CHEMICAL ENGINEERING TRANSACTIONS
VOL. 117, 2025
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Fabrizio Bezzo, Flavio Manenti, Gabriele Pannocchia, Almerinda di Benedetto
Copyright © 2025, AIDIC Servizi S.r.l.
ISBN 979-12-81206-17-5; **ISSN** 2283-9216
DOI: 10.3303/CET25117170

Integration of Genetic Algorithm with Machine Learning for Properties Prediction

Rathachai Chawuthai[a], Siripan Murathathunyaluk[b], Nalin Amornratthamrong[b], Run Arunchaipong[b], Amata Anantpinijwatna[b*]

[a]Department of Computer Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand
[b]Department of Chemical Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand
amata.an@kmitl.ac.th

Numerous studies have demonstrated that machine learning (ML) provides more accurate estimations of properties for oxygenated organic derivatives compared to the conventional Quantitative Structure-Property Relationship (QSPR) method. Consequently, ML's predictive capabilities have been extended to encompass a broader range of properties, including Partition Coefficient, Boiling Point, and Solubility, among others, for oxygenated hydrocarbon derivatives. Algorithms such as Linear Regression, Support Vector Machine, Random Forest, and Gaussian Process are selected through trial-and-error to identify the most suitable approach. The models are trained and validated using experimental data from published literature. Despite the accuracy of these property predictions, they have limited practical utility in industry, where specific property ranges are essential for processes. To address this, Genetic Algorithms (GA) are employed to design chemical compounds that meet industrial requirements. Integrating GA with ML could yield alternative chemical compounds, enhancing overall production processes by increasing economic potential, sustainability, and reducing environmental impact.

1. Introduction

Hydrocarbons, organic compounds composed solely of carbon and hydrogen atoms, are categorized into alkanes, alkenes, and alkynes. These compounds, with various carbon chain lengths, exhibit diverse properties influenced by their molecular structure (Wilkes and Schwarzbauer, 2010). Each type of hydrocarbon structure imparts distinct properties, such as boiling temperature and solubility parameter (Wilkes, 2010). Hydrocarbon compounds are widely utilised in industry. Currently, the selection of substances with complex structures or new substances is tailored for specific processes (Hall et al., 2003). However, optimal substance selection necessitates experimental trials to determine the most suitable compound, as certain properties of some substances remain unknown without experimentation. These trials incur financial and temporal costs.
Traditional methods of assessing the properties of these hydrocarbons often lack sufficient accuracy, rendering them unsuitable for reliable references or practical application (Dablander et al., 2023; Muratov et al., 2020). Consequently, experiments are required to verify these properties, but they are time-consuming and costly. Thus, models have been developed to predict these properties. For instance, research by Oprisiu et al. (2013) employed Quantitative Structure Models to predict boiling temperatures and utilised non-linear models such as Associative Neural Networks (ASNNs) along with other linear regression models. Boiling point data focused on organic compounds containing carbon atoms greater than two and other atoms including C, H, F, Cl, Br, N, and O. Additionally, Chorbngam et al. (2021) studied and developed models to predict boiling temperatures, while Thakaew and Sribut (2021) investigated the water solubility properties of hydrocarbons and developed models to predict solubility.In developing models to predict substance properties, algorithm selection and hyperparameter tuning, along with the training dataset, significantly impact accuracy. Research by Chorbngam et al. (2021), Saengsuradech and Nukaew (2021), and Thakaew and Sribut (2021) enhanced prediction

Paper Received: 31 December 2024; Revised: 14 April 2025; Accepted: 24 April 2025
Please cite this article as: Chawuthai R., Murathathunyaluk S., Amornratthamrong N., Arunchaipong R., Anantpinijwatna A., 2025, Integration of Genetic Algorithm with Machine Learning for Properties Prediction, Chemical Engineering Transactions, 117, 1015-1020 DOI:10.3303/CET25117170

efficiency and model application in industry. In general industry practices, property ranges of substances are chosen to fit processes. However, the aforementioned models predict properties based on structure. For real-world applications, new models need to be constructed by defining the prediction target as the substance structure (SMILES). This research aims to enhance predictive models using machine learning by conducting cross-validation to compare model performance and applying Genetic Algorithms (Singh et al., 2010). Genetic Algorithms, techniques for finding desired values or problem solutions based on biological principles or natural selection (Schiano di Visconte et al., 2019), are employed to identify appropriate property ranges.

