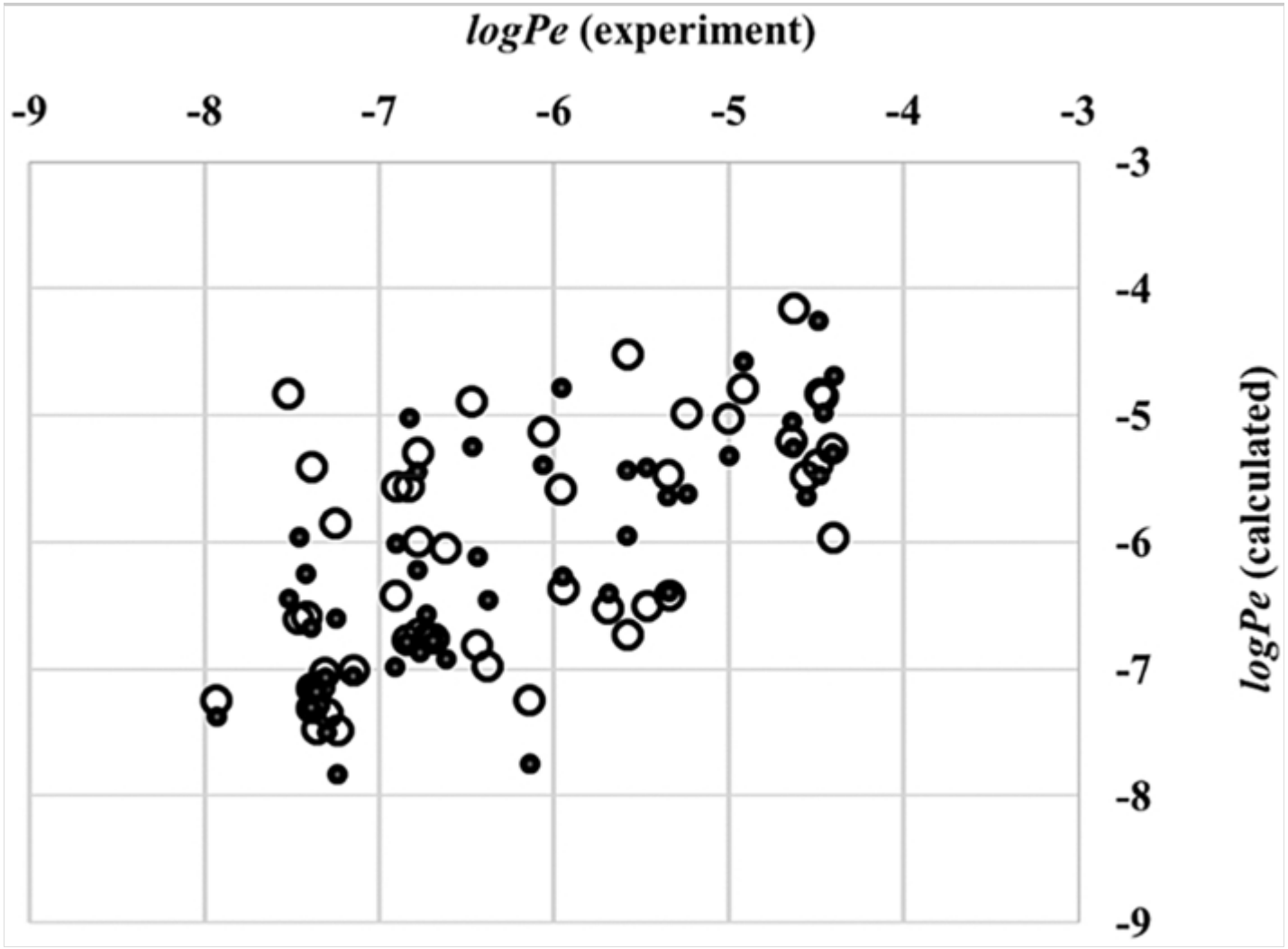


Table 3

Mean absolute error (MAE) and RMSE in predicting logP_e (in log unit) for the test sets of neutral/amphoteric and ionic compounds at *p*H 5.0 and *p*H 9.0 using *model2a*

