Theor Chem Acc (2016) 135:66
DOI 10.1007/s00214-016-1812-1

REGULAR ARTICLE

Quantitative structure–activity relationships to predict sweet and non-sweet tastes

Cristian Rojas[1,2] · Davide Ballabio[3] · Viviana Consonni[3] · Piercosimo Tripaldi[4] · Andrea Mauri[3] · Roberto Todeschini[3]

Received: 13 October 2015 / Accepted: 18 January 2016

Abstract The aim of this work was the calibration and validation of mathematical models based on a quantitative structure–activity relationship approach to discriminate sweet, tasteless and bitter molecules. The sweet-tasteless and the sweet-bitter datasets included 566 and 508 compounds, respectively. A total of 3763 conformation-independent Dragon molecular descriptors were calculated and subsequently reduced through both unsupervised reduction and supervised selection coupled with the *k*-nearest neighbors classification technique. A model based on nine descriptors was retained as the optimal one for sweet and tasteless molecules, while a model based on four descriptors was calibrated for the sweetness-bitterness dataset. Models were properly validated through cross-validation and external test sets. The applicability domain of models was investigated, and the interpretation of the role of the molecular descriptors in classifying sweet and non-sweet tastes was evaluated. The classification and the performance of the models presented in this paper are simple but accurate. They are based on a relatively small number of descriptors and a straightforward classification approach. The results presented here indicate that the proposed models can be used to accurately select new compounds as potential sweetener candidates.

Keywords QSAR · *k*-Nearest neighbors · Classification · Sweetness

Published as part of the special collection of articles "CHITEL 2015 - Torino - Italy".

Electronic supplementary material The online version of this article (doi:10.1007/s00214-016-1812-1) contains supplementary material, which is available to authorized users.

✉ Cristian Rojas
crojasvilla@gmail.com

[1] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (CCT La Plata-CONICET, UNLP), Diag. 113 y 64, C.C. 16, Sucursal 4, 1900 La Plata, Argentina

[2] Decanato General de Investigaciones, Universidad del Azuay, Av. 24 de Mayo 7-77 y Hernán Malo, Apartado Postal 01.01.981, Cuenca, Ecuador

[3] Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1, 20126 Milan, Italy

[4] Laboratorio de Química-Física de Alimentos, Facultad de Ciencia y Tecnología, Universidad del Azuay, Av. 24 de Mayo 7-77 y Hernán Malo, Apartado Postal 01.01.981, Cuenca, Ecuador

1 Introduction

Taste perception is a subject of study in many disciplines such as psychology, chemistry, biochemistry, pharmacology, anatomy and physiology, and especially the food industry. There appears to be an increase in research related to taste perception in recent years. The taste mechanism in humans is complex and includes apparent contradictions. For example, a chemist might assume that taste is related to chemical reasons, while a psychologist might presume that taste might be related to some brain stimulus. While both the chemist and the psychologist are, in some ways correct, studies of anatomy and physiology have shown that taste is related to soluble molecules which interact with receptors on the taste buds in the tongue. The sweetness, bitterness, saltiness and sourness are the most important primary tastes that define a substance [1]. The human perception of these primary tastes, which may vary from person to person, may be related to subtle differences in psychology, anatomy, or receptor function.

Published online: 25 February 2016

Sweetness is one of the most important tastes in a variety of foods and produces a pleasant sensation. Sucrose is a common standard as a sweetener since it is easily obtained from renewable sources, such as sugar cane and sugar beets [2, 3]. However, the food industry has an increasing interest in discovering new sweeteners that might have beneficial properties. Food technologists and food chemists face the challenge of discovering compounds which exhibit a pure sweet taste resembling sucrose. For example, the development of low-calorie sweeteners without bad after-tastes may be useful in producing new products for diabetic individuals, especially those with type II diabetes. There are thousands of sweet molecules but, for reasons of safety and quality perception, few substances are permitted to be used as additives in the food industry [4]. Moreover, only sugars and their hydrogenated derivatives (e.g., polyols) give a clean sweet taste without aftertastes [4]. On the other hand, bitterness is usually perceived as an unpleasant taste although in some cases it is considered desirable (e.g., coffee, beer, tonic water, olives, etc.). Bitterness is exhibited by alkaloids and heavy metal salts. In fact, quinine is an alkaloid used as a component of some soft drinks to imprint bitter taste and it is the substance most frequently used as a standard for testing bitterness. Some changes in the chemical structure of a substance may change the sweetness to either tastelessness or bitterness. For example, the sweet compound, saccharin, turns bitter with the introduction of a chloride or a methyl group in the *meta* position, while the replacement of the imino group by a methyl, ethyl, or bromoethyl radical produces the loss of the sweet taste, that is, the compound becomes tasteless [5].

