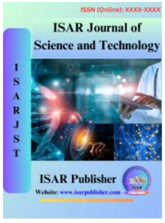
OPEN ACCESS



ISAR Journal of Science and Technology

Volume 3, Issue 3, 2025, Page: 13-22

Abbriviate Title- ISAR J Sci Tech

ISSN *(Online)*- 2584-2056

https://isarpublisher.com/journal/isarjst

Prediction of Melting and Boiling Points of Fatty Acids and Their Derivatives Using Quantitative Structure-Activity Relationship (QSAR) Methodology

Manar Amer Jebur[1*], Zahra Pahlavan Yali[2]

Faculty of Chemistry, Department of Analytical Chemistry, University of Mazandaran, Babolsar, Iran.

***Corresponding Author**

Manar Amer Jebur

Faculty of Chemistry, Department of Analytical Chemistry, University of Mazandaran, Babolsar, Iran.

Article History

Received: 11.12.2024
Accepted: 19.01.2025
Published: 10.03.2025

Abstract: Lipids contain fatty acids as fundamental components and their boiling and melting temperatures matter significantly for industrial uses. The researchers employed the Quantitative Structure-Property Relationship (QSPR) method for predicting melting and boiling points in fatty acids along with their derived substances. Measuring properties of chemical structures with molecular descriptors enabled the development of QSPR models that received validation by experimental results. During the training phase the predictive accuracy reached high levels as the coefficients of determination (R^2) values turned out to be 0.948 for melting points and 0.938 for boiling points. The cross-validation validation produced R^2 values at 0.925 for estimating melting points and also 0.925 for estimating boiling points which shows robust predictive capability. Due to their reliable nature QSPR models exhibit strong performance in predicting thermal characteristics of fatty acids for uses in biodiesel production as well as food processing and cosmetics industries.

Keywords: Fatty Acids, Quantitative Structure-Activity Relationship, Multiple linear regression, Multiple linear regression (MLR).

Cite this article:

Jebur, M. A., Yali, Z. P., (2025). Prediction of Melting and Boiling Points of Fatty Acids and Their Derivatives Using Quantitative Structure-Activity Relationship (QSAR) Methodology. *ISAR Journal of Science and Technology*, *3*(3), 13-22.

1. Introduction

Lipids represent organic compounds which fail to dissolve in water but contain oil and fats together with carbon hydrogen and oxygen as their main chemical substances [1]. Lipids join nucleic acids carbohydrates along with proteins to become one of the key macromolecules which exist in the human body. Lipids differ from other macro macromolecules by lacking polymer structure and monomer composition which separates them from typical structural characteristics. Hydrocarbon manacles that contain (-CH2–CH2–CH2–) within their molecular structure show hydrophobic properties because they appear frequently in biochemical structures [2]. Lipids perform their essential biological functions as building blocks for cell membranes and energy reservoirs as well as signal transmitters in various crucial biological operations [3].

Comprising elongated hydrocarbon chains between 4 and 36 carbons and a single carboxyl group, fatty acids are the most basic type of lipids. Many complicated lipids consist of these molecules, which are hence crucial components. In organic settings, fatty acids usually show an even amount of carbon particles; 16–18 carbon greasy acids are most shared [4]. As a result, the body might produce saturated fatty acids (SFAs), which are fats without double bonds. Foods produced from animals, such as red meat, poultry, and full-fat dairy products, are the main dietary sources. The word saturated describes a molecule in which every carbon atom has as many hydrogen atoms as feasible. Many saturated fatty acids possess both a common name and a chemically descriptive systematic name [5].

Molecules that contain unsaturated fatty acids develop one or more (bends) from their hydrocarbon chain because they incorporate one or more double bonds. Unsaturated natural fatty acids show a cis double bond configuration as their basic geometric structure. The molecular grouping of these substances proves to be ineffective. Intermolecular interactions between molecules possess much weaker strength than those observed in saturated molecules. Unsaturated fatty acids present lower melting points in comparison to other types of fatty acids [6]. These fatty acids stay in liquid form when the environment reaches normal temperature [7]. The melting points of fatty acids experience changes from two key factors which are chain length and amount of unsaturation found in the hydrocarbon chains. Temperature conditions typical for human rooms transform saturated fatty acids between 12:0 and 24:0 into waxy solid masses. Similar fatty acids with the same carbon chain structure exist in liquid form because their molecular arrangements differ somewhat among the fatty acid molecules. Unrestricted rotation of carbon-carbon bonds throughout saturated fatty acids allows their hydrocarbon chains to become highly flexible thereby lowering the steric hindrance. The crystalline formation of molecular structures occurs from Van der Waals forces but unsaturated fatty acid cis double bonds create chain flexion which blocks dense packing [8]. Their decreased molecular interaction with similar chain length saturated fatty acids leads to reduced melting temperatures throughout their structure [36]. Animals fats demonstrate higher saturated fatty acid content compared to vegetable oils which produces increased melting points.

