chemistry proceedings

MDPI

Proceedings Paper

Quantitative Structure-Property Relationship for the Retention Index of Volatile and Semi-Volatile Compounds of Coffee †

Cristian Rojas [1,*], Christian D. Alcívar León [2], Elizabeth Contreras Aguilar [3], Paola V. Mazón Ayala [2] and Doménica Muñoz [1]

[1] Grupo de Investigación en Quimiometría y QSAR, Facultad de Ciencia y Tecnología, Universidad del Azuay, Av. 24 de Mayo 7-77 y Hernán Malo, Cuenca, Ecuador; domenicav.munoz@gmail.com (D.M.)
[2] Facultad de Ciencias Químicas, Universidad Central del Ecuador, Francisco Viteri y Gilberto Sobral s/n, Ciudad Universitaria, Quito, Ecuador; cdalcivar@uce.edu.ec (C.D.A.L.); pvmazon@uce.edu.ec (P.V.M.A.)
[3] CEQUINOR (CONICET-UNLP), Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Bv. 120 No 1465, La Plata 1900, Argentina; econtrerasaguilar7@gmail.com (E.C.A.)
* Correspondence: crojasvilla@gmail.com
† Presented at the 25th International Electronic Conference on Synthetic Organic Chemistry, 15–30 November 2021; Available online: https://ecsoc-25.sciforum.net/.

Abstract: This study describes the development of a quantitative structure-property relationship to predict the retention index of volatile and semi-volatile compounds identified in Arabica coffee samples from different geographical origins. The analytical method utilized rapid headspace solid-phase microextraction (HSSPME)-gas chromatography-time-of-flight mass spectrometry (GC-TOFMS) data measured in the divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) fiber. A total of 102 molecules were optimized with the PM6/ZDO level of theory, in order to calculate several molecular descriptors. The ordinary least squares were coupled to the genetic algorithms supervised variable subset selection to find the best three descriptors. For model validation, the dataset was split into a training set (70%) and a test set (30%). The quality of the model was evaluated by means of the coefficient of determination and the root-mean-square error.

Keywords: Coffee; VOCs; SVOCs; PM6/ZDO; molecular descriptors; QSPR

Citation: Rojas, C.; Alcívar León, C.D.; Contreras Aguilar, E.; Mazón Ayala, P.V.; Muñoz, D. Quantitative Structure-Property Relationship for the Retention Index of Volatile and Semi-Volatile Compounds of Coffee. *Chem. Proc.* **2021**, *3*, x. https://doi.org/10.3390/xxxxx

Academic Editor: Julio A. Seijas

Published: 15 November 2021

1. Introduction

Brewed coffee is a popular, versatile and widely consumed beverage throughout the world. It is estimated that around 400 billion cups of coffee are consumed annually [1]. Hence, it is one of the most important agricultural products in the international trade of coffee producing countries, and adds approximately 35 billion US dollars per year to the overall world economy [2]. According to The International Coffee Organization (ICO) there was an increase in coffee consumption of 4.5 % between the years 2016–2021 [3]. In addition to the coffee culture and pleasure of the aroma and taste, bioactive components of coffee that include alkaloids, terpenes, phenolic compounds, and other secondary metabolites have been associated with antioxidant and anti-inflammatory properties that are associated with improved human health [4].

The aroma of coffee is one of its most relevant organoleptic characteristics, which is associated with volatile organic compounds (VOCs) and semi-volatile organic compounds (SVOCs) [5]. VOCs are strong-smelling molecules that contribute to the definition of the aroma and flavor profile of coffee, while the semi-volatile organic compounds (SVOCs) are a subgroup of VOCs exhibiting both high molecular weight and high boiling point temperatures. These chemicals are fundamental for defining the organoleptic quality of coffee developed during the production process [6,7]. In order to study the composition of these compounds, the rapid headspace solid-phase microextraction (HS-SPME)-

gas chromatography-time-of-flight (high-speed data acquisition rate option) mass spectrometry (GC-TOFMS), using the DVB/CAR/PDMS (divinylbenzene/carboxen/polydimethylsiloxane) fiber, were used for analysis [8]. This fiber was shown to be the most suitable stationary phase for analyzing complex components with different polarities in food matrices, due to its capability to deal with high temperatures. In this context, our research group developed a QSPR model to predict the retention indices (*I*) of diverse volatiles detected in the headspace of rice using the DVB/CAR/PDMS fiber in the solid-phase microextraction-gas chromatography-mass spectrometry (SPME-GC-MS) analysis [9].

