Research Article

Chemical Review and Letters
journal homepage: www.chemrevlett.com
ISSN (online): 2645-4947 (print) 2676-7279

Prediction of hydration enthalpy of low molecular weight organic molecules with machine learning regression based on COSMO-SAC

Iman Sabeeh Hasan[1], Alhussein Arkan Majhool[2], Mustafa Humam Sami[3] and Ahmed Kareem Obaid Aldulaimi[4*]

[1]*Department of Pharmacy, Al-Zahrawi University College, Karbala, Iraq*
[2]*College of Applied Medical Sciences, University of Kerbala, Kerbala, Iraq*
[3]*Department of Pharmacy, Al-Noor University College, Nineveh, Iraq*
[4]*College of Food Sciences, Al-Qasim Green University, Babylon, Iraq*

ARTICLE INFO

Article history:
Received 26 May 2023
Received in revised form 10 June 2023
Accepted 11 June 2023
Available online 25 June 2023

Keywords:
Enthalpy of hydration
COSMO-SAC
Machine learning
FreeSolv
Random Forest regression

ABSTRACT

COSMO-SAC modeling is a reliable method to determine the activity coefficient of the mixtures and is used to predict low molecular weight organic materials hydration enthalpy. A dataset of 96 organic molecules' activity coefficients in the different solvents (water, ethanol, methanol, toluene, and benzene) mixtures have been obtained in full range composition with COSMO-SAC. The created database has been merged with the FreeSolv dataset to include the hydration enthalpy of these materials as input of machine learning training besides the Van der Waals diameter, other important molecular descriptive. The support vector regressor, random forest regressor, and gradient boosting decision tree regressor have been used for data training and prediction of hydration enthalpy of the organic and pharmaceutical materials. Variation of training and testing rates is most effective parameter in the prediction of enthalpy of hydration. The random forest regression is the most accurate method in the prediction of the enthalpy of hydration with 1.5 % RMSD with a train: test ratio of 0.25:0.75 between the studied methods.

1. Introduction

Predicting the chemical thermodynamic properties of pure materials and mixtures is a vital matter for industrial purposes [1], [2]. Converting scientific data to engineered products in the industry requires reliable methodologies that could provide necessary data where those data are not available [3]. Many efforts were taking place to develop models that could predict the thermodynamic properties [4]–[6]. Local composition models [7], [8], cubic models[9], [10], statistical thermodynamics models such as statistical associated fluid theory (SAFT) are successful models that are used extensively [11]–[13]. Recently, developed models based on quantum and statistical mechanics for equilibrium thermodynamics such as COSMO-RS have also shown good performance [14], [15]. All of these cases involve complex calculations with a long computational time that may be an obstacle for a researcher in the process of conducting applied studies in chemistry and chemical engineering.

COSMO-SAC model is commonly used for the activity coefficient calculation of mixtures. [16] It works based on the statistical thermodynamics that gets σ-profiles from quantum mechanics calculations as input. Generally, dmol3 was used for geometry optimization and minimization of molecule energy, and evaluation of σ-profiles.[17] Also, the COSMO-SAC model provides good results for the activity coefficient with a low deviation from experimental results. Indeed, it has a good reputation and is considered a reliable method in the prediction of the activity coefficient of organic materiTaals. Also, it has been shown that the COSMO-SAC thermodynamic properties depends on the chemical family rather than the size of the molecule that makes it powerful tool for the purpose of this study.[18]

Various validated databases in chemistry are developed for the military, industrial, pharmaceutical, and educational purposes. One of the most reliable thermochemical databases is created by the National Institute of Standards and Technology (NIST) that is the most reliable database for materials thermochemical information [19], [20]. However, some valid datasets are

[*] Corresponding author.; e-mail: ahmedaldulaimi1@gmail.com

also available as open-source physicochemical information of materials that are created by different research teams, such as the MGCDB84, GMTKN55, and Minnesota Database databases [21]–[23] that are the result of quantum computing. The FreeSolv dataset was published as open-source that contains the free energy of the low molecular weight organic molecules including different functional groups mainly pharmaceutical substances [24]. Generally, it includes experimental data combined with quantum computational data (DFT) and molecular dynamics that make it an appropriate dataset for machine learning uses.

Machine learning (ML) is an artificial intelligence branch that could be used for the prediction of a variable with an automated process without the need for explicit programming [25]. The ML is based on the data analysis that began with access to data and uses it for learning. The learning process begins with observations of data to find instructions with specific patterns in the data. In this respect, the data are divided into two parts: training data and test data. There are many different algorithms for machine learning, and they are typically categorized as supervised learning, unsupervised learning, and semi-supervised learning [26].

Enthalpy of hydration is mainly an essential parameter for estimating different thermodynamic and chemical engineering variables such as solubility, required heat for processing, etc. [27], [28]. Generally, measurement of this quantity requires a precise and expensive microcalorimeter or it is impossible to measure it in the determined conditions due to degradation of the material [29], [30]. Prediction of the enthalpy of hydration with acceptable accuracy for organic materials, especially medicinal products is a vital matter that could be used effectively to accelerate the engineered processes [31]–[33]. As previously mentioned, the time-consuming and complex quantum computational methods are limiting factors for industrial purposes. In this regard, machine learning can help speed up the calculations to obtain the required enthalpy of hydration of material with proper initial inputs.

