Published in final edited form as:

Fluid Phase Equilib. 2012 October 25; 332: 85–93. doi:10.1016/j.fluid.2012.07.001.

A NON-LINEAR STRUCTURE-PROPERTY MODEL FOR OCTANOL-WATER PARTITION COEFFICIENT

Krishna M. Yerramsetty, Brian J. Neely^a, and Khaled A. M. Gasem

School of Chemical Engineering, 423 Engineering North, Oklahoma State University, Stillwater, OK 74078

Abstract

Octanol-water partition coefficient (K_{ow}) is an important thermodynamic property used to characterize the partitioning of solutes between an aqueous and organic phase and has importance in such areas as pharmacology, pharmacokinetics, pharmacodynamics, chemical production and environmental toxicology. We present a non-linear quantitative structure-property relationship model for determining K_{ow} values of new molecules *in silico*. A total of 823 descriptors were generated for 11,308 molecules whose K_{ow} values are reported in the PhysProp dataset by Syracuse Research. Optimum network architecture and its associated inputs were identified using a wrapper-based feature selection algorithm that combines differential evolution and artificial neural networks. A network architecture of 50-33-35-1 resulted in the least root-mean squared error (RMSE) in the training set. Further, to improve on single-network predictions, a neural network ensemble was developed by combining five networks that have the same architecture and inputs but differ in layer weights. The ensemble predicted the K_{ow} values with RMSE of 0.28 and 0.38 for the training set and internal validation set, respectively. The ensemble performed reasonably well on an external dataset when compared with other popular K_{ow} models in the literature.

Keywords

mathematical models; neural networks; octanol-water partition coefficient; QSPR; solubility

1. INTRODUCTION

Within the context of the pharmaceutical industry, the octanol-water partition coefficient (K_{ow} or $\log K_{ow}$) is a thermophysical property that expresses the ratio of concentrations of a compound in coexisting phases comprised of octanol and water, which denotes the lipophilicity of a compound. Lipophilicity is involved in the prediction of a compound's partitioning, hydrophobicity and other pharmacokinetic and pharmacodynamic characteristics, which are important aspects of successful drug design. Leo *et al.* [1] were the first authors to comprehensively review the octanol-water partition system and its applications. Several researchers have studied the effects of lipophilicity on the biological activity of drugs [2-6] and on their transport properties [7-9]. Their findings and many other

© 2012 Elsevier B.V. All rights reserved.

^aSchool of Chemical Engineering, 423 Engineering North, Oklahoma State University, Stillwater, OK 74078, Phone (405) 744-5280, Fax: (405) 744-6338, neelyb@okstate.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

studies indicate the importance of lipophilicity in the evaluation of new drugs or prodrugs [10].

The experimental procedures for measuring the K_{ow} values are detailed by Sangster [11]; however, only a relatively small percentage of existing commercial chemicals have been tested experimentally for their K_{ow} values [10]. This is particularly so for highly hydrophobic compounds with K_{ow} values $> 10^6$. The low solubility of these compounds in the aqueous-rich phase renders K_{ow} measurements difficult to undertake, and therefore, few values exist at this range. Further, K_{ow} experiments are, in general, time and labor intensive, and they are impractical to carry out for the large number of potential drugs identified in the developmental stages of drug discovery. As such, a need exists for reliable predictive models to determine accurately K_{ow} values without the need for experimentation.

Although rudimentary predictive models for K_{ow} were established nearly half a century ago, advances in computational capabilities has led to the more recent development of a diverse variety of models. An article by Mannhold *et al.* [12] lists and compares the state-of-the-art models available currently for K_{ow} . These models for K_{ow} can be broadly classified as:

1. Fragment-based methods that divide the molecule into various fragments (either at the molecular or atomic level) and then sum the contributions of these individual fragments to provide the final value. Examples include KowWIN based on the algorithms developed by Meylan and Howard [10], CLOGP [13] and Ghose-Crippen models [14-16].
2. Molecular-property based methods that utilize characteristics of the entire molecule to predict for K_{ow} . These characteristics are usually referred to as molecular descriptors and are normally calculated from the three-dimensional structure of the molecule or from the topology of the molecule. Common examples of this model classification are QLOGP based on the algorithm by Bodor and Buchwald [17], GBLOGP by Totrov [18] and ALOGPS [12, 19-21].

Fragment based methods have been highly successful in developing accurate models for K_{ow} . After comparing the performance of currently available methods on three different datasets, Mannhold *et al.* [12] ranked the best methods, which included fragment-based methods like AB/LogP [22], CLOGP [13] and KowWIN [10]. The fragment-based methods rely solely on the two-dimensional structure of the molecules; as such, thousands of molecules typically generated in a virtual molecule design system can be processed in a relatively short period of time. A disadvantage of the fragment-based methods is the lack of parameter values when a structure cannot be decomposed to sub-structures for which the fragment values are available [23]. Additionally, these methods cannot be used to attach any physical significance to the structural factors affecting the value of the partition coefficient.

The molecular-property based methods do not need additional correction factors and provide better physical insight into the factors affecting the partition coefficient. However, finding the optimal three-dimensional structure of the molecule is a time consuming task, which limits the ability of these methods to handle large numbers of molecules in a reasonable amount of time. Most molecular-property based methods do not use all available molecular descriptors, but limit themselves to a small subset of descriptors, which have already been proven to be effective like E-state indices [24, 25] by ALOGPS [19, 20] and VLOGP [26], topological descriptors by TLOGP [27] and molecular size and H-bonding descriptors by QLOGP [17]. As a result, these models provide little indication in the identification of specific molecular properties that affect the partition coefficient.

