 doi:10.20944/preprints202002.0248.v1

Article

Predictive Modeling of Henry's Law Constant in Chemical Structures Using LSSVM and ANFIS Algorithms

Qikai Wang [1,*], Aiqin Yao [1], Manouchehr Shokri [2], Adrienn A. Dineva [3,*]

[1] School of Information and Communication Engineering, North University of China, Shanxi, 030051, China

[2] Institute of Structural Mechanics, Bauhaus-Universität Weimar, D-99423 Weimar, Germany; manouchehr.shokri@uni-weimar.de

[2] Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam

* Correspondence: adrienndineva@duytan.edu.vn

Abstract: Henry's constants for different existing compounds in water have great importance in transfer calculations. Measurement of these constants face different difficulties including high costs of experiment and low accuracy of measurement apparatus. Due to these facts, proposing a low cost and accurate approach becomes highlighted. To this end, adaptive neuro-fuzzy inference system (ANFIS) and least squares support vector machine (LSSVM) have been used as Henry's constant predictor tools. The molecular structure of compounds has been used as inputs of models. After training the models, the visual and mathematical studies of outputs have been done. The coefficients of determination of LSSVM and ANFIS algorithms are 0.999 and 0.990 respectively. According to the comprehensiveness of databank and accurate prediction of algorithms, it can be concluded that LSSVM and ANFIS algorithms are accurate methods for prediction of Henry's constant in wide range of chemical structure of compounds in water.

Keywords: Henry's Law; chemical structure; Artificial intelligence; LSSVM; ANFIS

1. Introduction

The fates of different organic materials in environment extensively depend on various processes especially transfer of chemical materials between aqueous and air phases [1]. For the compounds in water, the Henry's law constant is known as one of the utmost important process's parameter [2]. This constant for various compounds in water has vital role in different areas of chemistry such as geochemistry, toxicological chemistry, environmental chemistry and chemical engineering. The H is defined by the ratio of chemical's concentrations in water to air. Due to this fact, reliable source of data for H is highly required to check the fates of chemical compounds in environment.

Overall, it is clear that the precise determination of H is costly due to the adsorption of low amounts of solute on the apparatus and also there are some limitations in the analytical detection of very hydrophobic compounds at low concentrations. Consequently, the prediction of H has fundamental value in several scientific phenomenon [3,4].

In the literature, there are some approaches to predict H of organic compounds in water based on chemical structure directly. Additionally, a number of indirect approaches for prediction of H based on vapor-liquid equilibrium data including activity coefficient, however their applications for prediction of the H are not exactly assessed [5,6]. Consequently in this paper, we focus on those approaches which can predict the H directly. There are two main types of correlation for prediction

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0248.v1

of the H. The first type belongs to the correlations of the physical properties such as aqueous solubility and vapor pressure for prediction of the H. One of the popular approaches in this type is the correlation suggested in [7]. In this correlation, there are some significant disadvantages including the degree of accuracy is a function of required physical properties or approaches applied to predict the characteristics and properties. Moreover, when a required property is missed, the prediction of the H is not possible.

The second type of correlations is known as quantitative structure property relationships which utilize molecular parameters for estimation of the H. The applicable correlations of this type are suggested by Yafe et al., Lin and Sandler, Yao et al. , English and Carrol, Dearden et al., Katritzky et al, Abraham et al. Meylan and Howard, and Hine and Mookerjee[8-18].

In the current study, two new computational methods are presented to predict H of organic compounds in terms of existing functional groups. To this end, adaptive neuro-fuzzy inference system and least squares support vector machine have been employed and finally, different statistical and graphical comparison methods have been applied to determine precision of these algorithms.

2. Methodology

2.1. Experimental Data Gathering

The generalization of molecular-based prediction method is highly function of comprehensiveness of databank of materials used to its preparation. Due to this fact, the diversity of chemical families and the number of available compounds in databank have become highlighted. Investigation of literature reveals that the most reliable databank for H of compounds has been collected by Yaws so that 1940 H values for pure compounds can be found in Yaws' work [19]. It is worthy to mention that H values have been gathered in terms of atm.m^3.mol^{-1} and shown in a decimal log(H) at temperature of 25^{o}C. Their ranges are between -13.461 to 6.238.according to the previous works, this databank is known as the most comprehensive and reliable databank has been applied for estimation of the H values of organic compounds in water. After gathering the databank, the chemical structural analysis of these data has shown that 107 functional groups exist in the structure of under-studied compounds. The number of the functional groups in the structure of compounds is used as inputs of models.

2.2. Adaptive neuro-fuzzy inference system

The development of fuzzy logic was proposed by Zade. Applying ANN and fuzzy logic methods simultaneously makes a new form of artificial intelligence method called ANFIS.

