Using Machine Learning to predict partition co-efficient (Log P) and distribution co-efficient (Log D) with molecular descriptors and liquid chromatography retention time

Journal:	*Journal of Chemical Information and Modeling*
Manuscript ID	ci-2022-01373z.R2
Manuscript Type:	Article
Date Submitted by the Author:	14-Feb-2023
Complete List of Authors:	Win, Zaw-Myo; Centre for Eye and Vision Research Limited, CHEONG, Allen; The Hong Kong Polytechnic University Hopkins, W.; University of Waterloo, Department of Chemistry

Using Machine Learning to Predict Partition Coefficient (Log P) and Distribution Coefficient (Log D) with Molecular Descriptors and Liquid Chromatography Retention Time

Zaw-Myo Win [1,2,3], *Allen M Y Cheong**[1,2], *W. Scott Hopkins**[1,3,4,5]

[1] Centre for Eye and Vision Research, Hong Kong Science Park, New Territories, 999077, Hong Kong

[2] School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong

[3]Department of Chemistry, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada.

[4]Waterloo Institute for Nanotechnology, University of 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada

[5]WaterMine Innovation, Inc., Waterloo, Ontario, N0B 2T0, Canada

Abstract

During pre-clinical evaluations of drug candidates, several physicochemical (p-chem) properties are measured and employed as metrics to estimate drug efficacy in vivo. Two such p-chem properties are the octanol-water partition coefficient, Log P, and distribution coefficient, Log D, which are useful in estimating the distribution of drugs within the body. Log P and Log D are traditionally measured using the shake-flask method and high-performance liquid chromatography (HPLC). However, it is challenging to measure these properties for species that are very hydrophobic (or hydrophilic) owing to the very low equilibrium concentrations partitioned into octanol (or aqueous) phases. Moreover, the shake-flask method is relatively time-consuming and can require multistep dilutions as the range of analyte concentrations can differ by several orders of magnitude. Here, we circumvent these limitations by using machine learning (ML) to correlate Log P and Log D with liquid chromatography (LC) retention time (RT). Predictive models based on four ML algorithms, which used molecular descriptors and LC RTs as features, were extensively tested and compared. The inclusion of RT as an additional descriptor improves model performance (MAE = 0.366 and R^2 = 0.89), and SHAP analysis indicates that RT has the highest impact on model accuracy.

KEYWORDS: Physicochemical properties, Machine learning, Cheminformatics

Introduction

In early-stage drug discovery, candidate molecules must be screened to identify pharmacological activity and physicochemical (p-chem) properties.[1,2,3,4] This process commonly begins with a computer-aided molecular design and *in silico* property predictions which are then experimentally refined for a selected subset of molecules.[4] With regard to p-chem properties, medicinal chemists are concerned with the tangible physical attributes that are related to molecular interactions with different media and environments.[4] Some of the most important parameters in this context are the pKa (acid dissociation constant), PSA (polar surface area), solubility, and lipophilicity (as defined by the partition and distribution coefficients, Log P and Log D).[5,6,7–9] Log D is a measure of the concentration ratio of an ionizable compound following equilibrated distribution between water and a hydrophobic solvent (*e.g.*, octanol), whereas the partition coefficient (Log P) refers to the concentration ratio of un-ionized compounds. For non-ionizable compounds, Log P is equal to Log D for all pH, whereas for ionizable compounds Log D considers the partition of both ionized and non-ionized forms. For example, drug molecules must possess a suitable water solubility for transport in aqueous media like body fluids, yet must also exhibit lipophilicity suitable for the environments of drug action and transport (*e.g.*, through membranes).[4] Identifying molecules that lie in this "goldilocks zone" of compromise while simultaneously exhibiting appropriate activities, pharmacodynamic responses, and pharmacokinetic exposures (while also minimizing toxicity and off-target activity) lies at the heart of medicinal chemistry.[10,11]

For a compound to be a drug candidate, it should have a lipophilicity (Log P/Log D) value between 1 and 3.[12,13] If the value is lower than this range, the compound will have low membrane permeability, whereas species with lipophilicity > 3 exhibit poor absorption.[14] Traditionally, lipophilicity is determined by using methods such as water-octanol shake flask and high-performance liquid chromatography (HPLC) to partition and measure both un-ionized and ionized molecular concentration in the aqueous and hydrophobic phases.[15] Although relatively time-consuming (measurements can take several hours), the shake flask method provides a direct and accurate determination of Log P and Log D.[16] However, it is necessary to conduct measurements with purified solvents and analytes to prevent confusion in analysis and unintended matrix effects.[17] Analyte purification post-synthesis can be accomplished using column chromatography. In fact, reaction mixtures are commonly characterized using LC (typically coupled with mass spectrometry; MS) to confirm the presence of the desired product and estimate product yield. This serves the dual purposes of separating the components of the mixture and determining analyte affinity for the column packing material (measured as retention time; RT), which is also a metric that may be employed for compound characterization and identification. Using this RT information for the added purpose of determining Log P and Log D (or validating subsequent measurements) would be desirable.

Recently, we have reported correlations between gas phase dynamic ion-solvent clustering behaviour and solution phase p-chem properties[18–20]. We hypothesize that the clustering process effectively samples the ion-solvent interaction potential, and that the gas phase interactions correlate strongly with the condensed phase interactions that give rise to p-chem properties such as lipophilicity. To use these data to create general predictive models for p-chem properties of interest, it is necessary to treat dynamic clustering data with supervised machine learning (ML). A variety of ML algorithms have been used for regression and classification purposes in drug discovery studies,[21–27,28,29] the most common of which are support vector machine (SVM),[23] random forest (RF),[30] and multi-layer perceptron (MLP; *i.e.*, a feedforward artificial neural network).[31] Regarding Log P predictions, several methods have been reported that employ physicochemical descriptors to represent molecules.[12,32–38] Given our success with gas phase ion-solvent clustering, we hypothesize that incorporating LC RTs as a feature will improve ML model accuracy since this parameter will provide an experimental measurement of the condensed phase analyte-solvent interaction potential. We test this hypothesis in this work.

In this study, we investigate 2070 molecules from the small molecule retention time (SMRT) data set for which p-chem properties could be found in the ChEMBL database.[39,40] These molecules are represented as a vector of molecular descriptors as determined from SMILES codes via the RDKit software.[41] To these feature vectors, we append RT as an additional feature. We then evaluate the accuracy of predictive models based on the supervised ML algorithms: SVM, MLP, RF, and extreme gradient boosting (XGB). We show that the use of RT as a descriptor improves the accuracy of the ML models, that one can approach the inherent accuracy of the target variable (*e.g.*, Log P, Log D) with only 2070 molecules in the training set, and that including additional molecules in the training set further improves model accuracy.

Methods

Molecules from the METLIN SMRT dataset were screened against the ChEMBL database to identify compounds for which RT, Log P and Log D were known.[39] The METLIN small molecule RT (SMRT) dataset contains 80,038 small molecules that were measured with single reversed-phase method LC-MS using a Zorbax Extend-C18 reverse-phase column (2.1 x 50 mm, 1.8μm, Agilent Technologies, Santa Clara,CA).[39] RT variability was measured at 36 s. Further details of the dataset are provided in the supporting information. This process yielded a data set of 2070 compounds. The SMILES notations for these compounds were then used in conjunction with the open-source software RDKit (Version: 2021.03.4)[41] to generate a vector of 204 molecular descriptors, which included properties such as atom-type, molecular weight, and number of rotatable bonds. Features with low variance (< 0.05), high degrees of correlation with another feature (<0.95), missing values, and zero values (for all analytes) were removed from the data set prior to scaling each feature to a mean value of 0 and variance of 1. Following pruning, a total of 125 RDKit descriptors were carried forward for ML treatment. The data set and a detailed description of the molecular descriptors[42] is provided in the supporting information that accompanies this manuscript (supporting information, Table S1 and Table S2). The original SMRT dataset consists of data for molecules of seven chemical classes: organoheterocyclic compounds, organic acids, organic nitrogen compounds, benzenoids, organic oxygen compounds, organosulfur compounds, and “other compounds”, which includes lipids, lignans, nucleosides, nucleotides, phenylpropanoids, and polyketides. Our pruned dataset includes compounds from each of these classes.

