Journal of Computer-Aided Molecular Design (2019) 33:965–971
https://doi.org/10.1007/s10822-019-00253-5

# Prediction of P-glycoprotein inhibitors with machine learning classification models and 3D-RISM-KH theory based solvation energy descriptors

**Vijaya Kumar Hinge[1] · Dipankar Roy[1] · Andriy Kovalenko[1,2]**

Received: 24 September 2019 / Accepted: 14 November 2019 / Published online: 19 November 2019
© Crown 2019

**Abstract**
Development of novel in silico methods for questing novel PgP inhibitors is crucial for the reversal of multi-drug resistance in cancer therapy. Here, we report machine learning based binary classification schemes to identify the PgP inhibitors from non-inhibitors using molecular solvation theory with excellent accuracy and precision. The excess chemical potential and partial molar volume in various solvents are calculated for PgP± (PgP inhibitors and non-inhibitors) compounds with the statistical–mechanical based three-dimensional reference interaction site model with the Kovalenko–Hirata closure approximation (3D-RISM-KH molecular theory of solvation). The statistical importance analysis of descriptors identified the 3D-RISM-KH based descriptors as top molecular descriptors for classification. Among the constructed classification models, the support vector machine predicted the test set of Pgp± compounds with highest accuracy and precision of ~97% for test set. The validation of models confirms the robustness of state-of-the-art molecular solvation theory based descriptors in identification of the Pgp± compounds.

**Keywords** P-glycoprotein (PgP) · PgP inhibitors · Multidrug resistance (MDR) · 3D-RISM-KH · Solvation free energy · Excess chemical potential · Partial molar volume (PMV)

## Introduction

Multidrug resistance (MDR) is a cellular drug resistance developed in cancer cells that involves reduced drug accumulation in intracellular space. The most common cellular response associated with MDR is the overexpression of membrane transporter proteins belonging to ATP-binding cassette superfamily. Among these transporter proteins, P-glycoprotein (PgP) is overexpressed in many cancer cell-line models [1]. PgP is also known as an ATP-binding cassette sub-family B member 1 or multidrug resistance protein 1 (MDR1) [2, 3]. The PgP is widely distributed in the hepatocytes of bile duct, apical membranes of intestinal mucosal cells, renal proximal tubular cells of kidney, and capillary endothelial cells of the brain and testis. This transmembrane glyco-protein of 1280 amino acids is important for intestinal absorption, drug metabolism, and blood brain barrier (BBB) penetration, and is expressed by MDR1 gene [4]. The PgP consists of two transmembrane domains, each containing six transmembrane α-helices which make drug binding domains to transports the drugs. Two ATP binding domains located on the cytoplasmic side of membrane are crucial for the transport of toxins by hydrolysis of ATP [3].

PgP shows broad ligand specificity, and translocates its ligands out of the cell against the concentration gradient using the energy derived by ATP hydrolysis. Overexpression of PgP lowers intracellular concentrations of drugs to sub-therapeutic levels by increased ATP dependent efflux leading to MDR. The progress in understanding of the MDR have made the inhibition of PgP a viable and attractive therapeutic

Vijaya Kumar Hinge and Dipankar Roy have equally contributed to this work.

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s10822-019-00253-5) contains supplementary material, which is available to authorized users.

✉ Andriy Kovalenko
andriy.kovalenko@ualberta.ca

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

[2] Nanotechnology Research Centre, 11421 Saskatchewan Drive, Edmonton, AB T6G 2M9, Canada

approach to overcome MDR [5–7]. In the past decades, several inhibitors designed to target the PgP inhibition failed in clinical trials [8]. The known inhibitors of PgP are broadly classified into four generations. The first and second generations of inhibitors showed uncertain pharmacokinetics [9] and interaction with oxidizing enzyme [10, 11], respectively. The third-generation of inhibitors improved significantly but were unsuccessful in clinical trials due to their toxicity [12, 13]. The fourth-generation of inhibitors are natural products, show less toxicity and low molecular weight, and can potentially lead to a next generation of PgP inhibitors [14, 15]. The quest for novel PgP inhibitors for the reversal of MDR in cancer patients is currently of research interest. This require development of new QSAR models with novel descriptors to identify the PgP inhibitors with the highest accuracy and precision.