In this work, we propose a non-linear quantitative structure-property relationship (QSPR) model for predicting the octanol-water partition coefficient. The basic premise of a QSPR

methodology asserts that a thermophysical property to some degree is a function of its structural attributes [28]. QSPR models have been able to predict successfully for a various number of thermophysical properties such as normal boiling point [29-31], melting point [32-34], refractive index [35, 36] and glass transition temperature [37]. All the molecular-property based models for K_{ow} discussed previously represent different types of QSPR models, where a variety of structural descriptors are employed, including constitutional, topological, geometrical, electrostatic, quantum-chemical and/or thermodynamic descriptors.

Our model utilizes all the descriptors (including three-dimensional descriptors) of the molecule generated using CODESSA [38]. Nearly 800 descriptors belonging to various classes like constitutional, topological, geometrical, electrostatic, quantum-chemical and thermodynamic were generated for each molecule. Using a wrapper-based algorithm, we determined that 50 descriptors resulted in an accurate model for the partition coefficient; nevertheless, the pruning of descriptors from 800 to 50 is not a trivial task, particularly in non-linear QSPR modeling. A recent review article by Dudek *et al* [39] summarizes the different types of descriptor pruning techniques in use. In general, these methods fall into two categories:

1. Filter-based methods: These methods are implemented before the mapping of the structural attributes to the property of interest. Some examples include methods based on mutual information [40].
2. Wrapper-based methods: In these methods, the selection of best descriptors is undertaken along with the mapping or the learning step. The error in the mapped model is used as the judging criterion for the selection process. Some common examples include sequential forward selection, sequential backward elimination, genetic algorithms (GA), simulated annealing (SA) and ant colony optimization.

The filter-based methods are quite fast but may not result in the selection of the best subset of descriptors. A more reliable descriptor subset selection procedure is observed with wrapper-based methods, but these methods are slow in the final stages of the algorithm. This is particularly true of the stochastic methods such as GA, SA and ant colony optimization. In this work, we propose a novel wrapper-based algorithm for the selection of the best subset of descriptors using an evolutionary algorithm called differential evolution (DE) [41], which uses artificial neural networks as the non-linear mapping functions. DE has been proven to be as effective as or better than GA and also is easier to implement [42, 43].

2. METHODS

The development of a QSPR model involves the following series of steps: (a) data set generation, (b) descriptor calculation, (c) descriptor reduction and model training, (d) model validation and (e) model interpretation.

2.1. Data set generation

Experimental octanol-water partition coefficient (log K_{ow}) values were taken from the PhysProp database by Syracuse Research. This database had experimental log K_{ow} values listed for 13,553 compounds. Since the molecule optimization routines employed in the study are tailored for organic compounds and to avoid skewing the resulting model, 350 compounds including metal containing compounds, inorganics (compounds without a carbon atom) or isomers of another molecule were not considered for the modeling effort.

Also, for 33 compounds, ChemBioDraw Ultra 11.0 [44] was unable to generate two-dimensional (2-D) structures from the molecule names in the PhysProp database and were

excluded. As such, a total of 13,170 molecules were selected for further analysis. However, only 11,308 molecules could be optimized for their most favorable (lowest energy) three-dimensional conformation using our automated procedure. The structure optimization in this work was performed in two stages. In Stage 1, an automated program that utilizes Chem3D was used, and AMPAC 6.0 was employed as an additional optimization step in Stage 2. Due to the differences in the molecular structure file formats required for these two programs, the molecules had to be converted from .mop format to .int format. For 14% of the compounds, this file conversion led to changes in the molecular structure information contained in the file. Therefore, these compounds were not included in the model development process.

Further, there is no certain way to know if the optimized structures have been truly optimized, unless an exhaustive search of all possible 3D conformations is conducted. This can quickly become impractical for large molecules with more than 10 atoms. Instead, in the current work, we have tried to rotate each freely rotating bond eight times and each of these resultant conformations was used as starting points in the MM2 optimization routine in the Chem3D software. The conformation with the least overall energy was then used as the most stable 3D conformation.

While characterization of this large database is beyond the scope of this work, Hansch and Leo [45] have stated that the Log K_{ow} values can be experimentally determined to an average deviation of ± 0.05 for most solutes. For solutes that have a Log K_{ow} value lower than -3 and greater than 6 as well as solutes that are relatively insensitive to gas chromatography, the average deviation expected is ± 0.1 .

2.2 Descriptor calculation

Descriptor calculation requires a series of steps common to all QSPR models. ChemBioDraw Ultra 11.0 [44] was used to generate the 2-D structures for the molecules in the data set. These 2-D structures were then used to generate three dimensional (3-D) structures. Each 2-D structure can be translated into a large number of 3-D conformations; however, only the conformation with the lowest conformational energy is considered representative of the natural state of the molecule. When considering the multiple minima of the total-energy curve, finding this 3-D conformation is not a trivial task. Chem3D Pro 11.0 [44] is a commercial software used commonly to minimize the total-energy of a 3-D conformation; however, the software is not guaranteed to find the global minimum energy conformation. Therefore, an optimization using several initial 3-D conformations will have an improved chance of locating the global minimum. This process is not integrated into Chem3D Pro 11.0 and requires the 3-D structure to be manually reinitialized to a different starting conformation each time before optimization. This operation not only places an increased time and effort burden on the user, but it is not a reliable method of locating the global minimum. To alleviate this problem, we have used an automated strategy for identifying the 3-D conformation with the least total energy. Chem3D Pro 11.0 was used as the optimizing engine, but it was controlled using its Component Object Model (COM) interface with Microsoft Visual Studio 8 [46] as the back-end. The 3-D structures were further optimized using the AM1 semi-empirical quantum method in AMPAC 6.0 [47]. As opposed to the more theoretical and rigorous *ab initio* and DFT techniques, semi-empirical optimization is often employed in QSPR model development with the main benefit of a reduction in the computational burden. Studies have found that electron densities and heats of formation calculated by semi-empirical techniques agree favorably with the more theoretical techniques [48-51]. Further, QSPR models developed using both optimization techniques for such properties as vapor pressure [52], toxicity [53] and acid dissociation constant [54] resulted in either similar or acceptable levels of accuracy.