Machine learning models

Four ML algorithms (*i.e.*, SVM, MLP, XGB, RF) were used to develop the descriptor-based models. These ML algorithms are implemented in the scikit-learn package (Version: 0.24.2) of Python (Version: 3.9.6 × 64).[43] The complete list of hyper-parameters optimized per ML method is provided in the supporting information, Table S6.

Multilayer perceptron (MLP) is a fully connected artificial neural network (ANN) which trains using backpropagation.[44] It consists of five layers of nodes: an input layer, three hidden layers, and an output layer. Each neuron in MLP uses a nonlinear activation function except the input layer. The following hyper-parameters were optimized: the size of the hidden layers, the maximum number of iterations, the neuron activation function, the solver, regularization, and the learning rate. Details for all algorithms tested are provided in the supporting information. Other hyper-parameters were set at default values as described in the SKLearn software package. Extreme gradient boosting (XGB) is a decision tree-based method that minimized errors associated with bias and variance.[45] In the training of XGB, the following hyper-parameters were optimized: learning rate, number of boosting stages, subsample, minimum number of samples for internal node, and minimum number of samples for leaf node. Support vector machine (SVM) is a popular machine learning method based on statistical learning theory.[23,24,46] Here, a linear kernel that was used for which the regularization parameter (C) was optimized in the range of 0.1 - 100. Random forest (RF) is a decision tree-based learning algorithm (like XGB) which employs ensemble learning for classification and regression.[47] The following hyper-parameters were optimized: number of trees in the forest, maximum depth of the tree, minimum number of samples, minimum impurity decrease, and number of features.

Hyper-parameter tuning

To identify the ideal hyper-parameters for the ML algorithms, the hyperopt package (Version: 0.2.5)[48] was employed to optimize algorithm hyperparameters (*e.g.*, number of trees in the RF treatment). Python scripts are available via our research group GitHub site (https://github.com/jamesleocodes/p_chem_CEVR.git). To evaluate model performance, the dataset was randomly split into training (80%), validation (10%), and testing (10%) sets. Following model training, predictions ($\hat{y}_i$) were made using the out-of-the-bag test set and were subsequently compared to the literature values (y_i) to calculate the mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). These metrics are described by equations 1 – 4, respectively.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (3)$$

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y}_i)^2} \qquad (4)$$

Where *n* is total number of instances in the test set and $\bar{y}_i$ represents the mean of the values.

Results & Discussion

Distribution of parameters

Figure 1 shows the distribution of values for Log P, Log D, and RT of the compounds in the data set used for this study. Distributions were approximately Gaussian in nature for the distribution coefficients, which span *ca*. 12 orders of magnitude, and partition coefficients, which span *ca*. 5 orders of magnitude. RT values ranged from *ca*. 200 – 1400 s. As an initial test to assess whether RT might be a useful feature for predicting p-chem properties, Spearman's correlation coefficients (ρ) were computed for Log P, Log D, and RT. Spearman's correlation assess monotonic relationships (whether linear or not), rather than simply assessing linear relationships as does Pearson correlation. Figure 2 shows a heatmap of the computed ρ values; Log P and Log D values in our dataset are positively correlated with RT. Out of curiosity, we also computed the correlation coefficients of polar surface area (PSA) and pKa, which were weakly anti-correlated. The correlation between RT and Log P/Log D values suggested that including RT as a descriptor in a ML model might improve prediction accuracy and/or reduce the number of instances required for model training.

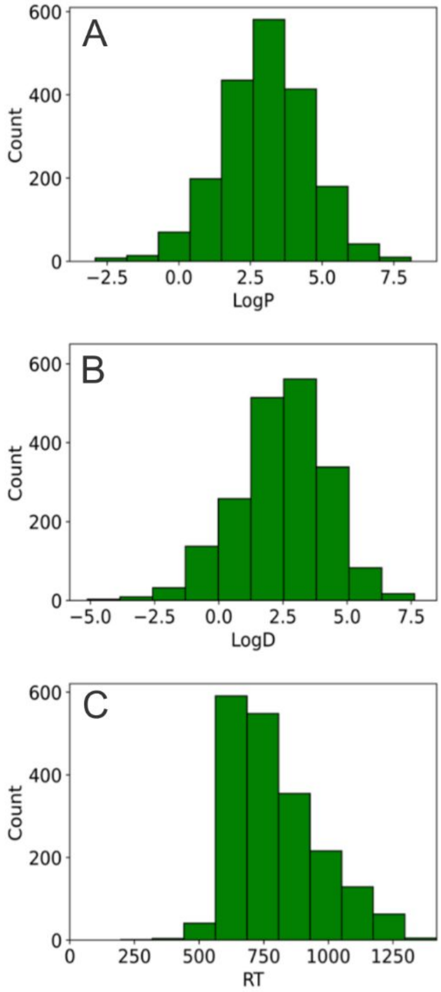


Figure 1. Distribution of values of (A) Log P, (B) Log D, and (C) RT of compounds in the data set used for this study.

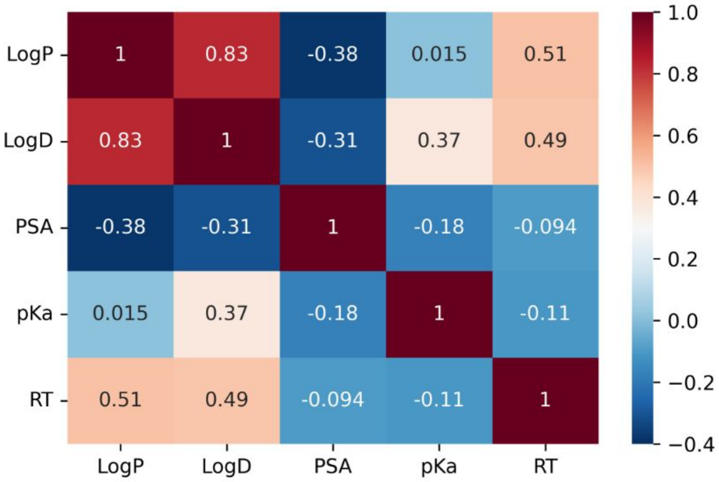


Figure 2. The heat map of Spearman's correlation coefficient (ρ) for experimental retention time and physicochemical properties in the dataset. Values of 1, 0, and –1 indicate perfect correlation, no correlation, and anti-correlation, respectively.

ML Model Construction

For ML model training and testing purposes, the extracted data were randomly divided into training (80%), validation (10%), and testing (10%) sets as per conventional protocol.[49,50] The hyperopt optimization technique, which uses a form of Bayesian optimization to identify the best parameter for a given model,[48] was employed in conjunction with the validation set to optimize model hyper-parameters. Fifty independent runs with different random seeds for data splitting were performed and the average taken to ensure results are statistically valid. Since all the models in this study were for regression tasks, they were evaluated predominantly by mean absolute error (MAE), but other metrics (as described above) were also assessed to ensure statistical validity.

