Journal of Chromatography A 1730 (2024) 465146

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Prediction of retention data of phenolic compounds by quantitative structure retention relationship models under reverse-phase liquid chromatography

Roberto Laganà Vinci [a], Katia Arena [a], Francesca Rigano [a,*], Francesco Cacciola [b], Paola Dugo [a,c], Luigi Mondello [a,c]

[a] Messina Institute of Technology c/o Department of Chemical, Biological, Pharmaceutical and Environmental Sciences, former Veterinary School, University of Messina, Viale G. Palatucci snc 98168 – Messina, Italy
[b] Dipartimento di Scienze Biomediche, Odontoiatriche e delle Immagini Morfologiche e Funzionali, Università degli Studi di Messina, Via Consolare Valeria, Messina 98125, Italy
[c] Chromaleont s.r.l. c/o Department of Chemical, Biological, Pharmaceutical and Environmental Sciences, former Veterinary School, University of Messina, Viale G. Palatucci snc 98168 – Messina, Italy

ARTICLE INFO

Keywords:
QSRR
Flavonoids
Phenolic compounds
Molecular descriptors
Bergamot juice

ABSTRACT

Quantitative Structure-Retention Relationship models were developed to identify phenolic compounds using a typical LC- system, with both UV and MS detection. A new chromatographic method was developed for the separation of fifty-two standard phenolic compounds. Over 5000 descriptors for each standard were calculated using AlvaDesc software and then selected through Genetic Algorithm. The selected descriptors were used as variables for models construction and to obtain a better understanding of the retention behaviour of phenols during reverse-phase separation. Three distinct molecule sets, including fifty-two phenolic compounds (Set 1), 32 flavonoids (Set 2) and 15 mono-substituted flavonoids were divided into training and validation sets to build Partial Least Square, Multiple Linear Regression and Partial Least Square-Artificial Neural Network models. To assess the predictivity of the models, these were tested on a bergamot juice sample. Partial Least Square and Partial Least Square-Artificial Neural Network exhibit the lowest prediction error, and the latter showed the best predictive power in real sample recognition. The building and implementation of such predictive models showed to be a powerful tool to identify phenolic compounds based on retention data and avoiding the use of expensive and sophisticated detectors such as tandem MS.

1. Introduction

In recent years, the growing interest in the "functional foods" field enhanced the necessity of investigating the composition of several foods and spices to obtain a better understanding, by *silico* or clinical trials, of the health benefits that different classes of compounds can provide.

Among them, phenolic compounds are of utmost importance. They are plants secondary metabolites, made by at least one aromatic ring and one hydroxyl group, which are mainly synthesised by the phenyl-propanoid metabolic pathway [1,2]. Currently, about 8000 polyphenols have been identified and classified into five main families: phenolic acids, flavonoids, stilbenes, lignans and tannins. In plants, these compounds play a defence role against UV radiation and pathogens' aggression, while in foods they participate in the organoleptic profile, contributing to bitterness, flavour, odour, colour and oxidative stability [2,3]. The well-known scientific interest in these molecules for human health is due to their antioxidant power and the observed correlation between the consumption of polyphenol-rich foods and the decrease in the risk of developing chronic diseases. In particular, these molecules showed cardio- and neuro-protective effects and seem to participate in the protection against ageing, cancer and diabetes [3].

Due to their enormous complexity in the plant kingdom, their separation and identification represents an intriguing challenge, especially when is not supported by literature data. For this purpose, several analytical methods, in particular reversed phase HPLC (RP-HPLC) and multidimensional LC techniques, hyphenated with PDA, MS or MS/MS,

* Corresponding author.
E-mail address: frigano@unime.it (F. Rigano).

https://doi.org/10.1016/j.chroma.2024.465146
Received 26 April 2024; Received in revised form 3 July 2024; Accepted 4 July 2024
Available online 9 July 2024

were developed for the characterisation of phenolic compounds in different matrices, however, most of them require to be optimised depending on the matrix analysed, due to the huge polarity range covered by this molecule family [4,5].

Briefly, the main issue for a systematic separation and identification of phenolic compounds is due to the large number of molecules, often with similar polarity, belonging to this class; moreover, most of these possess the same phenolic precursor bonded with acyl, glycoside or phenolic groups.

Among them, most compounds belong to flavonoids, which can polymerize with other flavonoids (condensed tannins) or be functionalized, in specific bonding sites, with hydroxy-, methoxy-, acyl- and glycosyl- groups, but also with hydroxybenzoic or hydroxycinnamic acids. Other phenolic compounds such as hydroxybenzoic and hydroxycinnamic acids can be found bonded with the abovementioned functional groups and polyols too, but even polymerized with one or more phenolic acids in a larger variety of bonding sites [6,7] On this basis, the understanding of the retention behaviour based only on structure observation becomes quite challenging.

In this context, a powerful and useful tool can be the Quantitative Structure-Retention Relationship (QSRR) theory, which finds correlations, through computational data processing, between molecular descriptors and retention times (t_R). In other terms, molecules' physico-chemical properties could be used to predict molecules' retention time [8]. Currently, to the best of the author's knowledge, only few papers were centered on applying the QSRR theory to predict retention times of phenolic compounds and study the physico-chemical properties involved in their reversed-phase separation. In particular, among the studies conducted on t_R prediction, two were focused on flavonoid aglycone positional isomers [9,10], one on hydroxycinnamic acid amides [11] and one on phenolic compounds belonging to different classes, including small phenolic acids and larger flavonoid molecules [12]. However, all these research works combined the use of prediction models with spectral databases (i.e. MS/MS libraries), thus increasing total analysis cost and the required level of operator expertise.

The scope of this work was to develop an RP-HPLC method for the separation of fifty-two phenolic compounds, aiming to find a correlation between structure and retention data through the use of molecular descriptors of chemical structures optimized through density functional theory (DFT). Moreover, the construction of QSRR models allowed t_R prediction for unknown compounds, avoiding the use of sophisticated and expensive detectors such as MS/MS ones for their univocal and reliable identification. Fig. 1 summarizes the typical workflow for the application of QSRR models. The optimization of the geometrical structure of the molecule is the first step, essential to determine the most stable conformation of the molecule capable of specific interaction with both stationary and mobile phases depending on its energy, electronic density, steric hindrance, dipole moment. In this regard, the DFT approach is more efficient than faster semi-empirical optimization, which takes into account only correlation between valence electrons, neglecting other electron correlations [13].

In parallel to geometry optimization leading to the determination of molecular descriptors, an analytical method needs to be developed to achieve an efficient separation of target analytes. Then, experimental retention times undergo a statistical treatment to find correlations with previously found molecular descriptors. Such correlations represent the QSRR models from which t_R of not tabulated compounds can be predicted and the separation mechanism fully elucidated.

2. Materials and methods

2.1. Chemicals and samples

All solvents were purchased from Merck KGaA (Darmstadt, Germany). LC-MS grade water, acetonitrile (ACN) and formic acid were used for HPLC-PDA/ESI-MS analyses.