The final optimized structures were provided to CODESSA PRO [55] for descriptor calculation. CODESSA PRO has the capability to generate over 800 descriptors; however, due to structural complexity, this number may be lower for a particular structure and for such structures the missing descriptors were assigned a zero value.

2.3. Descriptor reduction and model training

A non-linear reduction and training strategy was employed in the current work to discover the best descriptor set and the optimum neural network architecture simultaneously based on the generated descriptors. An evolutionary algorithm, differential evolution, was used to search for the best descriptors and network architecture, and artificial neural networks (ANNs) were used to build the non-linear models based on these descriptors. A flowchart for the algorithm is provided as Figure 1.

In brief, DE is a stochastic optimization algorithm that finds the global extremes (minimum or maximum) of a function. The method begins with an initial random population of individuals (or independent variables) that are transformed through mutation and crossover operations over a number of successive generations. Later generations are, on average, closer to the global minimum (or maximum) of the dependent variable. For a detailed overview of this process, see the description of the algorithm provided by Storn [41]. Two essential features of any evolutionary algorithm are (a) genetic representation and (b) the objective function, which are described below.

2.3.1 Genetic representation—A good genetic representation of the solution domain is an important step in developing an efficient DE algorithm. In the current work, the solution space is comprised of all possible molecular descriptors and all possible one or two hidden-layered neural network architectures. Since a two hidden-layer network is capable of reasonable approximation of any non-linear function, the maximum number of hidden layers was limited to two [56]. The number of descriptors used for building the models was fixed at fifty. These fifty descriptors form an individual in the solution space and each individual in the population is characterized by this set of fifty descriptors. As in most evolutionary algorithms, arrays are the most suitable form for representing an individual in the solution space. In the present work, a fifty-three element integer array, D_i , was used to represent the i^{th} individual in the initial population. Each of the first fifty elements of the array stores an integer that represents a descriptor from the set of all possible descriptors. Therefore, $D_{i,j}$ for $j = 1$ to 50 represents the j^{th} element (or descriptor) of the i^{th} individual in the initial population. D_i , M_i , and T_i , are used to distinguish between the individuals in the initial population, mutated population and the trial population, respectively. In addition to the descriptors, the algorithm also optimizes the network architecture. Therefore, the genetic representation includes elements that denote the number of hidden layers (1 or 2) and also the number of neurons in each of these hidden layers. Employing an analogous representation scheme as used for the descriptors, the 51st element in the arrays D_i , M_i , and T_i , was used to represent the number of hidden layers in the network. If this element was set to a value of one, then the corresponding network had a single hidden-layer, and if the value of the element was zero, then the network contained two hidden-layers. Further, the last two elements of the arrays were used to represent the number of neurons in the one or two hidden-layer architectures, respectively. The maximum number of neurons in either hidden layer was set at 35, because using more than 35 hidden neurons in a single layered network with 50 inputs results in a degree of freedom ratio of less than two (denoted as max_neurons in the subsequent discussion).

2.3.2 Objective function—Another major aspect of a DE algorithm is choice of a suitable objective function. In this work, the root-mean squared error (RMSE) between the

experimental and the predicted K_{ow} values was chosen as the objective function. The algorithm searches for the set of descriptors and the network architecture that result in a neural network with the least RMSE.

If DE is considered the searching mechanism (or the descriptor reduction mechanism) of the algorithm, ANNs are the non-linear optimization tools of the algorithm. In general ANNs are used to map complex non-linear relations between inputs (or the independent variables) and the outputs (the dependent variables). In this work, a feed-forward neural network with fifty neurons in the input layer and one neuron in the output layer was used to model the relationships between the fifty descriptors and the output K_{ow} . The number of neurons in the hidden layers was set based on the last two elements of the representation arrays, D_i and T_i . Early stopping (Levenberg-Marquardt) was used as the optimization algorithm with 7122 data points in the training-set and 4186 data points in the validation-set. Random sphere exclusion algorithm [57] with a dissimilarity level of 0.8 was used to divide the data into training and validation sets. In addition to limiting the maximum number of neurons in each hidden layer to less than 35, another constraint was imposed on the number of neurons in the hidden layers by maintaining a degree of freedom ratio of at least two. The target values for the network outputs were the logarithmic transformations of the experimental K_{ow} values. In addition, all inputs and target values were normalized to have zero mean and a standard deviation of one. These transformations avoid any bias in the model due to extremely low or high target or input values.

Several authors [58, 59] have demonstrated the advantages of an ensemble of neural networks over individual networks. An ensemble is created by combining the predictions of several individual networks using a simple average or other more complex methods like weighted averages. The individual networks generalize better in different regions of the input space of the problem due to the differences in their inputs, their architecture or the weights between the layers. Therefore, an ensemble that combines all these different networks is expected to perform better than any of the individual networks in the entire input space in the data. Hansen *et al.* [58] have demonstrated the advantages of using an ensemble of networks on different case studies. In the current work, an ensemble of neural networks was created using networks with the same architecture and inputs but differing in the values of the weights between the different layers.

2.4. External validation

Tropsha *et al.* [60] have emphasized the need to validate QSPR models using external data sets. Also, Mannhold *et al.* [12] have recently compared the various K_{ow} models in the literature using an external public database by Avdeef [61] which consisted of 266 molecules in total. Of these, 214 molecules were similar to the molecules in the PhysProp database and were classified as star set molecules. The remaining 43 molecules were classified as non-star set molecules. Since these molecules were not reported in the PhysProp dataset, they can be used as an external validation set for comparing our model with other models in the literature. The performance of the current model on this new dataset would provide some preliminary indication of the generalization capability of the final model.