Long-chain fatty acids have a markedly low vapor pressure, which escalates as the chain length diminishes. Vegetable oils mostly consist of triglycerides containing long-chain fatty acids, resulting in very low vapor pressures; for instance, soybean and olive oils have vapor pressures of 0.001 and 0.05 mm Hg at 254 and 308°C, respectively [9]. Fatty acids have significant volatility; monoglycerides possess a considerably greater vapor pressure. Consequently, these hydrolytic cleavage products provide a source of smoke derived from fried oil waste to solve this problem, chemometrics calculation methods can be useful. The statistical and mathematical analysis of chemical data is usually referred to as chemometrics. In other words, chemometrics is an efficient method for summarizing useful information from a specific data series and predicting other data series. In fact, the goal of chemometrics is to improve measurement processes and extract more useful chemical information from physical and chemical measured data [10]. Chemometrics is used in various branches of chemistry, some of these applications include process control, analysis and recognition of patterns, signal processing and optimizing conditions. One of the important fields of application chemometrics is in studies that relate the properties of molecules to their structural characteristics [11]. The purpose of QSAR studies is to find the relationship between the physicochemical behavior of a molecule and its structural parameters. The results of these studies, in addition to clarifying the relationship between the properties of molecules and their structural characteristics, help researchers predict the behavior of new molecules based on their behavior, as similar molecules help [12].

Lemaoui et al. [13] established a molecular-based method to forecast eutectic solvent pH values during their investigation for efficient green solvent development. This research follows a similar predictive approach to the work presented in our research. The research field of sustainable solvents matters to both academic researchers and business operations. The increasing scientific comprehension of typical organic solvent dangers has led experts to create multiple environmentally conscious safer solvent replacements. This research employed two prediction algorithms through multiple linear regression (MLR) and artificial neural network (ANN) to determine pH levels of ESs while utilizing chemical descriptors from COSMO-RS database. A total of 648 experimental points were used for adequate data representation because they included 41 chemically different ESs derived from combinations of 9 HBAs with 21 HBDs at various temperatures. The analysis indicates that both prediction methods show powerful capabilities in new ESs pH forecasting though the ANN method provides stronger predictive strength and the MLR method offers better interpretability. These predictive models can reduce time and expenses by forthcoming the characteristics of designed solvents based on provided molecular sketches. Fitranda et al. [14] studied antibacterial properties of castor oil and its derivatives and their physicochemical characteristics. The obtained substances included K-soap (solid white form with melting point range 168-175°C) and free fatty acids (liquid yellow substance that boils at 210°C with density of 0.98 g/mL and refractive index 1.46 and viscosity 693.22 cSt and containing 145.88 (mgKOH/g) acids, 294.52 (mgKOH/g) saponification, and 148.64 (MgKOH/g) ester values) along with fatty acids methyl esters (liquid yellow material having 170°C boiling point). The researchers designed a precise melting temperature estimation model by using molecular weight and carbon-carbon double bond counting as descriptive elements in the HSVR framework. The development process for the HSVR-based model consists of two distinct parts. The testing phase for SVR model evaluation uses descriptors consisting of double carbon bond counts and molecular weights within a test-set-cross validation environment. In the second step researchers conduct more SVR training and testing through utilization of melting point predictions computed during the initial phase. The proposed hybrid system achieves better generalizing and forecasting abilities than traditional SVR would perform. The HSVR-based model achieves greater precision in determining the melting points of sixty-two fatty acids than existing predictive models such as Guendouzi and Guijie et al. models [15].

2. Experimental work

2.1. QSPR method

Quantitative structure property relationship (QSPR) method, utilized in computational chemistry [16], enables the prediction and estimation of molecular properties based on their structural features. For fatty acids, QSPR can predict their melting points (mp) and boiling points (bp) by analyzing their molecular structures [17]. To achieve this, a dataset containing information on various fatty acids, including their molecular structures and experimentally measured melting and boiling points, is collected.

The molecular structures are then converted into numerical representations called molecular descriptors, which quantify different features of the molecules. These descriptors serve as input variables for developing a QSPR model. Statistical and machine learning techniques are employed to establish the relationship between the molecular descriptors and the melting and boiling points of fatty acids in the dataset. The QSPR model is trained using this data and validated using a separate set of fatty acids not used during model development to ensure accurate predictions for unseen fatty acids. Once validated, the QSPR model can predict the melting and boiling points of new or unmeasured fatty acids based on their molecular structures [18].

The outcome of these predictions depends on three elements: high-quality dataset selection and appropriate descriptors along with reliable statistical or machine learning algorithms. The QSAR/QSPR approach uses CODESSA 3.3.1 software package to predict melting and boiling points through its function as an encoded developed tool according to this research. A QSPR modeling within CODESSA performs multilinear regression analysis with up to 50 separate molecular descriptors that cover constitutional as well as morphological and topological and electrostatic and quantum chemical and thermodynamic factors. The electrical descriptors describe molecular dipole moment together with the internal distribution of negative charges whereas topological descriptors reveal atomic quantities and their types and connection patterns.

2.1.1. The methodology based on QSPR

The procedure of utilizing QSPR method was summarized in Figure 1.