Fixed molecular representations were used as input data to build predictive models for molecular properties.[51–57] Fixed representations such as Morgan fingerprints are widely known and commonly used as molecular descriptions.[58] Here, the RDKit python library was used to generate a Morgan Fingerprint that contains 204 molecular descriptors.[41,42]

The best performing model was the MLP neural network. Hyper-parameter optimization resulted in an MLP configuration with three hidden layers having 120, 80, 40 neurons, respectively, maximum iteration of 100, a hyperbolic tangent (tanh) activation function, an alpha parameter of 0.0001, and an constant learning rate. Weights and bias updates are handled via momentum, specifically through the Adam routine, implemented as described in the work of Kingma and Ba.[59] Using the MLP algorithm, MAE = 0.366 ± 0.057 was achieved for Log P and MAE = 0.470 ± 0.030 was achieved for Log D predictions (see Tables 1 and 2). From the learning curves for the MLP model for Log P (shown in Figure 3a) and Log D (shown in Figure 3c), it is apparent that the testing set error decreases with increasing size of training set, but that the error has not plateaued to a constant value, indicating that including additional instances in the data set will further improve model performance. Correlation plots of the predicted versus experimental values of Log P and Log D are presented in Figures 3b and 3d, respectively. Correlations were quite strong, with correlation coefficients of R^2 = 0.88 and R^2 = 0.85 for Log P and Log D, respectively. Furthermore, as can be seen in Figure 3b and 3d, much of the prediction error occurs at low values of Log D and Log P, where there were relatively few experimental measurements in the data set (see Figure 1) and where there tends to be more error associated with experimental measurements.

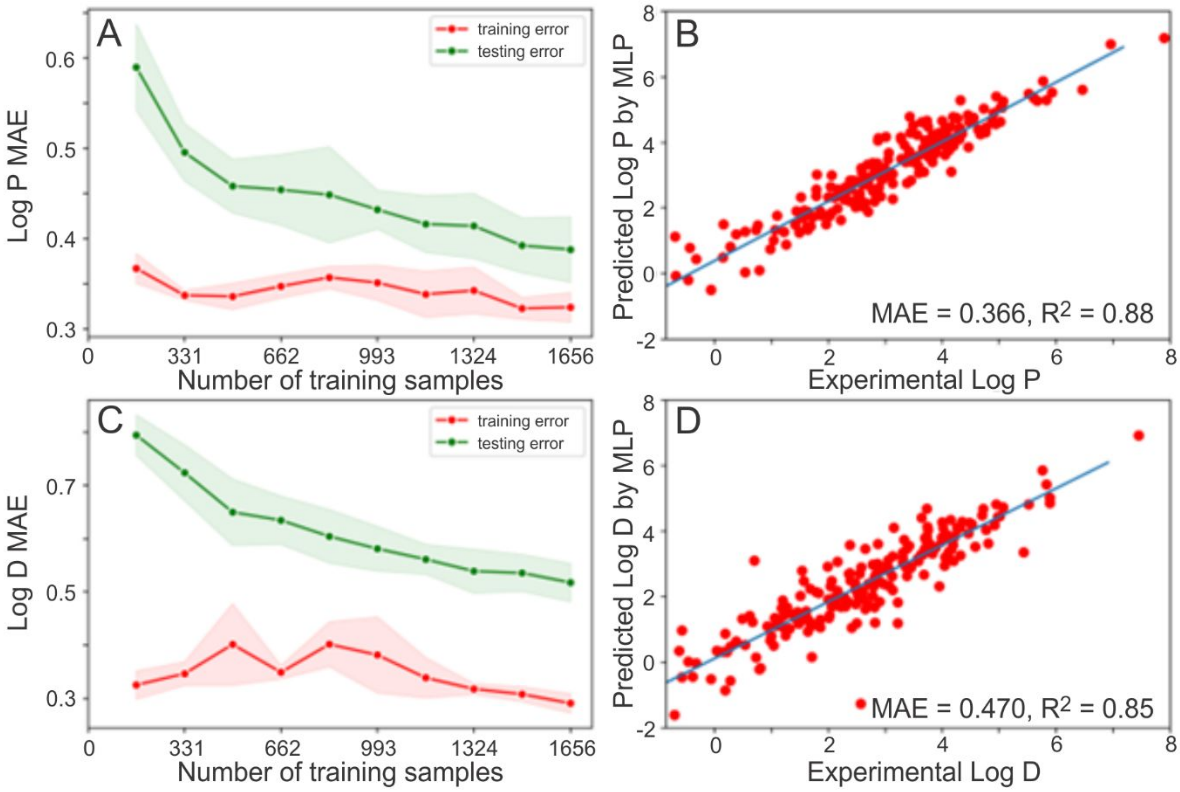


Figure 3. (A) Mean absolute error (MAE) for Log P predictions for the training and test set as a function of training set size. (B) Correlation plot of experimentally determined Log P values and Log P values predicted by the MLP model. (C) Mean absolute error (MAE) for Log D predictions for the training and test set as a function of training set size. (D) Correlation plot of experimentally determined Log D values and Log D values predicted by the MLP model.

The XGB, SVM, and RF models were less accurate than the MLP model. Details for these other models, including learning curves and correlation plots, are provided in the supporting information. Table 1 presents the performance results for the MLP model of Log P and Table 2 provides the results for the MLP model of Log D. Table 3 shows the three largest deviations between the literature Log P values (from ChEMBL) and those predicted by the MLP algorithm. The complete list of deviations predicted by other methods is provided in the supporting information, Tables S8-10.

To evaluate the models, fifty independent runs with different random seeds were conducted for generating the training, validation, and test sets at a splitting ratio of 80% : 10% : 10%. Model accuracies and precisions, as reported in Tables 1 and 2, indicate that there is a slight overfitting of the training set data. Nevertheless, test set accuracies indicate that our MLP model, which includes RTs, improves on other recently published models. For example, the MLP model reported by Datta *et al.* uses the DeepChem database and performs with an estimated accuracy of MAE = 0.477 on Log P predictions, 30% larger the error that we achieved when including RT as a feature in the dataset.[38] Samarjeet *et al.* also recently developed a predictive model for Log P, which yields RMSE = 0.61 for a set of 11 drug-like molecules provided by SAMPL6, 20% larger than the error that we obtain from our treatment.[32] Ulrich *et al* developed a deep neural network (DNN) model with RMSE = 0.50, similar to the performance of our model.[35] The dataset used by Ulrich *et al.* is classified depending on the number of nonhydrogen atoms (NHA). They are molecules with an NHA of 1-10, molecules with an NHA of 11-20, molecules with an NHA of 21-30, and molecules with above an NHA of 30. [35] Ulrich *et al.* converted SMILES codes into molecular graphs for input features and required nearly 14,000 instances to achieve the same error as our model does with ~2,100 instances. Unfortunately, learning curves were not provided, so we can't comment as to if Ulrich's model will further improve with additional data.

Table 1. Performance of the MLP model for Log P predictions. Statistics were assessed for fifty independent, randomized runs. Mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the correlation coefficient (R^2) for predicted versus measured values are provided. The precision of the statistical metrics are reported as $\pm\sigma$ (i.e., 68% confidence interval).