The FreeSolv dataset includes the free energy of hydration for low molecular weight organic molecules [24]. There are experimental, DFT calculations, and molecular dynamics data including different thermodynamic properties and molecular descriptive in the FreeSolv [24]. The FreeSolv dataset has been merged with a produced COSMO-SAC dataset including infinite dilution activity coefficient of the low molecular weight organic materials in various solvents such as water, ethanol, methanol, benzene, and toluene. Different machine learning methods such as support vector machine, random forest, and gradient boosting decision tree are used to predict the enthalpy of hydration.

2. Materials and methods

2.1. COSMO-SAC model

The procedures have been implemented with python in the Jupyter environment. Different two PCs with different configurations have been used to evaluate the results, and the results were identical in the two configurations which are important for the repeatability of the process. Accordingly, the FreeSolv dataset and VT2005 σ-profiles dataset has been used as initial data.[24], [34] It should be noted that there were 96 exact matches according to the IUPAC names of the materials between the two datasets, and it was a limitation of this work. The activity coefficients of 96 organic materials with different solvents such as methanol, ethanol, benzene, toluene, and water in full range composition (mole fractions of solute = 0, 0.1, …, 0.9, 1) at 298.15 K have been calculated by the open-source benchmark of the COSMO-SAC implemented by Bell et al. A detailed information is available in the corresponding paper. Also, it is accessible from the GitHub repository [35].

2.2. Machine learning

2.2.1. Support vector machine regressor

The model generated by the support vector machine (SVM) classifier depends only on a subset of the training data where the cost function for constructing the model does not matter to the training data that are beyond the margin. Similarly, the model generated by SVR depends only on a subset of the training data, because the cost function ignores samples whose prediction is close to the target. Selecting the appropriate kernel for subset tuning will be the main issue using this method. In this research, Gaussian, sigmoid, and polynomials kernels of SVR have been used [36] for prediction of enthalpy of hydration of organic materials.

2.2.2. Random forest regressor

Random Forest is a meta-estimator that fits classification decision tree sets on different sub-set of the dataset and uses averaging to improve forecasting accuracy and over-fitting control. Decision trees are a non-parametric supervised learning method used for classification and regression. Therefore, the random forest has been used as a white-box model with simple interpretation while the black-box models (for example, in an artificial neural network), and interpretation of the results may be more difficult. It is possible to validate the model using statistical tests, which increases the reliability of the model. Also, it has a good performance and no major difference will be created even if its assumptions are partially violated by the actual model from which the data is derived [37].

2.2.3. Gradient boosting decision tree regression (GBDTR)

The GBDTR makes a cumulative step-by-step model and makes it possible to optimize arbitrary distinct cost functions. At each step, a regression tree is proportional to the negative gradient of the cost function is established. The BBDTR is a generalized model of boosting to arbitrary distinguishable loss functions from the decision tree. It is an accurate and effective method that can be used for regression and classification problems in various fields such as search space ranking. Another advantage of this method is the ability to construct a mathematical formula from a regression problem, which allows providing a comprehensive formula for the regression performed, in which case the importance of the properties can also be examined [38].

2.3. Assessment efficiency of machine learning prediction

In statistics, the mean absolute error (MAE) is a measure of the errors between pairwise observations that express a phenomenon. The sample of Y versus X include a comparison between the predicted value versus the real value of the label that is calculated as follows:

$$MAE(y,\bar{y}) = \frac{1}{n}\sum_{i=0}^{n-1}|y_i - \bar{y_i}| \quad (1)$$

The means squared error function has been used to evaluate the performance of the machine learning method. In statistics, the mean squared error of the estimator measures the mean squared error. There is a risk function between the estimated values and the actual value of the variable that corresponds to the expected value of the square error. Information that can provide a more accurate estimate, which is calculated as follow:

$$MSE(y,\bar{y}) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \bar{y_i})^2 \quad (2)$$

Also, the root mean square error is evaluated using following relation:

$$MSE(y,\bar{y}) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \bar{y_i})^2}$$

(3)

These three statistical variables are the criteria for the assessment of the ML methods accuracy and reliability in the prediction.

3. Results and Discussion

3.1. COSMO-SAC model for infinite dilution activity coefficient

Basically, the COSMO-SAC model uses the quantum mechanics data through a statistical mechanic approach to evaluate the thermodynamic properties of a system. This aim starts with σ-profiles and continues with a series of equation to reach the activity coefficients [16]. Also, the evaluated activity coefficients could be used through the thermodynamic relation to calculate the Gibbs free energy.

The activity coefficient of the organic compounds in different solvents has been predicted using the COSMO-SAC model. This model use σ-profiles of the materials to evaluate the thermodynamic properties based on the statistical thermodynamic relations [16]. The corresponding σ-profiles for the studied materials are available in VT2005 dataset [24], [34]. The predicted activity coefficient data for binary mixtures of thiophene in the studied solvents have been illustrated in Fig 1 as an example. Also, the infinite dilution activity coefficients of these materials are given in Tables 1 in different solvents such as water, ethanol, methanol, benzene, and toluene.