In this method, the configuration have 5 different layers**.** The Gaussian membership function is optimized to reach most accurate answers [20-24]:

$$O_i^1 = \mu_{Ai}(x), \qquad \forall\, i \in \{1,2\} \qquad Eq.(1)$$

$$O_i^1 = \mu_{B_{i-1}}(y), \qquad \forall\, i \in \{3,4\} \qquad Eq.(2)$$

$$\mu_A(x) = e^{-\frac{(x-c_i)^2}{2\sigma_i^2}} \qquad Eq.(3)$$

Where O_i^j denotes i-th output for j-th layer, x and y denote input parameters.

The 2nd layer which have constant nodes, can be expressed as below:

$$O_i^2 = \omega_i = O_i^1 = \mu_{Ai}(x)\mu_{Bi}(y), \qquad \forall\, i \in \{1,2\} \qquad Eq.(4)$$

In 3rd layer which called normalized layer, the firing strength outputs is normalized:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0248.v1

$$O_i^3 = \overline{\omega}_i = \frac{\omega_1}{\omega_1 + \omega_2}, \qquad \forall\, i \in \{1,2\} \qquad Eq.(5)$$

The 4th layer belong to linguistic expressions for outputs as following:

$$O_i^4 = \overline{\omega}_i f_i = \overline{\omega}_i (p_i x + q_i y + r_i), \qquad \forall\, i \in \{1,2\} \qquad Eq.(6)$$

The last step applied all rules together as below:

$$O_i^5 = \sum_i \overline{\omega}_i f_i = \frac{\sum_i \omega_i f_i}{\sum_i \omega_i} \qquad Eq.(7)$$

In the current work, particle swarm optimization is used for optimization of ANFIS algorithm as shown in **Figure 1**.

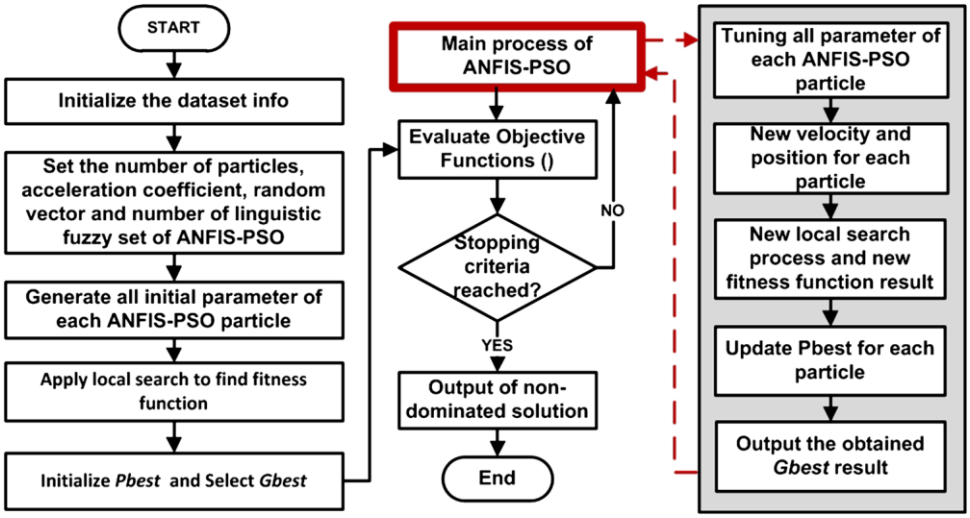


Figure 1. Flow chart diagram of ANFIS-PSO

1.1. Least squares support vector machine

Regression problems assume a model such as following equation in original space:

$f(x) = \omega^T \emptyset(x) +$

b

Eq.(8)

In which $\emptyset$, f(x) and x represent nonlinear transformations, output and input sample. The LSSVM algorithm in the original space can be explained as below optimization formulation:

$$min = \frac{1}{2}\|\omega\|^2 + \frac{1}{2}\gamma \sum_{i=1}^{n} e_i^2 \qquad Eq.(9)$$

Subject to below equality:

$$y_i = \langle \omega, \emptyset(x_i) \rangle + b + e_i \qquad Eq.(10)$$

Where ω can become an infinite dimensional vector, therefore the first form of optimization problem may not be solved. Due to this fact, the aforementioned problem must be reformulated to a dual optimization. The function of output can be formulated for original dimensional feature space as below:

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x) + \beta \qquad Eq.(11)$$

Where K is known as kernel function. In the current work, radial basis function has been chosen as kernel function:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0248.v1

$$K(x_k, x) = \exp\left(-\frac{||x_k - x||^2}{\sigma^2}\right) \quad \text{Eq. (12)}$$

σ^2 is the radial basis function width [25-33].