The aim of the work presented here is the development of a computational model based on the quantitative structure-property relationship (QSPR) for the study 102 VOCs and SVOCs detected in coffee samples from different origins. The specific property endpoint is the calculation of a retention index quantified in the DVB/CAR/PDMS (divinylbenzene/carboxen/polydimethylsiloxane) fiber by means of a coupled system of rapid headspace solid-phase microextraction (HS-SPME)-gas chromatography-time-of-flight (high-speed data acquisition rate option) mass spectrometry (GC-TOFMS). Compounds were optimized according to quantum chemical calculations, and represented by several molecular descriptors and fingerprints. Then, the unsupervised V-WSP variable reduction was used to obtain a pool of the most relevant descriptors to be submitted to the Genetic Algorithms-variable subset selection (GAs-VSS) coupled to the ordinary least squared (OLS) to search for an optimal model. The QSPR model was evaluated by diverse internal and external validation approaches, as well as the applicability domain assessment. As a complement, the mechanistic of action of each molecular descriptors used to predict the *I* of the VOCS and SVOCs was provided. The model was developed following the five principles stated by the Organization for Economic Co-operation and Development [10].

2. Materials and Methods

2.1. Database Description

For this study, we considered the retention index (*I*) of 102 volatile and semi-volatile organic compounds identified in Arabica coffee samples from different geographical origins (Brazil, Colombia, Costa Rica, Guatemala, Ethiopia and Indonesia) [8]. The experimental *I* was obtained through rapid headspace solid-phase microextraction (HS-SPME)-gas chromatography-time-of-flight (high-speed data acquisition rate option) mass spectrometry (GC-TOFMS). A system of the CombiPAL SPME autosampler with a gas chromatograph 6890 coupled to a Pegasus III mass spectrometer was used for the automatization of the SPME procedure. In addition, the mass spectrometer was equipped with a time-of-flight mass analyzer. The SLB-5 column (10 m × 180 μm × 0.18 μm) was used for the separation of volatile and semi-volatile compounds. This column is constituted by 5% diphenyl and 95% dimethylpolysiloxane. Subsequently, a fiber optimization experiment was performed in order to evaluate diverse commercially available fiber coatings under the same GC and MS conditions. The divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) 50/30 μm was proved optimal for the quantification of the retention indices of the volatile and semi-volatile organic compounds present in coffee samples. Refer to Table S1 for details of the retention indices of the 102 compounds. To the best of our knowledge, no computational modeling has been previously conducted with this dataset.

2.2. Molecular Representation and Geometry Optimization

The chemical name of each of the compounds **[1–102]** reported in Table S1 was used for retrieve the PubChem CID and the CAS registry number in the PubChem open library [11]. In addition, the .sdf file format of each compound was also obtained for molecule visualization and optimization. Initially, the alvaMolecule software [12] was used for the molecule standardization and curation. These processes include the verification of each query through several filters, which performed the standardization of benzene rings into aromatic form, converted unusual covalent bonds to ionic forms, added charge to

quaternary nitrogen atoms, removed/added exceeding/missing hydrogens, and standardized nitro, azide and diazo groups. The canonical SMILES (simplified molecular input line entry system) notation of each compound was also generated in alvaMolecule. Then, an initial optimization of all molecules was development with the Avogadro program using the UFF force field and the algorithm steepest descent [13]. The optimized conformations were then used to performed quantum chemical calculations in the ground state (gas phase) of the compounds with the program package Gaussian 09, Rev D.01 [14]. The final optimization was performed with the semi-empirical method PM6/ZDO. In addition, the frontier orbital energies (HOMO and LUMO) and global reactivity parameters associated with chemical reactivity were calculated as 3D molecular descriptors. In this way, the gap of energy between HOMO and LUMO (ΔE) allowed the estimation of the stability of the molecules. The electron affinity (A) and ionization potential (I) can be defined as $A = -E_{LUMO}$ and $I = -E_{HOMO}$, respectively. Moreover, descriptors associated with the electronic structure and chemical reactivity as the electronegativity ($\chi = (I + A)/2$), chemical potential ($\mu = -\chi$), chemical hardness ($\eta = \Delta E/2$), chemical softness ($\sigma = 1/2\eta$), electrophilicity index ($\omega = \mu 2/2\eta$) and nucleophilicity index ($N = 1/\omega$) were also calculated [15,16].