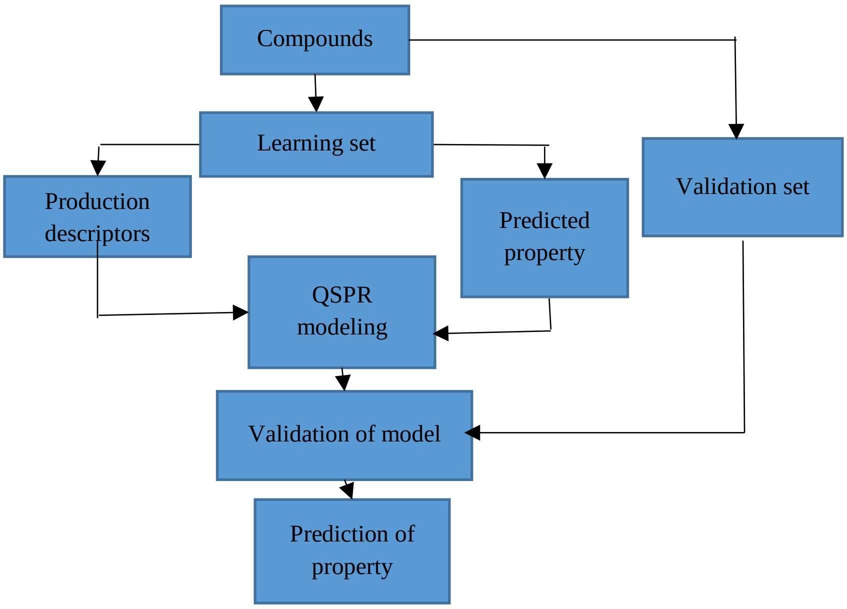


Figure 1. The procedure of the present work to estimate boiling point and melting point of different fatty acids.

2.2. Datasets

The data set provided for the boiling points were presented in Table 3.1. In the cases of melting point, we utilized the dataset provided in the Ref. [19]. An example from these datasets were presented in Table 1 and Table 2

Table 1. The experimental boiling points (bps) of 26 studied fatty acids [20-21].

Class	Fatty acid	Boiling points (oC)
A	Isoamy Laurate	132.10
B	Caproic	205.8
C	Caprylic	239.7
A	Methyl laurate	255.14
B	Capric	260
C	Lauric	298.9
A	Ethyl palmitate	309.13
B	Ethyl linoleate	319.16
C	Myristic	326.2
A	Ethyl oleate	331.52
B	Ethyl ricinoleate	344.01
C	Palmitic	351.5
A	Stearic	371.1
B	Triolein	414.91
C	Tripalmitin	412.69
A	Methyl laurate	263
B	Methyl myristate	296
C	Methyl palmitate	338
A	Methyl stearate	351

B	Methyl oleate	351
C	Methyl linoleate	351
A	Methyl linolenate	351
B	Methyl arachidate	370
C	Methyl behenate	387
A	Methyl erucate	406
B	Methyl lignocerate	407

Table 2. The experimental melting points (mp) of different fatty acids. More experimental data was extracted from Ref. [19]

Class	Fatty acid	Melting points (oC)
A	3-7-11-15-Tetramethylhexadecanoic acid	-65.0
B	Cis-cis-cis-cis-6-9-12-15-Octadecatetraenoic acid	-57.0
C	Cis-cis-cis-cis-5-8-11-14-Eicosatetraenoic acid	-49.0
A	Cis-cis-cis-cis-cis-cis-4-7-10-13-16-19-Docosahexaenoic acid	-45.0
B	Pentanoic acid	-33,0
C	3-Methylbutanoic acid	-29.0
A	Cis-cis-cis-9-12-15-Octadecatrienoic acid	-11.0
B	Cis-cis-9-12-Octadecadienoic acid	-7.0
C	Heptanoic acid	-7,0
A	Butanoic acid	-5.0
B	cis-9-Tetradecenoic acid	-4,0
C	Cis-cis-5-13-Docosadienoic acid	-4,0
A	Hexanoic acid	-3,0
B	cis-9-Hexadecenoic acid	0,0
C	12-Hydroxy-cis-9-octadecenoic acid	5,0
A	Nonanoic acid	12,0
B	cis-9-Octadecenoic acid	13.0
C	cis-11-Octadecenoic acid	15.0
A	Octanoic acid	16.0
B	cis-trans-9-11-Octadecadienoic acid	20,0
C	trans-cis-10-12-Octadecadienoic acid	23.0
A	cis-11-Eicosenoic acid	24.0
B	cis-9-Eicosenoic acid	24,0
C	9-Decenoic acid	26.0
A	cis-5-Eicosenoic acid	26.0
B	Undecanoic acid	28,0
C	cis-6-Octadecenoic acid	29.0
A	Decanoic acid	31.0
B	cis-12-13-Epoxy-cis-9-octadecenoic acid	32.0
C	trans-trans-cis-9-11-13-Octadecatrienoic acid	32,0

The initial and most crucial step in QSPR modeling involves the collection and selection of a desirable data set that can be determined from a chemical family with the laboratory-measured property of interest, under consistent conditions of pressure and temperature, and with high precision measurements. It is essential to ensure that the experimental errors are not significant, as an accurate and reliable model relies on precise measurements, and lower measurement errors improve the predictability of the model [17].