In order to determine hyper-parameters of LSSVM, PSO algorithm has been implemented as shown in **Figure 2**.

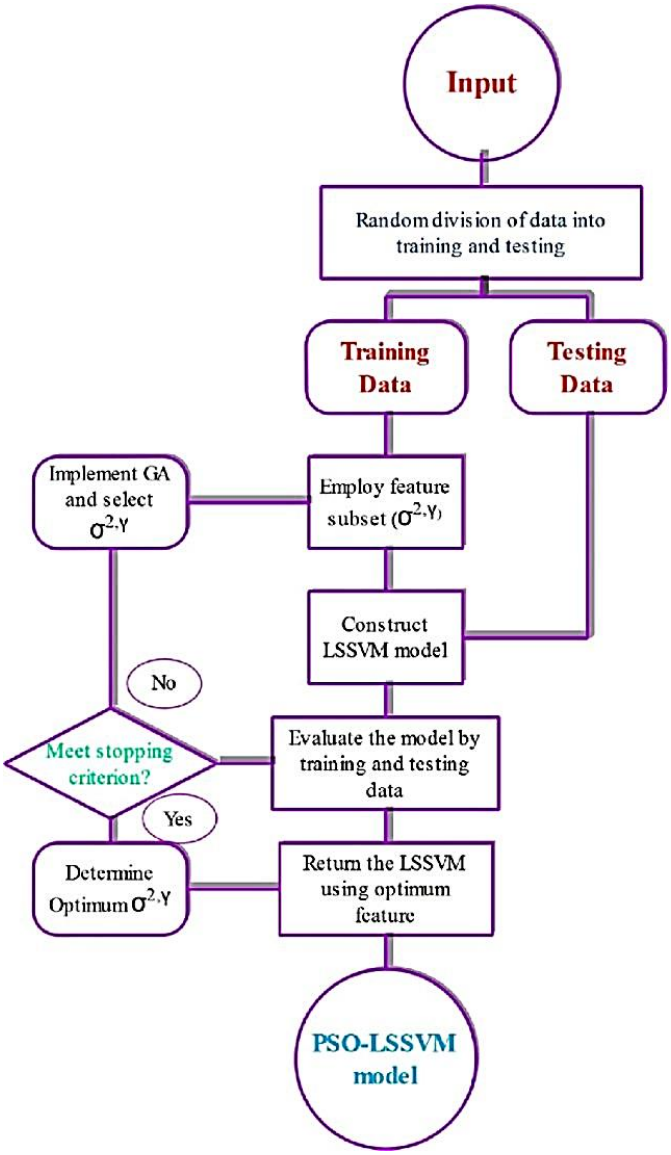


Figure 2. Flow chart diagram of LSSVM-PSO

3. Results and discussion

In order to determine the H of pure compounds in aqueous solutions, two new computational methods including LSSVM and ANFIS algorithms have been used. It is obvious that the main step of development of a model is evaluation of accuracy so this section has implemented different statistical parameters including:

- R-squared (R^2):

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0248.v1

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(X_i^{actual} - X_i^{predicted})^2}{\sum_{i=1}^{N}(X_i^{actual} - \overline{X^{actual}})^2} \quad \text{Eq. (13)}$$

- Mean squared error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(X_i^{actual} - X_i^{predicted})^2 \quad \text{Eq. (14)}$$

- Standard deviations (STD)

$$STD_{error} = \left(\frac{1}{N-1}\sum_{i=1}^{N}(error - \overline{error})\right)^{0.5} \quad \text{Eq. (15)}$$

- Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i^{exp.} - X_i^{predicted})^2} \quad \text{Eq. (16)}$$

- Mean relative error

$$MRE = \frac{100}{N}\sum_{i=1}^{N}\left(\frac{X_i^{actual} - X_i^{predicted}}{X_i^{actual}}\right) \quad \text{Eq. (17)}$$

The above parameters have been reported in **Table 1** for LSSVM and ANFIS algorithms.

Table 1. Statistical parameters for prediction of Henry's law constant

		R^2	MRE(%)	MSE	RMSE	STD
ANFIS	Train	0.990	7.650	0.0601	0.2452	0.2004
	Test	0.990	6.391	0.0550	0.2344	0.1865
	Total	0.990	7.335	0.0588	0.2344	0.1970
LSSVM	Train	0.999	2.685	0.0072	0.0847	0.0647
	Test	0.998	2.797	0.0097	0.0984	0.0780
	Total	0.999	2.713	0.0078	0.0984	0.0683