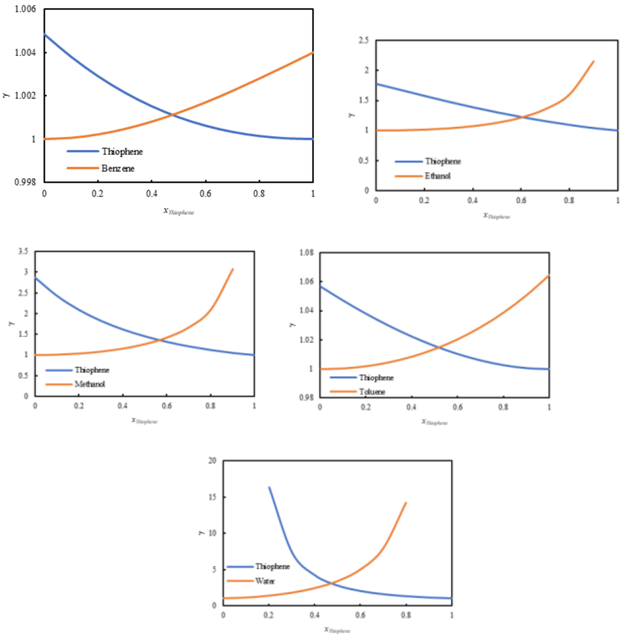

Figure 1. The activity coefficients of the binary mixture's components including thiophene in different solvents (water, ethanol, methanol, benzene, and toluene) versus the mole fraction of thiophene using COSMO-SAC under 0.1 MPa pressure at 298.15 K.

The pioneers and the developers of the COSMO-SAC model have shown that the model is quite reliable.[16]–[18], [34], [39]–[41] On the other hand, the integrity of the evaluated data with the COSMO-SAC is more important rather than the accuracy of the data, and no data comparison with experimental results has been carried out. The COSMO-SAC uses the quantum mechanics data as primary data and evaluates chemical thermodynamic data.[42] These two types of data might be in contradiction due to their different microscopic and macroscopic approaches. Accordingly, the evaluated data has been used without validation in the machine learning process.Base on the thermodynamics rules the activity coefficient of a chemical directly depended on the Gibbs free energy while it is relationship with enthalpy is much more complex. Accordingly, a simple regression could not used to predict the enthalpy with activity coefficient. At this point, the machine learning regression could be used in the prediction of the enthalpy of hydration based on the infinite dilution activity coefficient of a chemical in different solvents at given temperature and pressure.

3.2. Machine learning prediction of enthalpy of hydration

A pre-processing step was required to match whole variables based on the units, significant digits, and other parameters. The pre-processing has been carried out with python encoding of the evaluated COSMO-SAC dataset and FreeSolv dataset. The datasets have been merged and prepared for the machine learning regression process. It should be noted that this step of the procedure is crucial before executing machine learning and a little conflict might cause significant errors in results. Accordingly, the dataset has been checked manually for any defection after all automated procedures.

Table 1. The infinite dilution activity coefficient (γ^{∞}) of some organic materials calculated with COSMO-SAC model under 0.1 MPa at 298.15 K.

Compound Name	γ^{∞}				
	Water	Benzene	Toluene	Ethanol	Methanol
TOLUENE	4607.74	1.024963	1	2.831673	5.66182
1-NITROBUTANE	1502.968	1.086386	1.211993	2.39232	4.183599
2-NITROPROPANE	357.1037	1.109689	1.264321	2.20636	3.391647
THIOPHENE	482.2757	1.004869	1.056671	1.775255	2.868932
ETHYLENE	67.63741	1.001869	1.001967	1.693681	2.384917
2-BUTOXYETHANOL	1284.476	11.54071	12.95789	1.136087	2.080568
CYCLOHEXENE	36110.42	1.376674	1.185683	3.554677	7.698908
PIPERAZINE	2.555234	3.16513	3.756248	0.117807	0.074906
O-CRESOL	399.1619	1.763261	2.013672	0.360845	0.748153
PYRROLE	8.400408	3.95112	4.91893	0.085345	0.152817
INDANE	29574.31	1.119021	1.027768	3.745279	8.91272
ISOBUTANE	6704.744	1.795927	1.463579	3.909756	8.331183
PYRROLIDINE	41.07241	1.621331	1.617396	0.345506	0.363241
P-XYLENE	22676.63	1.100848	1.020159	3.546547	8.21727
PYRENE	1205668	1.067918	0.953135	4.377152	13.97494
NITROBENZENE	1703.4	1.094654	1.226589	2.474726	4.259157
1-METHYLNAPHTHALENE	53959.37	1.013451	0.993618	3.548059	8.59341
NAPHTHALENE	15133.97	0.988594	1.003646	2.925318	6.286829
ACETONE	9.59154	0.954739	1.158579	1.285666	1.318272
METHYLCYCLOHEXANE	125171.2	2.386617	1.792137	6.530031	17.88058
ACETONITRILE	7.085093	2.824741	3.829731	2.714958	2.591342
2-METHYLPYRIDINE	149.8361	1.01337	1.10866	0.931551	1.139583
BENZENE	1041.27	1	1.021475	2.336048	4.054609
METHANE	67.42368	1.297832	1.179531	1.878916	2.695403
P-CRESOL	275.5887	6.094782	6.852709	0.218295	0.494917
3-METHYLHEXANE	516316.5	2.512464	1.836551	7.869931	24.33865
CYCLOPENTENE	2688.153	1.273184	1.127777	2.952746	5.813084
3-METHYLPYRIDINE	105.622	0.992501	1.099709	0.847358	0.980889
METHANOL	3.068168	29.72688	36.09604	1.017971	1
CYCLOHEXANOL	428.6342	9.52625	9.839024	1.257546	1.831696
ETHANOL	9.129828	13.42828	15.70898	1	1.03061
IODOBENZENE	8319.385	0.957857	0.929615	2.270389	4.725659
MORPHOLINE	8.486411	1.629723	1.955081	0.467688	0.402831
CHLOROFORM	313.3233	0.863334	0.795296	0.184994	0.408778
2-CHLOROBUTANE	4244.71	1.083709	1.017008	2.891984	5.831206
2-ETHOXYETHANOL	80.15009	12.98919	15.81315	0.969956	1.361784
2-BROMOPROPANE	1326.125	1.014375	0.998383	2.472567	4.465382
BENZONITRILE	596.9878	1.263966	1.502918	2.397146	3.663027
ANTHRACENE	536203.1	1.05441	0.961368	4.149903	12.34225
M-CRESOL	86.54336	2.908726	3.314322	0.047316	0.118839
ANILINE	72.91311	1.751454	2.174237	0.109579	0.218697
CYCLOHEXANONE	96.35549	0.757274	0.842338	1.31383	1.69736