Another vital consideration in dataset selection is to ensure that it is sufficiently extensive and diverse. Larger datasets lead to the development of more robust predictive models, while greater diversity in compounds enables the model to effectively predict a wider range of substances. Therefore, key criteria for a satisfactory QSPR model include [22]:

a: Diversity of the dataset together with adequate size.

b: Measurements which often were conducted under consistent and reproducible conditions.

The QSPR modeling of forecasting the boiling point and melting point of fatty acids utilizes relevant data from Tables 1 and 2. This document demonstrates how the graphical user interface of ChemBioDraw Ultra version 12.0 produced three-dimensional molecular representations.

2.3. Multiple linear regression (MLR) analysis

In this work, the multiple linear regression (MLR) analysis was utilized for investigating the relationship that stands among the response variable and predictor variables and also for estimating the response variable according to the predictor variables. MLR fits a linear model of the form as following Eq (1)[23]:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots b_kX_k + e \qquad (1)$$

The dependent variable appears as Y while the independent variables appear as X1, X2,..., Xk together with e being the random error and b0, b1, b2,..., bk indicating the estimable regression coefficients. During the MLR approach the selection process chooses regression coefficients which minimize the square value of estimation-observation discrepancies. The primary task in multilinear regression involves obtaining the optimal regression coefficient estimations (b0, b1, b2, . . ., bk) to reduce the error sum (e) while achieving the most accurate data match. The calculation happens through several statistical methods with the least squares method being among them.

Multiple Linear Regression serves numerous research domains including economics and social sciences and engineering and data science to discover predictive patterns from several predictor variables [24-26]. Scientists can use this method to determine independent variable contributions separately as well as understand their combined influence on the dependent variable [27].

2.4. Validation and verification

Regarding validation and verification of the used models in this study, some analysis was incorporated in verifying the correctness of the correlations and using them to estimate bp and mp attributes. It usually consists of elements similar to other various training sets as well as well recognized qualities based on experimental evidence. By means of a comparison between the projected data against experimental information derived from the literature, one may investigate and assess the predictive power of the correlation .

One may evaluate the predictive capacity of the correlations by using two regression correlation coefficients of the cross-validation R2cv as well as R2.Popularly used and quite useful for evaluating the dependability of statistical methods was the cross validation R2cv. In this process, updated data sets were created by eliminating one or group of objects in every case as well as for every statistics set using an input-output model based on the applied approach. Its precision helps one to project the reactions of the residual data, so influencing also.

The R2adj (attuned coefficient of a manifold linear regression determination) model has been introduced as terms of the coefficient of determination as the following relation [28]:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \qquad (2)$$

Considering a data set, where ni stands for the number of compounds, and pi shows the number of descriptions.

3. Results

Many non-empirical molecular descriptors may be obtained using the Codessa software applied in current work. We considered the preliminary regression analysis and derived the complete original Codessa descriptors. Furthermore used was the BMLR regression, thus the pool of the descriptors gets even smaller. With regard to BMLR correlations for all the included substances, the primary goal was to find the ideal number of descriptors which fit the instance of data given in Table 1 and Table 2. At last, suitable equations for varying numbers of descriptors were discovered. Respectively, Figure 4.1 and 4.2 exhibits the behavior of descriptors on the R2 as well as R2adj for bp andmp parameters. As advised in the Ref. [19], the R2 values lower than 0.02 is chosen to serve as a breakpoint criteria to prevent from the over parameterization .

Table 3 and Table 4 present the several used elements in the instance of bp andmp prediction derived from the QSPR approaches. The models for MP and bp were constructed using training sets of 62 and 30 fatty acids, respectively. The optimal QSPR equations are characterized by the subsequent equations:

The molecular descriptors in both models were considered as follows:

A2 and B2 (PPSA-3 atomic charge)

A3 and B3 (The number of aromatic bonds),

A4 and B4 (Mean complementary information content),

A5 and D5 (Balaban index),

A6 and D6 (YZ Shadow) and

A7 and D7 (Min atomic orbital, electronic population, and Boiling points (property)),

Table 3. Effective parameters in the use of QSPR for 30 data related to mp (X shows regression coefficients, ΔX error, t-Test, and p-Values).

M: $R^2 = 0.9312,\ AdjustedR^2 = 0.928,\ \ Std.E.E: 14.0281$				
Factors	X	X2-X1	t-Test	p-Value
Intercept	202.895	21.25	4.235	
A2 (cofficianent)	-2.125	≈ 0.001	-13	$< 10^{-3}$
A3	16.5	≈ 0.001	24.23	$< 10^{-3}$
A4	-1.114	≈ 0.001	-24	$< 10^{-3}$
A5	0.256	≈ 0.001	12.23	$< 10^{-3}$

Table. 4. Effective parameters in the use of QSPR for 30 data related to bp

M: $R^2 = 0.9482,\ AdjustedR^2 = 0.9387,\ \ Std.E.E: 12$				
Factors	X	X2-X1	t-Test	p-Value
Intercept	142	26.523	5.654	
B3 (cofficianent)	-8.256	≈ 0.002	32.251	$< 10^{-4}$
B4	-12.589	≈ 0.002	-4.985	$< 10^{-4}$
B5	-8.54	≈ 0.002	-23.251	$< 10^{-4}$

Figure 2 shows the R2 obtained for different number of descriptors in the case of bp prediction. As can be seen, when the number of descriptors increased to 7, the R2 was obtained around 0.938 which can be sufficient and appropriate. As same Figure was provided to mp and the max R2 was obtained around 0.948.

