QSPR Models for the Molar Refraction, Polarizability and Refractive Index of Aliphatic Carboxylic Acids Using the ZEP Topological Index

Zoiţa Mărioara Berinde

Department of Chemistry and Biology, North University Center at Baia Mare,
Technical University of Cluj-Napoca, 430122 Baia Mare, Romania; zoita.berinde@cb.utcluj.ro

Abstract: The molar refraction, polarizability, and refractive index for a series of monocarboxylic, dicarboxylic, and unsaturated monocarboxylic acids, having a symmetric or asymmetric structure, were investigated by the application of quantitative structure property relationship (QSPR) technique. We used a linear regression method and a single molecular descriptor, the ZEP topological index, calculated in a simple manner, with the help of weighted electronic distances, and also calculated on the basis of the chemical structure of the molecules. The high-quality performance and predictive ability of the QSPR models obtained were validated by means of specific validation techniques: y-randomization test, the leave-one-out cross validation procedure, and external validation. The investigated properties are well modeled (with $r^2 > 0.99$) by the ZEP index, using the regression analysis as a statistical tool for developing reliable QSPR models. Our approach provides an alternative technique to the existing additive methods for predicting the molar refraction and polarizability of carboxylic acids, which is essentially based on the summation of atom and/or functional group contributions or bond contributions, and of some correction increments.

Keywords: aliphatic carboxylic acids; molar refraction; polarizability; refraction index; topological index

1. Introduction

Carboxylic acids form a family of organic compounds that contain the characteristic carboxyl functional group (-COOH or -CO$_2$H), and they constitute an important class of chemicals that are very important in industry and also occur in many other processes. Among the most significant uses of carboxylic acids are the following: in making soaps, detergents, and shampoos; in food industry; in pharmaceutical industry; in the manufacturing of rubber; in making dye stuffs, perfumes, and rayon. Moreover, this series of fatty carboxylic acids is extremely good for human health. In the last years, the use of the properties of carboxylic acids, as independent variables in QSAR models, has been steadily increasing [1–5]; however, QSPR models that involve the properties of carboxylic acids as dependent variables are very few. This lack of interest seems to be due to the specific structure of carboxylic acids, which strongly influences their properties. The carboxyl functional group is generally considered to be a highly polar organic functional group. Due to the sp^2 hybridization state of the carbon atom and of the oxygen double bond, the carboxyl functional group has a planar structure which favors the p-π conjugacy and creates a strong permanent dipole. The dipoles present in carboxylic acids allow them to form strong hydrogen bonds between acid molecules, and also between acid molecules and water or other molecular solvents. These aspects influence the essential relationship between the structural attributes and the properties of carboxylic acids.

In our study, we considered three properties of carboxylic acids: molar refraction, molar polarizability, and refractive index. These properties are interrelated and are influenced by the electronic interactions and the polarity of carboxylic acids. The molar refraction, R_m (cm^3 mol^{-1}) is a constitutive-additive molecular property of substances [6]. The molar

refraction is related to the polarizability of the molecules that make up the medium, by the Lorentz–Lorenz [7] equation:

$$R_m = \frac{n_D^2 - 1}{n_D^2 + 2} V_m = \frac{4\pi}{3} = N_A P \tag{1}$$

where n_D is the refractive index of the given substance at optical wavelengths, usually at 589 nm (sodium D-line), V_m is the molar volume, N_A is Avogadro's constant, and P is the mean polarizability of molecules. For a radiation of infinite wavelength, $R_m = V_m$ and, therefore, the molar refractivity can be used as a measurement of the real volume of the molecules, a very important fact for chemists and biologists. The molar refraction can also be evaluated by means of refractive index, molecular weight, and density, by replacing the molar volume in Equation (1) with the ratio of molecular weight (MW) and density (d): $V_m = MW/d$. On the other hand, molar refraction is a measure of the total polarizability of a mole of substance, see [6] for more details.

The refractive index (n_D) characterizes the capacity of a substance to refract the light. Light traversing a substance has a velocity different from the case when light is traversing a vacuum. The ratio of the velocity of light in a vacuum to that in a substance is the refractive index or the index of refraction of the substance. The refractive index is often used to identify a particular substance, to confirm its purity, or to measure its concentration, see [8] for more details.

Molar refraction and molecular polarizability being additive properties can be calculated by summing up the contributions of a variety of atoms and/or functional groups, bond contributions, and various corrections factors. The most developed way to obtain molar refraction uses Crippen's fragmentation methods [9,10]. Alternatively, attempts have been made by various QSAR researchers to model molar refractivity by using topological indices [11,12]. Verma, Kuo, and Hansch [13] studied the polarizability effects on ligand–substrate interactions, in terms of the number of valence electrons (NVE), and proposed various linear QSAR models. Verma and Hansch [2] performed a comparison regarding the use of NVE and calculated molar refractivity (CMR) in QSARs for studying chemical–biological interactions, while Hansch and Kurup [14] found that the simple summation of the valence electrons (H = 1, C = 4, O = 6, etc.) in a molecule is a measure of its polarizability. They also showed that this parameter correlates with the nerve toxicity of a wide variety of chemicals acting on the nerves of frogs, rabbits, cockroaches, and humans. Fast empirical models to predict molecular polarizability were also developed by Wang, Xie, Hou, and Xu [15], using two different approaches. The refractive index, molar refractivities, and molar polarizability constant of heterocyclic compounds were studied by Sonar and Pawar [16], while Granados, Gracia-Fadrique, Amigo, and Bravo [17] studied the refractive index, surface tension, and density of aqueous mixtures of carboxylic acids.

Starting from this background, the main aim of the present study was to develop linear monovariable QSPR models that are able to predict molar refraction, polarizability, and refractive index in the class of carboxylic acids by using the ZEP topological index.

2. Materials and Methods

In order to develop predictive QSPR models for molecular refraction, refractive index, and polarizability values of carboxylic acids, we followed the following steps:

(i)   the selection of the data set;
(ii) generation of molecular ZEP index for carboxylic acids used in this work;
(iii) building QSPR models within the selected data set;
(iv) validation of the obtained QSPR models using the y-randomization test and the internal and external validation strategies.