ACETALDEHYDE	6.825299	1.158496	1.425218	1.636963	1.67869
BENZALDEHYDE	434.4909	1.137095	1.331367	2.283804	3.363647
ACENAPHTHENE	164042.2	1.144738	1.029774	4.383909	11.98834
PHENOL	17.98833	2.9975	3.554209	0.03095	0.069694
M-XYLENE	21057.34	1.090546	1.016146	3.515932	8.088803
2-METHOXYETHANOL	20.32428	15.06353	19.21088	0.955238	1.166887
METHYLCYCLOPENTANE	43985.54	2.147471	1.661206	5.420284	13.57042
CYCLOHEXANE	36110.42	2.199196	1.699449	5.382837	13.243
ACETAMIDE	1.968434	815.2329	1218.74	1.487955	1.062032
1-CHLOROBUTANE	5695.793	1.092603	1.019959	2.950836	6.101385
1-BROMOHEPTANE	746649.9	1.401003	1.160729	5.798153	18.39757
2-PHENYLETHANOL	1044.872	6.176531	7.283265	0.957373	1.727539
DIBROMOMETHANE	182.2753	0.853986	0.873482	0.441019	0.771225
PHENANTHRENE	401476.4	1.030535	0.950743	3.876605	11.22711
PIPERIDINE	158.414	1.599406	1.527754	0.47179	0.572053
PYRIDINE	35.5597	1.132025	1.305483	0.884268	0.922824
HYDRAZINE	0.034717	31.08734	47.28629	0.225602	0.075435
BROMOBENZENE	5003.623	0.959659	0.936149	2.178889	4.347208
N-PENTANE	39791.7	2.12482	1.648148	5.31261	13.18544
O-XYLENE	15565.94	1.061277	1.006018	3.350859	7.478149
3-METHYLHEPTANE	2455148	2.834881	1.993342	10.13725	35.83899
ETHANE	404.6412	1.543673	1.334523	2.575437	4.308259
DICHLOROMETHANE	96.67328	0.858001	0.884986	0.435087	0.714049
STYRENE	6558.368	0.99514	1.018289	2.728007	5.47113
SEC-BUTYLBENZENE	219925.6	1.201203	1.059403	4.72396	13.35115
PROPANE	1792.976	1.678297	1.405363	3.221053	6.127064
HEXACHLOROETHANE	204436	1.527744	1.213086	4.326106	12.2774
1-NITROPROPANE	382.3169	1.175002	1.363469	2.255589	3.444005
1-BROMOBUTANE	7556.534	1.09394	1.019452	3.023963	6.401466
FLUOROBENZENE	1622.244	0.961736	0.960846	1.98872	3.59366
1-ETHYLNAPHTHALENE	197526	1.052595	0.995706	4.245522	11.59623
TETRAHYDROFURAN	57.02092	0.818083	0.84287	1.045901	1.282599
CYCLOPENTANONE	33.86651	0.799311	0.928438	1.262986	1.471186
HEXACHLOROBENZENE	7492273	2.067729	1.50015	8.42132	32.56266
1-BROMOPROPANE	1703.173	1.034672	1.001359	2.492793	4.617258
ISOBUTYLBENZENE	274990.8	1.243559	1.079586	4.899099	14.14769
AMMONIA	0.035364	6.879251	9.123908	0.114445	0.048416
NITROMETHANE	30.56728	2.504135	3.226734	2.040181	2.397103
FORMALDEHYDE	6.761536	1.595174	1.942322	2.140616	2.208623
4-METHYLPYRIDINE	98.57728	0.98488	1.094641	0.817215	0.935092
GLYCEROL	3.514602	218.6327	355.2905	1.512473	1.263066
TERT-BUTYLBENZENE	139255.5	1.161734	1.041633	4.409316	11.9489
QUINOLINE	972.5715	1.080042	1.20573	1.375235	1.975477
SULFOLANE	34.83991	1.450592	2.140087	3.159531	3.481822
N-BUTANE	8366.406	1.885813	1.520299	4.127596	8.960831
OCTAFLUOROCYCLOBUTANE	108346.3	1.962956	1.489617	4.495705	12.18102
2-METHYLTHIOPHENE	2267.019	1.005446	1.007084	2.390142	4.465463
2-METHYLHEXANE	613698.5	2.539177	1.849543	8.103687	25.442
QUINONE	143.3082	1.739593	2.321355	2.783526	3.623203
DIIODOMETHANE	612.2563	0.851717	0.854786	0.602674	1.136215
CYCLOPENTANE	10647.03	1.944532	1.555703	4.333652	9.604323
CHLOROBENZENE	4031.099	0.967464	0.938451	2.148591	4.223395