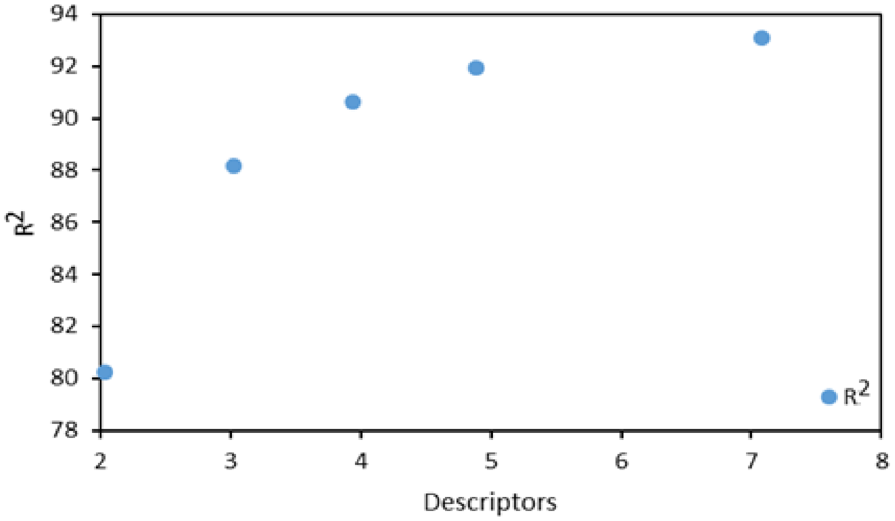


Figure 2. The R2 obtained for different number of descriptors in the case of mp prediction.

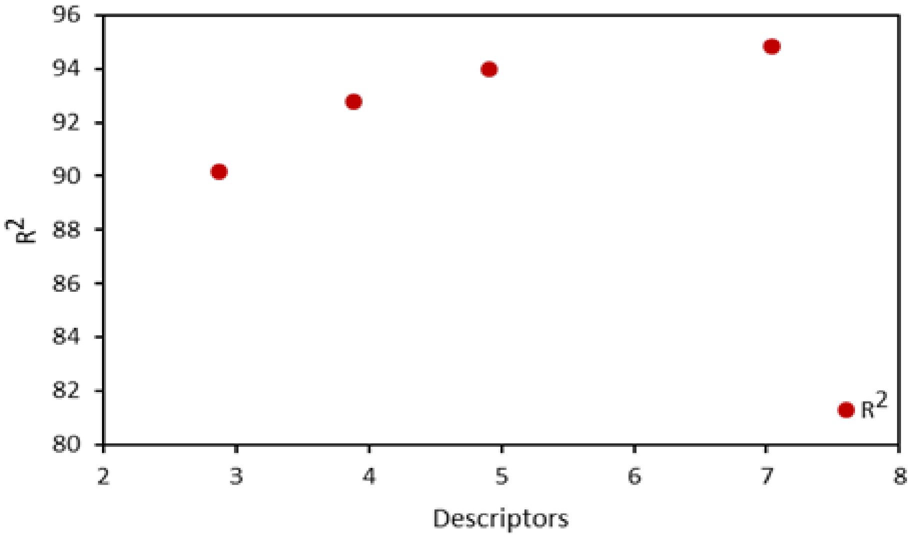


Figure 3. The R2 obtained for different number of descriptors in the case of bp prediction.

As can be seen, we have considered a same parameter to evaluate the commonalities and differences between parameters. According to the t-test values, in the case boiling point and melting point, A3 and B3 showed to have the most influences. It is found that the most important factor for both mp and bp is the number of aromatic bonds. Then, in the case of boiling point A5 has a more effect and for melting point B5 is the most important parameters. In the case of both bp and mp factors, the value of linear correlation coefficient was determined to be less to 0.5 (among two different descriptors). Hence, one can presume that the considered descriptors A2, A3, A4 and A5 as well as B2, B3, B4 and B5 are independent, mutually, considering the utilized QSPR method. Therefore, it is found that the 7-parameter model for mp and bp demonstrated satisfactory statistical data correlation coefficient.

In previous studies, two different methods have been proposed to validate the QSPR model [64]. The initial approach involves utilizing a subset of the available data to develop the model, taking into account the data in the bp and mp scenarios referred to as external validation. The subsequent approach entails employing the entirety of the data points to construct the model while reserving the validation method for internal cross-validation procedures. As suggested by in Ref. [45], the second approach was utilized based on the following stages:

(1) Have ordered the obtained points for both bp an mp data.

(2) In the case of mp, 30 data points were considered and ordered in three subsets (A–C) and the same was considered for bp with 30 data.

(3) Based on these information, we provided new datasets, considering whole combinations of the binary sums, including (A + B), (A + C) and (B + C). This dataset was utilized for training tasks due to the limitation of data for both bp and mp.