Metric	Training	Validation	Test
MSE	0.172 ± 0.012	0.255 ± 0.050	0.260 ± 0.057
RMSE	0.414 ± 0.014	0.503 ± 0.048	0.507 ± 0.054
MAE	0.307 ± 0.011	0.363 ± 0.022	0.366 ± 0.024
R2	0.927 ± 0.005	0.889 ± 0.019	0.885 ± 0.030

Table 2. Performance of the MLP model for Log D predictions. Statistics were assessed for fifty independent, randomized runs. Mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the correlation coefficient (R^2) for predicted versus measured values are provided. The precision of the statistical metrics are reported as $\pm\sigma$ (i.e., 68% confidence interval).

Metric	Training	Validation	Test
MSE	0.144 ± 0.017	0.446 ± 0.071	0.423 ± 0.065
RMSE	0.378 ± 0.022	0.666 ± 0.052	0.649 ± 0.051
MAE	0.498 ± 0.017	0.487 ± 0.033	0.470 ± 0.031
R2	0.953 ± 0.006	0.854 ± 0.023	0.857 ± 0.026

Table 3. The three largest deviations between literature and predicted Log P values.

Molecule	Literature Log P (ChEMBL)	Predicted Log P	Deviation
3-(4-oxo-1,2,3-benzotriazin-3-yl)propanoic acid	1.73	-0.26	1.96
(6S)-2-nitro-6-[[4-(trifluoromethoxy)phenyl] methoxy]-6,7-dihydro-5H-imidazo[2,1-b][1,3]oxazine	4.14	2.28	1.86
N-butyl-N-ethyl-2,5-dimethyl-7-(2,4,6-trimethylphenyl)pyrrolo[2,3-d]pyrimidin-4-amine	7.36	5.53	1.83

To explore the importance of the various features in our model, we conducted a SHAP (*i.e.*, Shapley Additive exPlanations) analysis.[60] The SHAP explanation method computes Shapley values from coalition game theory. The feature values of a data instance act as players in a coalition and the Shapley values describe how to distribute the prediction among the various features. In other words, the Shapley value of a given feature describes its importance to the accuracy of the prediction. In our data, each instance (*i.e.*, analyte molecule) is described by a vector of 125 features (see Table S2). Figure 4 shows a SHAP summary plot for the ten most important features in the dataset with respect to Log P predictions. Each point on the summary plot is a Shapley value for a feature and an instance; the position on the x-axis is determined by the Shapely value and the position on the y-axis is determined by the feature. The heatmap of Figure 4 shows the feature values from low to high. Points that overlap are jittered in the y-axis direction to provide a sense of the distribution of the Shapley values per feature. The features are ordered according to their importance, indicating that the most important feature for the predictive MLP model of Log P is the experimental RT.

One can interpret Shapley values as “forces” that either increase or decrease the prediction from a baseline value that corresponds to the average of all predictions. We find that high values of the RT “push” the predicted Log P to higher values and low values of RT push the predicted Log P to lower values. In other words, analytes that interact more strongly with the C18 column exhibit higher water:octanol partition coefficients.

The partition coefficient is defined by equation 5:

$$LogP_{octanol/water} = \log_{10}\left(\frac{[Solute]_{octanol}^{un-ionized}}{[Solute]_{water}^{un-ionized}}\right) \tag{5}$$

Species that bind more strongly to the C18 column. C18 is octyldecylsilane, which contains a linear eighteen carbon linear alkyl chain bound to the silica support. Molecules that bind to the C18 column have an affinity for the non-polar environment and are expected to exhibit likewise exhibit an affinity to the hydrophobic octanol environment. Consequently, we observe strong correlation between RT and the Log P value for a given analyte.

Similar, albeit less impactful contributions, are observed for the next three most important features: PEOE_VSA7, VSA_Estate6, and PEOE_VSA6. These three features describe van der Waals surface area (*i.e.*, VSA) charges for the analytes as estimated using the Molecular Operating Environment (MOE).[61] The correlation of these features with Log P also aligns with expectation since the van der Waals surface area of a molecule should correlate with molecular polarizability, and highly polarizable species are expected to bind relatively strongly with the C18 alkyl chains and exhibit relatively high solubility in octanol. Interestingly, the fifth-most important feature, the topological polar surface area (TPSA), exhibits a trend opposite to the four features of most importance; high values of TPSA push Log P predictions to lower values. This correlation indicates that polar species exhibit relatively high affinity for the polar aqueous phase (and thus relatively low affinity for the hydrophobic octanol environment), in accordance with chemical intuition, The features ranked from sixth to tenth in importance are two indices of van der walls surface area (Estate_VSA8 and SMR_VSA7), the BCUT descriptor for Log

P,[62] molecular mass (MolMR), and the number of aromatic rings (NumAromaticRings). These correlations can also be explained based on the molecular interactions described above. Below the five features of highest importance, feature impact on model performance diminishes significantly. In fact, including only the top 20 features in the dataset yields a model accuracy of MAE = 0.483.

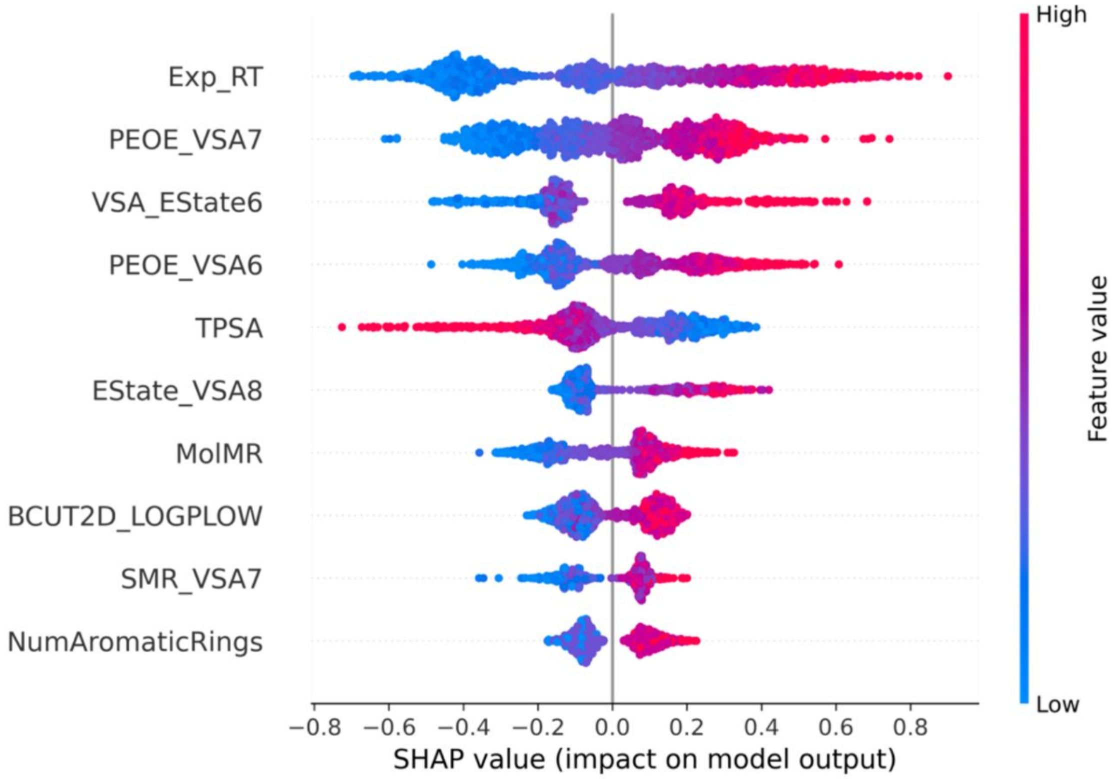


Figure 4. SHAP summary plot showing the impact of descriptors (the top 10) on the model. One dot represents one molecule, and the dots stack up to show its density.