It can be seen that the statistical parameters are determined as R^2=0.999, MRE=2.713, MSE=0.0078, STD=0.0683 and RMSE=0.0984 for LSSVM algorithm and also for ANFIS algorithm they are reported as R^2=0.990, MRE=7.335, MSE=0.0588, STD=0.1970 and RMSE=0.2344. Due to these results, LSSVM algorithm has better performance in calculation of the H of different compounds in aqueous solution. It is worthy to compare outputs of algorithms and experimental values of the H so the simultaneous demonstration of predicted H values and actual H values in logarithm scale are shown in **Figure 3**. This comparison expresses the high degree of agreement between models outputs and actual henry's constant. After that, the cross plots of actual logH versus predicted logH are shown in **Figure 4** for both algorithms. The interesting compaction of data points around y=x line gives information about the high accuracy of these computational algorithms in estimation of Henry's law constant.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0248.v1

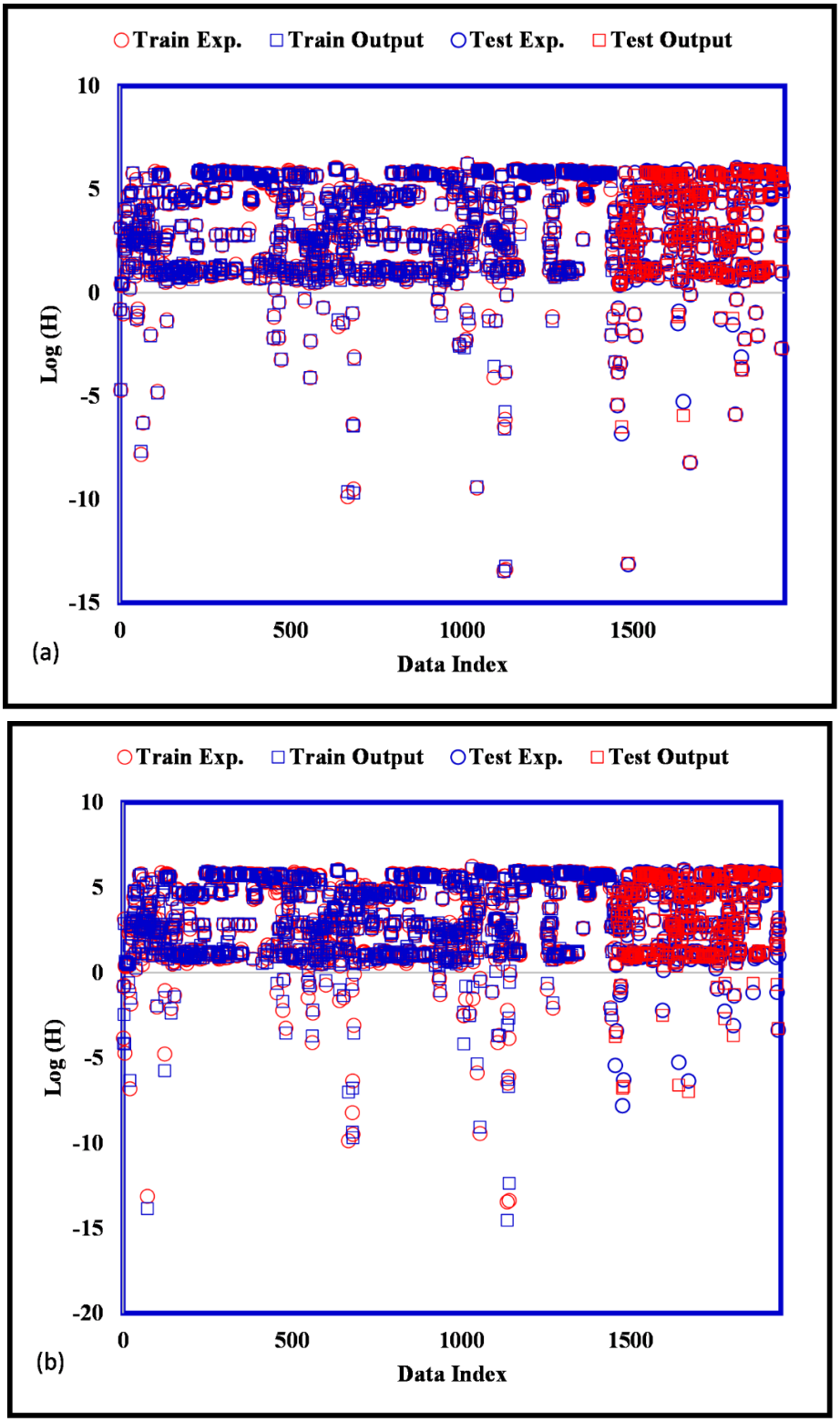


Figure 3. Simultaneous demonstration of predicted and experimental log (H) for a) ANFIS b) LSSVM

Finally, the relative deviation of each estimated H from the actual value is determined and reported in **Figure 5**. The low range of deviation of LSSVM and ANFIS Henry's law constants form this large dataset expresses the fact these two algorithms have high performance in prediction of the target.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0248.v1

Due to these explanation, LSSVM and ANFIS algorithms are robust tools in chemical engineering and chemistry especially for prediction of Henry's law constant.