2.5. Model interpretation (relative significance of descriptors)

The relative significance of the inputs to the output in a complex neural network is not readily apparent. Gevrey *et al.* [62] have compared several methods for quantifying neural network input significance and report that the partial derivatives ('PaD') method is the most reliable when compared to input perturbation and stepwise methods. Also, Olden *et al.* [63] have done a similar comparison study and they have concluded that the overall connection

weight approach by Olden *et al.* [64] is the most stable method when compared to the partial derivatives method, Garson's algorithm and the input perturbation method. In most of the above cited studies, the methods have been applied only to relatively small neural networks, i.e. less than 10 network inputs. In this article, we discuss three popular methods for analyzing neural network input saliency.

2.5.1. Partial derivatives—This method, which makes use of the partial derivatives in the network, was proposed by Dimopoulos *et al.* [65]. In brief, the method involves calculating the partial derivatives of the output values with respect to each input variable. The descriptors with large partial derivatives are considered the most significant. Consider a three layered feed-forward neural network with two hidden layers (with tangent-sigmoid transfer functions) and one output layer (with linear transfer functions). Let N_n be the number of inputs to the network, N_e be the number of samples in the training set, N_h and N_g be the number of neurons in the first and second hidden layers respectively, then the gradient of the output with respect to each input for the e^{th} sample in the training set data is calculated in the following way:

$$pd_{n,e} = \sum_{h=1}^{N_h} W_{h,n}^1 * S_{h,e}^1 \quad (1)$$

where, $pd_{n,e}$ is the partial derivative of the n^{th} input to the network calculated for the e^{th} sample in the data set used for training the network. $W_{h,n}^1$ is the weight between the n^{th} input and the h^{th} neuron in the first hidden layer, and $S_{h,e}^1$ is the sensitivity of the output of the h^{th} neuron in the first hidden layer with respect to its input, calculated for the e^{th} sample in the training set data. This sensitivity for the neurons in the first hidden layer are calculated based on the back propagation method used for training neural networks as shown below:

$$S_{h,e}^1 = \left(1 - (a_{h,e}^1)^2\right) * \left\{ \sum_{g=1}^{N_g} W_{g,h}^2 * S_{i,e}^2 \right\} \quad (2)$$

where, $a_{h,e}^1$ is the output from the h^{th} neuron in the first hidden layer calculated for the e^{th} sample in the training set data $W_{g,h}^2$ is the weight between the h^{th} neuron in the first hidden layer and the g^{th} neuron in the second hidden layer, and $S_{i,e}^2$ is the sensitivity of the output of the i^{th} neuron in the second hidden layer with respect to its input, calculated for the e^{th} sample in the training set data. The sensitivities for the second hidden layer are calculated again using the back propagation algorithm as follows:

$$S_{g,e}^2 = \left(1 - (a_{g,e}^2)^2\right) * \{-W_{1,g}^3\} \quad (3)$$

where, $a_{g,e}^2$ is the output from the g^{th} neuron in the first hidden layer calculated for the e^{th} sample in the training set data $W_{1,g}^3$ is the weight between the g^{th} neuron in the first hidden layer and the final output from the network.

Finally, the sensitivity of the output to each of the network inputs is calculated by summing the gradients calculated in equation (1) for the entire training set data:

$$Sn_n = \sum_{e=1}^{N_e} (pd_{n,e})^2 \quad (4)$$

The relative value of Sn_n with respect to the Sn values for the other inputs gives the estimate for the significance of the n^{th} input when compared to the other inputs to the network.

2.5.2. Randomized connection weight approach—A randomized connection weight approach was proposed by Olden and Jackson [64] to eliminate null-connection weights from the network that do not differ significantly from random values. In brief, the procedure involved randomly generating the output values and constructing neural networks for estimating these random output responses using the same input values that were used to build the optimal final network. All the individual weights were recorded and the connection weights and the overall connection weights for each input were calculated. Consider a three layered feed-forward neural network with two hidden layers (with tangent-sigmoid transfer functions) and one output layer (with linear transfer functions). Let N_n be the number of inputs to the network, N_e be the number of samples in the training set, N_h and N_g be the number of neurons in the first and second hidden layers respectively, then the overall connection weight from the n^{th} input to the network output is calculated the following manner:

$$OCW_n = \sum_{h=1}^{N_h} \sum_{g=1}^{N_g} (W_{h,1}^1 * W_{g,h}^2 * W_{1,g}^3) \quad (5)$$

where, OCW_n is the overall connection weight for the n^{th} input and the remaining terms on the right hand side of the equation denote the weights as describes in the previous section.

The above procedure was carried out 999 times and the proportion of the overall connection weights (including the original optimal network) which were larger in magnitude than the corresponding overall connection weights of the optimal network were recorded. An input is considered significant if the proportion of random connection weights that are greater than the corresponding connection weight is lower than 0.05 (for a confidence level of 95 %) and insignificant otherwise.

2.5.3. Mutual information—A concept from information theory called the measure of mutual information has been previously applied for feature selection in non-linear modeling [66]. Mutual information index measures the amount of information contained in a single or a group of independent variables about the dependent variable. Mutual information is usually estimated using histogram or kernel-based pdf distributions. However, these methods are prone to systematic errors [67] and also need a large number of samples in the modeling space [68]. Kraskov *et al.* [67] have developed a method for estimating the mutual information bases on k-nearest neighbor statistics. The details of the method and its implementation are provided elsewhere [66-68].