Performance	***Neutral molecules* (number of data points = 19)**		***Ionic molecules* (number of data points = 26)**	
	***p*H 5.0**	***p*H 9.0**	***p*H 5.0**	***p*H 9.0**

	GBM	XGBoost	GBM	XGBoost	GBM	XGBoost	GBM	XGBoost
MAE	0.47	0.37	0.98	0.58	0.52	0.21	0.47	0.46
RMSE	0.56	0.48	1.43	0.74	0.49	0.27	0.60	0.62
r^2	0.66	0.74	0.50	0.52	0.55	0.72	0.71	0.65

Binary predictions of PAMPA permeability

The binary classification of PAMPA permeability was initiated in the same way as the logP_e permeability modeling. The initial feature selection for binary permeability resulted in the selection of *X*log*P*, *apol*, *bpol*, and *TPSA* from the 2D-descriptor pool. While for flexible polar molecules the 3D-PSA is arguably a better descriptor for a molecular shape, the 2D-descriptor TPSA was shown to produce identical results [64] and hence used in this study. The selected 3D-RISM-KH descriptors were excess chemical potentials in octanol, DMSO, and water as solvents (Fig. 3). The initial model (*model1b*) contained all the descriptors, both 2D- and 3D-RISM-KH descriptors and is detailed in the ESM. The best model with the 3D-RISM-KH descriptors (*model2b*) performed impressively with 93% maximum accuracy and up to 94% precision. This model contained *A*log*P*, *TPSA*, *nHBAcc_Lipinski* from the 2D-descriptors and excess chemical potentials in DMSO, octanol, and water from the 3D-RISM-KH calculations. Removing *A*log*P* from the equation resulted in *model3b* with slight deterioration in the classification. The all 2D-descriptor based QSAR model (*model4b*) performed similarly to *models3b*. In fact, combining the 3D-RISM-KH computed parameters with the select 2D-descriptors provided the best results for both the prediction methods. The best performing machine learning model was based on the SVM. The other three methods, viz. GBM, GLM, and kNN, were less accurate and specific in their classification performance. The normalized performance indices for the three models obtained using different machine learning methods are provided in Table 4.

Fig. 3

Performance indices of different machine learning schemes used to classify PAMPA permeable and impermeable compounds