Several in-silico QSAR (quantitative structure–activity relationship) methodologies are known in the literature for identifying the drug molecules for PgP inhibition. Most of these methods were able to identify the PgP inhibitors using pharmacophore description models with the help of advanced machine learning algorithms. These models were broadly classified as binary classification models [16–26], correlation models [27–30], and pharmacophore based models [22, 23, 26, 31–35]. The SAR (structure–activity relationships) based methods confirm that lipophilicity (log *P*) [36–38], molecular weight [15, 17, 39], aromaticity [20, 22, 40], and hydrogen bond acceptor [20, 22, 35, 41] were important molecular properties for the identification of such inhibitors. These studies further support that lower log *P* as well as molecular weight are crucial physicochemical factors for ideal PgP inhibitors. The QSAR modeling studies on a small set of PgP inhibitors support that ideal compounds should possess log *P* greater than 2.92, high $E_{HOMO}$ (energy of highest occupied molecular orbital), and at least one tertiary basic nitrogen atom [36]. Chen et. al. developed QSAR classification models using fingerprints and molecular property descriptors for a diverse set of 973 PgP inhibitors with an accuracy of 81% [17]. These authors reported solubility, log *D*, and molecular weight as important descriptors for classification of inhibitors from non-inhibitors. Schyman et. al. have used variable-nearest neighbor (v-NN) method and predicted the PgP inhibitors for a diverse and large set of 2,276 compounds with an accuracy of 87% [25].

The present study focuses on development of the machine learning based binary classification schemes to identify the PgP inhibitors from non-inhibitors using 3D-RISM-KH based solvation free energy descriptors. This work is aimed as a proof of concept that molecular solvation theory can be successfully used to identify PgP inhibitors. We have used the 3D-RISM-KH molecular solvation theory to calculate the solvation free energy and solvation free energy based descriptors for PgP± compounds. The 3D-RISM-KH theory is a first principle statistical mechanics based solvation model that uses rigorous descriptions of direct correlation functions to calculate thermochemical properties of pure liquids and solutions in the form of excess and total chemical potentials, partial molar volume, solvent distribution function around solute, etc. The applicability of these descriptors have been tested by developing the classification schemes to identify the PgP± compounds. The machine learning methodologies have primarily estimated the importance of 3D-RISM-KH based descriptors in predicting the PgP± compounds, and these have been further used to develop the models with the classification schemes to identify the PgP± compounds.

## Computational methods

### Database preparation

The database of the PgP inhibitor and non-inhibitor (PgP±) compounds were taken from the published work of Broccatelli et al. [20]. Their extensive literature search from more than 60 references yielded a large data set of 1274 PgP± compounds. The details of data curation, experimental methods, and IC50 values used to classify PgP± compounds are given in the data collection section and supporting material published by Broccatelli et al. [20]. The duplication of PgP± compounds was not observed in the data set. The SMILES strings of all PgP± compounds are imported to the Molecular Operating Environment (MOE2018) drug discovery software platform [42] with the help of database preparation module. The addition of hydrogens and generation of 3D-Cartesian coordinates for PgP± compounds were carried out in the MOE. For all the calculations, we have used the neutral form of the species. The PgP± compounds were subjected to gas phase geometry optimization at the semi-empirical AM1 level using the Gaussian16 software package [43, 44]. The molecular descriptors of all the molecules were generated using the MOE2018 drug discovery software.

### 3D-RISM-KH based descriptors generation

The 3D-RISM-KH based excess chemical potential and partial molar volume (used as descriptors in prediction) were calculated for the PgP± compounds using our in-house 3D-RISM-KH code. A working version of this code (executed as rism1d and rism3d.snglpnt) is implemented in the AMBERTOOLS suite of programs [45]. We used five solvents, viz. chloroform, cyclohexane, n-hexadecane, n-octanol, and water for 1D-RISM susceptibility calculations for pure liquids. The parameters for these solvents were validated against experimental solvation free energy datasets, as reported by us previously [46, 47]. We have

employed UFF [48] parameters with AM1 charges for all the solutes. The 3D-RISM-KH calculations for solute molecules were performed using a uniform cubic 3D-grid of $128 \times 128 \times 128$ points in the box of size $64 \times 64 \times 64$ Å$^3$ to represent a solute with a few solvation layers with convergence accuracy set to $10^{-5}$ in the modified direct inversion in the iterative subspace (MDIIS) solver [49]. The detailed workflow chart describes the calculations of 3D-RISM-KH based descriptors given in the ESM (Fig. S1).