2.3. Molecular Descriptors Calculation and Reduction

For the development of the QSAR model, a new set of 5663 molecular descriptors (MDs) and 166 MACCS (Molecular ACCess System) fingerprints were computed in the alvaDesc software [17]. The molecular descriptor is a useful number (or the result of a standardized experiment) obtained from a well-defined mathematical algorithm applied to a symbolic representation of molecules [18]. In a first attempt to reduce the data dimensionality, descriptors with constant values and near constant values were excluded along with descriptors with missing values. These descriptors were merged with the 3D ones calculated in Gaussian. Subsequently, the unsupervised variable reduction based on the algorithm of Wootton, Sergent and Phan-Tan-Luu (V-WSP) was applied [19]. The V-WSP method uses a correlation threshold (defined by the user) in order to reduce the presence of descriptors with redundancy, multicollinearity and noise, in such a way as to obtain an optimal pool of descriptors with minimal correlation in multidimensional space.

2.4. Molecular Descriptor Selection

For the model development, a crucial step is the selection of a pool of the most important molecular descriptors for the multiple linear regression (MLR) based on the ordinary least squares (OLS). Among the diverse methods available for the variable selection, the Genetic Algorithms-Variable Subset Selection (GA-VSS) technique [20] was coupled with the OLS method in order to find the optimal subset of molecular descriptors. This supervised variable selection starts with an initial random population of models (i.e., chromosomes) constituted by binary vectors indicating the presence (or absence) of each descriptor in the model. Then, new models are created through an evolutionary process by the combination of chromosomes (models) of the initial population (crossover), as well as by randomly including (or excluding) descriptors (mutation). During the evolution of the population, the root-mean-square error (RMSE) in cross-validation of venetian blinds is optimized. The model development along with the descriptor selection was performed in the alvaModel software [21].

2.5. Validation of the Model

The merit of a QSPR model is related to the ability to correctly predict the property of an external set of VOCs and SVOCs, commonly not considered during the calibration of the model. Consequently, the dataset should be split into a training set and a test set in such a way as to guarantee a structure-property representation of the test set molecules into the space defined by the volatile and semi-volatile compounds of the training set. To this end, the Balanced Subsets Method (BSM) [22] was used for the partition of 112

compounds in the dataset. The essence of the BSM is to create groups based on the *k*-means cluster analysis, in order to identify similar volatile and semi-volatile compounds according to the descriptors (structure) and the retention index (property). During the genetic algorithms supervised selection, the training set was used to calibrate the MLR model, while the test set was used to measure the predictive ability of the QSPR model. The computational model was further validated by the leave-one-out and the leave-many-out cross-validation approaches. The former procedure excludes one compound from the training set at a time, while the latter randomly excludes a user-defined percentage of the molecules at each iteration. In the leave-many-out cross-validation, we applied the venetian blinds approach, the Monte Carlo random sub-sampling validation, as well as the Bootstrap method [23].

2.6. Applicability Domain

Since QSPR models are reductionist models associated with the reliability of the predictions of certain molecules, it is necessary to define the theoretical region (chemical space) defined by the MDs within the model that performs reliable predictions for new molecules. In this work, the leverage approach was used to determine if a compound of the test set lies within (or outside) this theoretical space [24]. This approach quantifies the distance of each test molecule with respect to the centroid of the training compounds defined by the Hat matrix (**X** matrix of descriptors only). Thus, it is possible to define a threshold value when the warning leverage is equal to $h^* = 3p/n$ (p is the number of parameters in the QSPR model and n is the number of training set molecules). Then, if the leverage value of each volatile or semi-volatile organic compound in the test set (h_{ii}) is lower than the defined threshold, the predicted I could be considered reliable. The applicability domain analysis was performed in the alvaModel software [21].