2.1. Data Set

The properties of aliphatic carboxylic acid selected in this study are molecular refractivity, denoted by R_m; refractive index, denoted by n_D; and polarizability, denoted by P. The data set includes 80 acids: 50 saturated aliphatic monocarboxylic acids, 17 unsaturated acids, and 13 aliphatic dicarboxylic acids. Molecular refractivity values for these acids, as well as the refractive index values for 33 saturated aliphatic monocarboxylic acids and polarizability values for 21 saturated aliphatic monocarboxylic acids, were taken from the literature [1,18,19]. The values of polarizability for the 20 other monocarboxylic acids were calculated by means of the relation:

$$P = \frac{3R_m}{4\pi N_A} = 0.3964308 \cdot 10^{-24} R_m \tag{2}$$

where N_A is Avogadro's constant. These values are given in Table 1 and are indicated by the superscript [b].

Table 1. The ZEP index of the training and test sets of saturated aliphatic monocarboxylic acids and their molecular refraction (R_m), refractive index (n_D), and polarizability (P).

Acid	ZEP	R_m			n_D			P		
		Exp.	Pred.	Error	Exp.	Pred.	Error	Exp.	Pred.	Error
propanoic	14.9568	17.51	17.76	−0.25						
butanoic	18.3959	22.14	22.33	−0.19	1.411 [b]	1.414	−0.003	8.77	8.86	0.01
2-methylpropanoic	18.2240	22.10	22.10	0.00	1.410 [b]	1.414	−0.004	8.76	8.78	−0.02
pentanoic	21.8600	26.77	26.94	−0.17	1.420	1.419	0.001	10.61	10.68	−0.07
2-methylbutanoic	21.6042	26.73	26.60	0.13	1.417 [b]	1.418	−0.001	10.59	10.55	0.04
2,2-dimethylpropanoic	21.3494	26.74	26.26	−0.47	1.419	1.418	0.001	10.60	10.42	0.18
hexanoic	25.3241	31.41	31.55	−0.14	1.427	1.420	0.007	12.45	12.50	−0.05
2-methylpentanoic	25.2312	31.36 [b]	31.43	−0.07	1.425	1.421	0.004	12.43 [b]	12.46	−0.03
3-methylpentanoic	25.1775	31.36	31.36	0.00	1.425	1.421	0.004	12.43	12.43	0.00
4-methylpentanoic	25.0790	31.36	31.23	0.13	1.425	1.421	0.004	12.43	12.38	0.05
2,2-dimethylbutanoic	24.9786	31.37 [b]	31.09	0.28				12.44 [b]	12.32	−0.12
2,3-dimethylbutanoic	25.0456	31.32 [b]	31.18	0.14				12.42	12.36	0.06
2-ethylbutanoic	25.3088	31.36	31.53	0.17	1.425	1.421	0.004			
heptanoic	28.7882	36.04	36.16	−0.12	1.429	1.425	0.003	14.28	14.32	−0.04
2-Methylhexanoic	28.6954	35.93 [b]	36.04	−0.11				14.24 [b]	14.24	0.00
3-methylhexanoic	28.6254	35.99 [b]	35.95	0.04				14.26 [b]	14.24	0.02
2,2-dimethylpentanoic	28.4177	36.00 [b]	35.67	0.33	1.423 [b]	1.424	−0.001	14.27 [b]	14.13	0.14
3,3-dimethylpentanoic	28.3330	36.01	35.56	0.45	1.431 [b]	1.424	0.007			
2,3-dimethylpentanoic	28.6098	35.94	35.95	−0.01				14.24	14.23	0.01
2,4-dimethylpentanoic	28.4434	35.94	35.71	0.23	1.429	1.426	0.003	14.19	14.14	0.05
2-ethylpentanoic	28.7567	35.95	36.12	−0.17				14.25	14.31	−0.06
2,2,3-trimethylbutanoic	28.5568	35.96 [b]	35.86	0.10				14.26 [b]	14.20	0.06
octanoic	32.2523	40.67	40.77	−0.10	1.433 [b]	1.428	0.005	16.12	16.14	−0.02
2-methylheptanoic	31.9740	40.62 [b]	40.41	0.21	1.430 [b]	1.428	0.002	16.10 [b]	16.00	0.10
2,4-dimethylhexanoic	32.0167	40.59	40.46	0.13	1.432 [b]	1.428	0.004	16.07	16.02	0.05
3,3-dimethylhexanoic	31.7721	40.61	40.14	0.47						
2-ethylhexanoic	32.2208	40.63	40.73	−0.10	1.435	1.428	0.007	16.10	16.12	−0.02
2,3,4-trimethylpentanoic	31.8740	40.53 [b]	40.27	0.26				16.07 [b]	15.94	0.13
2-propylpentanoic	32.2047	40.63	40.71	−0.08	1.435	1.428	0.007	16.10	16.12	−0.02
2,2,3,3-tetramethylbutanoic	31.4091	40.61 [b]	39.65	0.96				16.10 [b]	15.70	0.40
nonanoic	35.7164	45.30	45.39	−0.09	1.439 [b]	1.432	0.007	17.96	17.96	0.00
2-methyloctanoic	35.6236	45.25 [b]	45.26	−0.01	1.439	1.432	0.007	17.94 [b]	17.91	−0.03
2,3-dimethylheptanoic	35.5218	45.21	45.13	0.08	1.438 [b]	1.432	0.006			
2-ethylheptanoic	35.6849	45.25 [b]	45.34	−0.09				17.94 [b]	17.94	0.00
2,3,4-trimethylhexanoic	35.4382	45.16 [b]	45.02	0.14				17.90 [b]	17.81	0.09
decanoic	39.1805	49.94	49.99	−0.05	1.443	1.435	.0008	19.79	19.78	0.01
2-methylnonanoic	39.0877	49.88 [b]	49.87	0.01	1.441	1.435	0.006	19.77 [b]	19.73	0.07
2,2-dimethyloctanoic	38.8101	49.89 [b]	49.50	0.39	1.442	1.435	0.007			
2-ethyloctanoic	39.1490	49.88 [b]	49.96	−0.08	1.441	1.435	0.006	19.77 [b]	19.76	0.01
2-propylheptanoic	39.1329	49.88 [b]	49.93	−0.05	1.441	1.435	0.006			
undecanoic	42.6446	54.57	54.61	−0.04	1.445	1.439	0.006	21.63	21.60	0.03
2-methyldecanoic	42.5518	54.51 [b]	54.48	0.03	1.444	1.439	0.005	21.61 [b]	21.55	0.06
2,2-dimethylnonanoic	42.2742	54.52 [b]	54.12	0.40	1.445	1.438	0.007			
2-ethylnonanoic	42.6131	54.51 [b]	54.57	−0.06	1.445	1.438	0.007	21.61 [b]	21.58	0.03
2-propyloctanoic	42.5970	54.51 [b]	54.54	−0.03	1.444	1.439	0.005	21.61 [b]	21.57	0.04
dodecanoic	46.1087	59.18	59.22	−0.04				23.42	23.42	0.00
tridecanoic	49.5728	63.81	63.83	−0.02						
tetradecanoic	53.0369	68.44 [b]	68.44	0.00				27.13 [b]	27.05	0.08
pentadecanoic	56.5010	73.07 [b]	73.05	0.02	1.468 [b]	1.453	0.015	28.97 [b]	28.87	0.10
hexadecanoic	59.9651	77.70 [b]	77.66	0.04				30.80 [b]	30.69	0.11