### Machine learning and statistical modeling

The machine learning predictive models for PgP± compounds were developed with descriptors. The full list of descriptors for the entire dataset are provided in the ESM. The statistical importance analysis of descriptors, machine learning calculations and performance indices of models were performed using the Rstudio version 3.4.4 [50]. The R packages were used to perform the calculations briefly described in the S1 of ESM [51–57]. The definitions of machine learning methodologies and performance indices are given in the ESM (Tables S4–S7 and S2 in the ESM). The analysis of statistical importance of descriptors was performed with the GBM (gradient boosting machines) and RF (random Forest) methods to identify the crucial descriptors to use in predictive models. The database of PgP± compounds is divided into a training (75% of compounds) and a test set (25% of compounds) by randomly assigning the molecules. The GBM, GLM (Generalized linear models), SVM (support vector machines), and weighted-kNN (weighted κ-nearest neighbor) machine learning schemes were used to identify PgP± compounds. The performance indices (accuracy, precision, sensitivity, specificity, and F1-score) were calculated with R package by generating the confusion matrix for each classification run.

## Results and discussion

The current study aims at developing the binary classification models to identify the PgP inhibitors from non-inhibitors with precision and accuracy using the 3D-RISM-KH molecular solvation theory. The 3D-RISM-KH molecular solvation theory-based solvation parameter, the excess chemical potential in solvents as descriptors were calculated for PgP± compounds. The machine learning based binary classification schemes are developed with 3D-RISM-KH based descriptors along with other descriptors and used for classification of PgP± compounds. To achieve the objectives, we prepared the database of PgP± compounds and generated the descriptors for PgP± compounds as described in the computational methods. The analysis of statistical importance of descriptors was performed on descriptors (total 354) with the GBM (gradient boosting machines) method, and identified 23 descriptors as important ones for preliminary model building activities. The list of descriptors is given in Fig. 1 and Table S1 in the ESM. The pool of these descriptors consists of ten 3D-RISM-KH based descriptors and thirteen 2D-descriptors. Among all the 23 descriptors, the 3D-RISM-KH based descriptors contributed relative importance of 48.1% in total (Fig. 1). The top 5 crucial descriptors contributed 62% of relative importance, and the remaining 18 crucial descriptors contributed 38%.

The 23 descriptors obtained from the initial GBM calculations were subjected to further analysis of the descriptors importance using the GBM and RF methods, with the aim to reduce number of descriptors in the prediction model while keeping the accuracy intact. The descriptor list is given in Fig. 1b, c and in the ESM (Tables S2, S3). The GBM identified excess chemical potential in water and in octanol, number of aromatic atom, sum of atomic polarization, and topological polar surface area (*TopoPSA*) as crucial descriptors. The 3D-RISM-KH based descriptors excess chemical potential in water shows a highest relative importance of 38.8%. The RF method identified excess chemical potential in water and in octanol, number of aromatic atom, and sum of atomic polarization as crucial descriptors. The top descriptor, excess chemical potential in water shows a highest relative importance of 38.8% and 41.2% in the GBM and RF methods, respectively. The analysis of the crucial descriptor revealed that four descriptors found common with three classification methods. These are excess chemical potential in water and in octanol, number of aromatic atoms, and sum of atomic polarization. These findings are in line with the previous literature pointing out to lipophilicity and molecular weight as important descriptors for such a classification [15, 31–41].

We developed three descriptors models based on relative importance of descriptors from the statistical importance analysis: (i) *model-23d* (maximum descriptor model) (ii) *model-5d*, and (iii) *model-4d* (minimum descriptor model). *Model-23d, model-5d,* and *model-4d* were developed with 23, 5, and 4 descriptors, respectively, as suggested in the naming scheme of these models. The choice of descriptors for *model-4d* was guided by the RF method, and for *model-23d* and *model-5d* by the GBM method. The models were used to classify PgP± compounds with machine learning schemes as described in the computational methods section. The performance indices of different classification schemes using three models for the test set of compounds are given in Fig. 2 and Tables S4–S7 in the ESM.