Conclusions

A major challenge in developing accurate ML models is identifying high quality, representative data. With regard to predicting molecular properties, attention has focused on employing structural features owing to the perspective that structure affects functionality and because one can quickly generate a large number of molecular features from a simple SMILES string. By including additional information that encodes molecular interactions, such as experimental LC retention times, one provides the ML algorithm with important information regarding how strongly analytes interact with their environments (*e.g.*, with solvents or substrates). Such information should correlate strongly with several condensed phase properties. Here, we have demonstrated that LC retention times correlate with partition and distribution coefficients, and that inclusion of retention time as a dataset feature improves the accuracy of predictive ML models for Log P and Log D. Moreover, SHAP analysis showed that LC retention time is the most important feature with respect to accuracy for these models. We expect that including LC retention time as a feature will improve model accuracy for other molecular properties (*e.g.*, Log S). An interesting open question is: will the inclusion of other experimentally determined parameters further improve model accuracy or enable accurate predictions of other molecular properties (*e.g.*, pKa)?

Clearly, incorporating retention time as a model feature is not as convenient as simply using the structural features generated from, for example, a SMILES string; one must first measure the retention time, which requires both a physical sample and appropriate instrumentation. Thus, the method that we report here cannot replace the purely *in silico* screening of early-stage drug discovery. However, LC-MS measurements are commonly employed in the characterization of newly synthesized drug candidates, so the method reported here does offer the possibility of re-purposing these LC measurements for property determination (or perhaps confirmation/validation of standard techniques). In addition to pre-clinical drug discovery, the LC RT method described here could find application in other areas. For example, distribution coefficients are important parameters used in mass transport models for environmental contaminants.[64] Given that LC-MS is a commonly used tool in environmental analysis, one could envision simultaneously identifying a trace contaminant and predicting its physicochemical properties such that one could determine the compound’s fate in the environment. One could envision LC-MS measurements of environmental contaminants followed by structural characterization based on the mass spectrometric data and the METLIN

RT data. With molecular structure in hand, one could then employ our LC-based ML model to predict Log P and Log D.

As a final note, the variation in LC retention time between instruments is a well-known challenge for site-to-site comparisons. One strategy to overcome such variation would be the introduction of a site (or instrument) label as a feature in a combined dataset and the use of transductive transfer learning.[63] This method, which assumes a single source domain and single task (*e.g.*, prediction of Log P), can account for sample selection bias / covariance shifts. We are currently exploring the use of transfer learning for combining dataset (or pre-training models) and will report results in due course.

Data and Software Availability

The dataset and python scripts described in this work can be publicly accessed at the open-source GitHub repository "p_Chem_CEVR" (https://github.com/jamesleocodes/p_chem_CEVR.git). The METLIN small molecule(SMRT) dataset is available on (https://doi.org/10.6084/m9 .figshare.8038913.v1). The software tools used in this study, including RDKit (https://www.rdkit.org/), scikit-learn (https://www.scikit-learn.org), NumPy (https://www.numpy.org), SciPy (https://www.scipy.org), and Pandas (https://www.pandas.pydata.org) are freely available at their website.

Supporting Information

Supporting information Available: [LC method. Table S1:Data set used in the article. Table S2:Molecular descriptors, Table S3-S5: Performance comparison of the models. Table S6:List of optimized hyper-parameters per ML method. Table S7:Illustration of SMILES-based molecular construction predicted values by MLP. Table S8-S10:Three most deviations between the experimental values and predicted values for XGB, SVM, and RF. Figure S1-S3:learning curve for XGB , SVM and RF model (Log P). Figure S4:The complete set of performance metrics on the test set using different machine learning models. Figure S5:The learning curves for four models for Log D. Figure S6:the complete set of performance metrics on the test set for Log D using different machine learning models. Figure S7: SHAP summary for Log D]

Corresponding Authors:

*Prof. W. Scott Hopkins, Email: scott.hopkins@uwaterloo.ca

*Prof. Allen M Y Cheong, Email: allen.my.cheong@polyu.edu.hk

Author Contributions

Study concept and design: W.S.H. Writing the code for training the models and evaluating them: W.S.H., Z.M.W. Drafting manuscript: Z.M.W. Revision manuscript: W.S.H., A.M.Y.C., Z.M.W. Approval of published manuscript: W.S.H., A.M.Y.C., Z.M.W.

Acknowledgment

This work was supported by the Government of the Hong Kong Special Administrative Region & Innovation and Technology Fund.

REFERENCES

(1) Wang, T.; Wu, M.; Zhang, R.; Chen, Z.; Hua, C.; Lin, J.; Yang, L.; Li, L.; Chen, W.; Chen, T.; Renand, J.; Xu, Y. Advances in Computational Structure-Based Drug Design and Application in Drug Discovery Structure-Based Discovery of PDEs Inhibitors Recent Advances in Computer-Assisted Structure-Based Identification and Design of Histone Deacetylases Inhibitors Structur. **2016**, 16.

(2) Marshall, S. F.; Burghaus, R.; Cosson, V.; Cheung, S.; Chenel, M.; DellaPasqua, O.; Frey, N.; Hamrén, B.; Harnisch, L.; Ivanow, F.; Kerbusch, T.; Lippert, J.; Milligan, P. A.; Rohou, S.; Staab, A.; Steimer, J. L.; Tornøe, C.; Visser, S. A. G. Good Practices in Model-Informed Drug Discovery and Development: Practice, Application, and Documentation. *CPT Pharmacometrics Syst Pharmacol* **2016**, 5, 93–122.

(3) Hughes, J. P.; Rees, S. S.; Kalindjian, S. B.; Philpott, K. L. Principles of Early Drug Discovery. *Br J Pharmacol* **2011**, 162, 1239–1249.

(4) Bunally, S. B.; Luscombe, C. N.; Young, R. J. Using Physicochemical Measurements to Influence Better Compound Design. *SLAS Discovery*. SAGE Publications Inc. September 1, **2019**, 791–801.

(5) Arnott, J. A.; Planey, S. L. The Influence of Lipophilicity in Drug Discovery and Design. *Expert Opinion on Drug Discovery*. October **2012**, 863–875.

(6) Waring, M. J. Lipophilicity in Drug Discovery. *Expert Opinion on Drug Discovery*. March **2010**, 235–248.

(7) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv Drug Deliv Rev* **1997**, 23 , 3–25.

(8) Lipinski, C. A. Lead- and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discov Today Technol* **2004**, 1, 337–341.

(9) John Comer, K. B. High-Throughput Measurement of Drug PKa Values for ADME Screening. *SLAS TECHNOLOGY: Translating Life Sciences Innovation* **2003**, 8.

(10) Gleeson, M. P.; Hersey, A.; Montanari, D.; Overington, J. Probing the Links between in Vitro Potency, ADMET and Physicochemical Parameters. *Nat Rev Drug Discov* **2011**, 10, 197–208.

(11) Meanwell, N. A. Improving Drug Candidates by Design: A Focus on Physicochemical Properties as a Means of Improving Compound Disposition and Safety. *Chemical Research in Toxicology*. September 19, **2011**, 1420–1456.

(12) Arnott, J. A.; Planey, S. L. The Influence of Lipophilicity in Drug Discovery and Design. *Expert Opin Drug Discov* **2012**, 7 , 863–875.