2. Methodology

In this study, the dataset from the work of Chorbngam et al. (2021) was utilised, comprising boiling point data for 560 hydrocarbons containing 1-12 carbon atoms. The input features employed were consistent with those utilised in previous research, including the number of carbon atoms ('C'), double bonds ('='), triple bonds ('#'), branches ('('), and cyclic structures ('1'). Supervised learning was conducted using K-nearest Neighbour (KNN), Decision Tree (DT), and Random Forest (RF) algorithms, with performance enhanced through the application of the cross-validation technique. A Genetic Algorithm was subsequently applied to perform the backward estimation of a compound's structure based on the target property.

2.1 Model Development and Cross Validation

The cross-validation technique was employed to identify the most suitable algorithm for prediction. Cross-validation, as described by Refaeilzadeh et al. (2009), is a method used to estimate a model's performance on unseen data. This approach involves partitioning the dataset into multiple segments, training the model on a subset of these segments, and subsequently testing it on the remaining segment, as illustrated in Figure 1. The procedure is repeated with varying subsets, and the results are averaged to obtain a robust measure of model performance. The primary objective is to mitigate overfitting, where a model demonstrates strong performance on training data but performs poorly on new data.
For this study, the dataset was initially divided into a training set and a test set, with 460 out of 560 data points randomly selected for training. Within the training set, the data were further partitioned into five segments to facilitate algorithm tuning and evaluation. The remaining 100 data points constituted the test set and were used to estimate the performance of the final model. The constructed model was subsequently employed in the Genetic Algorithm.

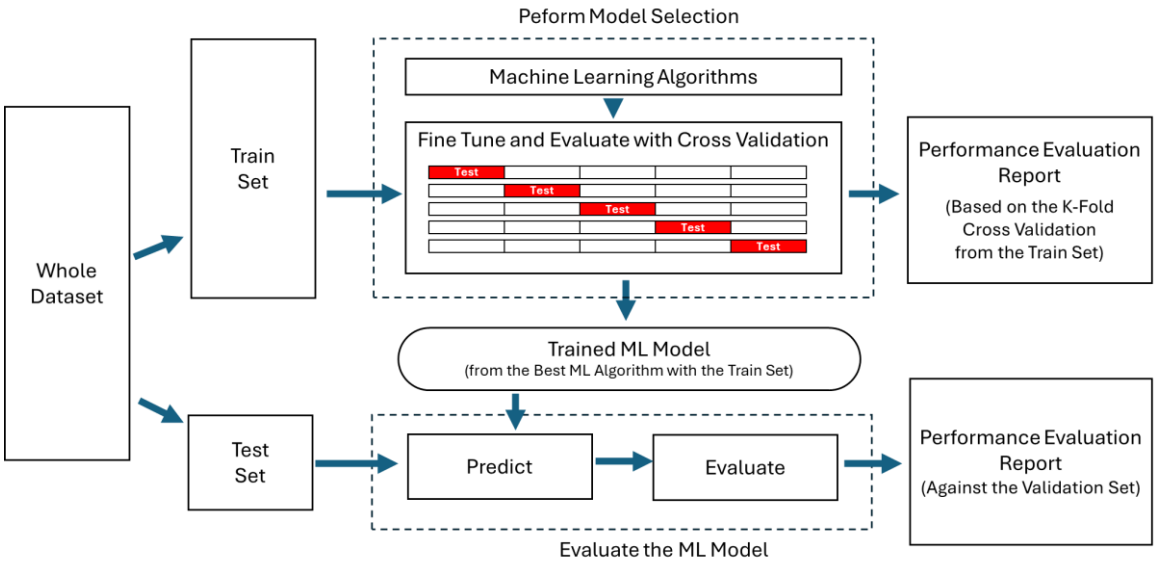

Figure 1: Cross-Validation

2.2 Genetic Algorithm

Genetic algorithm (GA), a subset of evolutionary algorithms (EAs), are adaptive search methods inspired by natural selection, widely used for optimization and search problems. These algorithms evolve a population of candidate solutions (individuals) through operators such as selection, crossover, and mutation (Holland, 1992).

The process begins with randomly generated individuals, whose fitness is evaluated using an objective function. The fittest individuals are selected to reproduce, forming successive generations. This iterative process continues until a satisfactory fitness level is achieved or a predefined number of generations is reached (Goldberg, 1989), as illustrated in Figure 2.