Let the inputs to the network be denoted by the vector X_e and the outputs by Y_e . Then, the e^{th} sample from the training set can be denoted in the input output space as

$$Z_e = (\{X_{1,e}, X_{2,e}, \dots, X_{n,e}, \dots, X_{N_n,e}\}, Y_e) \quad (6)$$

As can be seen, X_e is a N_n dimensional vector and Y_e is a one dimensional vector and both belong to the set of reals $\mathbf{R}$. Therefore, the distance between any two vectors can be calculated using the Euclidean norm denoted by $\|(\cdot)\|$.

Samples in the training set can be compared through the maximum norm defined as $\|Z_i - Z_j\| = \max(\|X_i - X_j\|, \|Y_i - Y_j\|)$, for the i^{th} and j^{th} samples in the training set

Now, calculate for all $\epsilon_i = \|Z_i - Z_{k(i)}\|$ for all $i=1$ to N_e , i.e. for all samples in the training set. Here, $k(i)$ is the k^{th} nearest neighbor to the i^{th} sample in the training set. The mutual information between the input set X and output Y can now be calculated in the following way:

$$MI(X, Y) = \Psi(k) - \frac{1}{N} * \sum_{e=1}^{N_e} \{ \Psi(n_e^X + 1) + \Psi(n_e^Y + 1) \} + \Psi(N_e) \quad (7)$$

where, $MI(X, Y)$ denotes the mutual information between X and Y , Ψ denotes the standard digamma operator, n_e^X and n_e^Y are the number of points whose distance from the e^{th} sample in the training set is less than ϵ_i in the input and output spaces, respectively.

A high value of mutual information between the input variables and the output variable implies that the input variables have the necessary information in them to describe the variance in the output variable. Conversely, a low value implies that the chosen input variables cannot accurately describe the variance in the output variable. The value of k chosen in this study was 6 as this value has been typically used in the literature [67, 68]. In this article, we have calculated the mutual information value between all the input variables and the output variable. Next, one input variable was deleted from the set of original input variables to the neural network and a modified set of input variables was created. The mutual information value between this modified set and the output variable was calculated. If the new mutual information value is significantly lower than the original mutual information value between all input variables (including the input variable that has been omitted) and the output variable, then the corresponding input variable that has been left out is considered significant. This procedure is carried out for all the input variables in the input set to the neural network.

3. RESULTS

3.1. Optimum Model

Using the DE + NN algorithm, a 50-33-35-1 neural network architecture was found to result in the least RMSE for the training set data. Neural networks with the same input data and architecture as the best network identified using the DE + NN algorithm were trained with different random initial weights. Of the networks generated, the five networks resulting in the least RMSE values in the training set were chosen, and their weights were recorded. The final predictions are calculated as a simple average of the individual predictions by these five networks. The addition of further networks did not improve significantly the overall RMSE of the training set. The RMSE values for the training set and validation set data for the five networks and the resultant average network are shown in Table 1. In subsequent discussions, the prediction results are from the ensemble network, which is the average of the predictions by the five best networks. Comparisons of the experimental and predicted K_{ow} values for the training and validation sets are shown in Figure 2 and Figure 3, respectively.

3.2. External Validation

Our model predictions for the star and the non-star sets from the Avdeef test set are shown in Figures 4 and 5.

The RMSE for the star-set molecules and the non-star set molecules were calculated to be 0.57 and 1.01, respectively. A comparison of the performance of the current model on the Avdeef set with those of the best models in the literature (as reported by Mannhold *et al.* [12]) is provided in Table 2.

3.3. Sensitivity Analysis

The partial derivative method was applied to each of the five best networks. Five common descriptors were found to be present in the top ten significant descriptors for each network in the ensemble. These descriptors are tabulated in Table 3 along with their average percentage contribution calculated from the percentage contributions from the five best networks.

The proportion of overall connection weights calculated using random target values was greater than 0.05 for all 50 inputs and therefore, no input is significant at a 95% confidence level (data not provided). Mutual information values with only 49 inputs (one input was successively removed from the original input set) were not significantly different from the mutual information value using all 50 inputs. This again proves that no descriptor is significant in itself.

4. DISCUSSION

The best network ensemble identified is a combination of two layered networks with fifty input descriptors each. This ensemble was able to account for 98% and 96% of the variation in the training (Figure 2) and validation (Figure 3) sets, respectively. The statistics presented in Table 1 justify the use of a neural network ensemble as opposed to individual networks. Network 1 in the ensemble had the lowest RMSE value of 0.329 for the training set, which is nearly 21% higher than the corresponding RMSE value of the entire ensemble (Table 1). The different networks in the ensemble had differing weights, which produces varying predictions in different regions of the input space. This further supports the advantage of ensemble networks over an individual network.

The DE + NN approach employed in the current work has been successful in identifying the best descriptors describing the octanol-water partitioning ability of molecules. This is evident from the performance of the model on the external validation set (Table 2) where the RMSE of the star-set was 0.57 and that of the non-star set was 1.01. Our model compares favorably with some of the popular K_{ow} models like ALOGPS [12, 19-21], CLOGP [13] and KowWIN [10]. However, this external validation data set is limited in its size and therefore cannot truly be used to test the generalization ability of the current model. In the future, we intend to test our model on much larger external data sets to clearly establish the generalization ability of the models built using the approach described herein. Also, the current model was built using only descriptors generated by CODESSA [55].

For developed QSPR models, knowledge of the underlying applicability domain (AD) provides a level of confidence in the application of the model to new molecules. In the current work, a distance-based methodology was employed to assess the AD. An average Mahalanobis distance (MD) based on the best molecular descriptors was calculated between each compound in the external Avdeef set and the three closest training set molecules. However, no correlation was found between the errors in the external test set and the distance values. Next, the percentage deviations were divided into ten equal sized grids in increasing order. Although, the mean percentage absolute deviation increases considerably across the grids, the mean MD values remain relatively constant. These results (not shown) indicate an inability to define an AD. This could be due to the large number of descriptors (50) used in the current K_{ow} model as opposed to the smaller number used for similar applicability domain models in the literature. In other words, the descriptor space in the current work is possibly too complex to be used to define the applicability domain.