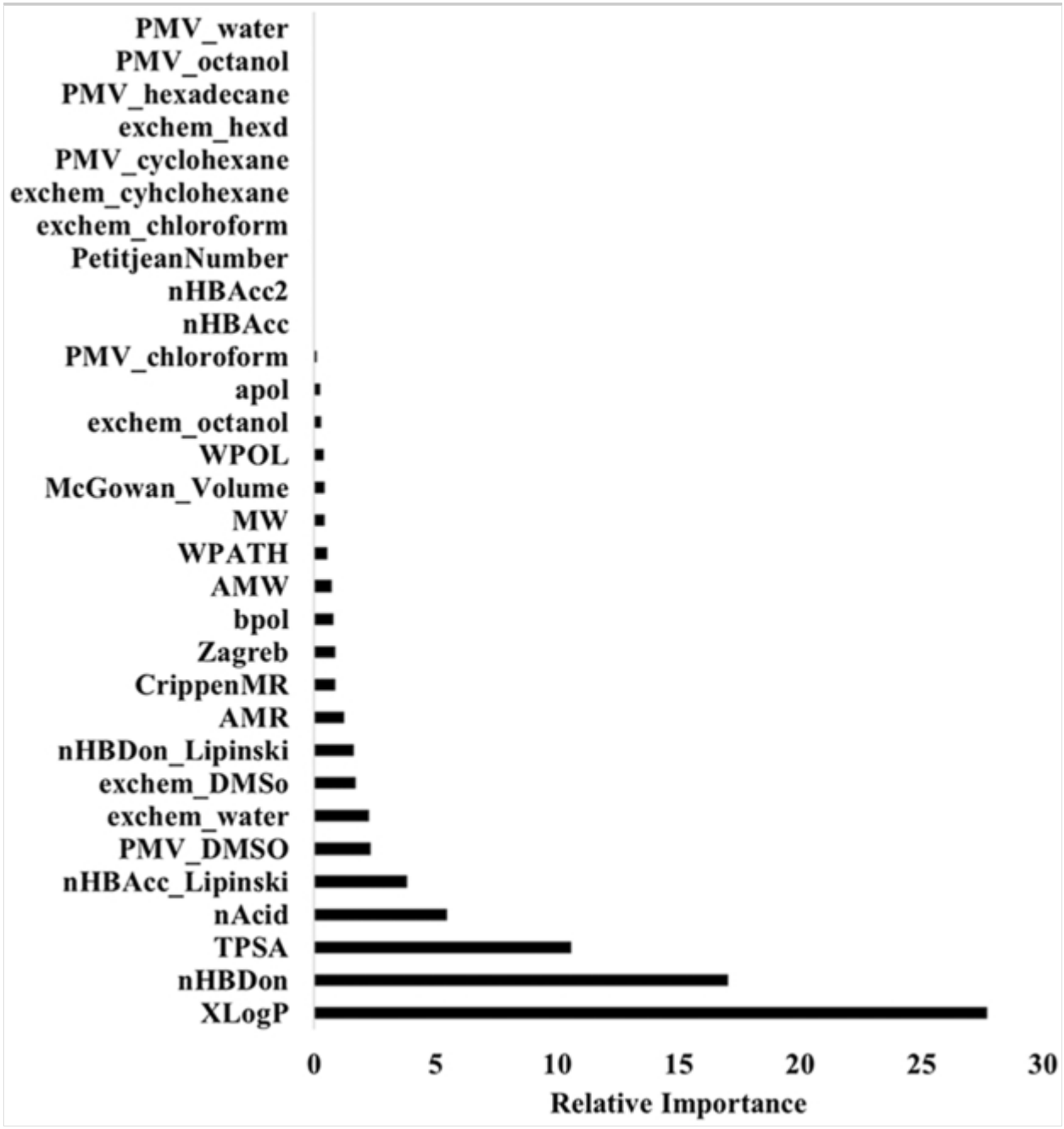

PMV_water
PMV_octanol
PMV_hexadecane
exchem_hexd
PMV_cyclohexane
exchem_cyhclohexane
exchem_chloroform
PetitjeanNumber
nHBAcc2
nHBAcc
PMV_chloroform
apol
exchem_octanol
WPOL
McGowan_Volume
MW
WPATH
AMW
bpol
Zagreb
CrippenMR
AMR
nHBDon_Lipinski
exchem_DMSo
exchem_water
PMV_DMSO
nHBAcc_Lipinski
nAcid
TPSA
nHBDon
XLogP
0
5
10
15
20
25
30
Relative Importance

Table 4

Normalized performance of different models in predicting the binary PAMPA permeability

Performance indices	*Model2b*				*Model3b*				*Model4b* (only 2D-descriptors)			
	GBM	GLM	SVM	kNN	GBM	GLM	SVM	kNN	GBM	GLM	SVM	kNN
Accuracy	0.72	0.79	*0.93*	0.78	0.72	0.79	0.88	0.78	0.79	0.76	0.90	0.72
Precision	0.74	0.80	*0.94*	0.83	0.71	0.80	0.91	0.83	0.78	0.77	0.87	0.72
Sensitivity	0.69	0.78	*0.92*	0.69	0.75	0.78	0.83	0.69	0.81	0.75	0.94	0.73
Specificity	0.75	0.81	*0.94*	0.86	0.69	0.80	0.92	0.86	0.78	0.78	0.86	0.72
F1-Score	0.71	0.79	*0.93*	0.76	0.73	0.79	0.87	0.76	0.79	0.76	0.91	0.72
The %-performance can be obtained by multiplying the indices by 100												