[b] The test set.

2.2. The ZEP Index

The molecular topological index ZEP used in this QSPR study was calculated using hydrogen-suppressed graphs of the carboxylic acids. The molecular topological index ZEP introduced by Berinde [20] is defined as:

$$\text{ZEP} = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \text{wed}\,(i,j) \right)^{\frac{1}{2}} \tag{3}$$

where wed (i,j) is the weighted electronic distance, also introduced by Berinde [20]:

$$\text{wed}(i,j) = \begin{cases} \dfrac{1}{b_{ij}} \cdot \dfrac{Z'_i + Z'_j}{v_i \cdot v_j}, & \text{if there is a bond between atom } i \text{ and atom } j \\[2ex] 0, & \text{if is not a bond between atom } i \text{ and } j \end{cases} \tag{4}$$

In (4) v_i, v_j denote the degrees of the vertices i and j, respectively; Z'_k denotes the formal degree of vertex k and is defined by $Z'_k = Z_k \cdot v_k$; and Z_k denotes the order number of atom k in Mendeleev's periodic system. The values of b_{ij} are 1, 2, 3, and 1.5 for a single bond, a double bond, a triple bond, and an aromatic bond, respectively. Alternatively, the topological index ZEP can be calculated by using the connectivity matrix, CEP [21]. In Table 2 are given the weighted electronic distances, the formal degrees of vertices, as well as the degrees of the vertices for common bonds in carboxylic acids. In order to emphasize the number of bonds of the carbon atom and oxygen atom, respectively, we kept the hydrogen atoms visible.

Table 2. Values of wed, Z'_k (upper row) and v_k (lower row).

Bond	$\overset{24}{\underset{4}{C}} = \overset{16}{\underset{2}{O}}$	$\overset{24}{\underset{4}{C}} - \overset{8}{\underset{1}{OH}}$	$\overset{6}{\underset{1}{CH_3}} - \overset{12}{\underset{2}{CH_2}}$	$\overset{6}{\underset{1}{CH_3}} - \overset{18}{\underset{3}{CH}}$	$\overset{6}{\underset{1}{CH_3}} - \overset{24}{\underset{4}{C}}$	$\overset{12}{\underset{2}{CH_2}} - \overset{12}{\underset{2}{CH_2}}$	$\overset{12}{\underset{2}{CH_2}} - \overset{18}{\underset{3}{CH}}$	$\overset{12}{\underset{2}{CH_2}} - \overset{24}{\underset{4}{C}}$
wed	2.5	8	9	8	7.5	6	5	4.5
Bond	$\overset{18}{\underset{3}{CH}} - \overset{18}{\underset{3}{CH}}$	$\overset{18}{\underset{3}{CH}} - \overset{24}{\underset{4}{C}}$	$\overset{24}{\underset{4}{C}} - \overset{24}{\underset{4}{C}}$	$\overset{12}{\underset{2}{CH_2}} = \overset{18}{\underset{3}{CH}}$	$\overset{12}{\underset{2}{CH_2}} = \overset{24}{\underset{4}{C}}$	$\overset{18}{\underset{3}{CH}} = \overset{18}{\underset{3}{CH}}$	$\overset{18}{\underset{3}{CH}} = \overset{24}{\underset{4}{C}}$	$\overset{24}{\underset{4}{C}} = \overset{24}{\underset{4}{C}}$
wed	4	3.5	3	2.5	2.25	2	1.75	1.5

In contrast to the usual topological distance, which is equal to 1 for any bond between two atoms, the weighted electronic distance, according to its definition, is able to differentiate between simple and multiple bonds, between covalent non-polar bonds and polar covalent bonds, and is also able to differentiate between the bonds depending on their branching degree and their neighboring bonds. It is also able to differentiate between the symmetric and asymmetric arrangements of atoms or groups of atoms with respect to a chemical bond. This property of differentiating is illustrated in Figure 1 in the case of four marked molecular graphs, which represent the structures of the following carboxylic acids: ethanoic, propanoic, 2-methylpropanoic, and 2,2-dimethylpropanoic. The carboxyl functional group is linked in each of the four mentioned cases to the remaining catena by a simple bond, but with a different branching and different neighboring bonds. Therefore, the weighted electronic distances for these bonds are different: 7.5; 4.5; 3.5; 3.0, i.e., the smallest value of the weighted electronic distance corresponds to the greatest branching, see Figure 1.
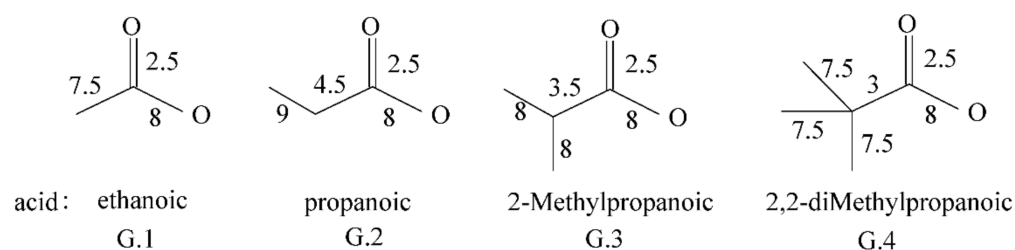
Figure 1. The wed for molecular graphs representing the skeleton of carboxylic acids.