In future work, we intend to employ other descriptors such as functional-group descriptors and WHIM descriptors (molecular descriptors obtained as statistical indices of the atoms projected onto the 3 principal components obtained from weighted covariance matrices of

the atomic coordinates) available in DRAGON [69]. The model in the current work is purely empirical and is therefore limited in the ability to generalize to systems beyond its applicability domain. To remedy this limitation, we intend to build a theory-based K_{ow} model that utilizes QSPR generalized activity coefficient models to predict the solubility of solute molecules in each of the two phases of the octanol-water system.

The most important descriptor identified by the partial derivatives method was '(1/6) X GAMMA polarizability (DIP),' which had an average percentage contribution of 63%. However, the results from the overall connection weight approach and the mutual information method indicate that none of the inputs to the optimal network are significant by themselves. Therefore, we cannot conclusively determine the significance of the inputs in the network. This could be due to the complex nature of the neural networks used in this article. The methods published in literature for calculating the significance of descriptors are tested on relatively small networks and the authors have not come across a successful application of these methods for larger networks (more than 30 inputs to neural networks).

5. CONCLUSIONS

The results from the current work lead to the following conclusions:

- A hybrid algorithm that combines differential evolution algorithms (DE) and artificial neural networks (ANNs) provides an accurate predictive model for octanol-water partition coefficient that compares favorably with viable literature models.
- An ensemble of neural networks that differ in the values of the weights between layers provides improved predictive generalizations compared to a single network and yielded predictions with an R^2 of 0.98 and 0.96 for the training and validation sets of a database involving 11,308 molecules.
- The sensitivity analysis methods used for calculating the significance of inputs to the neural network could not provide conclusive results, possibly due to the complex nature of the neural networks employed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support for this research was provided by the National Institute of Biomedical Imaging and Bioengineering (1R21EB005749). The authors would like to thank Dr. Dana Brunson and Vaibhav Deshpande from the High Performance Computing Facility at the Oklahoma State University, Stillwater, OK, USA for providing guidance on the usage of the CIMARRON cluster.

References

1. Leo A, Hansch C, Elkins D. Partition coefficients and their uses. *Chem Rev.* 1971; 71:525–616.
2. Corwin H, John MC. Lipophilic character and biological activity of drugs II: The parabolic case. *J Pharm Sci.* 1973; 62:1–21.
3. Corwin H, William JD III. Linear relationships between lipophilic character and biological activity of drugs. *J Pharm Sci.* 1972; 61:1–19. [PubMed: 4550859]
4. Glave WR, Corwin H. Relationship between lipophilic character and anesthetic activity. *J Pharm Sci.* 1972; 61:589–591. [PubMed: 5014317]

5. Renau TE, Sanchez JP, Shapiro MA, Dever JA, Gracheck SJ, Domagala JM. Effect of lipophilicity at N-1 on activity of fluoroquinolones against mycobacteria. *J Med Chem.* 1995; 38:2974–2977. [PubMed: 7636859]
6. Roland J. Parabolic relationship between lipophilicity and biological activity of aliphatic hydrocarbons ethers and ketones after intravenous injections of emulsion formulations into mice. *Acta Pharmacol Toxicol (Copenh).* 1975; 37:56–64. [PubMed: 1174281]
7. van Bree JB, de Boer AG, Danhof M, Ginsel LA, Breimer DD. Characterization of an “in vitro” blood-brain barrier: effects of molecular size and lipophilicity on cerebrovascular endothelial transport rates of drugs. *J Pharmacol Exp Ther.* 1988; 247:1233–1239. [PubMed: 3204515]
8. Waterhouse RN. Determination of lipophilicity and its use as a predictor of blood-brain barrier penetration of molecular imaging agents. *Mol Imag Biol.* 2003; 5:376–389.
9. Wils P, Warnery A, Phung-Ba V, Legrain S, Scherman D. High lipophilicity decreases drug transport across intestinal epithelial cells. *J Pharmacol Exp Ther.* 1994; 269:654–658. [PubMed: 8182532]
10. Meylan WM, Howard PH. Atom/fragment contribution method for estimating octanol-water partition coefficients. *J Pharm Sci.* 1995; 84:83–92. [PubMed: 7714751]
11. Sangster, J. Octanol-water partition coefficients: Fundamentals and physical chemistry. Wiley; New York: 1997.
12. Mannhold R, Poda GI, Ostermann C, Tetko IV. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *J Pharm Sci.* 2009; 98:861–893. [PubMed: 18683876]
13. Leo A, Hoekman D. Calculating log P(oct) with no missing fragments: The problem of estimating new interaction parameters. *Perspect Drug Discovery Des.* 2000; 18:19–38.
14. Arup KG, Avis P, Gordon MC. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships III: Modeling hydrophobic interactions. *J Comput Chem.* 1988; 9:80–90.
15. Arup KG, Gordon MC. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships I. Partition coefficients as a measure of hydrophobicity. *J Comput Chem.* 1986; 7:565–577.
16. Ghose AK, Crippen GM. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J Chem Inf Comput Sci.* 2002; 27:21–35. [PubMed: 3558506]
17. Bodor N, Buchwald P. Molecular size based approach to estimate partition properties for organic solutes. *J Phys Chem B.* 1997; 101:3404–3412.
18. Maxim T. Accurate and efficient generalized born model based on solvent accessibility: Derivation and application for Log P octanol/water prediction and flexible peptide docking. *J Comput Chem.* 2004; 25:609–619. [PubMed: 14735578]
19. Tetko IV, Bruneau P. Application of ALOGPS to predict 1-octanol/water distribution coefficients logP, and logD, of AstraZeneca in-house database. *J Pharm Sci.* 2004; 93:3103–3110. [PubMed: 15514985]
20. Tetko IV, Poda GI. Application of ALOGPS 2.1 to predict log D distribution coefficient for pfizer proprietary compounds. *J Med Chem.* 2004; 47:5601–5604. [PubMed: 15509156]
21. Tetko IV, Tanchuk VY. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J Chem Inf Comput Sci.* 2002; 42:1136–1145. [PubMed: 12377001]
22. Pranas J, Remigijus D, Alanas P. Fragmental methods in the design of new compounds. Applications of The advanced algorithm builder. *Quant Struct-Act Relat.* 2002; 21:23–37.
23. Lemke, TL.; Williams, DA.; Roche, VF.; Zito, SW. Foye’s principles of medicinal chemistry. Lippincott Williams & Wilkins; Philadelphia: 2008.
24. Hall LH, Kier LB. The E-State as the basis for molecular structure space definition and structure similarity. *J Chem Inf Comput Sci.* 2000; 40:784–791. [PubMed: 10850783]
25. Hall LH, Kier LB. Electrotopological state indices for atom types: A novel combination of electronic topological, and valence state information. *J Chem Inf Comput Sci.* 2002; 35:1039–1045.