Conclusion

In this work, we have demonstrated the applicability of the 3D-RISM-KH theory for calculating the excess chemical potentials and PMVs of solutes as descriptors in developing the QSAR models of PAMPA prediction. These descriptors work efficiently in predicting apparent permeability as well as classifying PAMPA permeable and impermeable compounds. In addition, the presence of DMSO, octanol, and water in the classification models can be rationalized. The three solvents together cover the majority of the solvent polarity spectrum, with dielectric constants (ε) ranging from 10.3 for $\varepsilon_{\text{octanol}}$ to 47.2 for $\varepsilon_{\text{DMSO}}$ and 80.1 for $\varepsilon_{\text{water}}$ (25 °C). The use of these diverse solvent models thus effectively mimics the transport of drugs through PAMPA. The choice of machine learning method(s) also plays an important role in achieving reasonable accuracy, sensitivity and specificity. Our models are easier to develop and deploy for a large class/set of molecules and could be potentially used in early drug development research. Furthermore, our findings provide an alternative to computational PAMPA permeability without the extensive use of molecular dynamics simulations. We would like to point that the models presented here for predicting PAMPA permeability work only for passive diffusion of molecules. For molecules (or drugs) which have a

transporter or a permeability modulator, these models should be modified to take these factors into consideration, too.

AQ5

AQ6

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements

This work was financially supported by the NSERC Discovery Grant (RES0029477), Alberta Innovates AARP VII Research Grant (RES0043948), and Alzheimer Society of Alberta and Northwest Territories AARP VII Research Grant (RES0043949). Generous computing time provided by WestGrid (www.westgrid.ca) and Compute Canada/Calcul Canada (www.computecanada.ca) is acknowledged. The authors thank Dr. Marcia LeVatte for assistance in editing the manuscript.

Authors contributions

All the authors contributed equally to writing the manuscript and approved the final version.

Compliance with ethical standards

Conflict of interest The authors have no conflicts of interest to declare.

Electronic supplementary material

Below is the link to the electronic supplementary material.Electronic supplementary material 1 (XLSX 224 kb)

References

1. Kansy M, Senner F, Gubernator L (1998) J Med Chem 41:1007–1010

2. Kansy M, Fischer H, Kratzat K, Senner F, Wagner B, Parrilla I (2001) In: Testa B, van de Waterbeemd H, Folkers G, Guy R (eds) Pharmacokinetic optimization in drug research: biological, physicochemical, and computational strategies. Verlag Helvetica Chimica Acta, Zürich/Wiley-VCG, Weinheim

3. Di L, Kerns EH (2003) Curr Opin Chem Biol 7:402–408

4. Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD (2002) J Med Chem 45:2615–2623

5. Kerns EH (2001) J Pharm Sci 90:1838–1858

6. Avdeef A, Bendels S, Di L, Faller B, Kansy M, Sugano K, Yamauchi Y (2007) J Pharm Sci 96:2893–2909

7. Bermejo M, Avdeef A, Ruiz A, Nalda R, Ruell JA, Tsinman O, González I, Fernández C, Sánchez G, Garrigues TM, Merino V (2004) Eur J Pharm Sci 21:429–441