(4) After that the standard modeling approach of QSPR method consisting of the most important multiple linear regression method (B-MLR) has been taken into consideration for the three datasets mentioned the previous stage. In addition, considering each training set, we drive the correlation equation with the same descriptors corresponding to introduced models

(5) The classical internal cross validation approach was then used to validation tasks.

Eq. (3) and Eq. (4) demonstrated the boiling point and melting point that are considerably related to sets of molecular descriptors, including topological and electrostatic properties. As mentioned different initial models firstly were evaluated based on statistical analysis the data provided previously and the mentioned relations were selected for analysis.

Between all data, for mp 30% of data was utilized to train the model and for bp this procedure was applied and the results were taken into account for external validation datasets. The QSPR efficiency was considered in the case of prediction tasks using the R2adj. The mean values of adjusted R2(Fit) and R2(Pred) were determined to be near to 0.913 and 0.925, respectively.

After initial investigations, it was determined that the following relations:

$$\mathrm{bp} = -20123.4 + 1462.2 \times \ \mathrm{A1} - 0.8 \times \ \mathrm{A2} - 5.2 \ \times \mathrm{A3} - 5.8 \ \mathrm{A4} + 0.3 \ \times \mathrm{A5} \text{ - } 5.4 \ \times \mathrm{A6} + 35878 \times \ \mathrm{A5} \tag{3}$$

$$mp = -72.58 + 25.89 \times B_3 + 42.23 \times B_4 - 32 \times B_5 - 1.89 \times B_6 + 258955 \times B_7 \tag{4}$$

Figure 4 and Figure 5 show the comparison between experimental against predicted data for bp and mp, respectively that can be used to detect any outliers and to prompt indication of the accuracy.

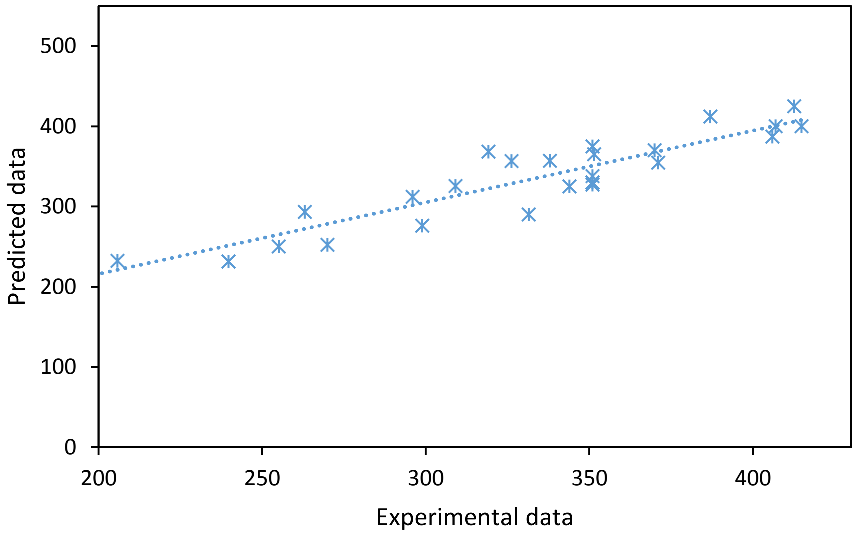


Figure 4. The comparison of experimental data against predicted data for bp

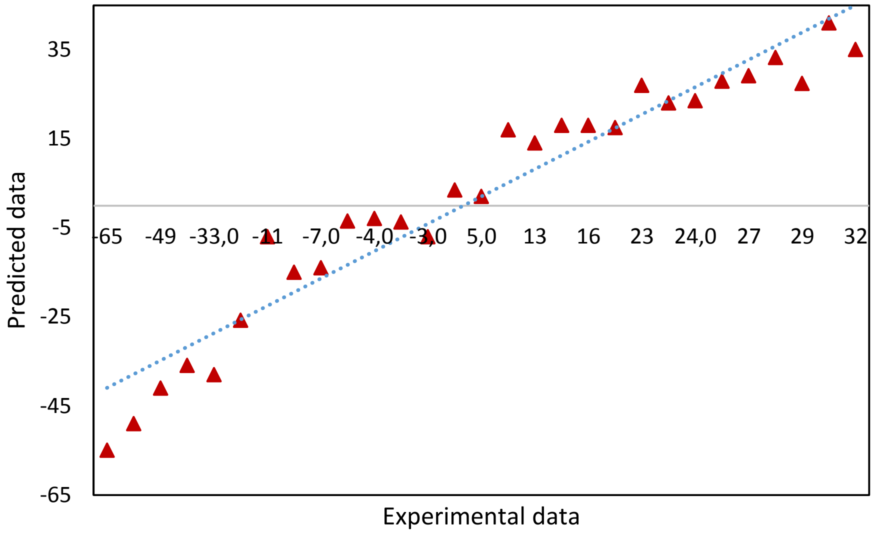


Figure 5. The comparison of experimental data against predicted data for mp

Tables 5 and 6 show the different QSPR models obtained using 5- parameter models for bp and mp, respectively. The models were established using training set consisted of 30 and 26 data in cases of bp and mp constituted, respectively.