(13) Waring, M. J. Lipophilicity in Drug Discovery. *Expert Opin Drug Discov* **2010**, 5, 235–248.

(14) Tsantili-Kakoulidou, A.; Demopoulos, V. J. Drug-like Properties and Fraction Lipophilicity Index as a Combined Metric. *ADMET DMPK* **2021**, 9, 177–190.

(15) Rappel, C.; Galanski, M.; Yasemi, A.; Habala, L.; Keppler, B. K. Analysis of Anticancer Platinum(II)-Complexes by Microemulsion Electrokinetic Chromatography: Separation of Diastereomers and Estimation of Octanol-Water Partition Coefficients. *Electrophoresis* **2005**, 26, 878–884.

(16) Schönsee, C. D.; Bucheli, T. D. Experimental Determination of Octanol-Water Partition Coefficients of Selected Natural Toxins. *J Chem Eng Data* **2020**, 65, 1946–1953.

(17) Chim Phys, J.; Dearden, J. C.; Bresnen, G. M. *QSAR and Strategies in the Design of Bioactiue Compounds*; Springer, **1988**; 7.

(18) Ieritano, C.; Hopkins, S. The Hitchhiker's Guide to Dynamic Ion-Solvent Clustering: Applications in Differential Ion Mobility Spectrometry. *Physical Chemistry Chemical Physics* **2022**, 24.

(19) Liu, C.; Yves Le Blanc, J. C.; Schneider, B. B.; Shields, J.; Federico, J. J.; Zhang, H.; Stroh, J. G.; Kauffman, G. W.; Kung, D. W.; Ieritano, C.; Shepherdson, E.; Verbuyst, M.; Melo, L.; Hasan, M.; Naser, D.; Janiszewski, J. S.; Hopkins, W. S.; Campbell, J. L. Assessing Physicochemical Properties of Drug Molecules via Microsolvation Measurements with Differential Mobility Spectrometry. *ACS Cent Sci* **2017**, 3, 101–109.

(20) Walker, S. W. C.; Anwar, A.; Psutka, J. M.; Crouse, J.; Liu, C.; le Blanc, J. C. Y.; Montgomery, J.; Goetz, G. H.; Janiszewski, J. S.; Campbell, J. L.; Hopkins, W. S. Determining Molecular Properties with Differential Mobility Spectrometry and Machine Learning. *Nat Commun* **2018**, 9.

(21) Fang, J.; Yang, R.; Gao, L.; Zhou, D.; Yang, S.; Liu, A.; Du, G. Predictions of BuChE Inhibitors Using Support Vector Machine and Naive Bayesian Classification Techniques in Drug Discovery. *J Chem Inf Model* **2013**, 53, 3009–3020.

(22) Sun, H. A Naive Bayes Classifier for Prediction of Multidrug Resistance Reversal Activity on the Basis of Atom Typing. *J Med Chem* **2005**, 48, 4031–4039.

(23) Zernov, V. v; Balakin, K. v; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. v. Drug Discovery Using Support Vector Machines. The Case Studies of Drug-Likeness, Agrochemical-Likeness, and Enzyme Inhibition Predictions. *J Chem Inf Comput Sci* **2003**, 43, 2048–2056.

(24) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. *J Chem Inf Comput Sci* **2003**, 43, 1882–1889.

(25) Korkmaz, S.; Zararsiz, G.; Goksuluk, D. MLViS: A Web Tool for Machine Learning-Based Virtual Screening in Early-Phase of Drug Discovery and Development. *PLoS One* **2015**, 10.

(26) Korkmaz, S.; Zararsiz, G.; Goksuluk, D. Drug/Nondrug Classification Using Support Vector Machines with Various Feature Selection Strategies. *Comput Methods Programs Biomed* **2014**, *117*, 51–60.

(27) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J Med Chem* **1998**, 41, 3325–3329.

(28) Li, J.; Tong, X. Y.; Zhu, L. Da; Zhang, H. Y. A Machine Learning Method for Drug Combination Prediction. *Front Genet* **2020**, 11, 1–9.

(29) Boobier, S.; Hose, D. R. J.; Blacker, A. J.; Nguyen, B. N. Machine Learning with Physicochemical Relationships: Solubility Prediction in Organic Solvents and Water. *Nat Commun* **2020**, 11.

(30) Zhang, Q. Y.; Aires-de-Sousa, J. Random Forest Prediction of Mutagenicity from Empirical Physicochemical Descriptors. *J Chem Inf Model* **2007**, 47, 1–8.

(31) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. *Neural Message Passing for Quantum Chemistry*; **2017**.

(32) Prasad, S.; Brooks, B. R. A Deep Learning Approach for the Blind LogP Prediction in SAMPL6 Challenge. *J Comput Aided Mol Des* **2020**, 34, 535–542.

(33) Lui, R.; Guan, D.; Matthews, S. A Comparison of Molecular Representations for Lipophilicity Quantitative Structure–Property Relationships with Results from the SAMPL6 LogP Prediction Challenge. *J Comput Aided Mol Des* **2020**, 34, 523–534.

(34) Lenselink, E. B.; Stouten, P. F. W. Multitask Machine Learning Models for Predicting Lipophilicity (LogP) in the SAMPL7 Challenge. *J Comput Aided Mol Des* **2021**, 35, 901–909.

(35) Ulrich, N.; Goss, K. U.; Ebert, A. Exploring the Octanol–Water Partition Coefficient Dataset Using Deep Learning Techniques and Data Augmentation. *Commun Chem* **2021**, 4, 1–10.

(36) Yoshimori, A. Prediction of Molecular Properties Using Molecular Topographic Map. *Molecules* **2021**, 26.

(37) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA Models for Predicting Physicochemical Properties and Environmental Fate Endpoints. *J Cheminform* **2018**, 10, 1–20.

(38) Datta, R.; Das, D.; Das, S. Efficient Lipophilicity Prediction of Molecules Employing Deep-Learning Models. *Chemometrics and Intelligent Laboratory Systems* **2021**, 213.

(39) Domingo-Almenara, X.; Guijas, C.; Billings, E.; Montenegro-Burke, J. R.; Uritboonthai, W.; Aisporna, A. E.; Chen, E.; Benton, H. P.; Siuzdak, G. The METLIN Small Molecule Dataset for Machine Learning-Based Retention Time Prediction. *Nat Commun* **2019**, 10, 1–9.

(40) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res* **2014**, 42.

(41) Landrum, G. Rdk. Open-Source Cheminformatics; 2012; http:// www.rdkit.org. Accessed: **2020**.

(42) RDKit: open-source cheminformatics software.

(43) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. *the Journal of machine Learning research* **2011**, 12, 2825–2830.

(44) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, 323, 533–536.

(45) Morgan, H. L. *1999 REITZ LECTURE GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE 1*; **2001**, 29.

(46) Czermiski, R.; Yasri, A.; Hartsough, D. Use of Support Vector Machine in Pattern Classification: Application to QSAR Studies. *Quantitative Structure-Activity Relationships* **2001**, 20, 227–240.

(47) Ho, T. K. Random Decision Forests. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*; IEEE Computer Society, 1995; 1, 278–282.

(48) Bergstra, J.; Yamins, D.; Cox, D. D. Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. In *Proceedings of the 12th Python in science conference*; Citeseer, **2013**; 13, 20.

(49) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J Med Chem* **2020**, 63, 8749–8760.

(50) Jiang, D.; Wu, Z.; Hsieh, C. Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models. *J Cheminform* **2021**, 13, 1–23.