The classification schemes identify the test set compounds as PgP-inhibitor (yes/1) or PgP-non-inhibitor (no/0), based on the models applied. The accuracy of the GBM, GLM, SVM, and Weighted kNN classification methods with *model-23d* is in the range of 84.0–86.5%, 48.1–71.4%,

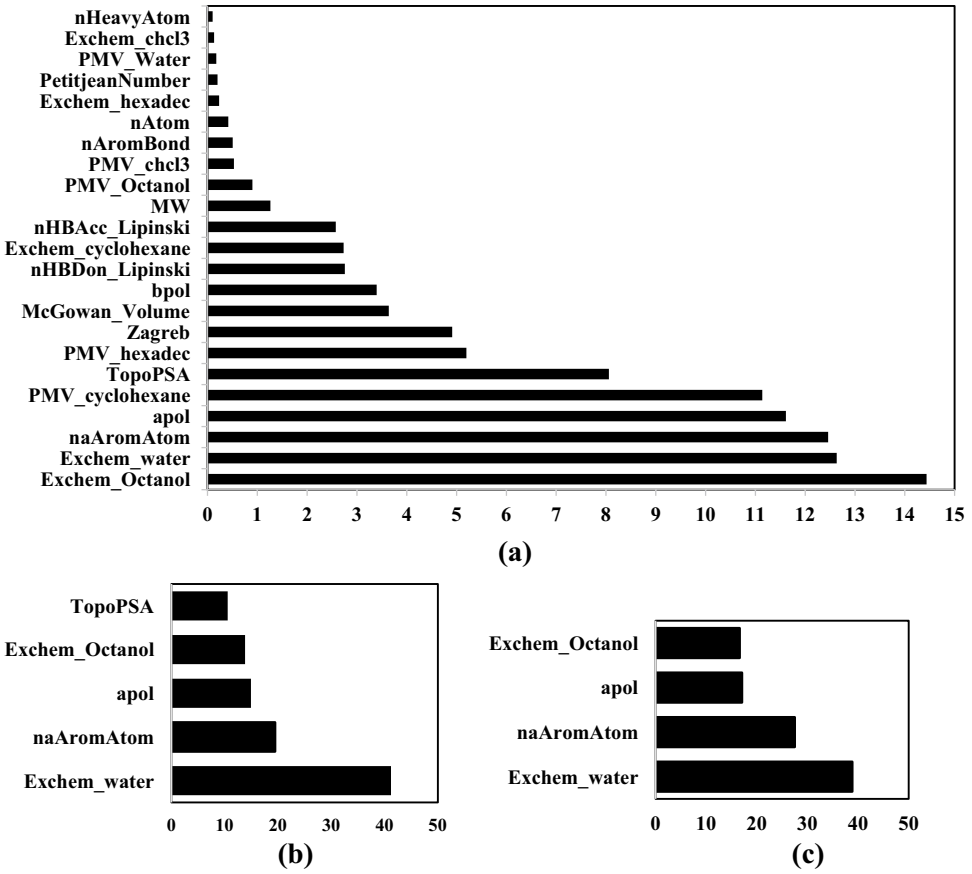


**Fig. 1** Relative importance of descriptors in the models obtained from statistical importance analysis descriptors with the GBM and RF methods. Upper panel: *Model-23d* was developed with 23 crucial descriptors obtained by statistical importance analysis of 354 descriptors. Lower panel: *Model-5d* (right) and *model-4d* (left) were developed with 5 and 4 crucial descriptors obtained by further statistical importance analysis of 23 crucial descriptors, respectively. Statistical importance of each descriptor (in percentage) given on X-axis