Table 5. QSPR validation for 5-parameter model related to bp prediction

Training data	N	R2	Test set	N	R2(pred)
A+B	18	0.932	C	8	0.941
A+C	17	0.937	B	9	0.952
B+C	17	≈0.94	A	9	0.962

Table 6. QSPR validation for 5-parameter model related to mp prediction

Training data	N	R2	Test set	N	R2(pred)
A+B	20	0.941	C	10	0.947
A+C	20	0.948	B	10	0.968
B+C	20	0.950	A	10	≈0.95

In Table 5, the results obtained from the QSPR validation tasks were evaluated in the case of the 5-parameter model efficiency to predict bp. As can be seen Table was divided into two sections, the first was Training data and the second was Test set showed the performance of the model for different combinations of training data (A, B, C) and the corresponding R2 values for each dataset.

(a): For the Training data section:

- Three combinations of training data are considered: A+B, A+C, and B+C.

- N stands for the number of data points in each training set.

- R2 indicated the coefficient of determination, which can be measured how well the model fits the training data. It quantifies the proportion of the variance in the dependent variable (bp) that is predictable from the independent variables (molecular descriptors).

- The R2 values for the three training sets are approximately 0.932, 0.937, and 0.940, respectively.

(b): For the Test set section:

- The table provides three different test sets: A, B, and C, each containing a different number of data points.

- N represents the number of data points in each test set.

- R2(pred) shows the predictive R2, which measures how well the model performs on unseen data, i.e., the ability to generalize to new data points not used during training.

- The R2(pred) values for the three test sets were determined to be 0.941, 0.952, and 0.962, respectively.

Overall, the 5-parameter model demonstrated good performance in predicting the boiling points of fatty acids. The R2 values for both training and test sets were relatively high, indicating a strong correlation between the molecular descriptors and the boiling points. The model's predictive capability, as evidenced by the R2 (pred) values, showed that it can generalize well to unseen data, providing reliable predictions.

Table 6 demonstrated the QSPR validation results for the 5-parameter model related to melting point (mp) prediction. Similar to the boiling point model, the 5-parameter model for melting point prediction also showed strong performance. The high R2 values for both training and test sets indicated a good correlation between the molecular descriptors and the melting points. The R2(pred) values demonstrated the model's ability to generalize well to new data, making it a reliable predictor for the melting points of fatty acids.

As a results, the 5-parameter models for both bp and mp prediction exhibited promising results, with high R2 values for both training and test datasets. The models showed good predictive capabilities, suggesting that they can be valuable tools for estimating the bp and mp of fatty acids based on their molecular descriptors. However, further external validation using independent datasets and considering additional molecular descriptors could enhance the models' reliability and generalization performance. The developed QSPR model based on non-empirical molecular descriptors can predict the boiling point and melting point of fatty acids. The model was trained and validated using a dataset of various fatty acids with known boiling and melting points. By providing the QSPR model with a chemical structure of a fatty acid, it can calculate the molecular descriptors for that compound and use them to predict its boiling and melting points. However, it is essential to consider that the accuracy of the predictions may be influenced by the similarity of the compound to those in the training dataset. For accurate predictions, it is recommended to validate the model's performance on external datasets or experimental measurements of new and unseen chemical structures . Therefore, the developed QSPR model can be utilized as a tool for predicting the boiling and melting points of fatty acids based on their molecular structures, but caution should be exercised when applying the model to compounds significantly different from those in the training data.

4. Conclusion

The Quantitative Structure-Property Relationship (QSPR) models developed in this study provide reliable predictions for the melting and boiling points of fatty acids and their derivatives. Using a combination of molecular descriptors and multiple linear regression (MLR), we successfully established predictive models that correlate the chemical structure of fatty acids with their thermal properties. The models demonstrated strong predictive power, with high coefficients of determination (R^2) for both training and test datasets, reaching up to 0.948 for melting points and 0.938 for boiling points. The cross-validation R^2 values further confirmed the robustness of the models, indicating their ability to generalize well to unseen data. These findings highlight the effectiveness of QSPR modeling in predicting the thermal properties of fatty acids, offering a valuable tool for the design and optimization of fatty acid derivatives in various industrial applications. Future work could further enhance the model's predictive accuracy by incorporating additional molecular descriptors or expanding the dataset to include a broader range of fatty acid derivatives.

References

1. Gandhi, V., Tiwari, B., & Sellamuthu, B. (2022). Conventional sources of lipids. In *Biomass, Biofuels, Biochemicals* (pp. 89-107). Elsevier.

2. Fahy, E., Cotter, D., Sud, M., & Subramaniam, S. (2011). Lipid classification, structures and tools. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 1811(11), 637-647..

3. Song, H., Hsu, F. F., Ladenson, J., & Turk, J. (2007). Algorithm for processing raw mass spectrometric data to identify and quantitate complex lipid molecular species in mixtures by data-dependent scanning and fragment ion database searching. Journal of the American Society for Mass Spectrometry, 18, 1848-1858.

4. Kumar, P., & Mina, U. (2023). Life Sciences, Fundamentals and Practice II, Seventh edition: Pathfinder Publication.

5. Quah, S. R. (2016). *International encyclopedia of public health*. Academic press.

6. Grumezescu, A. M. (Ed.). (2019). Biomedical applications of nanoparticles. William Andrew.

7. Jannin, V., Musakhanian, J., & Marchaud, D. (2008). Approaches for the development of solid and semi-solid lipid-based formulations. *Advanced drug delivery reviews*, *60*(6), 734-746.