(51) Sheridan, R. P.; Wang, W. M.; Liaw, A.; Ma, J.; Gifford, E. M. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. *J Chem Inf Model* **2016**, 56, 2353–2360.

(52) Zhang, J.; Mucs, D.; Norinder, U.; Svensson, F. LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity-Application to the Tox21 and Mutagenicity Data Sets. *J Chem Inf Model* **2019**.

(53) Stahura, F. L.; Godden, J. W.; Bajorath, J. Differential Shannon Entropy Analysis Identifies Molecular Property Descriptors That Predict Aqueous Solubility of Synthetic Compounds with High Accuracy in Binary QSAR Calculations. *J Chem Inf Comput Sci* **2002**, 42, 550–558.

(54) Svetnik, V.; Liaw, A.; Tong, C.; Christopher Culberson, J.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J Chem Inf Comput Sci* **2003**, 43, 1947–1958.

(55) Shen, J.; Cheng, F.; Xu, Y.; Li, W.; Tang, Y. Estimation of ADME Properties with Substructure Pattern Recognition. *J Chem Inf Model* **2010**, 50, 1034–1041.

(56) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J Med Chem* **2004**, 47, 4463–4470.

(57) Ren, Y. Y.; Zhou, L. C.; Yang, L.; Liu, P. Y.; Zhao, B. W.; Liu, H. X. Predicting the Aquatic Toxicity Mode of Action Using Logistic Regression and Linear Discriminant Analysis. *SAR QSAR Environ Res* **2016**, 27, 721–746.

(58) Morgan, H. L. *The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service*; Interscience Publishers, Inc.. New York. N. Y, **1964**, 4.

(59) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. **2014**.

(60) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st international conference on neural information processing systems*; **2017**, 4768–4777.

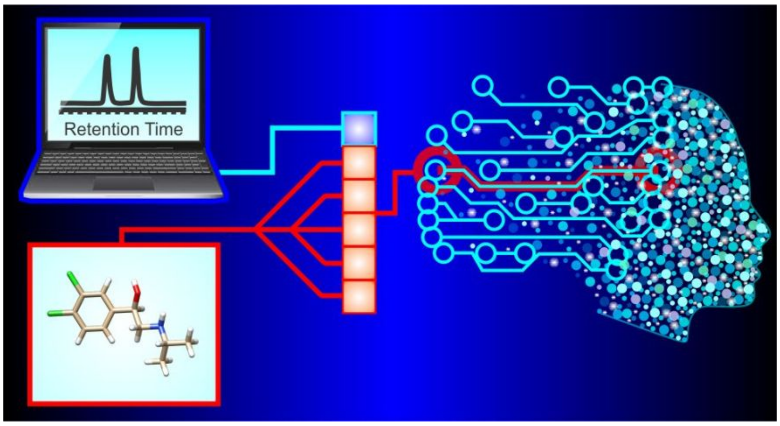
(61) Chemical Computing Group Inc., M. Q. Canada. Molecular Operating Environment (MOE), V. **2008**.

(62) Pearlman, R. S.; Smith, K. M. *Novel Software Tools for Chemical Diversity*; **1998**.

(63) Pan, S. J.; Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*. **2010**, 1345–1359.

(64) Teresa B. Culver, S. P. H. D. S. J. J. D. and J. A. S. Modeling the Desorption of Organic Contaminants from Long-Term Contaminated Soil Using Distributed Mass Transfer Rates. *Environ Sci Technol* **1997**, 31, 1581–1588.

For Table of Content Only



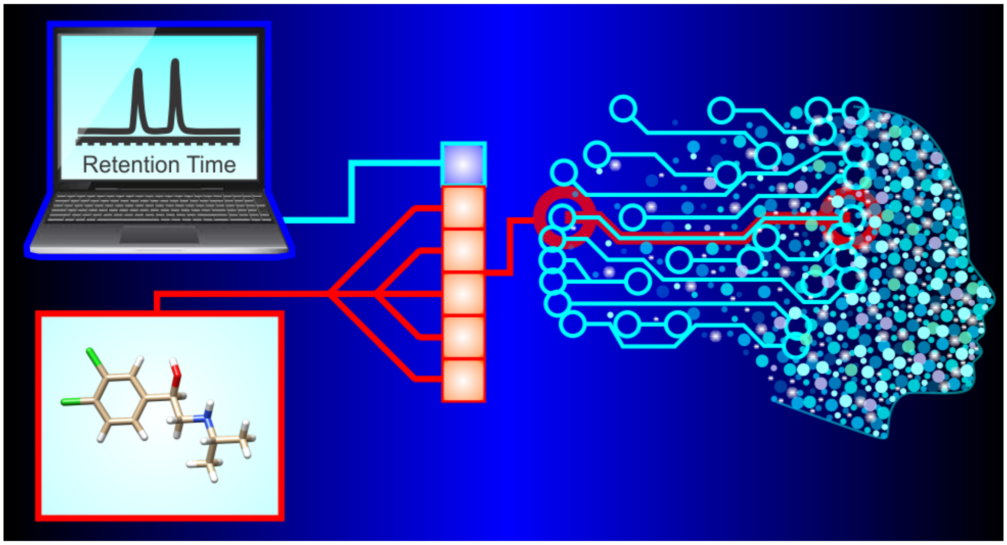


Table of contents graphic

82x44mm (300 x 300 DPI)

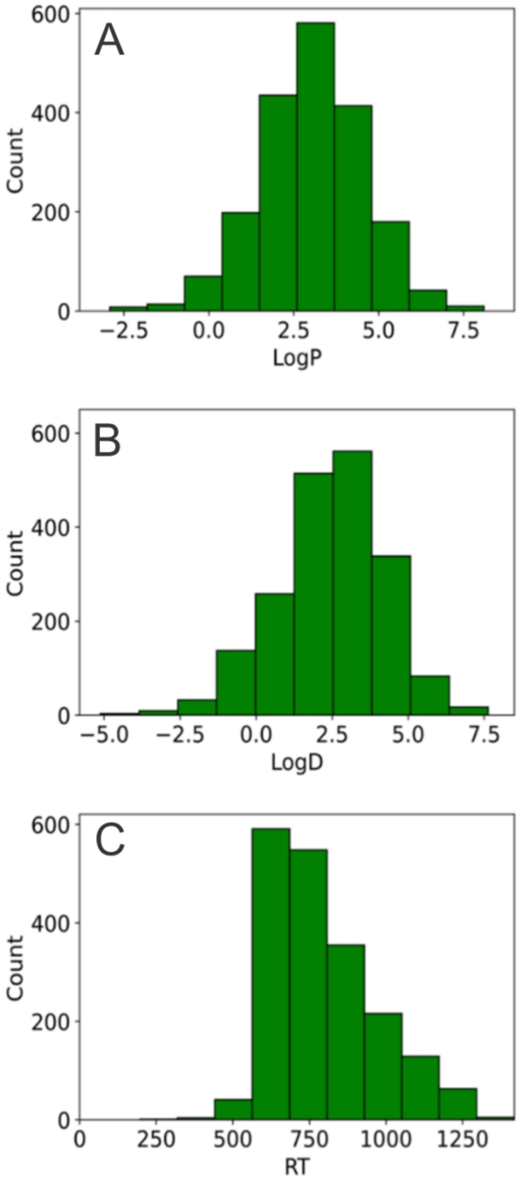


Figure 1. Distribution of values of (A) Log P, (B) Log D, and (C) RT of compounds in the data set used for this study.