Among a total of fifty-two phenolic compounds, twenty-seven standards, namely (-)Epicatechin, (+)Catechin, 1,5-Dicaffeoylquinic acid, 4-Methoxycinnamic acid, 4-O-caffeoylquinic acid, Apigenin, Caffeic acid, Daidzein, Ethyl gallate, Ferulic acid, Gallic acid, Genistein, Hesperidin, Naringenin, Naringin, N-trans-caffeoyltyramine, Oleuropein, p-coumaric acid, p-hydroxybenzoic acid, Protocatecuic acid, Pyrogallol, Quercetin, Salvianolic acid B, Sinapic acid, Syringol, Tyrosol and Verbascoside were purchased from Merck KGaA. Additional twenty-five phenolic standards, namely, (-) Gallocatechin, (+) Taxifolin, 3-O-caffeoylquinic acid, Daidzin, Eriocitrin, Eriodictyol, Eriodictyol-7-O-glucoside, Genistin, Isoquercetin, Isorhamnetin, Isorhamnetin-3-O-glucoside, Kaempferol, Kaempferol-3-O-glucoside, Luteolin, Luteolin-3′,7′-di-O-glucoside, Luteolin-7-O-glucoside, Myricetin, Narirutin, Neodiosmin, Nobiletin, Poncirin, Quercitrin, Rutin, Syringic acid and Tangeretin were purchased from Extrasynthese. The chemical structure of the employed standards is reported in Fig. S1. For each phenolic compound a stock solution (2000 mg L^{-1}) was prepared in methanol (MeOH), ethanol (EtOH), water, dimethylformamide (DMF) and/or dimethyl sulfoxide (DMSO) depending on their solubility (Table S1), then five multi-analyte solutions, containing 10/11 compounds each, were prepared to discriminate between molecule with the same molecular weight. Afterwards a final multi-analyte solution containing all

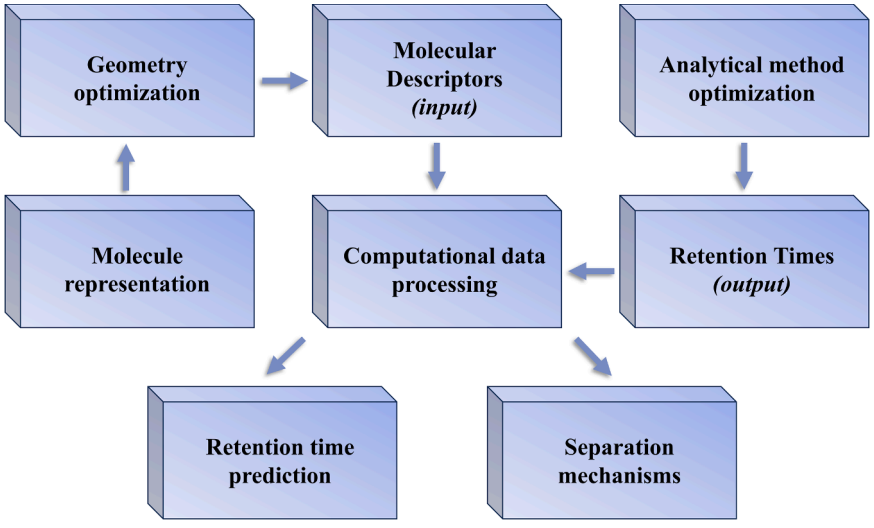

Fig. 1. QSRR Workflow.

the phenolic compounds (20 mg L^{-1} each) was prepared in MeOH and employed for optimising the separation and building the QSRR models. A commercial bergamot juice was bought from a local market and analysed to evaluate the models with a real sample. Before injection, the bergamot juice was only centrifuged and filtered through an Acrodisc filter 0.45 μm (Merck KGaA, Darmstadt, Germany), as previously described by Russo et al. [14].

2.2. HPLC-PDA-ESI/MS analysis

The HPLC-PDA/ESI-MS analysis of the phenolic standards mix was performed on a Shimadzu HPLC system (Shimadzu, Duisburg, Germany) equipped with a CBM-20A controller, two LC-30AD dual-plunger parallel-flow pumps, a CTO-20A column oven, a SIL-30AC autosampler and an SPD-M30A photodiode array detector coupled with an LCMS-2020 single quadrupole mass spectrometer equipped with an ESI interface (Shimadzu, Duisburg, Germany). Phenolic compounds were separated on an Ascentis Express C18 column (150 × 2.1 mm I.D., 2.7 μm d.p.; Merck KGaA, Darmstadt, Germany), employing 0.1 % formic acid in water (pH = 3; solvent A) and 0.1 % formic acid in acetonitrile (solvent B) under the following conditions: 0–10 min, 0–10 %B; 10–20 min, 10–11 %B; 20–30 min, 11–15 % B; 30–50 min, 15–18 %B; 50–65 min, 18–23 %B; 65–70 min, 23–100 %B. The flow rate was 0.5 mL/min. The injection volume was 2 μL and the oven temperature was set at 30 °C. The PDA acquisition were performed in the wavelength range 200–400 nm, with sampling frequency 12.5 Hz and time constant 0.16 s. The MS acquisitions were performed both in positive (+) and negative ionization modes (-), with the following parameters: interface and desolvation temperature were set at 350 °C and 300 °C, respectively; heat block temperature, 300 °C; nebulizing gas flow (N_2), 1.5 L min^{-1}; drying gas flow (N_2), 15 L min^{-1}; acquisition range, 100–1000 m/z (+/-). Data acquisition and processing were handled by the LabSolution ver. 5.97 software (Shimadzu, Duisburg, Germany).

2.3. Molecular structure optimization

To obtain a better understanding of the structural variables affecting the retention pattern of phenolic compounds, the Gaussian09W software [15] together with the GaussView 6.0 software were employed for molecular structure optimization and molecular visualization, respectively. The 3D structures of the employed standard compounds were all downloaded from PubChem Database [16], except for Salvianolic Acid B (not available on PubChem) which was drawn on GaussView; then all the structures were optimised employing two different algorithms: a PM6 semi-empirical method and a B3LYP (Becke, 3-parameter, Lee-Yang-Par) method with 6–311g(d) basis set, based on DFT.

2.4. QSRR models

The Quantitative Structure–Retention Relationship theory is a powerful tool based on finding statistical correlations between molecular descriptors and retention times. Firstly, the molecular structures obtained from PubChem were optimised using semi-empirical and DFT methods; afterwards, 5666 molecular descriptors (Table S2) for optimised and non-optimised structures were calculated using AlvaDesc software [17]. The obtained descriptor sets were then reduced by a variable selection performed with Genetic algorithm and employed as variables to build different QSRR models using Partial Least Square (PLS), Multiple Linear Regression (MLR) and Partial Least Square-Artificial Neural Network (PLS-ANN) methods to find the most reliable predictive model. All the statistical calculation were performed with PLS_Toolbox 9.2.1 [18] on MATLAB software (R2023a) [19].

3. Results and discussion

3.1. HPLC-PDA-ESI/MS analysis

The HPLC analysis was carried out using the most common set-up for polyphenols separation, with a C18 column, using water and acetonitrile, both acidified with 0.1 % of formic acid, as mobile phases.

Starting from a multi-analyte solution containing fifty-two phenolic compounds (Table 1), the chromatographic method was developed trying to reach the best compromise between resolution and analysis time. Nevertheless, the similar behaviour of certain analytes hindered their baseline separation. Fig. 2 reports the chromatogram obtained for the standard mixture. Five coelutions can be observed. However, the coeluted analytes can be easily discriminated by UV spectrum and ESI-qMS fragmentation, as visible from the inserts of Fig. 2. More specifically, all the coeluted analytes possess different UV spectra and different molecular weight, with the exception of Quercetrin (Quercetin 3-O-rhamnoside) and Kaempferol 3-O-glucoside, coeluted at t_R 39.47 min, which possess a similar UV spectrum and the same molecular

Table 1
Phenolic compounds and their retention time.