We can illustrate the calculation technique of ZEP index for the hydrogen-suppressed graph of propanoic acid (G.2) by using Formula (3):

$$\text{ZEP(G.2)} = 9^{\frac{1}{2}} + 13.5^{\frac{1}{2}} + 15^{\frac{1}{2}} + 2.5^{\frac{1}{2}} + 8^{\frac{1}{2}} = 14.9568$$

We note that the ZEP index has been studied by the author in various contexts, in order to check its correlation power with several properties, and it has provided good correlation parameters [20–22].

In this work, the values of ZEP index for 84 carboxylic acids were calculated (the four other acids will be used in the validation process of our QSPR models). Note that all these values are different from each other, which also indicates the fact that ZEP index also has a good discrimination power, see also [22].

2.3. QSPR Model Building

In order to build a QSPR model, the data set was randomly divided into two subsets, namely, the training set and the test set. The training set was used for developing QSPR models, while the test set was used for validating the predictive power of the obtained QSPR models. In the training set, using least-square regression and considering only one variable, i.e., the ZEP index, simple linear equations were developed. The statistical parameters used to test the goodness-of-fit between the model-predicted and experimental values were the correlation coefficient (R), the coefficient of determination (R^2), the standard deviation (s), and the Fischer statistic value (F). A model with high values of R^2 and F, and a low value of s is usually preferred. For the coefficient of determination, the following condition is recommended [23]: $R^2 > 0.6$. This condition shows that the model will have a better fitting ability, but it does not reflect at all on the predictive power of the model [24].

2.4. Model Validation

For evaluating the stability and the predictive ability of QSPR models developed in the present paper, we applied the following three validation strategies from the list of five basic validation procedures presented in [25–27]: y-randomization test, the internal validation, and external validation.

Y-randomization test. The main aim of the y-randomization test is to detect and quantify chance correlations between the dependent variable and descriptors [25]. This test is designed to ensure the robustness of a QSPR model [26]. When applying the y-randomization test, the dependent variable, in our case R_m, or n_D, or P, is randomly shuffled and a new QSPR model is developed using the independent variable, the ZEP index, but not randomly. The process is repeated several times. All QSPR models obtained are expected to have low R^2 values, otherwise the QSPR model developed cannot be used for the given data set. According to Kiralj et al. [27], if $R^2_{yi} < 0.2$, there is no risk of a chance correlation in the developed model.

Internal validation. In our study, the validity of the model was tested using the cross-validation (CV) method and 'leave-one-out' (LOO) procedure in the training set. As is well-known, the correlation coefficient leave-one-out cross-validation describes the stability of a regression model. According to Kiralj et al. [27], the criterion of robustness and predictive ability of the model assumes $R^2_{CV} > 0.5$. It is accepted that the minimal

acceptable statistics for a QSPR regression model are requirements $R^2_{cv} > 0.5$ and $R^2 > 0.6$, see [23]. It is also generally agreed that a large difference between R^2 and R^2_{cv} (exceeding 0.2–0.3) is an indicator of the overfitting of the QSPR model.

External validation. The purpose of the external validation is to test the true predictive ability of the QSPR model. For this purpose, we analyzed the test data set of compounds that were not included in the training set or used in the model development. We first applied the y-randomization test, then we calculated the statistical parameters R^2_{ext} and Q^2_{ext}, similarly to R^2 and R^2_{CV} for the training set. The external validation performance is given by R^2_{ext} and Q^2_{ext}. R^2_{ext} is a measure of fitting for the external validation set and can be compared to R^2 for the training data set [28].

3. Results and Discussion

We calculated the ZEP index for the 84 acids used in this study: the values obtained for the 50 saturated aliphatic monocarboxylic acids are listed in Table 1, the ZEP index values of the 13 aliphatic dicarboxylic acids and 17 unsaturated acids are listed in Tables 3 and 4, respectively, while the values of ZEP for the remaining four acids are listed in Table 5.

Table 3. The ZEP index values of aliphatic dicarboxylic acids, their molecular refraction (R_m).

Acid	ZEP	R_m			
		Exp.	Pred.	Error	Croos
Training set					
Propandioic	19.5651	19.07	19.05	0.02	19.04
Butandioic	23.0458	23.70	23.71	−0.01	23.71
Pentandioic	26.5099	28.34	28.34	0.00	28.34
Hexandioic	29.9740	32.97	32.97	0.00	32.97
Heptandioic	33.4381	37.60	37.60	0.00	37.60
Octanedioic	36.9022	42.24	42.23	0.01	42.23
Nonanedioic	40.3663	46.82	46.86	−0.04	46.86
Decanedioic	43.8304	51.50	51.49	0.01	51.49
Dodecanedioic	50.7587	60.77	60.76	0.01	60.76
Tetradecanedioic	57.6868	70.03	70.02	0.01	70.02
Test set					
Undecanedioic	47.2945	56.13 [b]	56.12	0.01	
Tridecanedioic	54.2228	65.39 [b]	65.39	0.00	
Pentadecanedioic	61.1509	74.66 [b]	74.65	0.01	

[b] The test set.

3.1. QSPR Model Building for Molecular Refraction

3.1.1. Saturated Aliphatic Monocarboxylic Acids

In order to build a QSPR model for the molecular refractivity, we applied the above mentioned procedure. Table 1 displays the experimental molecular refraction for the 50 saturated aliphatic monocarboxylic acids having asymmetric structure with respect to the carboxyl group or having an asymmetric carbon atom. They were divided into two subsets: one set with 26 acids that formed the training set used in the modelling process, and another set with 24 acids that formed the test set, which are marked with [b] and was used for testing the model in external validation.