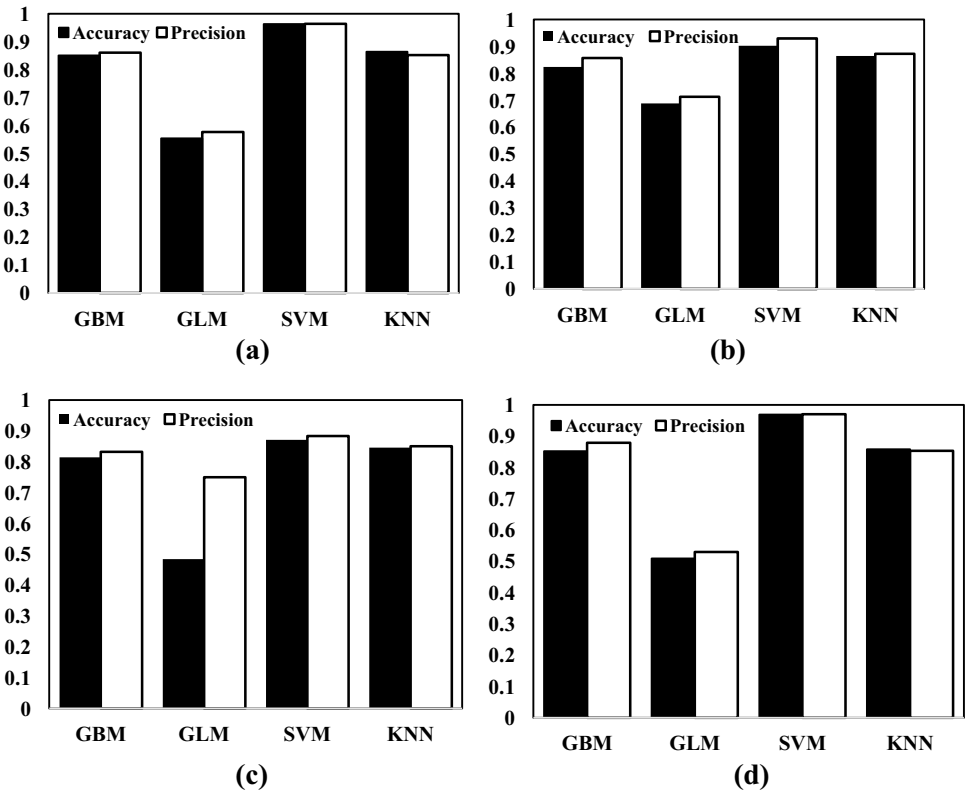


**Fig. 2** Performance indices (Tables S4–S7 and S2 in the ESM) of different machine learning schemes used for classification of PgP± compounds. **a** *model-23d* (average accuracy and precision for runs a to e, Table S4 in the ESM), **b** *model-5d*, **c** *model-4d*, **d** best accuracy and precision for different classification schemes used for *model-23d*

95.6–96.9% and 85.2–87.4%, respectively.[1] The SVM shows the best accuracy of ~97% among the four classification schemes with *model-23d*. The GLM method shows a low accuracy in identify the PgP± compounds using *model-23d*. The GBM and weighted-kNN methods show better accuracy than the GLM method. Similar trends in accuracy were observed in *model-5d* and *model-4d* with four classification schemes. The accuracy of the GBM, GLM, SVM, and Weighted kNN classification methods with *model-5d* is 82.4%, 68.9%, 90.3% and 86.4%, respectively. The accuracy of the GBM, GLM, SVM, and Weighted kNN classification methods with *model-4d* is 81.4%, 48.4%, 87.1% and 84.6%, respectively. The SVM method shows the best accuracy, whereas the GLM method shows a low accuracy with *model-5d* and *model-4d*. The GBM and weighted-kNN methods performed better than the GLM method with all the descriptor based models. Among all the different classification schemes used with the three models, the SVM identified the PgP± compounds with the best accuracy in the range of 87.1 to 96.9%. The stability of the statistical models was tested by randomly removing data points from the test set (50–100 points) and recalculating the statistical performance indices of the new test sets with a reduced number of data points. The best accuracy in identifying the PgP± compounds were achieved with the *model-23d* and SVM method. This model has a higher number of descriptors than the other two models. *Model-23d* was built with 10 of the 3D-RISM-KH based descriptors and 13 of the 2D-descriptors. We compared the performance of the current models with the literature known models. The literature references with their models are summarized in Table 1.

## Conclusions

In conclusion, we have applied our 3D-RISM-KH solvation theory based predictors to construct a PgP inhibitor model, using binary (1/0) values of PgP± compounds. This is the first report providing a proof of concept that 3D-RISM-KH solvation theory-based descriptors can be used successfully to predict the PgP± compounds in a binary fashion. Amongst different models tested here, the maximum descriptor model with the SVM classification scheme showed excellent performance.