8. Avdeef A, Artursson P, Neuhoff S, Lazorova L, Gråsjö J, Tavelin S (2005) Eur J Pharm Sci 24:333–349

9. Wohnsland F, Faller B (2001) J Med Chem 44:923–930

10. Sugano K, Hamada H, Machida M, Ushio H (2001) J Biomol Screen 6:189–196

11. Fujikawa M, Ano R, Nakao K, Shimizu R, Akamatsu M (2005) Bioorg Med Chem 13:4721–4732

12. Fujikawa M, Nakao K, Shimizu R, Akamatsu M (2007) Bioorg Med Chem 15:3756–3767

13. Verma RP, Hansch C, Selassie CD (2007) J Comput Aided Mol Des 21:3–22

14. Sun H, Nguyen K, Kerns E, Yan Z, Yu KR, Shah P, Jadhav A, Xu X (2017) Bioorg Med Chem 25:1266–1276

15. Oja M, Maran U (2015) Mol Inform 34:493–506

16. Oja M, Maran U (2016) SAR QSAR Environ Res 27:813–832

17. Oja M, Maran U (2018) Eur J Pharm Sci 123:429–440

18. Avdeef A, Tsinman O (2006) Eur J Pharm Sci 28:43–50

19. Diukendjieva A, Alov A, Tsakovska I, Pencheva T, Richarz A, Kren V, Cronin MTD, Pajeva I (2019) Phytomedicine 53:79–85

20. Karleson M, Karleson G, Tamm T, Tulp I, Jänes J, Tämm K, Lomaka A, Savchenko D, Dobchev A (2009) ARKIVOC Online J Org Chem 218–238

21. Hansen N, van Gunsteren WF (2014) J Chem Theory Comput 10:2632–2647

22. Skyner RE, McDonagh JL, Groom CR, van Mourik T, Mitchell JBO (2015) Phys Chem Chem Phys 17:6174–6191

23. Chandler D, McCoy JD, Singer SJ (1986) J Chem Phys 85:5971–5976

24. Chandler D, McCoy JD, Singer SJ (1986) J Chem Phys 85:5977–5982

25. Lowden LJ, Chandler D (1973) J Chem Phys 59:6587–6595

26. Kovalenko A (2015) Condens Matter Phys 18:32601

27. Palmer DS, Frolov A, Ratkova EL, Fedorov MV (2010) J Phys Condens Matter 22:492101

28. Kovalenko A, Hirata F (2005) Phys Chem Chem Phys 7:1785–1793

29. Kovalenko A (2013) Pure Appl Chem 85:159–199

30. Kovalenko A (2017) Multiscale modeling of solvation. In: Breitkopf C, Swider-Lyons K (eds) Springer handbook of electrochemical energy. Springer, Berlin