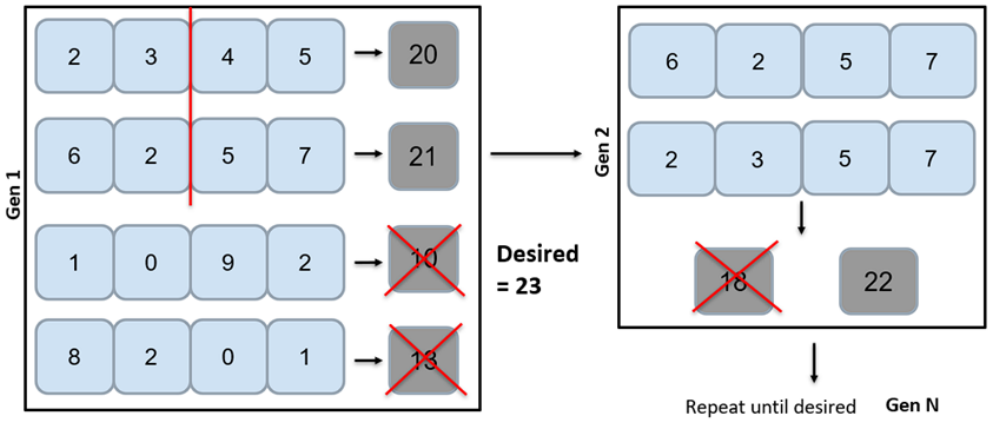

Figure 2: Genetic Algorithm

The application of the Genetic Algorithm to generate chemicals with desirable properties, as illustrated in Figure 3, comprises five steps: Initialization, Population, Selection, Crossover, and Mutation.

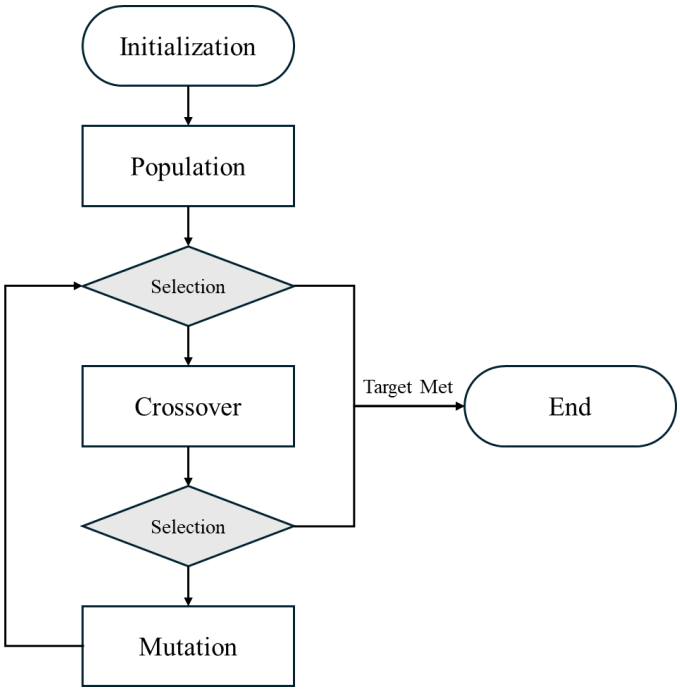

Figure 3: Genetic Algorithm Methodology for Chemical Structure Generation

The process begins with initialization, where the target property value is specified. During the population step, 60 chemical compounds with varied structures are randomly generated. In the subsequent selection step, 20 compounds with properties closest to the target are screened. If the target property is achieved during this step, the process terminates. If not, the procedure advances to the next phase.

In the crossover step, the structures of the selected compounds are combined with each other. While the number of carbon atoms remains fixed, other features are subject to crossover. Duplicate compounds are eliminated before the selection step is repeated. Depending on the outcomes of the selection step, the screened

compounds proceed to the mutation step, where only a single feature is allowed to change. Duplicates are again removed before returning to the selection step.
This iterative process continues until a compound with the desired property is identified, with the number of iterations capped at 100 to manage computational time.

2.3 Data Analysis

The performance of the predicted boiling point in this method was evaluated using root mean square error (RMSE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2), calculated using Equations (1)-(3), respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(T_i - T_i\right)^2} \tag{1}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{T_i - T_i}{T_i}\right| \times 100 \tag{2}$$

$$R^2 = 1 - \frac{\sum\left(T_i - T_i\right)^2}{\sum\left(T_i - \overline{T}\right)^2} \tag{3}$$

where T_i represents the actual boiling point, T_i denotes the predicted boiling point, and $\overline{T}$ indicates the average boiling point across the dataset.