Table 4. The ZEP index values of aliphatic unsaturated acids, their molecular refraction (R_m).

Acid	ZEP	R_m			
		Exp.	Pred.	Error	Croos
Training set					
trans-2-Pentenoic	20.9397	26.83	27.30	−0.05	27.34
4-Pentenoic	20.2379	26.50	25.93	0.57	25.87
trans-2-Hexenoic	24.3876	31.46	31.76	−0.30	31.79
trans-3-Hexenoic	24.1064	31.46	31.41	0.05	31.41
(E)-but-2-enoic	17.1824	22.20	22.48	−0.28	22.52
propenoic	13.3028	17.23	16.92	0.31	16.82
2-Octenoic	31.0076	40.73	40.28	0.45	40.19
2-Octynoic	30.8640	38.68	39.26	−0.58	39.36
2-Propynoic	12.5041	15.33	15.66	−0.33	15.78
2-Nonynoic	34.3281	43.31	43.73	−0.42	43.87
2-Nonenoic	34.4717	45.36	44.75	0.61	44.52
3-Butenoic	16.3394	21.87	21.44	0.43	21.37
Test set					
2-Methyl-propenoic	16.8167	21.96 [b]	21.73	0.23	
cis-2-Methyl-2-butenoic	20.9397	27.26 [b]	27.04	0.22	
trans-2-Heptenoic	27.5435	36.05 [b]	35.53	0.52	
trans-3-Heptenoic	27.5544	36.05 [b]	35.55	0.50	
3-heptynoic	26.8731	34.20 [b]	34.67	−0.47	

[b] The test set.

Table 5. Molar refraction of four external acids.

Acid	ZEP	R_m		
		Exp.	Pred.	Error
Octadecanoic	66.8933	86.96	86.88 Equation (6)	0.08
			87.15 Equation (9)	−0.19
			83.94 Equation (12)	3.02
Hexadecanedioic	64.6150	79.29	79.28 Equation (10)	0.01
			81.07 Equation (12)	−1.78
trans-5-octenoic	31.0453	40.4	40.04 Equation (11)	0.36
			38.81 Equation (12)	1.59
(Z)-9-Octandecenoic	65.6702	86.12	84.60 Equation (11)	1.52
			82.40 Equation (12)	3.72

By correlating the molecular refractivity with ZEP index for the 26 monocarboxylic acids, which were used as a training set, we obtained the following linear QSPR model:

$$R_m = -2.152\ (\pm 0.151) + 1.331\ (\pm 0.005)\ \text{ZEP} \tag{5}$$

N = 26 R = 0.9999 R^2 = 0.9998. R^2_{CV} = 0.9998 Q^2_{ext} = 0.9998 s = 0.2069 F = 73,732.2; MAE = 0.147; MAD = 0.11

The QSPR model (5) has a very good statistical quality for fitting the calculated R_m values to the experimental ones. The robustness of the model (5) and its internal predictive ability were evaluated by R^2_{CV}–cross validation coefficient based on leave-one-out (LOO); its value of 0.9998 being very good. Model (5) was also checked for reliability, robustness, and chance correlation by applying the y-randomization test. The y-randomization test was performed 10 times. Results of the y-randomization test are presented in Table 6.

Table 6. Results of the y-randomization test.

R_m Monocarboxylic Acids		n_D Monocarboxylic Acids		P Monocarboxylic Acids		R_m Dicarboxylic Acids		R_m Unsaturated Acids	
Iteration	R^2_{yi}	Iteration	Iteration	Iteration	R^2_{yi}	Iteration	R^2_{yi}	Iteration	R^2_{yi}
1	0.001	1	0.065	1	0.059	1	0.087	1	0.143
2	0.061	2	0.091	2	0.000	2	0.052	2	0.015
3	0.010	3	0.019	3	0.111	3	0.216	3	0.008
4	0.202	4	0.097	4	0.050	4	0.098	4	0.023
5	0.035	5	0.008	5	0.044	5	0.146	5	0.010
6	0.044	6	0.102	6	0.058	6	0.108	6	0.097
7	0.017	7	0.046	7	0.073	7	0.016	7	0.057
8	0.004	8	0.028	8	0.023	8	0.024	8	0.013
9	0.021	9	0.033	9	0.083	9	0.139	9	0.197
10	0.179	10	0.001	10	0.055	10	0.097	10	0.013

In each y-randomization run, $R^2_{yi} < 0.2$, which shows that the good results in our original model were not due to a chance correlation or structural dependency of the training set. The QSPR model (5) was statistically internally validated and this equation was used for the calculating values of the molecular refractivity for the training set, and also for the R_m predicted values of monocarboxylic acids in the test set. The results are presented in Table 1. The analysis of residuals of predicted molecular refractivity against the experimental values, in the training set, shows that the residuals only exceeded in three situations the standard deviation limits of ±2 s, in our case we had ±0.42. There were three small excesses that appeared for acids with similar structure: 2,2-dimethylpropanoic(−0.47 error), 3,3-dimethylpentanoic (+0.45 error), and 3,3-dimethylhexanoic (+0.47 error). The linear QSPR equation resulted from eliminating these three values from the correlation process is the following:

$$R_m = -2.279\ (\pm0.093) + 1.334\ (\pm0.003)\ \text{ZEP} \tag{6}$$

N = 23 R = 0.9999 R^2 = 0.9998 R^2_{CV} = 0.9998 s = 0.123 F = 199,918.5; MAE = 0.106; MAD = 0.08

By eliminating those values, the goodness of fit, the reliability, and the robustness of the QSPR model (6) are not significantly improved.

The capability of the linear model (5) to predict R_m values for monocarboxylic acids with unknown R_m, was investigated in the test set. The predicted R_m values for a series of the 24 monocarboxylic acids included in the test set were calculated with Equation (5) and are given in Table 1, together with their deviations from the corresponding experimental R_m values.