3. Results and Discussion

After the exclusion of non-informative alvaDesc descriptors, a pool of 3006 MDs was merged with the 11 quantum descriptors. Then, the V-WSP variable reduction was applied at a threshold value 0.95 (thr = 0.95). Thus, 1237 descriptors were submitted to the supervised selection through the GAs-VSS coupled with the OLS method to develop the QSPR model. For splitting the 102 VOCs and SVOCs, the BSM was used to define the training set (71 molecules) and the test set (31 molecules). Refer to Table S1 for the training set and test set assignments. Subsequently, the 71 training molecules were used for the GAs-VSS coupled with the OLS method in order to search for the optimal pool of MDs. During the supervised selection, the root-mean-square error (RMSE) in cross-validation of venetian blinds was optimized (minimized) to avoid overfitting the models. The coefficient of determination (R^2) was also considered as a parameter of the model. Thus, a three-variable model was retained as the QSPR model for further analysis:

$$I = 327.63 + 5.24\,MW + 375.44\,MDEN\text{-}23 - 385.26\,Mor13v \quad (1)$$

The statistical parameters for the training set ($R^2 = 0.920$ and $RMSE = 71.78$) and the test set ($R^2 = 0.897$ and $RMSE = 81.50$) reflected negligible differences and indicated the absence of potential overfitting in the model. Consequently, an appropriate QSPR model for predicting the I property was achieved. In addition, the model derived by Equation (1) was subjected to several cross-validation approaches: leave-one-out ($R^2 = 0.869$ and $RMSE = 91.74$), venetian blinds ($R^2 = 0.872$ and $RMSE = 90.62$), Monte Carlo 20% out with 1000 iterations ($R^2 = 0.870$ and $RMSE = 90.74$) and the Bootstrap with 1000 iterations ($R^2 = 0.867$ and $RMSE = 92.97$). Details of the predicted I by Equation (1) is available in Table S1, while the numerical values for the three MDs are available in Table S2.

Figure 1a shows the relationship between the experimental and predicted retention indices obtained by Equation (1). This figure suggested that this property has a linear relationship around the perfect fit line. Complementary, Figure 1b shows the dispersion of the residuals vs. the predicted I, which suggest a random distribution of the residuals

around the zero line. In this model, no training molecules lie outside the limit of ± 2 times the RMSE.