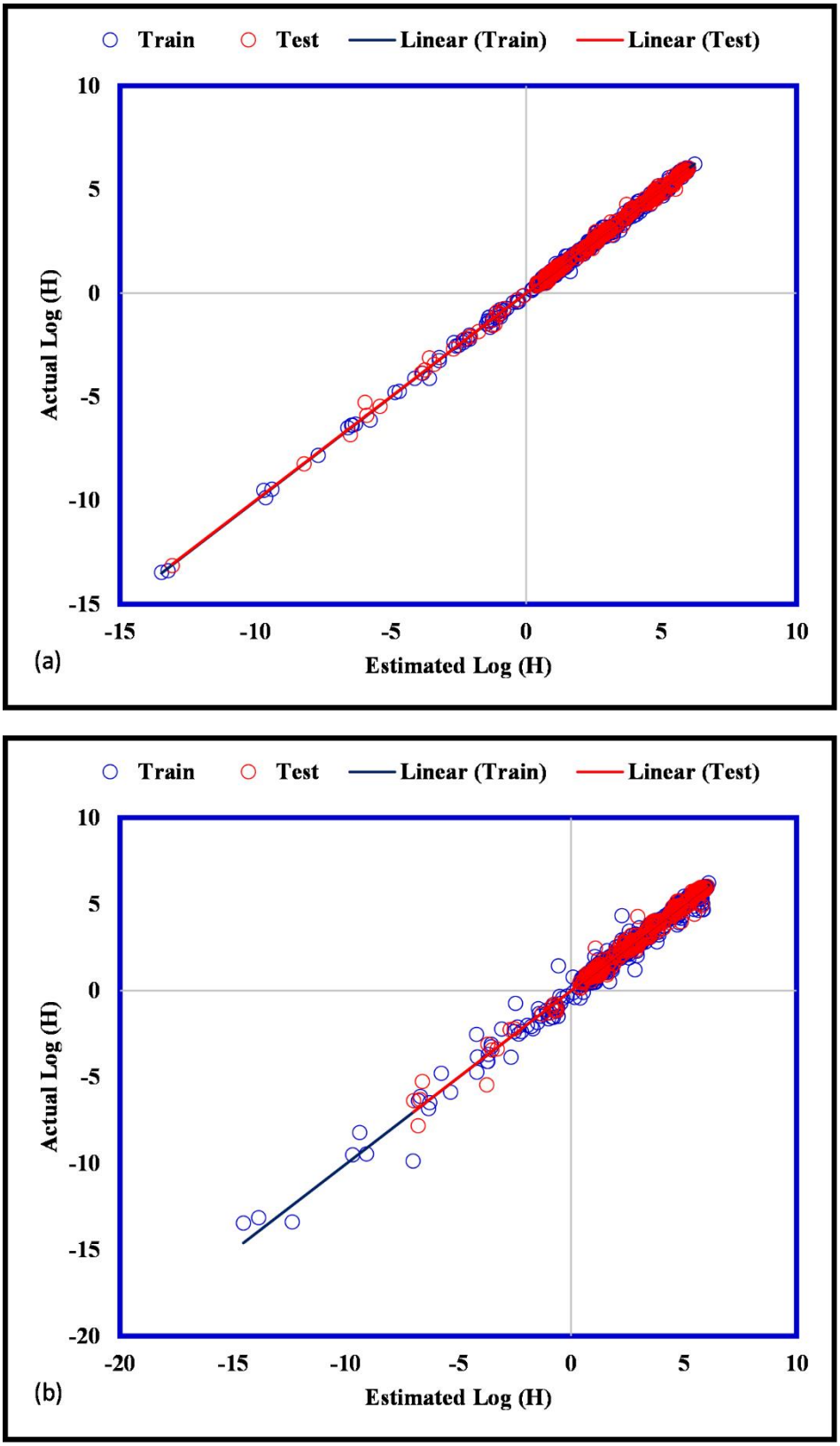
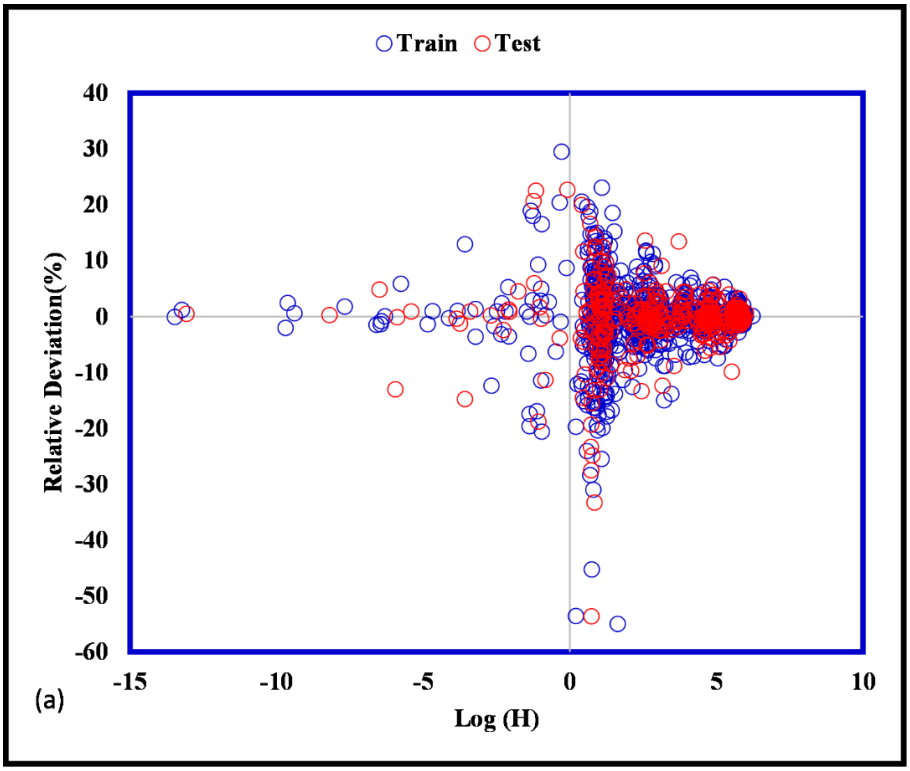