ETHYLBENZENE	17458.51	1.073951	1.010135	3.411851	7.709983
N-METHYLACETAMIDE	3.462217	30.44725	40.68785	0.84572	0.61492

The support vector, random forest, and gradient boosting decision tree regression methods have been applied to predict the enthalpy of hydration of organic materials with low molecular weight that are randomly selected with the automatic modules of Python scikit learn. The performance of studied ML methods in the predicting of the enthalpy of hydration is discussed for the studied methods with various train and test subsets ratios. The results of the applied ML methods have been evaluated with a different train and test ratio with 0.05:0.95, 0.10:0.90, 0.15:0.85, 0.20:0.80 and 0.25:0.75 to investigate the effect of the data training on the accuracy of the prediction.

The corresponding mean absolute error and mean square errors of the ML methods with different training and testing rates is summarized in Table 2 for the SVR to compare the kernels performance. Evidently, the reduced training rate led to increase the MAE and MSE values for the ML methods as demonstrated in Table 2. However, the results are not good enough that might be due to the origin of the SVR that comes from elastic net regression that is a type of linear regression.

Table 2. Mean absolute error, mean square error, and root mean square error in prediction of enthalpy of hydration of organic materials using SVR with different kernels under 0.1 MPa at 298.15 K.

Train: Test ratio	MAE	MSE	RMSE
SVR-Polynomial			
0.05: 0.95	3.11	15.94	3.99
0.10: 0.90	3.09	15.23	3.90
0.15: 0.85	4.04	25.40	5.04
0.20: 0.80	4.18	23.17	4.81
0.25: 0.75	8.43	72.04	26.88
SVR-Gaussian			
0.05: 0.95	3.12	15.98	3.99
0.10: 0.90	3.09	15.23	3.90
0.15: 0.85	4.15	27.27	5.22
0.20: 0.80	4.23	24.10	4.91
0.25: 0.75	3.01	13.91	3.73
SVR-Sigmoid			
0.05: 0.95	3.09	15.74	3.97
0.10: 0.90	3.05	15.04	3.88
0.15: 0.85	3.93	25.57	5.06
0.20: 0.80	4.18	23.48	4.85
0.25: 0.75	3.26	14.97	3.87

Overfitting and bias should be resolved in any regression problem. Avoiding bias and overfitting could be vanquished using another ML method named support vector regression (SVR) that is developed with an evolutionary process starting from linear regression, lasso, ridge, elastic net, and SVR. Also, the SVR includes different kernels that could be used to find the data distribution type and could be used to find the importance of the features in the regression [43]. In the SVR method, the degree of bias towards a particular result is much less than other methods. Since the degree of bias is low in this method, it has been used as reliability to interpret the observed linear relationship. In this respect, the SVR kernels including polynomial, Gaussian, and sigmoid have been used. The linear kernel shows the lowest regression error rate, indicating that linear relationships that were previously detected are still exist after all processing.

According to the results of the SVR machine learning, some other machine learning methods should be used to overcome these problems. The gradient boosting decision tree regression (GBDTR) is an effective method that is comparable to the random forest regression (RFR) [44–46]. The results for predicting enthalpy of hydration of low molecular weight molecules using RFR and GBDTR ML methods are given in Fig 2 for the train and test ratio of 0.80:0.20. Both methods have been able to accurately predict the enthalpy of hydration of the test data.

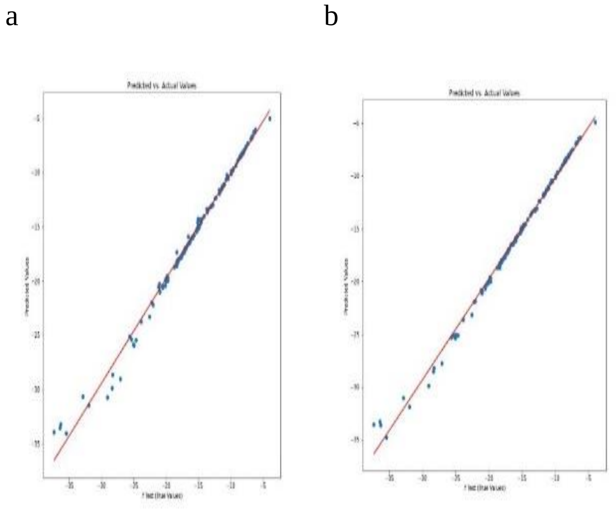

Fig 2. The scattering plot of predicted Enthalpy of hydration values of tested molecules versus the FreeSolv dataset values with a train and test ratio of 0.80:0.20: a) RFR, b) GBDTR.

Also, the corresponding deviations of the RFR and GBDTR have been given in Tables 3 and 4, respectively. As could be seen the results does not have significant differences while the RFR shows higher accuracy rather than the GBDTR. This could be related to the lower depth of decision trees in the GBDTR rather than the RFR. It means, the prediction of the enthalpy of hydration needs decision trees with higher depth with infinite dilution activity coefficient. However, the GBDTR is a faster method rather than the RFR due to its practical use in the higher data training rate. Accordingly, both RFR and GBDTR could be used to predict the thermodynamic properties with higher degree of complexity between the label and existing thermodynamic features.