N°	Name	t_R
1	Pyrogallol	1.97
2	Gallic acid	2.86
3	Syringol	4.79
4	Protocatechuic acid	6.28
5	(-)-Gallocatechin	7.82
6	p-hydroxybenzoic acid	8.82
7	Tyrosol	9.54
8	3-O-Caffeoylquinic acid	11.28
9 + 10	(+)-Catechin + 4-O-Caffeoylquinic acid	11.95
11	Caffeic acid	12.09
12	Syringic acid	13.19
13+14	Ethyl gallate + *p*-hydroxycinnamic acid	16.38
15	(-)-Epicatechin	17.20
16	Ferulic acid	21.22
17	Daidzin (Daidzein 7-O-glucoside)	23.70
18	Sinapic acid	24.01
19	(+)-Taxifolin	25.50
20	Eriodictyol-7-O-glucoside	29.85
20	Eriodictyol-7-O-glucoside	30.08
21	Luteolin-3′,7′-di-O-glucoside	30.72
22	Eriocitrin	30.97
23	Genistin (Genistein 7-O-glucoside)	31.87
24	Rutin	32.86
25	Isoquercetin	33.07
26	Luteolin-7-O-glucoside	33.87
27	N-trans-caffeoyltyramine	36.09
28	Narirutin	37.19
29	Verbascoside	38.21
30+31+32	Quercetrin + 1,5-Dicaffeoylquinic acid + Kaempferol-3-O-glucoside	39.47
33	Myricetin	39.77
34	Naringin	40.26
35	Isorhamnetin-3-O-glucoside	42.07
36	Hesperidin	43.59
37+38	4-Methoxycinnamic acid + Daidzein	47.00
39	Eriodictyol	48.81
40 + 41	Oleuropein + Neodiosmin (Diosmetin 7-O-neohesperidoside)	49.52
42	Quercetin	55.95
43	Luteolin	57.11
44	Salvianolic acid B	59.48
45	Naringenin	61.80
46	Genistein	63.11
47	Poncirin (isosakuranetin 7-O-neohesperidoside)	64.02
47	Poncirin (isosakuranetin 7-O-neohesperidoside)	64.33
48	Apigenin	66.42
49	Kaempferol	66.72
50	Isorhamnetin	67.07
51	Nobiletin	68.10
52	Tangeretin	68.45

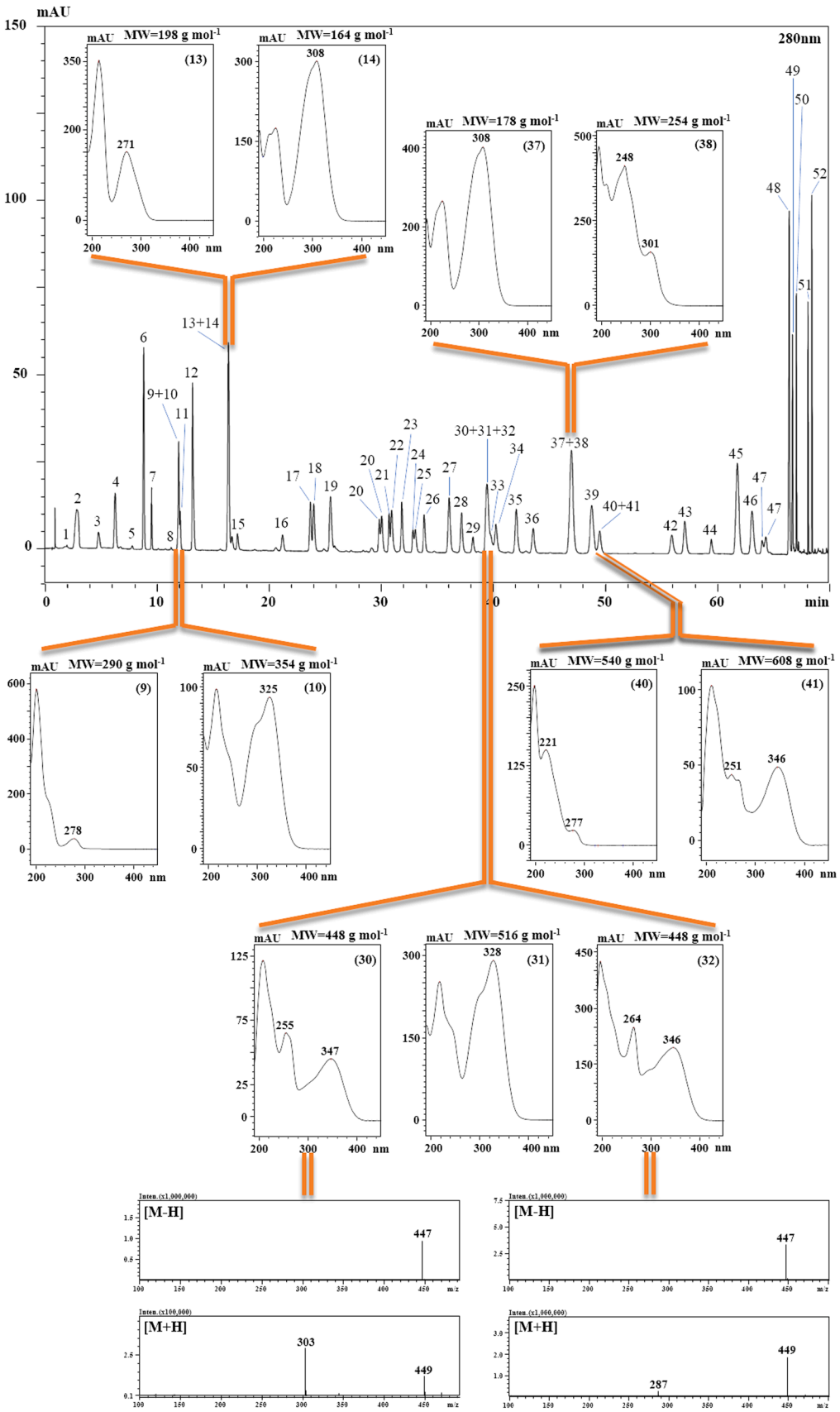

Fig. 2. Chromatogram of the mixture of 52 phenolic compounds. UV and ESI-qMS data were able to discriminate between coeluted compounds.

weight (448 g/mol); in this case, the aglycone fragments easily obtainable with an ESI-qMS detector operated in both negative and positive ionization mode are helpful for the identification.

3.2. Structure-Retention Relationship

The mechanisms involved in reverse phase separation are already well known. In particular, the main solute-column interactions involved in reverse-phase separation are hydrophobic interaction, steric resistance, hydrogen-bond basicity, hydrogen-bond acidity and ionic interactions [20]. On this basis, experimental data together with direct observation of the 2D and 3D phenolic molecular structure, firstly allowed us to confirm the already known variables that affect the retention order, such as class, hydroxyl, methoxy or glycosyl groups,. Then, using the DFT molecular optimization, it was possible to identify other factors that could influence the elution order on C18 columns. In general, for flavonoid aglycones, the retention pattern is mainly based on their class and respects the following order: Flavan-3-ols > Flavanones > Flavonols > Flavones [21]. Moreover, isoflavones elute before their corresponding flavones, due to the presence, in isoflavones, of the phenyl B-ring in C3 instead of C2 (Fig. 3), which reduce the conjugation degree, resulting in a less planar molecule, with higher steric hindrance and increased solubility [22]. Among the flavonoid classes, the retention time decreases with the increasing number of hydroxyl groups, while increases with the increasing number of methoxy groups [21]. Considering the general flavonoid structure, when flavones, flavanones or flavonols (who possess a keto group in C4) are functionalised with a hydroxyl group in C5, leading, through a hydrogen bond, to a planar and non-polar six-atom ring, the retention time increase [23]. Even the C3 hydroxyl group of flavonols can form a hydrogen bond with the C4 keto group, leading to a strained five-atom ring; however, this weaker hydrogen bond does not preclude the C3 hydroxyl group from the solvent [22].

Besides these already-known factors involved in the retention pattern of these molecules, the DFT optimization and the selected molecular descriptors (Table S3) allowed us to identify other factors involved in the elution order of phenolic compounds.