Figure 4. Cross plots of Actual and predicted log(H) for a)ANFIS b)LSSVM

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0248.v1



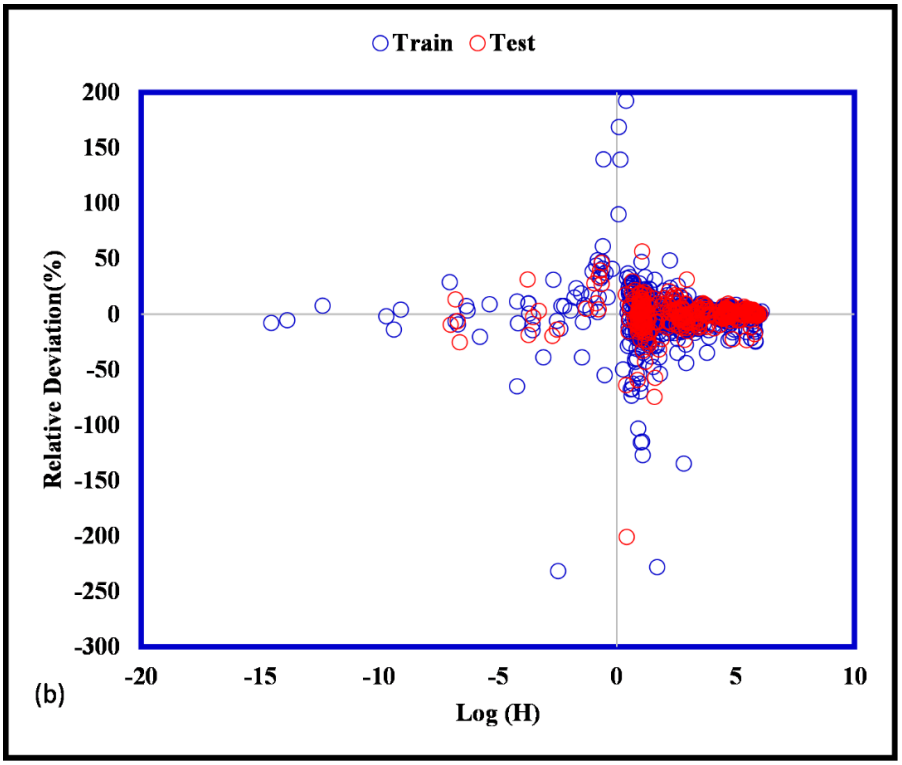


Figure 5. Relative deviations of predicted and actual log (H) for a) ANFIS b) LSSVM

4. Conclusions

In the current work, two novel molecular-based approaches were suggested for prediction of the henry's constant of various compounds in water. These models were constructed based on LSSVM and ANFIS algorithms. The models' variables and parameters include the existences of 107 classes for every compound. It is discussed that the majority of the classes are not existing in the compound.