Table 3. Mean absolute error, mean square error, and root mean square error in prediction of enthalpy of hydration of organic materials using RFR under 0.1 MPa at 298.15 K.

Train: Test ratio	MAE	MSE	RMSE
0.05: 0.95	0.61	0.44	0.66
0.10: 0.90	0.83	1.06	1.26
0.15: 0.85	0.87	1.20	1.32
0.20: 0.80	0.89	1.35	1.46
0.25: 0.75	0.93	1.50	1.52

Table 4. Mean absolute error, mean square error, and root mean square error in prediction of enthalpy of hydration of organic materials using GBDTR under 0.1 MPa at 298.15 K.

Train: Test ratio	MAE	MSE	RMSE
0.05: 0.95	0.71	0.62	0.79
0.10: 0.90	0.83	1.21	1.10
0.15: 0.85	0.90	1.69	1.30
0.20: 0.80	0.96	2.43	1.56
0.25: 0.75	1.22	2.58	1.60

4. Conclusion

Different machine learning methods have been utilized to predict the enthalpy of hydration of low molecular weight organic molecules that were common between the FreeSolv open-source dataset and VT2005 σ-profiles dataset. Since there is no linear relationship between the activity coefficients and enthalpy of hydration, machine learning approach has been used to predict the enthalpy of hydration of the low molecular weight organic molecules using infinite dilution activity coefficient evaluated from COSMO-SAC model. The SVR, RFR, and GBDTR machine learning methods used to predict of enthalpy of hydration. However, the RFR and GBDTR have more accuracy in the prediction of enthalpy of hydration rather than the SVR. This might be related to the bias in SVR method and corresponding overfitting or underfitting problems.

Acknowledgements

Hereby, we extend our gratitude to

Data and Software Availability

The datasets analyzed during the current study are available in the [MobleyLab/FreeSolv] repository, [https://github.com/MobleyLab/FreeSolv].

The free benchmark implementation of COSMO-SAC model usnistgov/COSMOSAC repository, [https://github.com/usnistgov/COSMOSAC].

References

[1] M. S. Mat Nor, Z. A. Manan, A. A. Mustaffa, and C. L. Suan, “An Evaluation of Thermodynamic Models for the Prediction of Solubility of Phytochemicals from Orthosiphon Staminues in Ethanol,” in *Computer Aided Chemical Engineering*, K. V. Gernaey, J. K. Huusom, and R. Gani, Eds., in 12 International Symposium on Process Systems Engineering and 25 European Symposium on Computer Aided Process Engineering, vol. 37. Elsevier, 2015, pp. 2087–2092. doi: 10.1016/B978-0-444-63576-1.50042-X.

[2] N. Chorbngam, R. Chawuthai, and A. Anantpinijwatna, “Novel method for properties prediction of pure organic compounds using machine learning,” in *Computer Aided Chemical Engineering*, M. Türkay and R. Gani, Eds., in 31 European Symposium on Computer Aided Process Engineering, vol. 50. Elsevier, 2021, pp. 431–437. doi: 10.1016/B978-0-323-88506-5.50068-1.

[3] O. Olatunji, S. Akinlabi, and N. Madushele, “Application of Artificial Intelligence in the Prediction of Thermal Properties of Biomass,” in *Valorization of Biomass to Value-Added Commodities: Current Trends, Challenges, and Future Prospects*, M. O. Daramola and A. O. Ayeni, Eds., in Green Energy and Technology. Cham: Springer International Publishing, 2020, pp. 59–91. doi: 10.1007/978-3-030-38032-8_4.

[4] G. J. Maximo, N. D. D. Carareto, and M. C. Costa, “Chapter 8 - Solid–Liquid Equilibrium in Food Processes,” in *Thermodynamics of Phase Equilibria in Food Engineering*, C. G. Pereira, Ed., Academic Press, 2019, pp. 335–384. doi: 10.1016/B978-0-12-811556-5.00008-9.

[5] I. Amaya, C. Jiménez, and R. Correa, “Phase Equilibrium Description of a Supercritical Extraction System Using Metaheuristic Optimization Algorithms,” in *Bioinspired Heuristics for Optimization*, E.-G. Talbi and A. Nakib, Eds., in Studies in Computational Intelligence. Cham: Springer International Publishing, 2019, pp. 43–60. doi: 10.1007/978-3-319-95104-1_3.

[6] J. L. de Medeiros and O. de Q. F. Araújo, “Thermodynamic Modeling of CO2-Rich Natural Gas Fluid Systems,” in *Offshore Processing of CO2-Rich Natural Gas with Supersonic Separator: Multiphase Sound Speed, CO2 Freeze-Out and HYSYS Implementation*, J. L. de Medeiros, L. de Oliveira Arinelli, A. M. Teixeira, and O. de Q. F. Araújo, Eds., Cham: Springer International Publishing, 2019, pp. 55–96. doi: 10.1007/978-3-030-04006-2_4.