For instance, regarding the diastereoisomers (+)-Catechin (2R, 3S) and (-)-Epicatechin (2R, 3R), their retention pattern is probably related to the distinct steric hindrance generated by the opposite orientation of the C3 bond and to the different number of hydrogen-bonding site available. Actually, different 3D descriptors seem able to discriminate between the two structures, but the most important difference was registered for CATS3D_02_DA descriptor. This descriptor is weighted by hydrogen-bonding donor-acceptor forces and is bigger for (-)-Epicatechin, suggesting that the higher t_R of this isomer is related to the possibility to have more point of interactions with the stationary phase.

Another interesting example is represented by the pairs Quercetin/

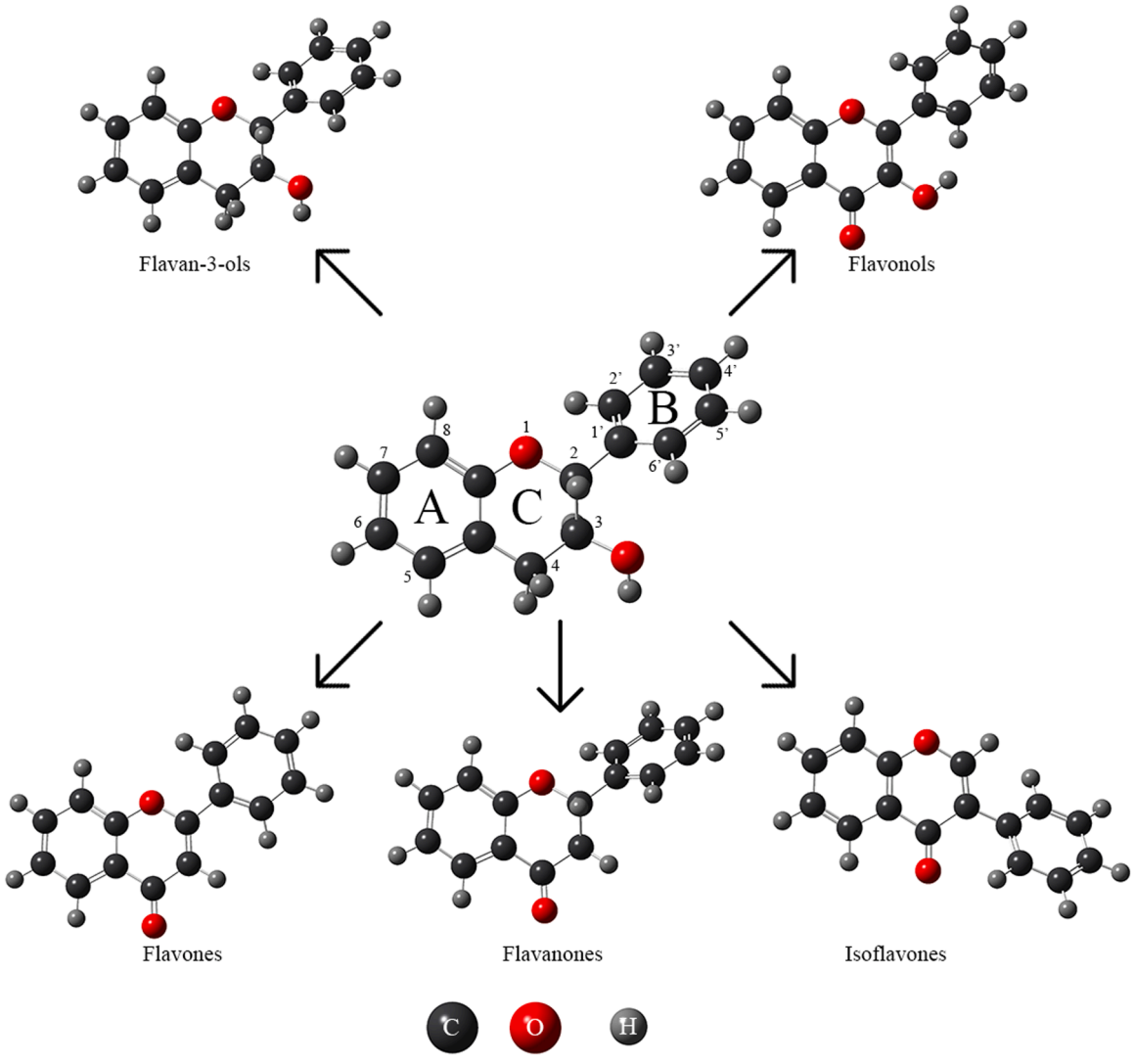

Fig. 3. General flavonoid structures.

Luteolin and Apigenin/Kaempferol, whose structure is displayed in Fig. 4. In both cases, the only difference consists in the presence of an extra hydroxyl group at the C3 position of the structure of Quercetin and Kaempferol, but while Quercetin elutes before Luteolin, Kaempferol elutes later than Apigenin. Considering that the hydroxyl group in C3 is capable to hydrogen-bonding with the keto group in C4, such intra-molecular interaction can be weakened by the solvent which is more accessible to quercetin and luteolin compared to apigenin and kaempferol, due to the presence of an additional hydroxyl groups in C3'. As a consequence, in the case of quercetin the solubility in the mobile phase is predominant in the three-terms interaction analyte/stationary phase/mobile phase, whereas the intra-molecular bond, which leads to an additional five-atom ring, becomes the main factor responsible for the stronger adsorption on the stationary phase of kaempferol with respect to apigenin.

Several 3D descriptors weighted by ionization potential seem to be related to the retention pattern of these pairs. In particular, Mor02i, a 3D-MoRSE (3D Molecular Surface Electrostatics) descriptor weighted by ionization potential, was inversely related to their elution order.

The intra-molecular interaction results also weak in the case of (+)-Taxifolin (2R,3R), which is the dihydroflavonol of Quercetin, where the hydrogenation of the C2-C3 double bond has led to a non planar molecule and a certain degree of bending of C2-C3 and C3-C4 bonds. This flavonoid elutes much earlier than the corresponding planar flavonol.

Regarding glycosyl flavonoids, an explicative case is represented by the retention behaviour of two naringenin derivatives, namely, rutinoside and neohesperioside glycosides, which are α-L-Rhamnopyranosyl-(1→6)-β-D-glucopyranose and α-L-Rhamnopyranosyl-(1→2)-β-D-glucopyranose, respectively. Their separation is probably due to the different disaccharide bonding. In fact, while in 1→2 bonding it is only the oxygen to connects the two monosaccharides, in 1→6 bonding it is a -CH_2O- portion. In this case, the DFT optimization highlighted that the presence of the CH_2 in the 1→6 disaccharide gives more flexibility to the molecule, giving to the rutinoside flavonoids an almost spherical shape, increasing the steric hindrance of the molecule and reducing the retention time compared to neohesperioside flavonoids (Fig. 5). This observation was confirmed by the ASP descriptor which define the non-sphericity of a molecule and that was bigger for all neohesperioside flavonoids than rutinoside flavonoids.

With regard to the relative retention pattern of rutinoside and glucoside flavonoids, Rutin (Quercetin 3-O-rutinoside) elutes before Isoquercetin (Quercetin 3-O-glucoside), while Eriocitrin (Eriodictyol 7-O-rutinoside) elutes later than Eriodictyol 7-O-glucoside. A possible explanation of this behavior can be ascribed to the different position of the sugars. The DFT optimization allowed to obtain a clear idea of the different possible interactions between glycosyl moieties and the flavonoid core; in particular, the results showed how Eriocitrin, Rutin and Isoquercetin possess smaller interatomic distance (1.74, 1.95 and 2.54 Å, respectively) capable of forming intramolecular hydrogen bonds, leading to an increased lipophilicity [24], differently from Eriodictyol 7-O-glucoside, characterized by higher interatomic distances (7.64 Å) and, consequently, eluted earlier. Moreover, comparing the descriptors of the two phenolic pairs (Eriodictyol 7-O-glucoside/Eriocitrin and Rutin/Isoquercetin) a correlation between their retention times and Mor28i 3D-MoRSE descriptor weighted by ionization potential was registered: the highest value was obtained for Eriodictyol 7-O-glucoside, followed by Eriocitrin, then Rutin and Isoquercetin. More in details, the higher retention of Eriocitrin with respect to Eriodictyol 7-O-glucoside is governed by intramolecular interactions made possible through the folding of the rutinoside portion in the C7 position of the flavonoid core. On the other hand, when the sugar is bound in the C3 position, no folding is observed and the elution order is governed by the polarity of the glycoside moiety.