Therefore, computing the input parameters based on chemical structure of any compound is simple. In order to prepare the models, 1940 different compounds were utilized so the proposed models can are implemented to estimate Henry's law constant for comprehensive range of chemical structures. According to the comparison of models and experimental data, it can be concluded that the statistical parameters are determined as R^2=0.999, MRE=2.713, MSE=0.0078, STD=0.0683 and RMSE=0.0984 for LSSVM algorithm and also for ANFIS algorithm they are reported as R^2=0.990, MRE=7.335, MSE=0.0588, STD=0.1970 and RMSE=0.2344. On the other hand, visual comparison of predicted and experimental Henry's law constants shows accuracy of models in the simplest manner. Therefore, the suggested algorithms have interesting accuracy in prediction of Henry's law constants for wide range of chemical structures in water. For the future work investigation on the hybrid machine learning models are encouraged.

Authors contributions:

Data curation, Qikai Wang and Manouchehr Shokri; Formal analysis, Qikai Wang and Adrienn Dineva; Funding acquisition, Manouchehr Shokri; Investigation, Qikai Wang ; Resources, Aiqin Yao and Manouchehr Shokri; Software, Aiqin Yao and Manouchehr Shokri; Supervision, Adrienn Dineva; Visualization, Aiqin Yao ; Writing – original draft, Adrienn Dineva; Writing – review & editing, Adrienn Dineva.

Declaration: Authors declare no conflict of interests.

References

1. Gharagheizi, F.; Abbasi, R.; Tirandazi, B. Prediction of Henry's law constant of organic compounds in water from a new group-contribution-based model. *Industrial & engineering chemistry research* **2010**, *49*, 10149-10152.
2. Schwarzenbach, R.P.; Gschwend, P.M.; Imboden, D.M. *Environmental organic chemistry*; John Wiley & Sons: 2016.
3. Altschuh, J.; Brüggemann, R.; Santl, H.; Eichinger, G.; Piringer, O.G. Henry's law constants for a diverse set of organic chemicals: Experimental determination and comparison of estimation methods. *Chemosphere* **1999**, *39*, 1871-1887.
4. Staudinger, J.; Roberts, P.V. A critical review of Henry's law constants for environmental applications. *Critical Reviews in Environmental Science and Technology* **1996**, *26*, 205-297.
5. Ramírez-Beltrán, N.D.; Vallés, H.R.; Estévez, L.A.; Duarte, H. A neural network approach to predict activity coefficients. *The Canadian Journal of Chemical Engineering* **2009**, *87*, 748-760.
6. Petersen, R.; Fredenslund, A.; Rasmussen, P. Artificial neural networks as a predictive tool for vapor-liquid equilibrium. *Computers & chemical engineering* **1994**, *18*, S63-S67.
7. Mackay, D.; Boethling, R.S. *Handbook of property estimation methods for chemicals: environmental health sciences*; CRC press: 2000.

8. Hine, J.; Mookerjee, P. The intrinsic hydrophilic character of organic compounds, Correlations in terms of structural contributions. *Chemischer Informationsdienst* **1975**, *6*, no-no.
9. Meylan, W.M.; Howard, P.H. Bond contribution method for estimating Henry's law constants. *Environmental Toxicology and Chemistry: An International Journal* **1991**, *10*, 1283-1293.
10. Meylan, W.M.; Howard, P. HENRYWIN, version 3.10. *Syracuse Research: Syracuse, NY* **2000**.
11. Abraham, M.H.; Andonian-Haftvan, J.; Whiting, G.S.; Leo, A.; Taft, R.S. Hydrogen bonding. Part 34. The factors that influence the solubility of gases and vapours in water at 298 K, and a new method for its determination. *Journal of the Chemical Society, Perkin Transactions 2* **1994**, 1777-1791.
12. Katritzky, A.R.; Mu, L.; Karelson, M. A QSPR Study of the Solubility of Gases and Vapors in Water. *Journal of chemical information and computer sciences* **1996**, *36*, 1162-1168.
13. Dearden, J.C.; Ahmed, S.A.; Cronin, M.T.; Sharra, J.A. QSPR prediction of Henry's law constant: improved correlation with new parameters. In *Molecular Modeling and Prediction of Bioactivity*, Springer: 2000; pp. 273-274.
14. Modarresi, H.; Modarress, H.; Dearden, J.C. QSPR model of Henry's law constant for a diverse set of organic chemicals based on genetic algorithm-radial basis function network approach. *Chemosphere* **2007**, *66*, 2067-2076.
15. English, N.J.; Carroll, D.G. Prediction of Henry's law constants by a quantitative structure property relationship and neural networks. *Journal of chemical information and computer sciences* **2001**, *41*, 1150-1161.
16. Yao, X.; Liu, M.; Zhang, X.; Hu, Z.; Fan, B. Radial basis function network-based quantitative structure–property relationship for the prediction of Henry's law constant. *Analytica Chimica Acta* **2002**, *462*, 101-117.
17. Lin, S.-T.; Sandler, S.I. Henry's law constant of organic compounds in water from a group contribution model with multipole corrections. *Chemical engineering science* **2002**, *57*, 2727-2733.
18. Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. A Fuzzy ARTMAP-based quantitative structure− property relationship (QSPR) for the Henry's law constant of organic compounds. *Journal of chemical information and computer sciences* **2003**, *43*, 85-112.
19. Yaws, C.L.; Gabbula, C. *Yaws" Handbook of Thermodynamic and Physical Properties of Chemical Compounds*; Knovel: 2003.
20. Keybondorian, E.; Taherpour, A.; Bemani, A.; Hamule, T. Application of novel ANFIS-PSO approach to predict asphaltene precipitation. *Petroleum Science and Technology* **2018**, *36*, 154-159.
21. Malmir, P.; Suleymani, M.; Bemani, A. Application of ANFIS-PSO as a novel method to estimate effect of inhibitors on Asphaltene precipitation. *Petroleum Science and Technology* **2018**, *36*, 597-603.