[7] T.-C. Liu and S.-T. Lin, “Exact Local Composition Model for Two-Dimensional Lattice Fluids,” *Ind. Eng. Chem. Res.*, vol. 58, no. 45, pp. 20779–20787, Nov. 2019, doi: 10.1021/acs.iecr.9b03218.

[8] T.-C. Liu and S.-T. Lin, “A new approach for developing exact local composition models for lattice fluids,” *Journal of the Taiwan Institute of Chemical Engineers*, vol. 96, pp. 63–73, Mar. 2019, doi: 10.1016/j.jtice.2018.11.023.

[9] G. Salgueiro, M. de Moraes, F. Pessoa, R. Cavalcante, and A. Young, “New volume translation functions for biodiesel density prediction with the Peng-Robinson Equation of state in terms of its raw materials,” *Fuel*,

vol. 293, p. 120254, Jun. 2021, doi: 10.1016/j.fuel.2021.120254.

[10] J. P. Hernández, L. A. Forero, and J. A. Velásquez, "Modelling low pressure LLE and VLE of methanol/alkane mixtures with a modified Peng-Robinson EoS and the Huron-Vidal mixing rules," *Fluid Phase Equilibria*, vol. 546, p. 113123, Oct. 2021, doi: 10.1016/j.fluid.2021.113123.

[11] I. Polishuk, A. Chiko, E. Cea-Klapp, and J. M. Garrido, "Implementation of CP-PC-SAFT and CS-SAFT-VR-Mie for Predicting Thermodynamic Properties of C1–C3 Halocarbon Systems. I. Pure Compounds and Mixtures with Nonassociating Compounds," *Ind. Eng. Chem. Res.*, vol. 60, no. 26, pp. 9624–9636, Jul. 2021, doi: 10.1021/acs.iecr.1c01700.

[12] M. Ascani and C. Held, "Prediction of salting-out in liquid-liquid two-phase systems with ePC-SAFT: Effect of the Born term and of a concentration-dependent dielectric constant," *Zeitschrift für anorganische und allgemeine Chemie*, vol. 647, no. 12, pp. 1305–1314, 2021, doi: 10.1002/zaac.202100032.

[13] I. Polishuk and J. M. Garrido, "Comparison of SAFT-VR-Mie and CP-PC-SAFT in predicting phase behavior of associating systems IV. Methanol–aliphatic hydrocarbons," *Journal of Molecular Liquids*, vol. 291, p. 111321, Oct. 2019, doi: 10.1016/j.molliq.2019.111321.

[14] W. Hu, Z. Shang, N. Wei, B. Hou, J. Gong, and Y. Wang, "Solubility of benorilate in twelve monosolvents: Determination, correlation and COSMO-RS analysis," *The Journal of Chemical Thermodynamics*, vol. 152, p. 106272, Jan. 2021, doi: 10.1016/j.jct.2020.106272.

[15] S. Balchandani and R. Singh, "Thermodynamic analysis using COSMO-RS studies of reversible ionic liquid 3-aminopropyl triethoxysilane blended with amine activators for CO2 absorption," *Journal of Molecular Liquids*, vol. 324, p. 114713, Feb. 2021, doi: 10.1016/j.molliq.2020.114713.

[16] R. Xiong, S. I. Sandler, and R. I. Burnett, "An Improvement to COSMO-SAC for Predicting Thermodynamic Properties," *Ind. Eng. Chem. Res.*, vol. 53, no. 19, pp. 8265–8278, May 2014, doi: 10.1021/ie404410v.

[17] S. Wang, S. I. Sandler, and C.-C. Chen, "Refinement of COSMO−SAC and the Applications," *Ind. Eng. Chem. Res.*, vol. 46, no. 22, pp. 7275–7288, Oct. 2007, doi: 10.1021/ie070465z.

[18] R. Fingerhut *et al.*, "Comprehensive Assessment of COSMO-SAC Models for Predictions of Fluid-Phase Equilibria," *Ind. Eng. Chem. Res.*, vol. 56, no. 35, pp. 9868–9884, Sep. 2017, doi: 10.1021/acs.iecr.7b01360.

[19] A. De Visscher, J. Vanderdeelen, E. Königsberger, B. R. Churagulov, M. Ichikuni, and M. Tsurumi, "IUPAC-NIST Solubility Data Series. 95. Alkaline Earth Carbonates in Aqueous Systems. Part 1. Introduction, Be and Mg," *Journal of Physical and Chemical Reference Data*, vol. 41, no. 1, pp. 013105-013105–67, Mar. 2012, doi: 10.1063/1.3675992.

[20] "ThermoLit: NIST Literature Report Builder for Thermochemical Property Measurements." https://trc.nist.gov/thermolit/main/home.html#home (accessed Feb. 09, 2022).

[21] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, and S. Grimme, "A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions," *Phys. Chem. Chem. Phys.*, vol. 19, no. 48, pp. 32184–32215, Dec. 2017, doi: 10.1039/C7CP04913G.

[22] R. Peverati and D. G. Truhlar, "Quest for a universal density functional: the accuracy of density functionals across a broad spectrum of databases in chemistry and physics," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 372, no. 2011, p. 20120476, Mar. 2014, doi: 10.1098/rsta.2012.0476.

[23] P. Morgante and R. Peverati, "ACCDB: A collection of chemistry databases for broad computational purposes," *Journal of Computational Chemistry*, vol. 40, no. 6, pp. 839–848, 2019, doi: 10.1002/jcc.25761.