3. Results and Discussions

3.1 Model Refinement with Cross-Validation

As stated in previous research (Chorbngam et al., 2021), a comparison of results with conventional group-contribution methods, such as the works of Marrero and Gani (2001), Constantinou and Gani (1994), and Joback and Reid (1987), revealed significantly lower errors in terms of both RMSE and MAPE for all chemical groups when the earlier method was applied without cross-validation.
The updated results from the refined model were compared with those from the earlier work, with a graphical comparison presented in Figure 4. Although the changes appear subtle, the predictions generated by the refined model align more closely with the 1:1 line, indicating slightly improved accuracy. Furthermore, Table 1 demonstrates that the incorporation of the cross-validation technique marginally reduces overall error. While cross-validation increases errors in terms of both MAPE and RMSE for the train set, it reduces errors for the test (unseen) set, enhancing the model's generalization performance.

Table 1: Comparison of the boiling points prediction results between the previous work and this work

Metrics	Previous Work			This Work (with Cross-Validation)		
	Train Set	Test Set	Total	Train Set	Test Set	Total
MAPE	1.647	1.942	1.700	1.671	1.722	1.680
RMSE	7.354	9.924	7.813	7.729	8.001	7.778
R^2	0.987	0.920	0.975	0.983	0.970	0.981

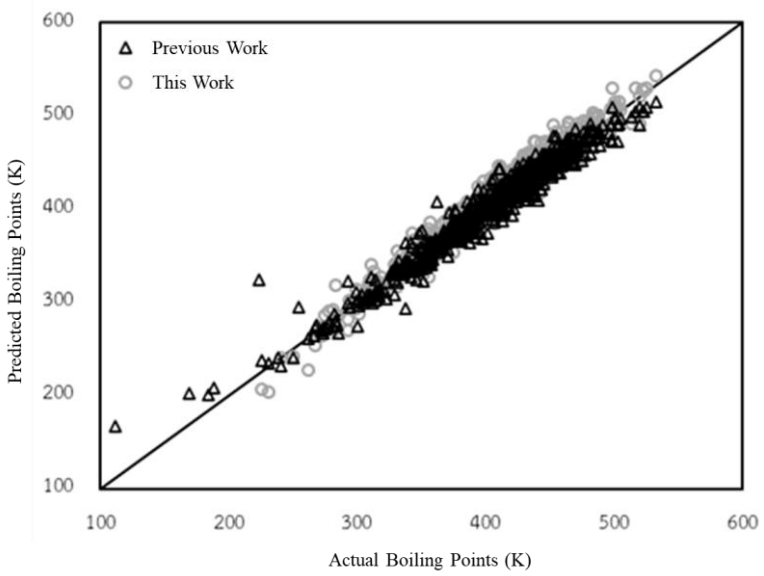

Figure 4: Plot of the predicted boiling points versus the actual boiling points

3.2 Compound Prediction with Genetic Algorithm

In this step, the predictive model developed in the previous section was employed to estimate the boiling points of chemical compounds generated through random population, crossover, and mutation. Six cases were investigated to generate chemical compounds with target boiling points of 200–205 K, 250–255 K, 300–305 K, 350–355 K, 400–405 K, and 450–455 K. Table 2 presents the results, including the target boiling temperature, the SMILES structure of the generated compounds, the predicted boiling temperature, the number of iterations utilized, and the deviation from the target values.

For the target boiling point ranges of 200–205 K and 250–255 K, the algorithm was unable to generate chemical structures within the specified range, reaching the maximum of 100 iterations and providing structures that were as close to the target as possible. Conversely, for the remaining four cases, the algorithm identified SMILES structures matching the target boiling points with only a minimal number of iterations. Upon detailed examination, the generated compound structures with boiling points between 300–350 K were identified as either 1-butyne or 2-butyne, hydrocarbons containing four carbon atoms and a single triple bond. Among these, 2-butyne exhibited a boiling point of 300 K.

The lack of success in the first two cases is attributed to the absence of feasible hydrocarbon structures within the given target range, rendering it impossible to generate a structure that satisfies the specified criteria, regardless of the number of iterations performed. Conversely, for the cases with feasible solutions, the genetic algorithm identified the desired structures with notable efficiency.