22. Mir, M.; Kamyab, M.; Lariche, M.J.; Bemani, A.; Baghban, A. Applying ANFIS-PSO algorithm as a novel accurate approach for prediction of gas density. *Petroleum Science and Technology* **2018**, *36*, 820-826.
23. Razavi, R.; Sabaghmoghadam, A.; Bemani, A.; Baghban, A.; Chau, K.-w.; Salwana, E. Application of ANFIS and LSSVM strategies for estimating thermal conductivity enhancement of metal and metal oxide based nanofluids. *Engineering Applications of Computational Fluid Mechanics* **2019**, *13*, 560-578.
24. Suleymani, M.; Bemani, A. Application of ANFIS-PSO algorithm as a novel method for estimation of higher heating value of biomass. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects* **2018**, *40*, 288-293.
25. Keybondorian, E.; Zanbouri, H.; Bemani, A.; Hamule, T. Estimation of the higher heating value of biomass using proximate analysis. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects* **2017**, *39*, 2025-2030.
26. Ardabili, S.; Mosavi, A.; Dehghani, M.; Várkonyi-Kóczy, A.R. Deep Learning and Machine Learning in Hydrological Processes Climate Change and Earth Systems a Systematic Review. In *Lecture Notes in Networks and Systems*, Springer: 2020; Vol. 101, pp 52-62.
27. Ardabili, S.; Mosavi, A.; Mahmoudi, A.; Gundoshmian, T.M.; Nosratabadi, S.; Várkonyi-Kóczy, A.R. Modelling Temperature Variation of Mushroom Growing Hall Using Artificial Neural Networks. In *Lecture Notes in Networks and Systems*, Springer: 2020; Vol. 101, pp 33-45.
28. Ardabili, S.; Mosavi, A.; Várkonyi-Kóczy, A.R. Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods. In *Lecture Notes in Networks and Systems*, Springer: 2020; Vol. 101, pp 215-227.
29. Fardad, K.; Najafi, B.; Ardabili, S.F.; Mosavi, A.; Shamshirband, S.; Rabczuk, T. Biodegradation of medicinal plants waste in an anaerobic digestion reactor for biogas production. *Comput. Mater. Continua* **2018**, *55*, 318-392, doi:10.3970/cmc.2018.01803.
30. Gundoshmian, T.M.; Ardabili, S.; Mosavi, A.; Várkonyi-Kóczy, A.R. Prediction of Combine Harvester Performance Using Hybrid Machine Learning Modeling and Response Surface Methodology. In *Lecture Notes in Networks and Systems*, Springer: 2020; Vol. 101, pp 345-360.
31. Mosavi, A.; Ardabili, S.; Várkonyi-Kóczy, A.R. List of Deep Learning Models. In *Lecture Notes in Networks and Systems*, Springer: 2020; Vol. 101, pp 202-214.
32. Mosavi, A.; Rabczuk, T. Learning and intelligent optimization for material design innovation. Kvasov, D.E., Sergeyev, Y.D., Battiti, R., Battiti, R., Kvasov, D.E., Sergeyev, Y.D., Eds. Springer Verlag: 2017; Vol. 10556 LNCS, pp 358-363.
33. Nosratabadi, S.; Mosavi, A.; Keivani, R.; Ardabili, S.; Aram, F. State of the Art Survey of Deep Learning and Machine Learning Models for Smart Cities and Urban Sustainability. In *Lecture Notes in Networks and Systems*, Springer: 2020; Vol. 101, pp 228-238.