Several theories regarding the relationship between chemical structure and sweet taste exist. Oertly and Myers [6] explained the sweetness production by the relationship between the *glucophores* and *auxoglucs* functional groups. Subsequently, Shallenberger and Acree [7] suggested that a sweet compound has to contain a hydrogen bond donor (AH) and hydrogen bond acceptor (B) separated by a distance of about 3 Å (AH-B theory). On the other hand, the B-X theory proposed by Lemont Kier [8] states that a sweetener must have a third binding site. Finally, the MultiPoint Attachment (MPA) theory was proposed by Nofre and Tinti [9] and suggests a total of eight interaction sites for the sweetness receptor although not all the sweeteners interact with all the sites.

Most of the approved artificial sweeteners for human intake, including natural compounds or food additives such as saccharin, cyclamate, and aspartame, were discovered by chance [4, 10]. However, several mathematical models based on the quantitative structure–activity/property relationships (QSAR/QSPR) approach have been established to discriminate sweet and non-sweet substances and support the view that the systematic selection of sweeteners is possible. The aim of the QSAR/QSPR theory is to build mathematical relationships between the chemical structures of molecules, described by means of molecular descriptors, and their activities/properties [11–14]. Therefore, QSAR models related to sweetness can be used in a screening step to predict compounds exhibiting sweet taste to be subsequently synthesized and tested.

QSAR models for the prediction of sweet taste of molecules have already been published and Table 1 summarizes, in a chronological order, the most important related studies. All the studies reported the Accuracy (AC) parameter as a measurement of the performance of the models. In 1980 Iwamura [15] proposed a quantitative structure–taste relationship study for 49 perillartine and aniline derivatives using five STERIMOL descriptors which characterize the molecular shape and size. A similar study was performed by van der Wel et al. [16]. The same year, Kier [17] used 20 bitter and sweet aldoximes derivatives to build a linear discriminant function based on two molecular connectivity indexes. Between 1982 and 1988, Takahashi et al. [18–21] and Okuyama et al. [22] used a series of perillartine derivatives, aspartyl dipeptides and carbosulfamates to calibrate QSAR models based on *k*NN and SIMCA approaches to discriminate sweet and non-sweet compounds. In addition, an extensive number of studies were performed by Spillane and coworkers [23–34] in order to differentiate sweet and non-sweet sulfamate derivatives by means of linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and classification trees (CART). Results of some of these studies are also available in published reviews [35–40].

According to the premises previously mentioned, the aim of this work was to build two QSAR models for the discrimination of sweet and non-sweet tastes, keeping in mind the five principles defined by the Organization for Economic Co-operation and Development (OECD) to make them applicable [41]. In summary, the modeled activity and the algorithm should be clearly defined, the model should be accompanied by a definition of its domain of applicability, the goodness-of-fit and predictivity of the model should be evaluated through appropriate strategies and, eventually, a mechanistic interpretation of model descriptors should be given, if possible. To this end, attention was paid to the curation of the experimental data, which led to the definition of two extended datasets consisting of 566 and 477 compounds, respectively. Two mathematical models based on a simple similarity-based algorithm (*k*-nearest neighbors, *k*NN) were developed and their applicability domains (AD) properly defined. The model predictive power was estimated by means of appropriate internal and external validation procedures. Moreover, the chemical information encoded by model descriptors was explained.

Table 1 Characteristics of QSAR classification models reported in the literature for the discrimination of sweet molecules