Table 2: Chemical Structure Generated using Genetic Algorithm

Targets (K)	Results (K)	Structure					Iteration	Error (%)
		C	=	#	(	1		
200-205	188.45	2	0	0	0	0	100	5.78
250-255	249.95	3	0	1	0	0	100	0.02
300-305	300.05	4	0	1	0	0	3	0.00
350-355	351.05	7	1	0	3	0	4	0.00
400-405	403.58	8	2	0	0	1	5	0.00
450-455	453.15	12	1	0	5	0	2	0.00

4. Conclusions

The application of machine learning and genetic algorithms demonstrates significant potential in predicting and optimising the properties of chemical compounds. The integration of cross-validation enhances model robustness, reducing errors in the prediction of unseen data. For instance, the incorporation of cross-validation

resulted in a reduction of RMSE from 9.92 K to 8.00 K and MAPE from 1.94% to 1.72% for the test set, indicating improved performance in property estimation.

When combined with a Genetic Algorithm, this methodology successfully generated chemical structures meeting specified property targets. In the boiling point prediction task, six cases were evaluated, with targets ranging from 200 K to 455 K. For four feasible target ranges (300–305 K, 350–355 K, 400–405 K, and 450–455 K), the algorithm identified structures such as 1-butyne and 2-butyne, achieving exact matches to the desired boiling points within an average of 3.5 iterations. In contrast, for infeasible target ranges (200–205 K and 250–255 K), the algorithm reached the maximum of 100 iterations without producing a structure within the specified range, highlighting the limitations of the input space.

Overall, the approach demonstrates improved predictive accuracy and efficient property-targeted compound generation. Further refinements to the dataset and algorithm optimisation may expand the applicability to a wider range of chemical properties and enhance prediction accuracy.

References

Chorbngam, N., Chawuthai, R., Anantpinijwatna, A., 2021. Computer Aided Chemical Engineering 431–437.

Constantinou, L., Gani, R., 1994. AIChE Journal 40, 1697–1710.

Dablander, M., Hanser, T., Lambiotte, R., Morris, G.M., 2023. J Cheminform 15, 47.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley.

Hall, C., Tharakan, P., Hallock, J., Cleveland, C., Jefferson, M., 2003. Nature 426, 318–322.

Holland, J.H., 1992. Adaptation in Natural and Artificial Systems. The MIT Press.

Joback, K.G., Reid, R.C., 1987. Chem Eng Commun 57, 233–243.

Marrero, J., Gani, R., 2001. Fluid Phase Equilib 183–184, 183–208.

Muratov, E.N., Bajorath, J., Sheridan, R.P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T.I., Baskin, I.I., Varnek, A., Roitberg, A., Isayev, O., Curtalolo, S., Fourches, D., Cohen, Y., Aspuru-Guzik, A., Winkler, D.A., Agrafiotis, D., Cherkasov, A., Tropsha, A., 2020. Chem Soc Rev 49, 3525–3564.

Oprisiu, I., Marcou, G., Horvath, D., Brunel, D.B., Rivollet, F., Varnek, A., 2013. Thermochim Acta 553, 60–67.

Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-Validation, in: Encyclopedia of Database Systems. Springer US, Boston, MA, pp. 532–538.

Saengsuradech, S., Nukaew, A., 2021. Prediction of Osmotic and Activity Coefficients of Alkyl Ammonium in Aqueous Solution with Machine Learning.

Schiano di Visconte, G., Spicer, A., Chuck, C.J., Allen, M.J., 2019. Applied Sciences 9, 4793.

Singh, S.P., Anantharaj, R., Banerjee, T., 2010. UNIFAC Group Interaction Prediction for Ionic Liquid-Thiophene Based Systems Using Genetic Algorithm, in: Deb, K., Bhattacharya, A., Chakraborti, N., Chakroborty, P., Das, S., Dutta, J., Gupta, S.K., Jain, A., Aggarwal, V., Branke, J., Louis, S.J., Tan, K.C. (Eds.), Simulated Evolution and Learning, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 195–204.

Thakaew, C., Sribut, C., 2021. Solubilities Prediction of Oxygenated Organic Compounds Using Novel Statistical-based Method.

Wilkes, H., 2010. Methods of Hydrocarbon Analysis, in: Handbook of Hydrocarbon and Lipid Microbiology. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 49–65.

Wilkes, H., Schwarzbauer, J., 2010. Hydrocarbons: An Introduction to Structure, Physico-Chemical Properties and Natural Occurrence, in: Handbook of Hydrocarbon and Lipid Microbiology. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–48.