Note that the number of acids in the test set (24 acids) is close to the number of compounds in the original training set (26 acids):

$$R_m = -1.832\ (\pm0.196) + 1.326\ (\pm0.005)\ \text{ZEP} \tag{7}$$

N = 24 R_{ext} = 0.999 R^2_{ext} = 0.998 R^2_{0ext} = 0.998 $R^2_{cv\ ext}$ = 0.998 s = 0.2356 F = 66,537.3; MAE = 0.1604; MAD = 0.105

The analysis of residuals shows a single compound, that is, 2,2,3,3-tetramethylbutanoic acid, falling outside the standard deviation limits of ±2 s. All the validation strategies show that the obtained model (5) is a valid QSPR model for the prediction of molecular refractivity of monocarboxylic acids. A general QSPR model for all the 50 monocarboxylic acids was also proposed:

$$R_m = -2.081\ (\pm0.115) + 1.334\ (\pm0.003)\ \text{ZEP} \tag{8}$$

N = 50 R = 0.9999 R^2 = 0.9998 R^2_{CV} = 0.9998 s = 0.2258 F = 158,423.74; MAE = 0.159; MAD = 0.105

The obtained result suggests that our QSPR model (5) is, indeed, very good.

3.1.2. Aliphatic Dicarboxylic Acids

Dicarboxylic acids contain in their structure two functional carboxylic groups. Therefore, in our study we have considered dicarboxylic acids with a linear and symmetric structure with respect to the two functional carboxylic groups. As a consequence of this fact, the polarizability of dicarboxylic acids and the electronic interactions are stronger than in the case of monocarboxylic acids. This is the reason why, in a first step, we developed separately a QSPR model for a set of 10 aliphatic dicarboxylic acids (as a training set). Table 3 presents the values of ZEP index calculated for these acids and the experimental values of molecular refractivity. By linear regression and using a single descriptor we obtained the following equation:

$$R_m = -7.105 \, (\pm 0.019) + 1.337 \, \text{ZEP} \tag{9}$$

$$N = 10 \, R = 0.9999 \, R^2 = 0.9998 \, s = 0.0178 \, F = 7,493,020.6 \, R^2_{CV} = 0.998; \, MAE = 0.011; \, MAD = 0.005$$

The coefficient of determination $R^2 = 0.9998$ and the standard error s = 0.0178 show a very good correlation between the ZEP index and molecular refractivity for aliphatic dicarboxylic acids. The model was validated by leave-one-out cross-validation and y-randomization. The results of y-randomization are presented in Table 6. These data show, for each iteration, values of $R^2_{yi} < 0.2$, which proves the stability of the model. On the other hand, the cross-validation coefficient, $R^2_{CV} = 0.998$, illustrates the reliability of the model. The leave-one-out cross-validation predicted values are presented in Table 3. Therefore, the obtained model (9) is indeed suitable for calculating the values of molar refractions in this class of dicarboxylic acids.

3.1.3. Unsaturated Carboxylic Acids

Unsaturated carboxylic acids contain in their structure double and triple bonds, alongside the carboxylic functional group. The multiple bonds are arranged asymmetrically with respect to the carboxyl group. The multiple bonds influence the polarizability and the electronic interactions of unsaturated carboxylic acids, but this influence is less significant than in the case of dicarboxylic acids. In our study, we developed a QSPR model for molar refraction (R_m), corresponding to a set of 12 unsaturated acids (as a training set). Table 4 presents the values of ZEP index calculated for these acids and the experimental values of molecular refractivity. By applying the linear regression method and using a single descriptor we obtained the following equation:

$$R_m = 0.083 \pm 0.673 + 1.287 \pm 0.027 \, \text{ZEP} \tag{10}$$

$$N = 12 \, R = 0.998 \, R^2 = 0.995 \, S = 0.717 \, F = 2195.62 \, R^2_{CV} = 0.993; \, MAE = 0.5233; \, MAD = 0.3765$$

The coefficient of determination $R^2 = 0.995$ and the standard error s = 0.717 show a very good correlation between ZEP index and molecular refractivity for aliphatic unsaturated acids. The model was validated by leave-one-out cross-validation and y-randomization techniques. The results of y-randomization are presented in Table 6. These results show, for each iteration, values $R^2_{yi} < 0.2$, which indicates the stability of the model. The value of cross-validated coefficient, $R^2_{CV} = 0.995$, which is very close to the coefficient of determination, illustrates the reliability of the model. The leave-one-out cross-validation predicted values are also presented in Table 4. In order to check the predictive ability of the model (10), we calculated, by using this equation, the values of the molar refraction for five unsaturated carboxylic acids. The obtained values were compared with the experimental values of molar refraction existing in the literature. The differences between the experimental and predicted values were not significant. Therefore, the QSPR model (10) was shown to be very good for the calculation of molar refraction for unsaturated carboxylic acids.

At the end of our study, we applied the linear regression method to the set of 80 acids obtained by the union of the set of 50 monocarboxylic acids, the set of 13 dicarboxylic acids, and the set of 17 unsaturated carboxylic acids. We thus obtained the following QSPR model for molar refraction:

$$R_m = -0.276 \, (\pm 0.629) + 1.259 \, (\pm 0.018) \, \text{ZEP} \tag{11}$$

$$N = 80 \; R = 0.992 \; R^2 = 0.984 \; S = 1.819 \; F = 4654.81 \; R^2_{CV} = 0.993; \; \text{MAE} = 0.175; \; \text{MAD} = 0.095$$

The obtained QSPR Equation (11), modelling the molar refraction of carboxylic acids, was used to compute the molar refraction for four other carboxylic acids, not previously considered in the QSPR study. The results obtained in this way are given in Table 5. As can be seen, the maximum standard error for R_m corresponds, as expected, to Equation (11), which comprises all carboxylic acids considered in the study.