26. Gombar VK, Enslein K. Assessment of n-Octanol/water partition coefficient: When is the assessment reliable? *J Chem Inf Comput Sci.* 1996; 36:1127–1134. [PubMed: 8941993]
27. Junghans M, Pretsch E. Estimation of partition coefficients of organic compounds: Local database modeling with uniform-length structure descriptors. *Fresenius J Anal Chem.* 1997; 359:88–92.
28. Balaban AT. From chemical topology to 3D geometry. *J Chem Inf Comput Sci.* 1997; 37:645–650.
29. Katritzky AR, Lobanov VS, Karelson M. Normal boiling points for organic compounds: Correlation and prediction by a quantitative structure-property relationship. *J Chem Inf Comput Sci.* 1998; 38:28–41.
30. Stanton DT, Egolf LM, Jurs PC, Hicks MG. Computer-assisted prediction of normal boiling points of pyrans and pyrroles. *J Chem Inf Comput Sci.* 1992; 32:306–316.
31. Wiener H. Structural determination of paraffin boiling points. *J Am Chem Soc.* 1947; 69:17–20. [PubMed: 20291038]
32. Bergstrom CAS, Norinder U, Luthman K, Artursson P. Molecular descriptors influencing melting point and their role in classification of solid drugs. *J Chem Inf Comput Sci.* 2003; 43:1177–1185. [PubMed: 12870909]
33. Katritzky AR, Lobanov VS, Karelson M. QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. *Chem Soc Rev.* 1995:279–287.
34. Katritzky AR, Maran U, Karelson M, Lobanov VS. Prediction of melting points for the substituted benzenes: A QSPR approach. *J Chem Inf Comput Sci.* 1997; 37:913–919.
35. Katritzky AR, Sild S, Karelson M. A general QSPR treatment of the refractive index of organic compounds. *J Chem Inf Comput Sci.* 1998; 38:840–844.
36. Xu J, Chen B, Zhang Q, Guo B. Prediction of refractive indices of linear polymers by a four-descriptor QSPR model. *Polymer.* 2004; 45:8651–8659.
37. Katritzky AR, Rachwal P, Law KW, Karelson M, Lobanov VS. Prediction of polymer glass transition temperatures using a general quantitative structure-property relationship treatment. *J Chem Inf Comput Sci.* 1996; 36:879–884.
38. Katritzky, AR.; Lobanov, VS.; Karelson, M. CODESSA user's manual. University of Florida; Gainesville: 1994.
39. Dudek AZ, Arodz T, Galvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. *Comb Chem High Throughput Screen.* 2006; 9:213–228. [PubMed: 16533155]
40. Tarca LA, Grandjean BPA, Larachi F. Feature selection methods for multiphase reactors data classification. *Ind Eng Chem Res.* 2005; 44:1073–1084.
41. Storn R. Differential evolution (DE) for continuous function optimization (an algorithm by Kenneth Price and Rainer Storn).
42. Mayer DG, Kinghorn BP, Archer AA. Differential evolution - an easy and efficient evolutionary algorithm for model optimisation. *Agric Syst.* 2005; 83:315–328.
43. Robi T, Filipi B. DEMO: Differential evolution for multiobjective optimization. *Evolutionary Multi-Criterion Optimization.* 2005:520–533.
44. ChemBioOffice. CambridgeSoft; Cambridge, MA: 2008.
45. Hansch, C.; Leo, A. Exploring QSAR. In: Heller, SR., editor. *Fundamentals and Applications in Chemistry and Biology.* Vol. 1. American Chemical Society; Washington, DC: 1995.
46. Visual Studio. Microsoft Corporation; Redmond, WA: 2005.
47. AMPAC. Semichem Inc.; Shawnee, KS: 1998.
48. Dewar MJS, Storch DM. Development and use of quantum molecular models. 75. Comparative tests of theoretical procedures for studying chemical reactions. *J Am Chem Soc.* 1985; 107:3898–3902.
49. Dewar MJS. Applications of quantum mechanical molecular models to chemical problems. Part 70. Quantum mechanical molecular models. *The Journal of Physical Chemistry.* 1985; 89:2145–2150.
50. Rauhut G, Clark T. Multicenter point charge model for high quality molecular electrostatic potentials from AM1 calculations. *J Comput Chem.* 1993; 14:503–509.