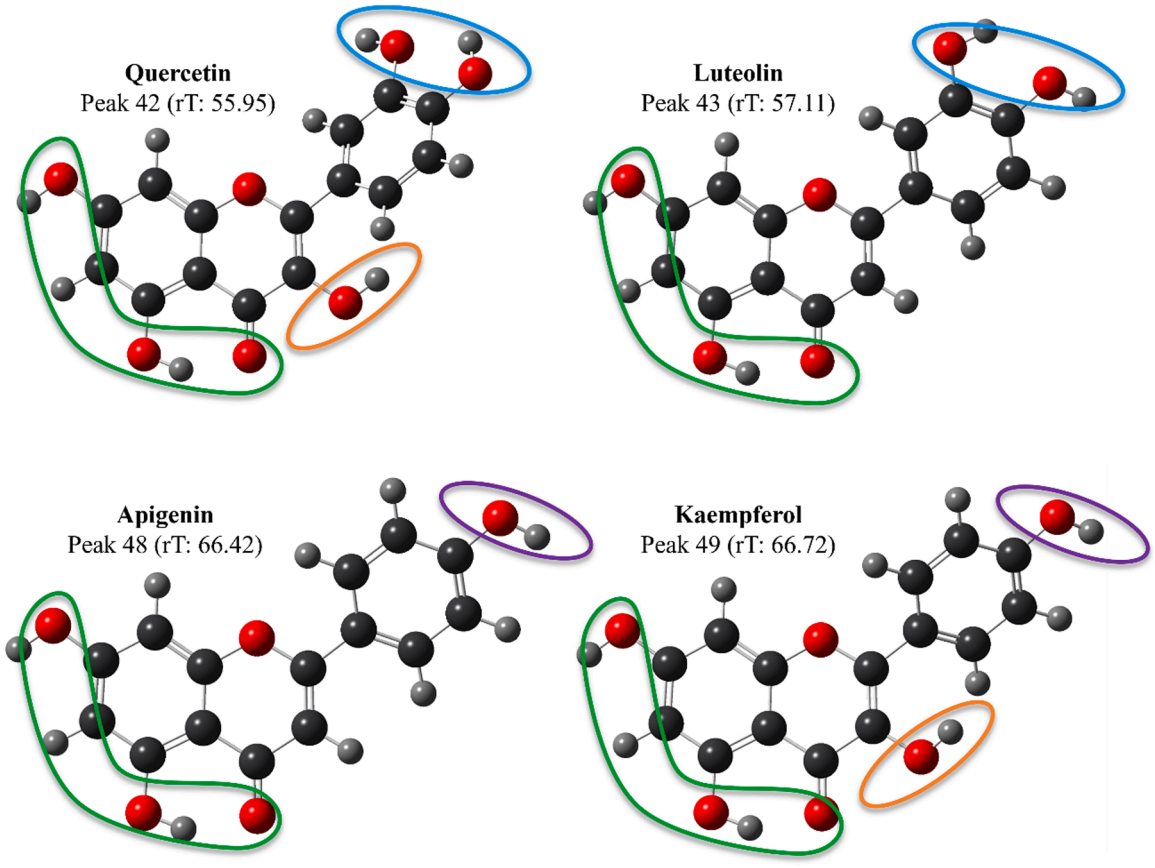

Fig. 4. DFT optimised structures of Quercetin, Luteolin, Apigenin and Kaempferol.

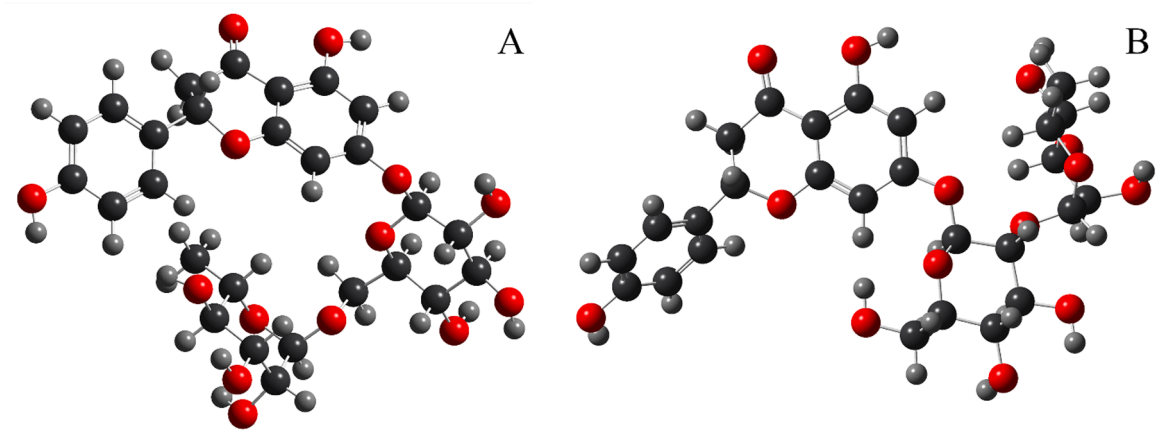

Fig. 5. DFT optimised structures of Naringenin-7-O-rutinoside (A) and Naringenin 7-O-neohesperidoside (B).

A correlation was also noticed with the SP13 descriptor, which was bigger for Eriodictyol 7-O-glucoside and Rutin with respect to Eriocitrin and Isoquercetin, respectively.

Regarding phenolic and hydroxycinnamic acids and their derivatives, the same observations made on flavonoids about the influence of hydroxy- and methoxy- groups on retention time can be made. Moreover, decarboxylated derivatives of phenolic acids, such as pyrogallol or syringol, elute before their corresponding acid, probably due to a stronger hydrogen-bond basicity interaction of the carboxyl group with the stationary phase. Hydroxycinnamic acids are more retained than their corresponding hydroxybenzoic acid, due to the longer carbon chain. Regarding Caffeic and Quinic acid derivatives, such as 3-O-Caffeoylquinic, 4-O-caffeoylquinic and 1,5-Dicaffeoylquinic acids (Fig. 6), the DFT optimization allowed to observe little difference in the number of possible intramolecular hydrogen bonding. More in details, a correlation between retention times and some H- and R-indices (HATS8m, HATS2s, R2u+, R2p+) of GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptors was registered.

It can be pinpointed as 1,5-Dicaffeoylquinic acid with the quinic moiety in the middle between two bent caffeic acids elutes much later than (mono)caffeoylquinic acids, probably due to the combination of intramolecular hydrogen-bonds and π-stacking interactions between the two caffeic acids aromatic rings.

3.3. QSRR models

Different QSRR predictive models were built by using three mathematical methods, namely PLS, MLR and PLS-ANN. Each model was firstly evaluated based on RMSEP (Root Mean Square Error of Prediction) values, which were higher for models using non-optimised and semi-empirical optimised structures, demonstrating the necessity to employ DFT optimised molecules for building QSRR models based on molecular descriptors. Three distinct sets of molecules were evaluated. Set 1: all the fifty-two phenolic compounds; Set 2: all the 32 flavonoids; Set 3: all the 15 *mono*-substituted flavonoids. Each set was then divided through Kennard and Stone algorithm into a training set (66 %) and a validation set (34 %) and the venetian-blind cross-validation was applied. The different models obtained with PLS, MLR and PLS-ANN are depicted in Fig. 7 and the validation metrics are reported in Table 2.