3.1.4. Building the QSPR Model for Polarizability

In Table 6 are presented the experimental values of molecular polarizability for 41 saturated aliphatic monocarboxylic acids divided into two subsets. One set containing 21 acids that will serve as training set in the QSPR modelling process, and another set containing 20 acids that form the test set, which are marked with the superscript [b] and which shall be used for testing the model by the method of external validation. The values for the polarizability of the 20 acids in the test set were obtained by conversion of molecular refractivity, using Equation (2). By correlating the molecular polarizability with ZEP index for these 21 monocarboxylic acids used as training set, the following linear QSPR model was obtained:

$$P = -0.792 \, (\pm 0.055) + 0.525 \, (\pm 0.002) \, \text{ZEP} \tag{12}$$

$$N = 21 \; R = 0.9999 \; R^2 = 0.9998 \; S = 0.0617 \; F = 81{,}178.039 \; R^2_{CV} = 0.9998 \; R^2_{CV \, ext} = 0.998; \; \text{MAE} = 0.438; \; \text{MAD} = 0.04$$

The QSPR model (12) has a very good statistical quality for fitting the calculated values of P to the experimental ones. The robustness of the model (12) and its internal predictive ability were evaluated using a R^2_{CV}–cross validation coefficient based on leave-one-out (LOO); its value of 0.9998 being very good. The model (12) was also checked for reliability, robustness, and chance correlation by applying the y-randomization test. The y-randomization test was performed 10 times. The results of the y-randomization test are presented in Table 6. In each y-randomization run, we obtained $R^2_{yi} < 0.2$, which shows that the good results in our original model were not due to a chance correlation or structural dependence of the training set.

The QSPR model (12) was statistically internally validated and then this equation was used for calculating the values of the molecular polarizability for the training set and also for the predicted values P of monocarboxylic acids in the test set. The results are presented in Table 1. Regression of the predicted polarizability against the observed molecular polarizability was $R^2_0 = 0.998$. The analysis of residuals of predicted molecular polarizability against the experimental values, in training set, showed that the residuals only once exceeded the standard deviation limits of ± 2 s, in our case ± 0.123. This corresponds to the compound 2,2-dimethylpropanoic (0.18 error). The capability of the linear model (12) to predict P values for monocarboxylic acids was investigated in the test set. The predicted values of P for a series of 20 monocarboxylic acids included in the test set, close to the number of acids in the training set, were calculated with Equation (12) and are given in Table 1, together with their deviations from the corresponding experimental values of P. The external predictive power was confirmed by $R^2_{CV \, ext} = 0.998$ and $R^2_{ext} = 0.9998$. The analysis of residuals shows a single compound, 2,2,3,3-tetramethylbutanoic acid, falling outside the standard deviation limits of ± 2 s. All the validation strategies show that the obtained model (12) is a valid QSPR model for the prediction of the polarizability of

monocarboxylic acids. A general QSPR model for all the 41 monocarboxylic acids was also proposed:

$$P = -0.806\ (\pm 0.047) + 0.527\ (\pm 0.001)\ \text{ZEP} \tag{13}$$

N = 41 R = 0.9999 R^2 = 0.9998 S = 0.083 F = 146,719.36 R^2_{CV} =0.998; MAE = 0.062; MAD = 0.04

The statistical results for Equation (13) suggest that our QSPR model (12) is very good. The obtained QSPR equations modelling the polarizability of carboxylic acids were used to compute the polarizability for four other carboxylic acids, not previously considered in the QSPR study. The results obtained in this way are given in Table 7.

Table 7. Polarizability of four external acids.

Acid	ZEP	P		
		Exp.	Pred.	Error
2-ethybutanoic	25.3088	12.43	12.49 Equation (13)	−0.06
			12.53 Equation (14)	−0.10
2,3-dimethylheptanoic	35.5218	17.92	17.86 Equation (13)	0.06
			17.92 Equation (14)	0.00
2-propilheptanoic	39.1329	19.77	19.75 Equation (13)	0.02
			19.82 Equation (14)	−0.05
tridecanoic	49.5728	25.30	25.23 Equation (13)	0.07
			25.32 Equation (14)	−0.02

Notably, the obtained values for polarizability and molar refraction increased relatively with the size and molecular weight of carboxylic acids. This fact is in agreement with the formula of Lorentz–Lorenz, which gives the relationship between polarizability, the molar refractivity, and volume [18].

3.1.5. Building QSPR Models for Refractivity Index

The molecular set considered here comprises 33 aliphatic monocarboxylic acids, with the corresponding n_D values (see Table 1), of which 22 acids were used as the training set in the modeling process and 11 acids were used as a test set for external validation, which are marked with the superscript [b]. The following QSPR model was obtained in this case:

$$n_D = 1.396\ (\pm 0.001) + 0.001\ \text{ZEP} \tag{14}$$

N = 22 R = 0.992 R^2 = 0.984 S = 0.001183 F = 1284.906 R^2_{CV} = 0.981; MAE = 0.062; MAD = 0.0015

The model was similarly validated by leave-one-out cross-validation and y-randomization techniques. The results of y-randomization are presented in Table 6. Using Equation (14), we calculated the values of the refractivity index for 11 saturated aliphatic monocarboxylic acids. The obtained values were compared with the experimental values of refractivity index existing in the literature. Therefore, the QSPR model (14) was shown to be very good for the calculation of refractivity index for saturated aliphatic monocarboxylic acids.

4. Conclusions

In this work, we presented various QSPR models, as an alternative technique to the existing additive methods, for predicting the molar refraction, polarizability, and refraction index of carboxylic acids. We used a linear regression method and a single molecular descriptor, the ZEP topological index. ZEP was calculated in a simple manner with the help of weighted electronic distances (wed), also calculated on the basis of the chemical structure of the molecules. The QSPR models developed were validated by means of a leave-one-out cross validation procedure, external validation, and y-randomization. The obtained results

show that the proposed models are simple and have a significant predictive potential. Therefore, all QSPR models thus developed, irrespective of the property by which they were constructed, i.e., for R_m, or P, or n_D, can also be applied for predicting the other two properties of carboxylic acids, in agreement with the Lorentz–Lorenz formula. This intercorrelation relationship also explains the fact that, for all QSPR models reported here, the correlation coefficients have closed values.