51. Beck B, Rauhut G, Clark T. The natural atomic orbital point charge model for PM3: Multipole moments and molecular electrostatic potentials. *J Comput Chem.* 1994; 15:1064–1073.
52. Clark, T. *Molecular Informatics: Confronting Complexity*. Bozen, Italy: 2002. DOES QUANTUM CHEMISTRY HAVE A PLACE IN CHEMINFORMATICS?.
53. Netzeva TI, Aptula AO, Benfenati E, Cronin MTD, Gini G, Lessigiarska I, Maran U, Vracko M, Schüürmann G. Description of the electronic structure of organic chemicals using semiempirical and ab initio methods for development of toxicological QSARs. *Journal of chemical information and modeling.* 2005; 45:106–114. [PubMed: 15667135]
54. Varekova RS, Geidl S, Ionescu CM, Skrehota O, Kudera M, Sehnal D, Bouchal T, Abagyan R, Huber HJ, Koca J. Predicting pKa Values of Substituted Phenols from Atomic Charges: Comparison of Different Quantum Mechanical Methods and Charge Distribution Schemes. *Journal of chemical information and modeling.* 2011; 51:1795–1806. [PubMed: 21761919]
55. CODESSA. Semichem Inc.; Shawnee, KS: 2007.
56. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Networks.* 1989; 2:359–366.
57. Golbraikh A, Tropsha A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J Comput Aided Mol Des.* 2002; 16:357–369. [PubMed: 12489684]
58. Hansen LK, Salamon P. Neural network ensembles. *IEEE Trans Pattern Anal.* 1990; 12:993–1001.
59. Krogh A, Sollich P. Statistical mechanics of ensemble learning. *Phys Rev E.* 1997; 55:811.
60. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci.* 2003; 22:69–77.
61. Avdeef, A. Prediction of Log P with Property-Based Method. In: Mannhold, R., editor. *Molecular Drug Properties*. 2003. p. 381–406.
62. Gevrey M, Dimopoulos I, Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol Model.* 2003; 160:249–264.
63. Olden JD, Joy MK, Death RG. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecol Model.* 2004; 178:389–397.
64. Olden JD, Jackson DA. Illuminating the “black box”: A randomization approach for understanding variable contributions in artificial neural networks. *Ecol Model.* 2002; 154:135–150.
65. Dimopoulos Y, Bourret P, Lek S. Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Processing Lett.* 1995; 2:1–4.
66. Verleysen M, Rossi F, François D. Advances in Feature Selection with Mutual Information. *Similarity-Based Clustering.* 2009:52–69.
67. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. *Phys Rev E.* 2004; 69:066138.
68. Rossi F, Lendasse A, François D, Wertz V, Verleysen M. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics Intellig Lab Syst.* 2006; 80:215–226.
69. Dragon. Milano Chemometrics; Milano, Italy: 2007.

Highlights

- > We model octanol-water partition coefficient values.
- > A non-linear quantitative structure-property relationship approach is employed.
- > Three dimensional molecular descriptors were used instead of molecular fragments.
- > Wrapper-based feature selection used with differential evolution and neural networks.
- > The model performed satisfactorily on an external dataset compared to other models.

Figure 1.
Flowchart for the differential evolution algorithm.

Figure 2.
Comparison between the experimental and predicted values of Log K_{ow} in the training set

Figure 3.
Comparison between the experimental and predicted values of Log K_{ow} in the validation set

Figure 4.

Comparison between the experimental and predicted values of Log K_{ow} in the star set of the external validation set

Figure 5. Comparison between the experimental and predicted values of Log K_{ow} in the non-star set of the external validation set

\$watermark-text

\$watermark-text

\$watermark-text

Table 1

Training and validation set root-mean squared error (RMSE) values for the five best networks and the ensemble (average network)

	Network 1	Network 2	Network 3	Network 4	Network 5	Ensemble
Training RMSE	0.329	0.343	0.345	0.348	0.350	0.279
Validation RMSE	0.447	0.450	0.446	0.452	0.462	0.380

\$watermark-text

\$watermark-text

\$watermark-text

Table 2

A comparison of the ensemble predictions with available literature models on the Avdeef test set as reported by Mannhold *et al*¹².

Method	Star set (223 molecules)			Non-star set (43 molecules)				
	RMSE	% of molecules within error range	> 1	RMSE	% of molecules within error range	> 1		
AB/LogP	0.41	84	12	4	1.00	42	23	35
S+logP	0.45	76	22	3	0.87	40	35	26
ACD/Log P	0.50	75	17	7	1.00	44	33	23
CLOGP	0.52	74	20	6	0.80	47	28	26
VLOGP OPS	0.52	64	21	7	0.91	47	28	26
ALOGPS	0.53	71	23	6	1.07	33	28	26
<i>This work</i>	0.57	71	21	8	1.01	37	35	28
MiLogP	0.57	69	22	9	0.86	49	30	21
XLOGP3	0.62	60	30	10	0.89	47	23	30
KowWIN	0.64	68	21	11	1.05	40	30	30

Table 3

The five most significant descriptors common to the five best models and their average percentage contribution towards prediction of log K_{ow}

Descriptor	Physical description	Average % contribution
(1/6)X GAMMA polarizability (DIP)	Describes the molecular size and reactivity of the molecule	63.4
Max n-n repulsion for a H-N bond	Describes the ability of the H-N bond to be involved in a chemical reaction	6.5
Max e-n attraction for a H-N bond	Describes the ability of the H-N bond to be involved in a chemical reaction	5.4
Number of atoms	Describes the number of atoms (may also be used to describe the size of the molecule)	2.6
Vibrational heat capacity (300K)		2.2