In general, to avoid overfitting, the difference between R^2 and Q^2 does not have to exceed 0.3 [25]. In this context all the models showed good R^2 and Q^2 values, but considering R^2CV and RMSEP, all the MLR models gave unsatisfactory performance, while PLS and PLS-ANN performed better.

Among them, PLS models always gave the best value of RMSEP and delivered the lowest value (1.78 min) of RMSECV (Root Mean Square Error of Cross-Validation) for set 3, while the PLS-ANN models provided the lowest values of RMSEC (Root Mean Square Error of Calibration) for

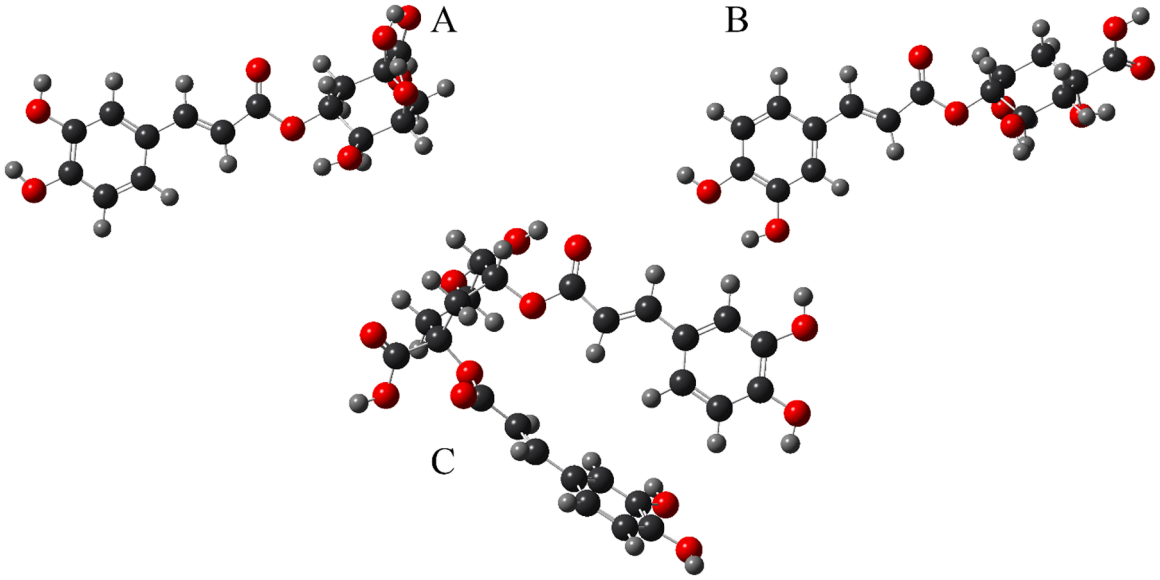

Fig. 6. DFT optimised structures of 3-O-Caffeoylquinic and 4-O-Caffeoylquinic acids (A and B) and 1,5-Dicaffeoylquinic acid (C).

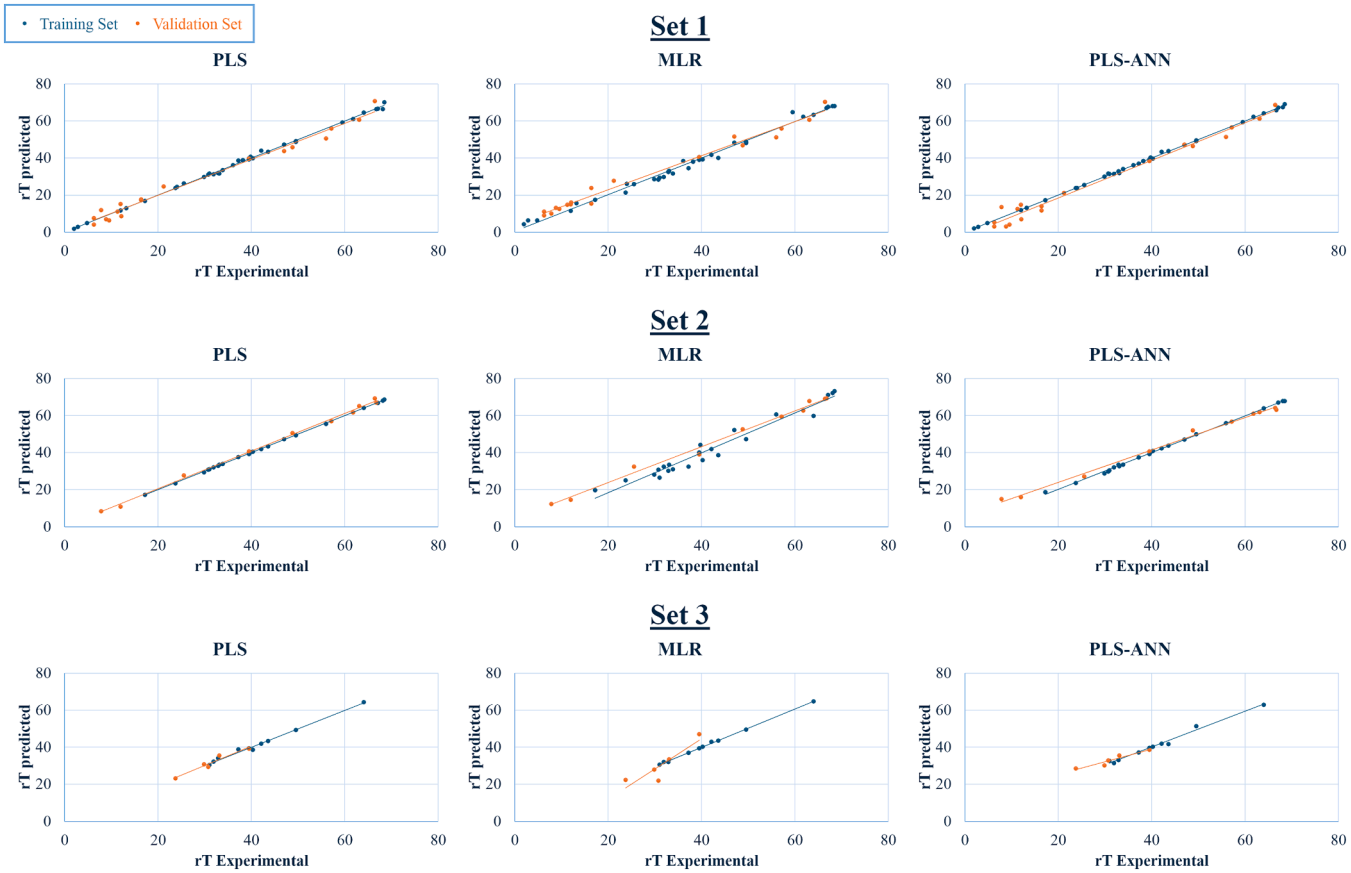

Fig. 7. Prediction results for the different tested models.

Table 2
Validation metrics of PLS, MLR and PLS-ANN models.

Set 1: 52 Phenolic compounds			
	PLS	MLR	PLS-ANN
RMSEC	0.76072	1.8213	0.49326
RMSECV	4.5335	4.9554	4.0857
RMSEP	2.8845	3.8932	3.4008
R^2	0.998	0.99	0.999
Q^2	0.983	0.982	0.981
Set 2: 32 Flavonoids			
	PLS	MLR	PLS-ANN
RMSEC	0.27148	3.4225	0.91534
RMSECV	4.168	7.004	4.3729
RMSEP	1.5463	3.6274	3.1537
R^2	1	0.957	0.996
Q^2	0.997	0.992	0.995
Set 3: 15 *mono*-substituted Flavonoids			
	PLS	MLR	PLS-ANN
RMSEC	0.88494	0.46305	0.058009
RMSECV	1.7805	7.4299	5.7853
RMSEP	1.3271	5.3063	2.6467
R^2	0.991	0.999	1
Q^2	0.946	0.8	0.913

set 3 (0.06 min), thus suggesting the possibility to use both PLS and PLS-ANN models for rT prediction of unknown.