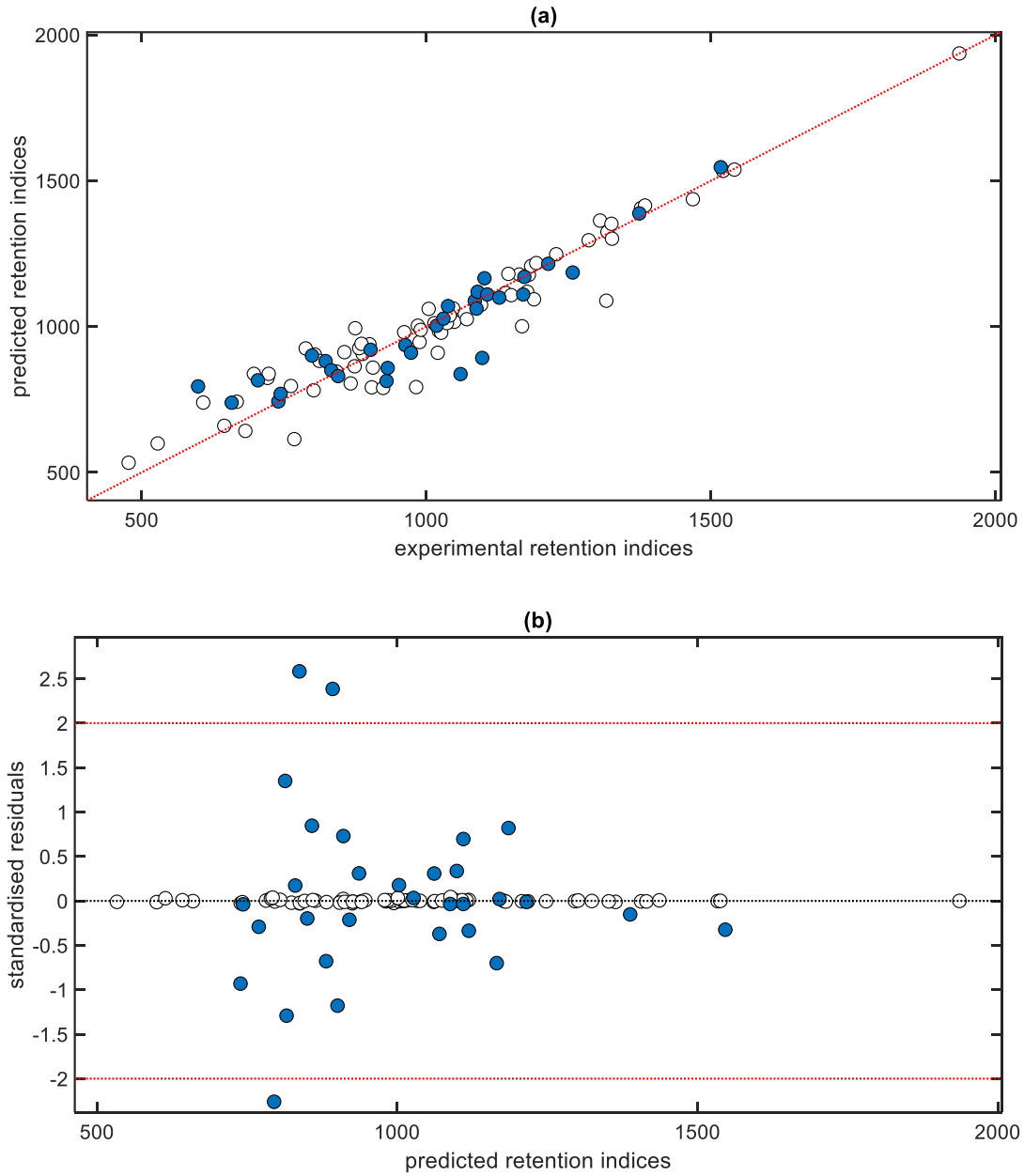

Figure 1. (**a**) experimental versus predicted retention indices for the volatile and semi-volatile organic compounds of coffee detected in the DVB/CAR/PDMS fiber in the HS-SPME- GC-TOFMS technique. (**b**) standardized residuals versus the predicted retention indices for the VOCs and SVOCs identified in coffee samples. Training set molecules are labeled in black and test set compounds are labeled in blue.

The mechanism of action of the retention index phenomenon presented in Equation (1) was constituted by two conformation-independent molecular descriptors: molecular weight (*MW*) and the molecular distance edge between all secondary and tertiary nitrogen atoms (*MDEN-23*). In addition, the signal 13/weighted by van der Waals volume (*Mor13v*) conformation-dependent descriptor complements the topological information of the retention phenomenon. The maximum correlation (R^2 = 0.189) between the *MW* and the *Mor13v* suggested a low internal correlation in the QSPR model. Consequently, each molecular descriptor contributed to particular aspects for the retention mechanism in the DVB/CAR/PDMS fiber in the HS-SPME-GC-TOFMS system. The standardized regression

coefficients of the model indicate the degree of contribution of each descriptor in predicting the *I* property: *MW* (0.75) > *Mor13v* (0.25) > *MDEN-23* (0.21).

The molecular weight is a zero-order constitutional index calculated as the sum of the atomic weights of all the atoms present in a molecule; that is, it describes the molecular size of both volatile and semi-volatile organic compounds [18]. Since this descriptor represents a linear group contribution (atomic masses) for predicting the retention index property, the larger the *MW* for a given compound, the greater the retention index of that compound. In a recent study, the synergistic effect of the *MW* for predicting the *I* property in the comprehensive two-dimensional gas chromatography combined with quadrupole-mass spectrometry (GC × GC/qMS) using the BPX5 and BP20 column coupled system was presented [25]. On the other hand, *MDEN-23* is a descriptor belonging to the Molecular Distance Edge (MDE) vector [26]. This descriptor measures the topological distances (2D) between nitrogen atoms; particularly, type 2 for secondary (-NH-) and type 3 for ternary (-N<) nitrogen atoms. Thus, the synergistic effect on the retention time could be related to the silanophile effect. In fact, when dealing with the *I* of pesticides in ultrahigh-performance liquid chromatography electrospray ionization quadrupole-Orbitrap (UHPLC/ESI Q-Orbitrap) mass spectrometry (MS), using the Hypersil Gold selectivity column (and the guard column Accucore aQ), it was shown that the slow elution of polar molecules (pesticides) through the stationary phase was caused by the high affinity of basic amine compounds for the silica surface of the column (which was constituted by active or acidic silanol groups) [27].