In the current study, the 2D descriptors show a significant contribution along with the 3D-RISM-KH based descriptors in predicting for the PgP± compounds. The previous reports also used lipophilicity (log P) [36–38], molecular weight [15, 17, 39], aromaticity [20, 40, 41], and hydrogen bond acceptor [20, 35, 41] as important 2D descriptors to distinguish inhibitors from non-inhibitors. These are not sufficient to reach a high accuracy in predicting the PgP± compounds. The current study specifies that the accounting of 3D-RISM-KH based descriptors along with 2D descriptors in the models show a higher accuracy in predicting the PgP± compounds. The presence of excess chemical potential in water and octanol in the *model-23d* point to the importance of the lipophilicity of molecules, being an important feature for such classification. For a molecule to involve in molecular recognitions in several cellular levels, it has to pass through a series of solvation-desolvation processes. The 3D-RISM-KH based solvation descriptors clearly capture this physical feature of the process. Octanol and cyclohexane are typical mimics of non-polar environment, something a drug molecule experiences on being absorbed from the plasma. The presence of the solvation free energy-based descriptors in our top descriptor list is also in agreement with the previous literature reports [17, 36–38]. The 3D-RISM-KH based solvation free energy descriptors were also used as crucial descriptors for prediction of blood–brain barrier (BBB) and skin permeability [46, 58].

The SVM-PgP± prediction model shows better accuracy in comparison with the literature reported predictive models. The maximum descriptor model may identify the PgP inhibitor compounds with high accuracy and precision. The models act as a tool for early phases of drug discovery to identify the PgP± compounds. The 3D-RISM-KH based descriptors may act as better descriptors for the prediction models to classify the inhibitors of other transporter proteins involved in the MDR.

[1] The performance range is based on five different runs with a different number of test data points, as some of the machine learning methods are known to be size-dependent.

**Table 1** Summary of the previously published predictive models for the PgP± compounds

| References | Type | Model | No. of compounds | Performance[a] |
|---|---|---|---|---|
| [21] | Classification | PLSDA | 325 | 72.40% (272) |
| [59] | Classification | NB | 609 | 82.20% (185) |
| [17] | Classification | RP, NB | 1273 | 81.20% (300) |
| [20] | Classification | PLS-D, LDA | 1275 | 85.0 & 86.0% (85 & 418) |
| [24] | Classification | KNN, SVM, RF | 1935 | 73.0% (334) |
| [60] | Classification | SVM | 1275 | 86.80% (418) |
| [16] | Classification | SVM, KNN, RF, DT, BQSAR | 1954 | 75.0 & 82 .0% (346 & 407) |
| [18] | Classification | Ensemble (FDA, RF, SVM) models | 2079 | 83.2%–86.7% (1040) |
| [61] | Classification | AECF | 2079 | 88.98% (460) |
| [28] | Correlation | BRNN | 57 | 72.8% (14) |
| [62] | Correlation | PLS | 58 | 60.0% (30) |
| [29] | Correlation | MLR, SVM | 70 | 81.0% (14) |
| [22] | Pharmacophore Correlation | Spearman's correlation | 40 | 70.0% (21) |
| [19] | Classification | RF, SVM, kNN, NB | 478 | 86.0% (187) |
| [23] | Classification | Pharmacophore | 272 | 84.2% (152)<br>100% (74) |
| [26] | Classification Pharmacophore | CHAID, QUEST, SVM, Bayesian network, C&R tree, neural network, C5.0, logistic regression, ensemble, decision list | 1690 | 92.00% (419)<br>100% (22) |
| [27] | Correlation | PhE/SVM | 130 | 87% (88)<br>96% (11) |
| [32] | Pharmacophore | CPH | 80 | 67.9% (20) |
| [63] | Classification | RT, C4.5, consensus | 59 | 80.0% (20) |
| [64] | Classification | ANN, SVM, ensemble | 135 | 89.0% (1120) |
| Present study | Classification | SVM | 1275 | 95.6–96.9% (320) |

*PLSDA* partial least squares discriminant analysis, *NB* naive Bayes, *RP* recursive partitioning, *LDA* linear discriminant analysis, *PLS-D* partial least-squares discriminant, *RF* random forest, *kNN* kappa nearest neighbor, *SVM* support vector machine, *DT* decision tree, *BQSAR* binary QSAR, *FDA* flexible discriminant analysis, AECF adaptive ensemble classification framework, *BRNN* Bayesian-regularized neural network, *MLR* multiple linear regression, *CHAID* Chi-square Automatic Interaction Detector, *QUEST* quick, unbiased, efficient statistical tree, *PhE/SVM* pharmacophore ensemble/support vector machine, *CPH* common pharmacophore hypothesis, *RT* random tree

[a]Predicted accuracy for test set compounds given in bracket

**Acknowledgements** This work was financially supported by the NSERC Discovery Grant (RES0029477), and Alberta Prion Research Institute Explorations VII Research Grant (RES0039402). Generous computing time provided by WestGrid (www.westgrid.ca) and Compute Canada/Calcul Canada (www.computecanada.ca) is acknowledged.