59x133mm (300 x 300 DPI)

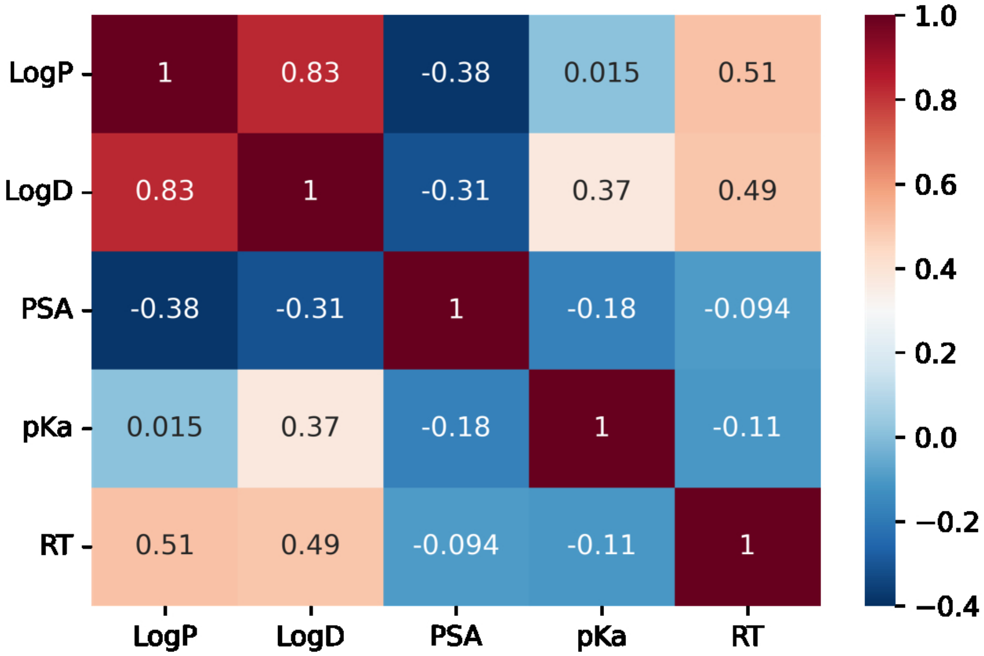


Figure 2. The heat map of Spearman’s correlation coefficient (□) for experimental retention time and physicochemical properties in the dataset. Values of 1, 0, and −1 indicate perfect correlation, no correlation, and anti-correlation, respectively.

96x68mm (300 x 300 DPI)

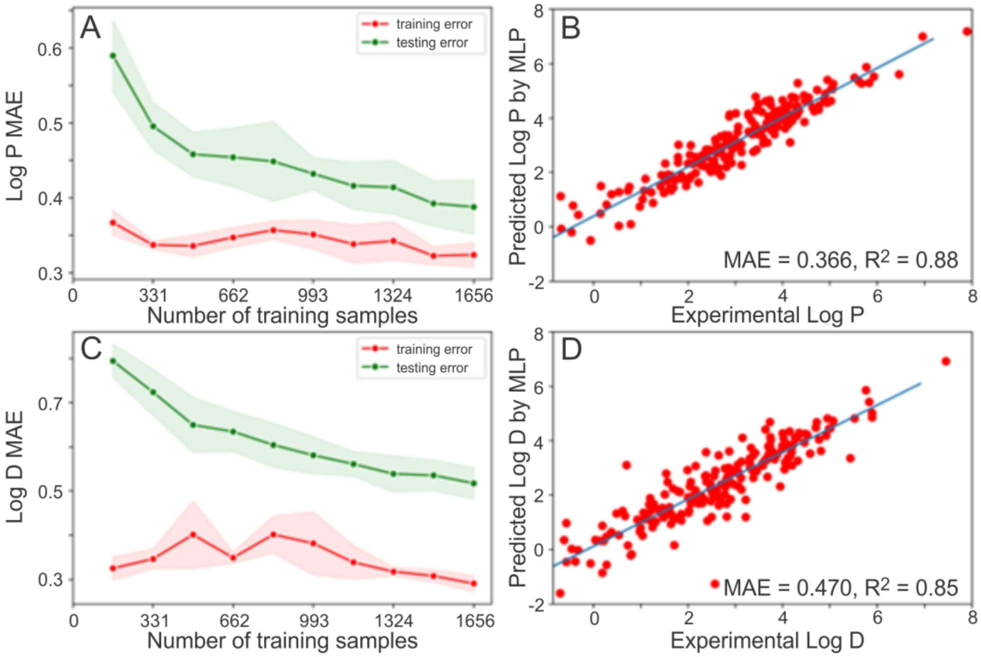


Figure 3. (A) Mean absolute error (MAE) for Log P predictions for the training and test set as a function of training set size. (B) Correlation plot of experimentally determined Log P values and Log P values predicted by the MLP model. (C) Mean absolute error (MAE) for Log D predictions for the training and test set as a function of training set size. (D) Correlation plot of experimentally determined Log D values and Log D values predicted by the MLP model.

159x107mm (300 x 300 DPI)

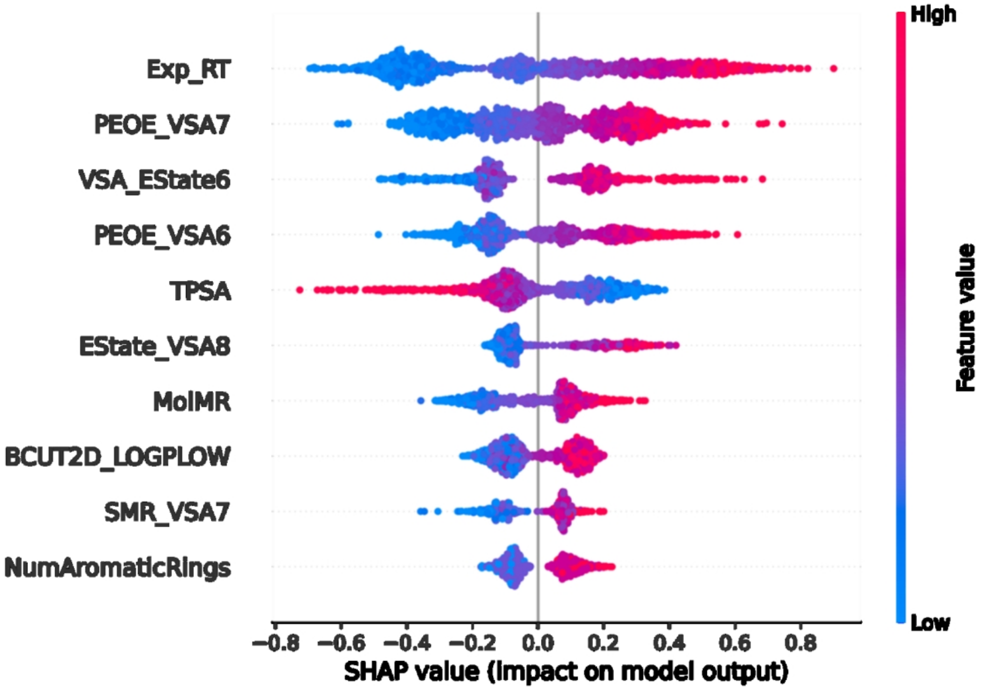


Figure 4. SHAP summary plot showing the impact of descriptors (the top 10) on the model. One dot represents one molecule, and the dots stack up to show its density.

150x105mm (300 x 300 DPI)