Indeed, excluding the MRL models, RMSEP values were between 1.33 and 3.40 min, representing a percentage error range of 1.9–4.9 %, which is similar or even lower compared to the ones observed in literature [9–12].

3.4. QSRR models application to a real sample

In the final stage of this research, in order to confirm such findings, all the models were tested for the identification of phenolic compounds in a bergamot juice, considering that a total of 12 phenolic compounds were identified in a similar sample by using MS and UV detection.

The structures of such 12 molecules were downloaded from PubChem database and processed as previously described for the standard compounds. A predicted t_R for each compound in each model was obtained, allowing to verify the real prediction capability of the different models.

Specifically, Fig. 8 depicted the chromatogram of the bergamot juice, while Table 3 reports peak identification for correctly predicted molecules. Only three flavonoids were tabulated in our database of fifty-two compounds and easily identified by experimental t_R, while for the other peaks the best predictive results were shown by the PLS-ANN of Set 1 and 3 (Table 3), which recognized, with an acceptable error (-4.50–8.07 %), 7 and 5 compounds, respectively. In particular, 2 hydroxycinnamic derivatives and 5 *mono*-substituted flavonoids were recognised. More in details, the analytes with the highest error (not reported in Table 3) correspond to chemical structures which do not have more than two structurally-related molecules in the model, such as di-substituted flavonoids. Specifically, Apigenin-6,8-diglucoside (X1) and Diosmetin-6,8-diglucoside (X2) were reliably identified by complementary information arising from UV and MS detection and were also confirmed by literature [14], but prediction models gave an error higher than 10 %, then unacceptable. On the contrary, compounds with less prediction error were the ones with a rutinoside or neohesperioside substituent, which were the most represented class in the model sets. This confirms the potential of QSRR models, together with the need for an higher number of representative standard compounds to be used for model building.

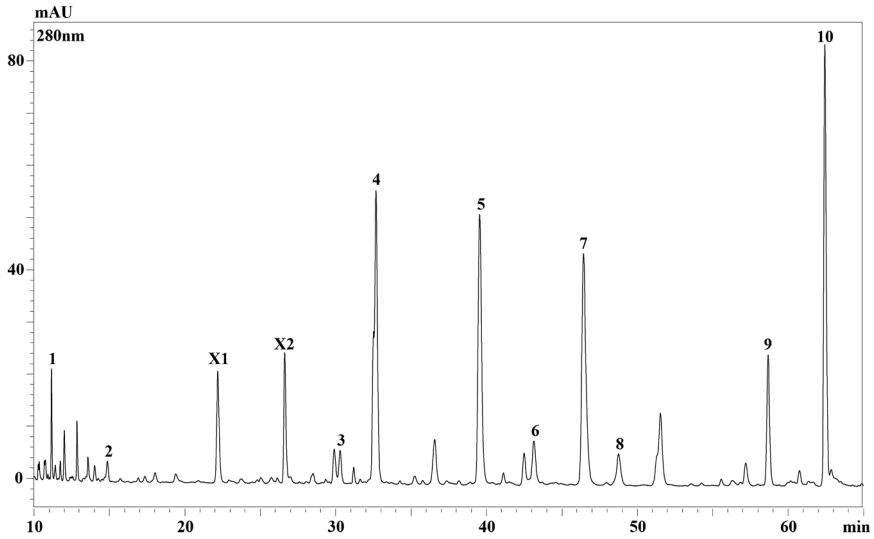

Fig. 8. Chromatogram of the bergamot juice extract in the 10–65 min elution range. Legend for compounds 1–10 is reported in Table 3. X1= Apigenin-6,8-diglucoside; X2= Diosmetin-6,8-diglucoside.

Table 3
Identified compounds using PLS-ANN of Set 1 and 3.

	Compounds	t_R	UV	MW	t_R Prediction of PLS-ANN Set 1	t_R Prediction of PLS-ANN Set 3	Prediction error Set 1 (%)	Prediction error Set 3 (%)
1	Ferulic acid 4-O-glucoside	11.17	290, 313	356	12.18	–	1.45	–
2	3-O-Sinapoyl D-glucose	14.84	288, 328	386	13.05	–	−2.54	–
3	Eriocitrin*	30.30	284, 332sh	596	31.37	32.56	1.53	3.24
4	Neoeriocitrin	32.53	284, 331sh	596	35.26	36.70	3.91	5.96
5	Naringin*	39.56	283, 328	580	39.62	40.32	0.08	1.08
6	Rhoifolin	43.16	266, 336	578	40.00	44.01	−4.50	1.22
7	Neohesperidin	46.50	283, 328	610	43.69	49.07	−4.01	3.68
8	Neodiosmin*	48.77	265, 346	608	49.06	51.27	0.41	3.57
9	Melitidin	58.68	283, 327sh	724	64.33	60.28	8.07	2.29
10	Brutieridin	62.44	283, 333sh	754	67.19	62.60	6.78	0.23

* Compounds used for model calibration/validation.

It is noteworthy that the use of such theoretical methods, in combination with recent advances in machine learning approaches, represents a valuable tool for analytical operators to reliably characterize complex matrices without the need for expensive and sophisticated instrumentations.

4. Conclusion

In the present work, fifty-two phenolic compounds were separated through a typical RP-HPLC set-up and retention data were correlated to molecular descriptors of optimized structures for a better understanding of their retention pattern. In general, the already known variable affecting flavonoids elution were confirmed. Additionally, the influence of substituents and their positions was assessed, as well as the effect of steric hindrance and planarity on the adsorption on the stationary phase. Then, QSRR models were built to predict the retention times of unknown compounds in real samples.

In details, PLS, MLR and PLS-ANN models were tested for different molecule sets, allowing to select two PLS-ANN models as the best predictive ones. Due to the non-linear method, MLR models shown unsatisfactory results, while PLS showed the lowest prediction error, but lacked in the real sample recognition. Both models, together with UV and MS information derived from a single quadrupole MS instrument, can be used for faster and more reliable identification of phenolic compounds, avoiding the employment of expensive tandem MS system. The future employment of bigger standard sets could enable to increase the knowledge about the retention behaviour of phenolic compounds. Moreover, as future perspective of this study, the use of different columns, mobile phases or miniaturised systems could allow optimising a linear gradient method, obtaining a more powerful predictive QSRR model. To achieve this task, no new optimization will be required d, but only new experiments and a new statistical correlation should be performed, thus avoiding the most time-consuming step. Furthermore, the model reliability could be improved by using relative t_R or linear retention indices to normalize observed shifts between theoretical and experimental retention data. Within this context, the internal standard or homologue series used for the calculation of relative t_R or linear retention indices should have similar retention behaviour than our analytes. This will presumably enhance the possibility to transfer our models on different columns, instrumentations or even by changing mobile phase composition.

CRediT authorship contribution statement

Roberto Laganà Vinci: Writing – original draft, Visualization, Validation, Investigation, Formal analysis, Data curation. **Katia Arena:**

Methodology, Investigation, Formal analysis. **Francesca Rigano:** Writing – review & editing, Conceptualization. **Francesco Cacciola:** Writing – review & editing, Supervision. **Paola Dugo:** Writing – review & editing. **Luigi Mondello:** Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors acknowledge Merck Life Science and Shimadzu Corporation for their continuous support.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2024.465146.