31. Kovalenko A, Gusarov S (2017) Phys Chem Chem Phys 20:2947–2969

32. Ratkova EL, Palmer DS, Fedorov MV (2015) Chem Rev 13:6312–6356

33. Roy D, Kovalenko A (2019) J Phys Chem A 123:4087–4093

34. Roy D, Blinov N, Kovalenko A (2017) J Phys Chem B 121:9268–9273

35. Roy D, Hinge VK, Kovalenko A (2019) ACS Omega 4:3055–3060

36. Roy D, Hinge VK, Kovalenko A (2019) ACS Omega 4:16774–16780

37. Roy D, Kovalenko A (2019) J Comput Aided Mol Des 33:905–912

38. Roy D, Kovalenko A (2020) J Phys Chem B 124:4590–4597

39. Truchon J-F, Pettit BM, Labute P (2014) J Chem Theory Comput 10:934–941

40. Misin M, Palmer DS, Fedorov MV (2016) J Phys Chem B 2016(120):5724–5731

41. Avdeef A (2002) In: van de Waterbeemd H, Lennernas H, Artursson P (eds) Drug bioavailability. Weinheim, Wiley-VCH

42. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) J Cheminform 3:33

43. Zhao Y, Truhlar DG (2008) Theor Chem Acc 120:215–241

44. Easton RE, Geisen DJ, Welch A, Cramer CJ, Truhlar DG (1996) Theor Chem Acc 93:281–301

45. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Petersson GA, Nakatsuji H, Li X, Caricato M, Marenich AV, Bloino J, Janesko BG, Gomperts R, Mennucci B, Hratchian HP, Ortiz JV, Izmaylov AF, Sonnenberg JL, Williams-Young D, Ding F, Lipparini F, Egidi F, Goings J, Peng B, Petrone A, Henderson T, Ranasinghe D, Zakrzewski VG, Gao J, Rega N, Zheng G, Liang W, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Throssell K, Montgomery JA Jr., Peralta JE, Ogliaro F, Bearpark MJ, Heyd JJ, Brothers EN, Kudin KN, Staroverov VN, Keith TA, Kobayashi R, Normand J, Raghavachari K, Rendell AP, Burant JC, Iyengar SS, Tomasi J, Cossi M, Millam JM, Klene M, Adamo C, Cammi R, Ochterski JW, Martin RL, Morokuma K, Farkas O, Foresman JB, Fox DJ (2016) Gaussian 16, Revision B.01. Gaussian, Inc., Wallingford CT

46. Case DA, Ben-Shalom IY, Brozell SR, Cerutti DS, Cheatham TE III, Cruzeiro VWD, Darden TA, Duke RE, Ghoreishi D, Gilson MK, Gohlke H, Goetz AW, Greene D, Harris R, Homeyer N, Izadi S, Kovalenko A, Kurtzman T, Lee TS, LeGrand S, Li P, Lin C, Liu J, Luchko T, Luo R, Mermelstein DJ, Merz KM, Miao Y, Monard G, Nguyen C, Nguyen H, Omelyan I, Onufriev A, Pan F, Qi R, Roe DR, Roitberg A, Sagui C, Schott-Verdugo S, Shen J, Simmerling CL, Smith J, Salomon-Ferrer R, Swails J, Walker RC, Wang J, Wei H, Wolf RM, Wu X, Xiao L, York DM, Kollman PA (2018) AMBER 2018. University of California, San Francisco

47. Rappe AK, Casewit CJ, Colwell KS, Goddard WA, Skiff WM (1992) J Am Chem Soc 114:10024–10035

48. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902–3909

49. MOPAC2016, Stewart JJP, Stewart Computational Chemistry, Colorado Springs, CO, USA

50. Yap CW (2011) J Comput Chem 32:1466–1476

51. Marvin 17.21.0, ChemAxon (https://www.chemaxon.com)

52. R Core Team (2017) R: A Language And Environment For Statistical Computing; R Foundation for Statistical Computing, Vienna, Austria

53. Robinson D, Gomez M, Demeshev B, Menne D, Nutter B, Johnston L, Bolker B, Briatte F, Arnold J, Gabry J (2017) Broom: Convert Statistical Analysis Objects into Tidy Data Frames

54. Wickham H (2007) ggplot2: elegant graphics for data analysis. Springer-Verlag, New York

55. Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 1(5):2008

56. Ridgeway G (2007) Generalized boosted models: a guide to the gbm package. R package vignette, http://CRAN.R-project.org/package=gbm. Accessed 12 May 2020

57. Liaw R, Wiener M (2002) Classification and regression by randomForest. R News 2:18–22

58. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A (2008) e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.5-18. https://cran.r-project.org/web/packages/e1071 /index.html. Accessed 12 May 2020

59. Schliep K, Hechenbichler K (2016) Weighted k-nearest neighbors for classification, regression and clustering. R package version 1.3. https://cran.r-project.org/web/packages/kknn/index.html. Accessed 12 May 2020

60. Eric S, Kalinic M, Ilic K, Zloh M (2014) SAR QSAR Environ Res 25:939–966

61. Kuhn M, Johnson K (2018) Applied predictive modeling. Springer, New York

62. James G, Witten D, Hastie T, Tibshirani R (2014) An introduction to statistical learning with applications in R. Springer, New York

63. Natekin A, Knoll A (2013) Front Neurorobot 7:21

64. Ertl P, Rhode B, Selzer P (2000) J Med Chem 43:3714–3717