The results reported in this paper could be used in QSAR (quantitative structure activity relationship) for the prediction of the biological or pharmaceutical activity of carboxylic acids. Thus, the molar refraction values could be used for the estimation and prediction of the lipophilicity of a homologous series of saturated fatty acids [3], while the refraction index could be used for the estimation and prediction of the toxicity of aliphatic carboxylic acids [1]. These aspects will be considered in a future work.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Maguna, F.P.; Nunez, M.B.; Okulik, N.B.; Castro, E.A. Improved QSAR Analysis of the Toxicity of Aliphatic Carboxylic Acids. *Russ. J. Gen. Chem.* **2003**, *73*, 1792–1798. [CrossRef]
2. Verma, R.P.; Kurup, A.; Hansch, C. On the role of polarisability in QSAR. *Bioorg. Med. Chem.* **2005**, *13*, 237–255. [CrossRef]
3. Pyka, A.; Bober, K. Selected traditional structural descriptors and RM values for estimation and prediction of lipophilicity of homologous series of saturated fatty acids. *J. Am. Oil Chem. Soc.* **2006**, *83*, 747–752. [CrossRef]
4. Sakuratani, Y.; Kasai, K.; Noguchi, Y.; Yamada, Y. Comparison of predictivities of log P calculation models based on experimental data for 134 simple organic compounds. *QSAR Comb. Sci.* **2007**, *26*, 109–116. [CrossRef]
5. Shafiei, F. Relationship between Topological Indices and Thermodynamic Properties and of the Monocarboxylic Acids Applications in QSPR. *J. Math. Chem.* **2015**, *6*, 15–28. [CrossRef]
6. Atkins, P.W. *Physical Chemistry*, 6th ed.; Oxford University Press: Oxford, UK; Melbourne, Australia; Tokyo, Japan, 1998; p. 654.
7. Glasstone, S. *Textbook of Physical Chemistry*; Macmillan: London, UK, 1948; p. 543.
8. Charles, K. *Introduction to Solid State Physics*, 8th ed.; John Wiley & Sons, Inc.: New York, NY, USA, 2005; p. 464.
9. Ghose, A.K.; Crippen, G.M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21–35. [CrossRef]
10. Miller, K.J.; Savchik, J.A. A new empirical Method to calculate Average Molecular Polarizabilities. *J. Am. Chem. Soc.* **1979**, *101*, 7206–7213. [CrossRef]
11. Padrón, J.A.; Carrasco, R.; Pellón, R.F. Molecular descriptor based on a molar refractivity partition using Randic-type graph-teoretical invariant. *J. Pharm. Pharmaceut. Sci.* **2002**, *5*, 258–266.
12. Naef, R.A. Generally Applicable Computer Algorithm Based on the Group Additivity Method for the Calculation of Seven Molecular Descriptors: Heat of Combustion, LogP$_{O/W}$, LogS, Refractivity, Polarizability, Toxicity and LogBB of Organic Compounds; Scope and Limits of Applicability. *Molecules* **2015**, *20*, 18279–18351. [CrossRef] [PubMed]
13. Verma, R.P.; Hansch, A.C. A comparison between two polarisability parameters in chemical-biological interactions. *Bioorg. Med. Chem.* **2005**, *13*, 2355–2372. [CrossRef]
14. Hansch, C.; Kurup, A. QSAR of Chemical Polarizability and Nerve Toxicity. 2. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1647–1651. [CrossRef]
15. Wang, J.; Xie, X.Q.; Hou, T.J.; Xu, X.J. Fast Approaches for Molecular Polarizability Calculations. *J. Phys. Chem. A* **2007**, *111*, 4443–4448. [CrossRef]
16. Sonar, A.N.; Pawar, N.S. Studies on viscosity, density and refractive index of substituted heterocyclic compounds in different media. *Rasayan J. Cem.* **2010**, *3*, 250–254.
17. Granados, K.; Gracia-Fadrique, J.; Amigo, A.; Bravo, R. Refractive Index, Surface Tension, and Density of Aqueous Mixtures of Carboxylic Acids at 298.15 K. *J. Chem. Eng. Data* **2006**, *51*, 1356–1360. [CrossRef]
18. Golovanov, I.B.; Zhenodarova, S.M. Quantitative structure-property relationship: XXVI. Toxicity of aliphatic carboxylic acids. *Russ. J. Gen. Chem.* **2006**, *76*, 40–44. [CrossRef]
19. Weast, R.C. *CRC Handbook of Physics and Chemistry*, 68th ed.; CRC: Bwa Raton, FL, USA, 1987.
20. Berinde, Z. *Applications of Molecular Topology in The Study of Physico-Chemical Properties of Organic Compounds*; Cub Press 22: Baia Mare, Romania, 2001. (In Romanian)
21. Berinde, Z.; Berinde, M. On a matrix representation of molecular structures. *Carpathian J. Math* **2004**, *20*, 205–209.

22. Berinde, Z.M. Comparing the molecular graph degeneracy of Wiener, Harary, Balaban, Randic and ZEP topological indices. *Creat. Math. Inf.* **2014**, *23*, 165–174. [CrossRef]

23. Tropsha, A.; Gramatica, P.; Gombar, V.K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69–77. [CrossRef]

24. Roy, P.P.; Roy, K. On some aspects of variable selection for partial least squares regression models. *QSAR Comb. Sci.* **2008**, *27*, 302–313. [CrossRef]

25. Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357. [CrossRef]

26. Topliss, J.G.; Costello, R.J. Chance Correlations in Structure-Activity Studies Using Multiple Regression Analysis. *J. Med. Chem.* **1972**, *15*, 1066–1068. [CrossRef] [PubMed]

27. Kiralj, R.; Ferreira, M.M.C. Basic Validation Procedures for Regression Models in QSAR and QSPR Studies: Theory and Application. *J. Braz. Chem. Soc.* **2009**, *20*, 770–787. [CrossRef]

28. Consonni, V.; Ballabio, D.; Todeschini, R. Evaluation of model predictive ability by external validation techniques. *J. Chemom.* **2010**, *24*, 194–201. [CrossRef]