## References

1. Goldstein LJ, Galski H, Fojo A, Willingham M, Lai SL, Gazdar A, Pirker R, Green A, Crist W, Brodeur GM, Lieber M, Cossman J, Gottesman MM, Pastan I (1989) J Natl Cancer Inst 81:116–124
2. Juliano RL, Ling V (1976) Biochim Biophys Acta 455:152–162
3. Chen CJ, Chin JE, Ueda K, Clark DP, Pastan I, Gottesman MM, Roninson IB (1986) Cell 47:381–389
4. Gottesman MM, Pastan I (1993) Annu Rev Biochem 62:385–427
5. Sikic BI, Fisher GA, Lum BL, Halsey J, Beketic-Oreskovic L, Chen G (1997) Cancer Chemother Pharmacol 40:S13–19
6. Germann UA, Harding MW (1995) J Natl Cancer Inst 87:1573–1575
7. Beketic-Oreskovic L, Duran GE, Chen G, Dumontet C, Sikic BI (1995) J Natl Cancer Inst 87:1593–1602
8. Chung FS, Santiago JS, Jesus MF, Trinidad CV, See MF (2016) Am J Cancer Res 6:1583–1598
9. Krishna R, Mayer LD (2000) Eur J Pharm Sci 11:265–283
10. Lum BL, Gosland MP (1995) Hematol Oncol Clin N Am 9:319–336
11. Amin ML (2013) Drug Target Insights 7:27–34
12. Szakacs G, Varadi A, Ozvegy-Laczka C, Sarkadi B (2008) Drug Discov Today 13:379–393
13. Kelly RJ, Draper D, Chen CC, Robey RW, Figg WD, Piekarz RL, Chen X, Gardner ER, Balis FM, Venkatesan AM, Steinberg SM, Fojo T, Bates SE (2011) Clin Cancer Res 17:569–580
14. Mohana S, Ganesan M, Agilan B, Karthikeyan R, Srithar G, Mary RB, Ambudkar SV (2016) Mol Biosyst 12:2458–2470