[24] D. L. Mobley and J. P. Guthrie, "FreeSolv: a database of experimental and calculated hydration free energies, with input files," *J Comput Aided Mol Des*, vol. 28, no. 7, pp. 711–720, Jul. 2014, doi: 10.1007/s10822-014-9747-x.

[25] S. Abaimov and M. Martellini, "Understanding Machine Learning," in *Machine Learning for Cyber Agents: Attack and Defence*, S. Abaimov and M. Martellini, Eds., in Advanced Sciences and Technologies for Security Applications. Cham: Springer International Publishing, 2022, pp. 15–89. doi: 10.1007/978-3-030-91585-8_2.

[26] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016, doi: 10.1109/COMST.2015.2494502.

[27] A. Malloum, J. J. Fifen, and J. Conradie, "Determination of the absolute solvation free energy and enthalpy of the proton in solutions," *Journal of Molecular Liquids*, vol. 322, p. 114919, Jan. 2021, doi: 10.1016/j.molliq.2020.114919.

[28] J. Wang *et al.*, "Hydration Energetics of a Diamine-Appended Metal–Organic Framework Carbon Capture Sorbent," *J. Phys. Chem. C*, vol. 124, no. 1, pp. 398–403, Jan. 2020, doi: 10.1021/acs.jpcc.9b08008.

[29] S. A. Potekhin, "High-Pressure Scanning Microcalorimetry – A New Method for Studying Conformational and Phase Transitions," *Biochemistry Moscow*, vol. 83, no. 1, pp. S134–S145, Jan. 2018, doi: 10.1134/S0006297918140110.

[30] M. Bonto, H. M. Nick, and A. A. Eftekhari, "Thermodynamic Analysis of the Temperature Effect on Calcite Surface Reactions in Aqueous Environments," *Energy Fuels*, vol. 35, no. 20, pp. 16677–16692, Oct. 2021, doi: 10.1021/acs.energyfuels.1c01652.

[31] D. Zheng and F. Wang, "Performing Molecular Dynamics Simulations and Computing Hydration Free Energies on the B3LYP-D3(BJ) Potential Energy Surface with Adaptive Force Matching: A Benchmark Study with Seven Alcohols and One Amine," *ACS*

Phys. Chem Au, vol. 1, no. 1, pp. 14–24, Nov. 2021, doi: 10.1021/acsphyschemau.1c00006.

[32] T. R. Rogers and F. Wang, “Accurate MP2-based force fields predict hydration free energies for simple alkanes and alcohols in good agreement with experiments,” *J. Chem. Phys.*, vol. 153, no. 24, p. 244505, Dec. 2020, doi: 10.1063/5.0035032.

[33] P. Sahu, S. Krishnaswamy, K. Ponnani, and N. K. Pande, “A thermodynamic approach to selection of suitable hydrate formers for seawater desalination,” *Desalination*, vol. 436, pp. 144–151, Jun. 2018, doi: 10.1016/j.desal.2018.02.001.

[34] I. H. Bell *et al.*, “A Benchmark Open-Source Implementation of COSMO-SAC,” *J. Chem. Theory Comput.*, vol. 16, no. 4, pp. 2635–2646, Apr. 2020, doi: 10.1021/acs.jctc.9b01016.

[35] *COSMO-SAC*. National Institute of Standards and Technology, 2022. Accessed: Jul. 01, 2022. [Online]. Available: https://github.com/usnistgov/COSMOSAC

[36] M. Awad and R. Khanna, “Support Vector Regression,” in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, M. Awad and R. Khanna, Eds., Berkeley, CA: Apress, 2015, pp. 67–80. doi: 10.1007/978-1-4302-5990-9_4.

[37] J. VanderPlas, *Python data science handbook: Essential tools for working with data*. O’Reilly Media, Inc., 2016.

[38] R. Ramanathan, M. Mathirajan, and A. R. Ravindran, Eds., *Big Data Analytics Using Multiple Criteria Decision-Making Models*. Boca Raton: CRC Press, 2017. doi: 10.1201/9781315152653.

[39] T.-C. Liu and S.-T. Lin, “A new approach for developing exact local composition models for lattice fluids,” *Journal of the Taiwan Institute of Chemical Engineers*, vol. 96, pp. 63–73, Mar. 2019, doi: 10.1016/j.jtice.2018.11.023.

[40] S. Balchandani and R. Singh, “Thermodynamic analysis using COSMO-RS studies of reversible ionic liquid 3-aminopropyl triethoxysilane blended with amine activators for CO2 absorption,” *Journal of Molecular Liquids*, vol. 324, p. 114713, Feb. 2021, doi: 10.1016/j.molliq.2020.114713.

[41] W. Hu, Z. Shang, N. Wei, B. Hou, J. Gong, and Y. Wang, “Solubility of benorilate in twelve monosolvents: Determination, correlation and COSMO-RS analysis,” *The Journal of Chemical Thermodynamics*, vol. 152, p. 106272, Jan. 2021, doi: 10.1016/j.jct.2020.106272.

[42] M. R. Shah and G. D. Yadav, “Prediction of Liquid–Liquid Equilibria for Biofuel Applications by Quantum Chemical Calculations Using the Cosmo-SAC Method,” *Ind. Eng. Chem. Res.*, vol. 50, no. 23, pp. 13066–13075, Dec. 2011, doi: 10.1021/ie201454m.