Tastes	Classes	Method	d	N_{train}	N_{test}	AC_{train}	AC_{cv}	AC_{test}	References
Sweet and bitter	2	SLR	3	49	–[a]	–	–	–	[15]
Sweet and bitter	2	LDA	2	20	9	0.850	–	0.778	[17]
Sweet and bitter	2	LDA	2	35	12	0.900	–	0.917	[23]
Sweet and bitter	2	LLA	3	22	–	1	–	–	[18]
		*k*NN	6	22	–	0.909	–	–	
Sweet and bitter	2	LDA	3	33	–	0.848	–	–	[24]
Sweet and bitter	2	LDA	3	22	9	1	–	0.778	[19]
			2			0.955	–	0.778	
Sweet and bitter	3	SIMCA	5	91	–	0.857	–	–	[21]
Sweet and non-sweet	2	SIMCA	7	50	–	0.740	–	–	[20]
Sweet and non-sweet	2	SIMCA		25	–	0.800	–	–	[22]
				20	–	0.750			
Sweet and bitter	2	LDA	3	23	–	0.870	–	–	[25]
Sweet and bitter	2	LDA	3	33	–	0.848	–	–	[26]
				23	–	0.870			
Sweet and non-sweet (bitter, sour and aniline- or hydrocarbon-like taste)	4	Plot	2	40	–	0.825	–	–	[27]
Sweet and bitter	3	DA	11[b]	50	–	–	–	–	[28]
	2	PCA			–	–	–	–	
Sweet and non-sweet	2	LDA	4	101	–	0.673	–	–	[29]
		QDA			–	0.801	–	–	
		CART	3		–	0.861	–	–	
Sweet and bitter	2	Plot	2	23	–	0.870	–	–	[30]
		LDA	4		–	0.870	–	–	
		QDA			–	0.913	–	–	
Sweet and non-sweet	2	LDA	4	132	–	0.697	–	–	[31]
		QDA			–	0.682	–	–	
		CART	3		–	0.795	–	–	
Sweet	3	LDA	8	75	8	0.547	0.413	0.500	[32]
		QDA				0.770	0.493	0.250	
		CART classification				0.770	–	–	
		CART regression ($R^2 = 0.627$)	6			0.813	–	0.750	
Sweet	3	CART classification	6	82	–	0.890	–	–	[33]
		CART classification	6	70	12	0.871	–	0.583	
		CART regression ($R^2 = 0.757$)	7	70	12	0.886	–	0.833	
Sweet and non-sweet (bitterness, blandness, or tastelessness)	2	LDA	2	58	–	0.655	0.600	–	[34]
	2	QDA	3	58	–	0.758	0.603	–	
	2	CART	6	48	10	0.958	–	0.700	
	3	CART	6	48	10	0.938	–	0.800	

SLR simple linear regression, *SIMCA* soft independent modeling by class analogy, *LLA* linear learning machine, *DA* discriminant analysis, *d* number of descriptors

[a] Not available, [b] number of descriptors considering for the DA and PCA analysis

2 Materials and methods

2.1 Experimental dataset

Experimental data on sweetness-tastelessness and sweetness-bitterness were retrieved from several referenced sources [1, 5, 42–55]. Details for both datasets are given in Tables 1S and 2S, respectively. The initial sets consisted of 620 and 589 molecules for the sweet-tasteless and sweet-bitter datasets, respectively. A qualitative experimental taste response is associated with each substance. In particular, molecules are identified as sweet, bitter or tasteless.

In the majority of the cases, a quantitative value of relative sweetness (RS) is associated with sweet compounds. As described in the literature, the experimental taste test consists in an initial training of panelists by means of sweet, sour, bitter and salty standards following a sip and spit methodology; then, a definite taste and its intensity is identified for each compound [27, 56]. In this study, three classes of taste were analyzed separately in two datasets (sweet *vs* tasteless and sweet *vs* bitter), in order to better understand the differences among the chemical structures of such tastes.

2.2 Data curation and filtering

In order to guarantee conformation-independent QSAR models, compounds containing disconnected structures (salts) were removed from the data sets due to the fact that calculations of molecular descriptors is limited for these kind of structures. Moreover, when dealing with stereoisomers exhibiting the same activity, only one of them was retained; stereoisomers belonging to different classes were excluded. Thus, 566 molecules were retained in the sweet-tasteless dataset (433 sweet and 133 tasteless), while 508 compounds were considered in the sweet-bitter dataset (427 sweet and 81 bitter). The simplified molecular input line entry system (SMILES) strings were obtained for each compound by using Babel software [57]. Data curation selection and filtering of the sweet-tasteless and sweet-bitter datasets were carried out by means of a KNIME workflow [58] which is summarized in Table 2.

2.3 Molecular descriptors

Molecular descriptors are used as the structural representation of molecules in order to develop QSAR models. Descriptors are the final result of a logical and mathematical procedure that transforms chemical information encoded within a symbolic representation of a molecule into a numerical quantity or into the result of some standardized experiment [14]. HyperChem [59] was used for the molecular design. Subsequently, 3763 conformation-independent molecular descriptors were calculated by means of Dragon software (version 6.0) [60]. Such descriptors were grouped into the following families: constitutional indices, functional group counts, atom-centered fragments, molecular properties, ring descriptors, topological indices, walk and path counts, connectivity indices, information indices, 2D matrix-based descriptors, 2D autocorrelations, Burden eigenvalues, P_VSA-like descriptors, edge adjacency indices, CATS 2D, 2D atom pairs, atom-type E-state indices, ETA indices. A two-dimensional structural representation was selected instead of geometrical representation to avoid irreproducible 3D structure optimizations; 3D descriptors could add valuable chemical information; however, these require a geometrical optimization and this can be an issue when applying QSAR models to new molecules, since the difference between the 3D conformers can affect the descriptor values. The optimization of the geometries to find the spatial position of all the atoms in the Cartesian space during the search of the minimum in the conformational energy hypersurface of molecules involve high computational costs and long times [61]. Moreover, some other potential problems regarding the use of 3D-descriptors have been previously reported [62, 63].