15. Syed SB, Arya H, Fu IH, Yeh TK, Periyasamy L, Hsieh HP, Coumar MS (2017) Sci Rep 7:7972
16. Klepsch F, Vasanthanathan P, Ecker GF (2014) J Chem Inf Model 54:218–229
17. Chen L, Li Y, Zhao Q, Peng H, Hou T (2011) Mol Pharm 8:889–900
18. Yang M, Chen J, Shi X, Xu L, Xi Z, You L, An R, Wang X (2015) Mol Pharm 12:3691–3713
19. Broccatelli F (2012) J Chem Inf Model 52:2462–2470
20. Broccatelli F, Carosati E, Neri A, Frosini M, Goracci L, Oprea TI, Cruciani G (2011) J Med Chem 54:1740–1751
21. Crivori P, Reinach B, PezzettaItalo D, Poggesi I (2006) Mol Pharm 3:33–44
22. Ekins S, Kim RB, Leake BF, Dantzig AH, Schuetz EG, Lan LB, Yasuda K, Shepard RL, Winter MA, Schuetz JD, Wikel JH, Wrighton SA (2002) Mol Pharmacol 61:974–981
23. Ferreira RJ, dos Santos DJ, Ferreira MU, Guedes RC (2011) J Chem Inf Model 51:1315–1324
24. Poongavanam V, Haider N, Ecker GF (2012) Bioorg Med Chem 20:5388–5395
25. Schyman P, Liu R, Wallqvist A (2016) ACS Omega 1:923–929
26. Ngo T-D, Tran T-D, Le M-T, Thai K-M (2016) SAR QSAR Environ Res 27:747780
27. Leong MK, Chen H-B, Shih Y-H (2012) PLoS One 7e33829.
28. Wang YH, Li Y, Yang SL, Yang L (2005) J Comput Aided Mol Des 19:137–147
29. Wu J, Li X, Cheng W, Xie Q, Liu Y, Zhao C (2009) Qsar Comb Sci 28:969–978
30. Ramu A, Ramu N (1994) Cancer Chemother Pharmacol 34:423–430
31. Pajeva IK, Wiese M (2002) J Med Chem 45:5671–5686
32. Tawari NR, Bag S, Degani MS (2008) J Mol Model 14:911–921
33. Pajeva IK, Globisch C, Wiese M (2009) ChemMedChem 4:1883–1896
34. Shukla S, Kouanda A, Silverton L, Talele TT, Ambudkar SV (2014) Mol Pharm 11:2313–2322
35. Langer T, Eder M, Hoffmann RD, Chiba P, Ecker GF (2004) Arch Pharm 337:317–327
36. Wang RB, Kuo CL, Lien LL, Lien EJ (2003) J Clin Pharm Ther 28:203–228
37. Tardia P, Stefanachi A, Niso M, Stolfa DA, Mangiatordi GF, Alberga D, Nicolotti O, Lattanzi G, Carotti A, Leonetti F, Perrone R, Berardi F, Azzariti A, Colabufo NA, Cellamare S (2014) J Med Chem 57:6403–6418
38. Pellicani RZ, Stefanachi A, Niso M, Carotti A, Leonetti F, Nicolotti O, Perrone R, Berardi F, Cellamare S, Colabufo NA (2012) J Med Chem 55:424–436
39. Pleban JK, Rinner U, Chiba P, Ecker GF (2012) J Med Chem 55:3261–3273
40. Globisch C, Pajeva IK, Wiese M (2006) Bioorg Med Chem 14:1588–1598
41. Parveen Z, Brunhofer G, Jabeen I, Erker T, Chiba P, Ecker GF (2014) Bioorg Med Chem 22:2311–2319
42. Molecular Operating Environment (MOE) (2018) 2013.08; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7
43. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902–3909
44. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Petersson GA, Nakatsuji H et al. (2016) Gaussian16, revision B.01. Gaussian Inc.: Wallingford (complete citation is provided in the ESM).
45. Case DA, Ben-Shalom IY, Brozell SR, Cerutti DS, Cheatham TEIII, Cruzeiro VWD, Darden TA, Duke RE, Ghoreishi D, Gilson MK et al. AMBER 2018, University of California, San Francisco. (complete citation is provided in the ESM)
46. Roy D, Hinge VK, Kovalenko A (2019) ACS Omega 4:3055–3060
47. Roy D, Kovalenko A (2019) J Phys Chem A 123:4087–4093
48. Rappe AK, Casewit CJ, Colwell KS, Goddard WA, Skiff WM (1992) J Am Chem Soc 114:10024–10035
49. Kovalenko A, Ten-no S, Hirata F (1999) J Comput Chem. 20:928–936
50. Core Team R (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
51. Robinson D, Gomez M, Demeshev B, Menne D, Nutter B, Johnston L, Bolker B, Briatte F, Arnold J, Gabry J, Broom (2017) Convert statistical analysis objects into tidy data frames
52. Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer-Verlag, New York
53. Kuhn M (2008) J Stat Softw 28(5):1–26
54. Ridgeway G (2007) Generalized boosted models: a guide to the gbm Package. R package vignette. https://CRAN.R-project.org/package=gbm
55. Liaw R, Wiener M (2002) R News 2:18–22
56. 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/package=e1071
57. Schliep, K.; Hechenbichler, K (2016) Weighted k-Nearest Neighbors for Classification, Regression and Clustering. R package version 1.3. https://cran.r-project.org/package=kknn
58. Hinge VK, Roy D, Kovalenko A (2019) J Comput Aided Mol Des 33:605–611
59. Sun HM (2005) J. Med. Chem. 48:4031–4039
60. Tan W, Mei H, Chao L, Liu TF, Pan XC, Shu M, Yang L (2013) J Comput-Aided Mol Des 27:1067–1073
61. Yang M, Chen J, Xu L, Shi X, Zhou X, Xi Z, An R, Wang X (2018) RSC Adv 8:11661–11683
62. Müller H, Pajeva IK, Globisch C, Wiese M (2008) Bioorg Med Chem 16:2448–2462
63. Rapposelli S, Coi A, Imbriani M, Bianucci AM (2012) Int J Mol Sci 13:6924–6943
64. Eric S, Kalinic M, Ilic K, Zloh M (2014) SAR QSAR Environ Res 25:939–966


**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.