References

[1] J. Wang, J. Xu, X. Gong, M. Yang, C. Zhang, M. Li, Biosynthesis, chemistry, and pharmacology of polyphenols from chinese salvia species: a review, Molecules 24 (2019), https://doi.org/10.3390/molecules24010155.
[2] W. Li, H. Chen, B. Xu, Y. Wang, C. Zhang, Y. Cao, X. Xing, Research progress on classification, sources and functions of dietary polyphenols for prevention and treatment of chronic diseases, J. Future Foods 3 (4) (2023) 289–305, https://doi.org/10.1016/j.jfutfo.2023.03.001.
[3] K.B. Pandey, S.I. Rizvi, Plant polyphenols as dietary antioxidants in human health and disease, Oxid. Med. Cell. Longev. 2 (2009) 897484, https://doi.org/10.4161/oxim.2.5.9498.
[4] F. Cacciola, S. Farnetti, P. Dugo, P.J. Marriott, L. Mondello, Comprehensive two-dimensional liquid chromatography for polyphenol analysis in foodstuffs, J. Sep. Sci. 40 (1) (2017) 7–24, https://doi.org/10.1002/jssc.201600704.
[5] F. Cacciola, K. Arena, F. Mandolfino, D. Donnarumma, P. Dugo, L. Mondello, Reversed phase versus hydrophilic interaction liquid chromatography as first dimension of comprehensive two-dimensional liquid chromatography systems for the elucidation of the polyphenolic content of food and natural products, J. Chromatogr. A 1645 (2021) 462129, https://doi.org/10.1016/j.chroma.2021.462129.
[6] O.M. Andersen, K.R. Markham, Flavonoids: chemistry. Biochemistry and Applications, 1st ed., CRC Press, 2005 https://doi.org/10.1201/9781420039443.
[7] S. Quideau, D. Deffieux, C. Douat-Casassus, L. Pouységu, Plant polyphenols: chemical properties, biological activities, and synthesis, Angew. Chem. Int. Ed. 50 (3) (2011) 586–621, https://doi.org/10.1002/anie.201000044.
[8] J. Krmar, B. Svrkota, N. Đajić, J. Stojanović, A. Protić, B. Otašević, S. C. Moldoveanu, V.D. Hoang, V. David, QSRR approach: application to retention mechanism in liquid chromatography. Novel Aspects of Gas Chromatography and Chemometrics, IntechOpen, Rijeka, 2022, https://doi.org/10.5772/intechopen.106245. Ch. 7.
[9] Z. Lei, L. Jing, F. Qiu, H. Zhang, D. Huhman, Z. Zhou, L.W. Sumner, Construction of an ultrahigh pressure liquid chromatography-tandem mass spectral library of plant natural products and comparative spectral analyses, Anal. Chem. 87 (14) (2015) 7373–7381, https://doi.org/10.1021/acs.analchem.5b01559.
[10] S. Sun, B. Cui, F. Kong, Z. Zhang, Y. Qiao, S. Zhang, X. Zhang, C. Sun, Construction and application of a QSRR approach for identifying flavonoids, J. Pharm. Biomed. Anal. 240 (2024) 115929, https://doi.org/10.1016/j.jpba.2023.115929.
[11] Z. Li, C. Zhao, X. Zhao, Y. Xia, X. Sun, W. Xie, Y. Ye, X. Lu, G. Xu, Deep annotation of hydroxycinnamic acid amides in plants based on ultra-high-performance liquid chromatography–high-resolution mass spectrometry and its in silico database, Anal. Chem. 90 (24) (2018) 14321–14330, https://doi.org/10.1021/acs.analchem.8b03654.
[12] J. Akbar, S. Iqbal, F. Batool, A. Karim, K.W. Chan, Predicting retention times of naturally occurring phenolic compounds in reversed-phase liquid chromatography: a Quantitative Structure-Retention Relationship (QSRR) approach, Int. J. Mol. Sci. 13 (11) (2012) 15387–15400, https://doi.org/10.3390/ijms131115387.
[13] S.I. Gorelsky, Ab initio and semiempirical methods, in: R.A. Scott (Ed.), Encyclopedia of Inorganic and Bioinorganic Chemistry, 2011, pp. 1–12, https://doi.org/10.1002/9781119951438.eibc0377.
[14] M. Russo, A. Arigò, M.L. Calabrò, S. Farnetti, L. Mondello, P.B. Dugo, Citrus Bergamia Risso) as a source of nutraceuticals: limonoids and flavonoids, J. Funct. Foods 20 (2016) 10–19, https://doi.org/10.1016/j.jff.2015.10.005.
[15] Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Scalmani, G.; Barone, V.; Petersson, G.A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A.; Bloino, J.; Janesko, B.G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J.V.; Izmaylov, A.F.; Sonnenberg, J.L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V.G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, Jr., J.A.; Peralta, J.E.; Ogliaro, F.; Bearpark, M.; Heyd, J.J.; Brothers, E.; Kudin, K.N.; Staroverov, V.N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J.C.; Iyengar, S.S.; Tomasi, J.; Cossi, M.; Millam, J.M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J.W.; Martin, R.L.; Morokuma, K.; Farkas, O.; Foresman, J.B.; Fox, D.J. Gaussian 09, Revision A.02, Gaussian, Inc., Wallingford CT, 2016. https://gaussian.com/g09citation/.
[16] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton, PubChem 2023 Update, Nucleic Acids Res. 51 (D1) (2023) D1373–D1380, https://doi.org/10.1093/nar/gkac956.
[17] A. Mauri, K. Roy, alvaDesc: a tool to calculate and analyze molecular descriptors and fingerprints. Ecotoxicological QSARs, Springer, US: New York, NY, 2020, pp. 801–820, https://doi.org/10.1007/978-1-0716-0150-1_32.
[18] PLS_Toolbox 9.2.1, Eigenvector Research, Inc., Manson, WA USA 98831; 2023 http://www.eigenvector.com.
[19] MATLAB version: 9.14.0 (R2023a), The MathWorks Inc.; Natick, Massachusetts; 2022. https://www.mathworks.com.
[20] L.R. Snyder, J.W. Dolan, P.W. Carr, The hydrophobic-subtraction model of reversed-phase column selectivity, J. Chromatogr. A 1060 (1) (2004) 77–116, https://doi.org/10.1016/j.chroma.2004.08.121.
[21] B.R. Kumar, Application of HPLC and ESI-MS techniques in the analysis of phenolic acids and flavonoids from green leafy vegetables (GLVs), J. Pharm. Anal. 7 (6) (2017) 349–364, https://doi.org/10.1016/j.jpha.2017.06.005.
[22] X. Dong, X. Li, X. Ruan, L. Kong, N. Wang, W. Gao, R. Wang, Y. Sun, M. Jin, A deep insight into the structure-solubility relationship and molecular interaction mechanism of diverse flavonoids in molecular solvents, ionic liquids, and molecular solvent/ionic liquid mixtures, J. Mol. Liq. 385 (2023) 122359, https://doi.org/10.1016/j.molliq.2023.122359.
[23] I. Novak, P. Janeiro, M. Seruga, A.M. Oliveira-Brett, Ultrasound extracted flavonoids from four varieties of Portuguese red grape skins determined by reverse-phase high-performance liquid chromatography with electrochemical detection, Anal. Chim. Acta 630 (2) (2008) 107–115, https://doi.org/10.1016/j.aca.2008.10.002.
[24] D. Chen, M. Zhao, W. Tan, Y. Li, X. Li, Y. Li, X. Fan, Effects of intramolecular hydrogen bonds on lipophilicity, Eur. J. Pharm. Sci. 130 (2019) 100–106, https://doi.org/10.1016/j.ejps.2019.01.020.
[25] D.M. Hawkins, The problem of overfitting, J. Chem. Inform. Comput. Sci. 44 (1) (2004) 1–12, https://doi.org/10.1021/ci0342472.