The model presented here also considers a conformation-dependent descriptor belonging to the 3D-MoRSE (3D Molecule Representation of Structures based on Electron diffraction). These descriptors transform the 3D coordinates of compounds into a molecular code applying a modified equation used in electron diffraction studies for preparing theoretical scattering curves [28,29]. Specifically, the *Mor13v* descriptor considers the scattering parameter s = 12 $Å^{-1}$ and weighs the atoms by the corresponding van der Waals volume. This weighting scheme considers a minimum effect of hydrogen atoms, decreases the contribution of nitrogen, oxygen and fluorine, and provides significant influence to silicon, phosphorus, bromine and iodine atoms. Since the coefficient of *Mor13v* is negative, this descriptor could be related to the contribution of the volume of the volatile and semi-volatile compounds expressed in terms of the van der Waals volume. Thus, higher retention indices are related to molecules having in their scaffold fragments (pairwise atoms) with high volume, which in turn interact with the stationary phase and delay the time elution.

Finally, the applicability domain of the QSPR model was analyzed in order to define the theoretical space where the model makes reliable predictions of the retention index of new VOCs or SVOCs (i.e., interpolations). The leverage approach (Figure 2) stablished a threshold limit $h^* = 0.1408$, which indicated that predictions were reliable to only VOCs or SVOCs with a leverage value below this threshold limit; that is, predictions are the result of interpolation of the model (i.e., reliable). In this work, no test set compounds fell outside the AD of the model, indicating that Equation (1) makes reliable predictions of the retention index property, and consequently, could be useful for eliciting this property of new VOCs and SVOCs of different coffee samples.

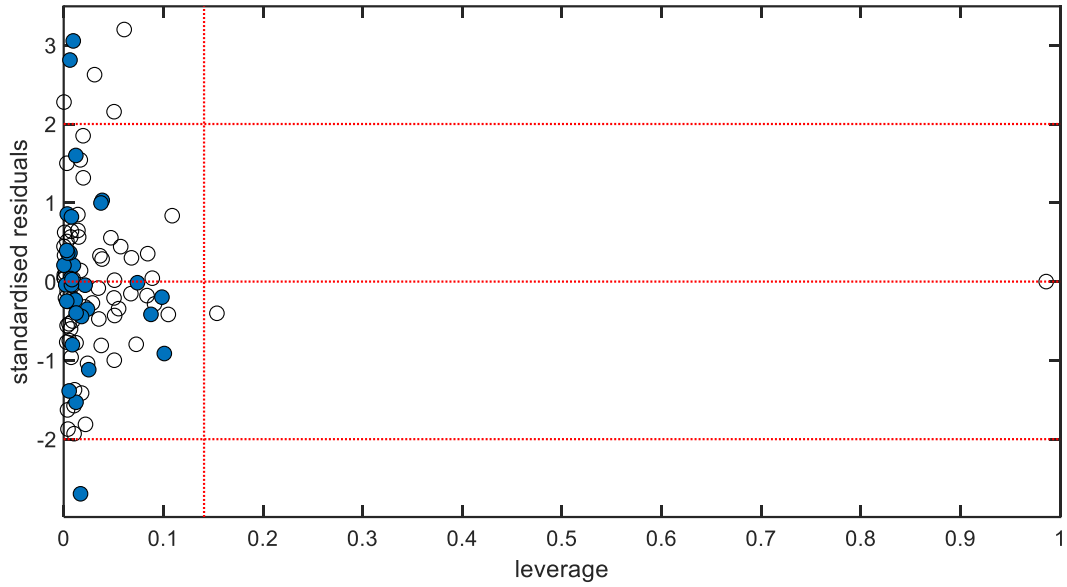

Figure 2. William plot for defining the applicability domain of the QSPR model. Training set molecules are labeled in black and test set compounds are labeled in blue.

4. Conclusions

In this study, we developed a computational model based on the quantitative structure-property relationship for the retention index of 102 volatile and semi-volatile organic compounds detected in coffee samples. A 3D chemical structure based on the PM6/ZDO of each molecule was used to calculate several molecular descriptors and fingerprints. To handle the large number of variables, the V-WSP unsupervised reduction was applied to obtain a pool of useful descriptors to be used to calibrate the multiple linear regression model coupled with the genetic algorithms variable subset selection. The three-descriptor predictive model was validated through several internal and external criteria, according to the five principles stated by the OECD to make it applicable to predict the retention index of new volatile and semi-volatile organic compounds present in coffee samples of diverse origin by means of the HS-SPME-GC-TOFMS quantified in the DVB/CAR/PDMS.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Table S1: Details of the dataset: chemical names, PubChem CID, CAS registry number, canonical SMILES, and the retention indices (*I*) for the 102 VOCs and SVOCs by the HS-SPME-GC-TOFMS, as well as the training set and test set assignments using the BSM technique, Table S2: Numerical values for the three molecular descriptors for each of the 102 VOCs and SVOCs in the computational model.