8. Gaba, B., Fazil, M., Ali, A., Baboota, S., Sahni, J. K., & Ali, J. (2015). Nanostructured lipid (NLCs) carriers as a bioavailability enhancement tool for oral administration. *Drug delivery*, *22*(6), 691-700.

9. Formo, M. W., (1979). Physical properties of fats and fatty acids, in Bailey's Industrial Oil and Fat Products, 4th ed., Vol. 1, D. Swern, Editor, Wiley: New York. p. 177–212.

10. Zappi, A., Marassi, V., Giordani, S., Kassouf, N., Roda, B., Zattoni, A., ... & Melucci, D. (2023). Extracting information and enhancing the quality of separation data: a review on chemometrics-assisted analysis of volatile, soluble and colloidal samples. Chemosensors, 11(1), 45.

11. Rácz, A., Bajusz, D., & Héberger, K. (2018). Chemometrics in analytical chemistry. Applied Chemoinformatics: Achievements and Future Opportunities, 471-499.

12. Herrera, S. E., Agazzi, M. L., Apuzzo, E., Cortez, M. L., Marmisollé, W. A., Tagliazucchi, M., & Azzaroni, O. (2023). Polyelectrolyte-multivalent molecule complexes: physicochemical properties and applications. Soft Matter, 19(11), 2013-2041.

13. Lemaoui, T., Abu Hatab, F., Darwish, A. S., Attoui, A., Hammoudi, N. E. H., Almustafa, G., ... & Alnashef, I. M. (2021). Molecular-based guide to predict the pH of eutectic solvents: promoting an efficient design approach for new green solvents. ACS Sustainable Chemistry & Engineering, 9(17), 5783-5808.

14. Fitranda, M. I., & Marfu'ah, S. (2020, May). Physicochemical properties and antibacterial activity of castor oil and its derivatives. In IOP Conference Series: Materials Science and Engineering (Vol. 833, No. 1, p. 012009). IOP Publishing ,Bnstol , England .

15. Owolabi, T. O., Zakariya, Y. F., Olatunji, S. O., & Akande, K. O. (2017). Estimation of melting points of fatty acids using homogeneously hybridized support vector regression. Neural Computing and Applications, 28, 275-287.

16. Ahmadi, S., Lotfi, S., Hamzehali, H., & Kumar, P. (2024). A simple and reliable QSPR model for prediction of chromatography retention indices of volatile organic compounds in peppers. *RSC advances*, *14*(5), 3186-3201.

17. Rybińska-Fryca, A., Sosnowska, A., & Puzyn, T. (2020). Representation of the structure—A key point of building QSAR/QSPR models for ionic liquids. *Materials*, *13*(11), 2500.

18. Liang, G., Xu, J., & Liu, L. (2013). QSPR analysis for melting point of fatty acids using genetic algorithm based multiple linear regression (GA-MLR). *Fluid Phase Equilibria*, *353*, 15-21.

19. Guendouzi, A., & Mekelleche, S. M. (2012). Prediction of the melting points of fatty acids from computed molecular descriptors: A quantitative structure–property relationship study. *Chemistry and Physics of Lipids*, *165*(1), 1-6.

20. Santander, C. M. G., Rueda, S. M. G., da Silva, N. D. L., de Camargo, C. L., Kieckbusch, T. G., & Maciel, M. R. W. (2012). Measurements of normal boiling points of fatty acid ethyl esters and triacylglycerols by thermogravimetric analysis. *Fuel*, *92*(1), 158-161.

21. Ruan, D. F., Chen, Z. H., Wang, K. F., Chen, Y., & Yang, F. (2014). Physical property prediction for waste cooking oil biodiesel. *Open Fuels Energy Sci J*, *7*(1), 62-68.

22. Toropov, A. A., Raška Jr, I., Toropova, A. P., Raškova, M., Veselinović, A. M., & Veselinović, J. B. (2019). The study of the index of ideality of correlation as a new criterion of predictive potential of QSPR/QSAR-models. *Science of the Total Environment*, *659*, 1387-1394.

23. Marill, K. A. (2004). Advanced statistics: linear regression, part II: multiple linear regression. *Academic emergency medicine*, *11*(1), 94-102.

24. Etemadi, S., & Khashei, M. (2021). Etemadi multiple linear regression. *Measurement*, *186*, 110080.

25. Roback, P., & Legler, J. (2021). *Beyond multiple linear regression: applied generalized linear models and multilevel models in R*. Chapman and Hall/CRC.

26. Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, *106*, 234-240.

27. Kaya, U., Özkan, H., Yazlık, M., Güngör, G., Çamdeviren, B., Karaaslan, İ., ... & Yakan, A. (2023). Determination of milk fatty acids and some phenotypic characters affecting total milk fat in dairy cows with multiple linear regression. *Veteriner Hekimler Derneği Dergisi*, *94*(2), 119-126.

28. Alexopoulos, E. C. (2010). Introduction to multivariate regression analysis. *Hippokratia*, *14*(Suppl 1), 23.