Author Contributions: C.R.: conceptualization, methodology, formal analysis, validation and writing—original draft preparation; C.D.A.L.: methodology, software, formal analysis and writing—original draft preparation; E.C.A.: methodology, software, formal analysis and writing—original draft preparation; P.V.M.A.: software, data curation and validation; D.M.: data curation and validation. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement:

Acknowledgments: We thank Wayne R. Hanson for his valuable revision of the manuscript and for providing some useful comments for improving the technical quality of it. Christian D. Alcívar León and Paola V. Mazón Ayala thank the financial support from the Central University of Ecuador (Grant, DI-CON-2019-007), Faculty of Chemical Sciences.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fernandes, A.P.; Santos, M.C.; Lemos, S.G.; Ferreira, M.M.; Nogueira, A.R.A.; Nóbrega, J.A. Pattern recognition applied to mineral characterization of Brazilian coffees and sugar-cane spirits. *Spectrochim. Acta B: At. Spectrosc.* **2005**, *60*, 717–724, https://doi.org/10.1016/j.sab.2005.02.013.
2. Dos Santos, É.J.; de Oliveira, E. Determination of mineral nutrients and toxic elements in Brazilian soluble coffee by ICP-AES. *J. Food Compos. Anal.* **2001**, *14*, 523–531, https://doi.org/10.1006/jfca.2001.1012.
3. Torga, G.N.; Spers, E.E. Perspectives of global coffee demand. In *Coffee Consumption and Industry Strategies in Brazil*; de Almeida, L.F., Spers, E.E., Eds.; Woodhead Publishing: Duxford, UK, 2020; pp. 21-49.
4. de Melo Pereira, G.V.; de Carvalho Neto, D.P.; Júnior, A.I.M.; do Prado, F.G.; Pagnoncelli, M.G.B.; Karp, S.G.; Soccol, C.R. Chemical composition and health properties of coffee and coffee by-products. *Adv. Food Nutr. Res* **2020**, *91*, 65–96, https://doi.org/10.1016/bs.afnr.2019.10.002.
5. Wang, T.; Shanfield, H.; Zlatkis, A. Analysis of trace volatile organic compounds in coffee by headspace concentration and gas chromatography-mass spectrometry. *Chromatographia* **1983**, *17*, 411–417, https://doi.org/10.1007/BF02262920.
6. Pimenta, C.J.; Angélico, C.L.; Chalfoun, S.M. Challengs in coffee quality: Cultural, chemical and microbiological aspects. *Ciênc. Agrotec.* **2018**, *42*, 337–349, https://doi.org/10.1590/1413-70542018424000118.
7. Sharma, H. A detail chemistry of coffee and its analysis. In *Coffee: Production and Research*; Castanheira, D.T., Ed.; IntechOpen: 2020.
8. Risticevic, S.; Carasek, E.; Pawliszyn, J. Headspace solid-phase microextraction–gas chromatographic–time-of-flight mass spectrometric methodology for geographical origin verification of coffee. *Anal. Chim. Acta* **2008**, *617*, 72–84, https://doi.org/10.1016/j.aca.2008.04.009.
9. Rojas, C.; Tripaldi, P.; Pérez-González, A.; Duchowicz, P.R.; Pis Diez, R. A retention index-based QSPR model for the quality control of rice. *J. Cereal Sci.* **2018**, *79*, 303–310, https://doi.org/10.1016/j.jcs.2017.11.004.
10. OECD. *Guidance Document on the Validation of (Quantitative)Structure-Activity Relationships [(Q)SAR] Models*; OECD: 2014.
11. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B. PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Res.* **2019**, *47*, D1102–D1109, https://doi-org.proxy.unimib.it/10.1093/nar/gky1033.
12. Alvascience. alvaMolecule (Software to View and Prepare Chemical Datasets) Version 1.0.4. 2020. Available online: https://www.alvascience.com (accessed on).
13. Hanwell, M.D.; Curtis, D.E.; Lonie, D.C.; Vandermeersch, T.; Zurek, E.; Hutchison, G.R. Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform.* **2012**, *4*, 17, https://doi.org/10.1186/1758-2946-4-17.
14. Millam, J.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J.; Martin, R.; Morokuma, K.; Farkas, O.; Foresman, J.; Fox, D. *Gaussian 16, Rev. C. 01*; Gaussian, Inc.: Wallingford, CT, USA, 2016.
15. Yılmaz, Z.T.; Odabaşoğlu, H.Y.; Şenel, P.; Adımcılar, V.; Erdoğan, T.; Özdemir, A.; Gölcü, A.; Odabaşoğlu, M. A novel 3-((5-methylpyridin-2-yl) amino) isobenzofuran-1 (3H)-one: Molecular structure describe, X-ray diffractions and DFT calculations, antioxidant activity, DNA binding and molecular docking studies. *J. Mol. Struct.* **2020**, *1205*, 127585, https://doi.org/10.1016/j.molstruc.2019.127585.
16. Chandran, K.; Seetharamiah, N.K.; Sambanthan, M.; Anandhan, M.; Venkatesan, R. Crystal Structure, Spectral investigations, DFT and Antimicrobial activity of Brucinium Benzilate (BBA). *J. Mol. Model.* **2021**, in press, https://doi.org/10.21203/rs.3.rs-343924/v1.
17. Alvascience. alvaDesc (Software for Molecular Descriptors Calculation) Version 2.0.9. 2021. Available online: https://www.alvascience.com (accessed on).
18. Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; WILEY-VCH: Weinheim, Germany, 2009.
19. Ballabio, D.; Consonni, V.; Mauri, A.; Claeys-Bruno, M.; Sergent, M.; Todeschini, R. A novel variable reduction method adapted from space-filling designs. *Chemom. Intell. Lab. Syst.* **2014**, *136*, 147–154, https://doi.org/10.1016/j.chemolab.2014.05.010.
20. Leardi, R. Genetic algorithms in chemistry. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, 2nd ed.; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier: Amsterdam, The Netherlands, 2020; Volume 1, pp. 617–634.
21. Alvascience. alvaModel (Software to Model QSAR Data) Version 2.0.0. 2021. Available online: https://www.alvascience.com (accessed on).
22. Rojas, C.; Duchowicz, P.R.; Tripaldi, P.; Pis Diez, R. QSPR analysis for the retention index of flavors and fragrances on a OV-101 column. *Chemom. Intell. Lab. Syst.* **2015**, *140*, 126–132, https://doi.org/10.1016/j.chemolab.2014.09.020.
23. Varmuza, K.; Filzmoser, P. *Introduction to Multivariate Statistical Analysis in Chemometrics*; CRC Press: Boca Raton, FL, USA, 2009.
24. Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **2012**, *17*, 4791–4810, https://doi.org/10.3390/molecules17054791.
25. Rojas, C.; Duchowicz, P.R.; Castro, E.A. Foodinformatics: Quantitative structure-property relationship modeling of volatile organic compounds in peppers. *J. Food Sci.* **2019**, *84*, 770–781, https://doi.org/10.1111/1750-3841.14477.
26. Liu, S.; Cao, C.; Li, Z. Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance-edge (MDE) vector, λ. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387–394, https://doi.org/10.1021/ci970109z.

27. Rojas, C.; Aranda, J.F.; Jaramillo, E.P.; Losilla, I.; Tripaldi, P.; Duchowicz, P.R.; Castro, E.A. Foodinformatic prediction of the retention time of pesticide residues detected in fruits and vegetables using UHPLC/ESI Q-Orbitrap. *Food Chem.* **2021**, *342*, 128354, https://doi.org/10.1016/j.foodchem.2020.128354.
28. Schuur, J.H.; Selzer, P.; Gasteiger, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334–344, https://doi.org/10.1021/ci950164c.
29. Devinyak, O.; Havrylyuk, D.; Lesyk, R. 3D-MoRSE descriptors explained. *J. Mol. Graph. Model.* **2014**, *54*, 194–203, https://doi.org/10.1016/j.jmgm.2014.10.006.