Table 2 Results from the data curation and filtering of the sweet-tasteless and sweet-bitter datasets

Reason for removal of molecules from the dataset	Number of removed molecules	
	Sweet-tasteless	Sweet-bitter
Disconnected structures (salts)	12	12
Stereoisomers replicates belonging to the same class	32	49
Stereoisomers exhibiting different classes	10	20
Total number of excluded molecules	54	81

2.4 Model development

2.4.1 QSAR classification model

Since the sweet-tasteless and sweet-bitterness datasets have discrete response (assignation of a substance to a sweet or a tasteless/bitter class), the nonparametric *k*NN classification technique [64] was used to establish a mathematical relationship between the chemical structure encoded in molecular descriptors and the modeled classes. The *k*NN classification rule is conceptually quite simple: a molecule is classified according to the majority of its *k* closest neighbors in the space defined by molecular descriptors. *k*NN is an appropriate classification method for nonlinear class separation. In this work, descriptors were pretreated by means of autoscaling and the Euclidean metric was used to measure distances between pairs of molecules. The optimal *k* value was selected according to the lowest error in cross-validation.

2.4.2 Molecular descriptor reduction and selection

Unsupervised descriptor reduction is a useful approach to reduce the presence of multicollinearity, redundancy, and noise in QSAR data. In this study, the V-WSP variable reduction method [65] was used for this purpose. This is a modification of the algorithm proposed by Wootton,

Sergent and Phan-Tan-Luu (WSP) for the selection of points from a pool of candidates in such a way as to be at a pre-fixed minimal Euclidean distance from each point in the defined multidimensional space. V-WSP selects a subset of representative descriptors instead of points: molecular descriptors are chosen in order to have a minimal correlation from each descriptor in the defined multidimensional space.

One of the fundamentals of QSAR is the supervised selection of descriptors in order to build a parsimonious model based on a pool of useful descriptors, which guarantees interpretability, stability and reliability of predictions. To this end, Genetic Algorithms-Variable Subset Selection (GA-VSS) technique [66] was coupled with the *k*-nearest neighbors (*k*NN) classification methodology to find the optimal subset of molecular descriptors. The essence of the GA-VSS is to start from an initial random population of chromosomes, that is, binary vectors indicating the presence or absence of descriptors. Then, an evolutionary process is performed in order to optimize an established fitness function, such as the classification non-error rate (*NER*) in cross-validation, and new chromosomes are constructed by the combination of chromosomes presented in the initial population through genetic operation (crossover and mutation). Finally, the optimal model is constructed by including the most frequently selected descriptor in each genetic evolution. The approach used in this paper took into account the repeatability of the selection, i.e., the selection of variables was performed by repeating GAs *t* times (runs) and then including the variables on the basis of the frequencies of selection of each variable in the best model of each run and on the basis of *NER* values as a function of the number of selected variables. Therefore, each run was independent from the others, and the selected descriptors were those which were included more times in the *t* runs.

2.4.3 Model validation

In order to avoid the problem of overfitting when dealing with GA, models were validated by means of an external test set, which confirmed the substantial comparability of the performance of the model on training and test sets. To this end, each dataset was randomly divided into training and test sets, containing 70 and 30 % of the molecules, respectively. The split of datasets was carried out in order to maintain the class proportions, that is, the number of the test molecules included in each class were proportional to the total number of training molecules included in the same class. This partition guarantees similar class representation, especially when one of the classes contains a larger number of molecules with respect to the other class. Molecules of the training set were used during the supervised selection of molecular descriptors and to calibrate the models. Test molecules were used just to evaluate the prediction ability of the training set model. A cross-validation protocol based on five cancelation groups divided in venetian blinds was used during the GA-VSS procedure [67].

Classification models were properly evaluated by means of specificity (*Sp*) and sensitivity (*Sn*) of classes. Sensitivity describes the model's ability to correctly recognize samples belonging to the class, while specificity characterizes the ability of the class to reject the samples of all the other classes [67]. When two classes are considered, specificity of the first class corresponds to the sensitivity of the other. The Non-Error Rate (*NER*) was calculated as the arithmetic mean of the class sensitivities. All these parameters were used to evaluate the classification performance of the datasets.

2.4.4 Descriptor interpretation

Another important issue to be addressed in QSAR studies is how descriptors included in the model are related to the activity of the molecules. Since the *k*NN classification approach does not provide coefficients to quantify the contribution of each descriptor, descriptor interpretation was carried out by means of principal component analysis (PCA) [68]. Principal component analysis (PCA) is a well-known multivariate technique for exploratory data analysis, which projects the data in a reduced hyperspace, defined by orthogonal principal components [69, 70]. These are linear combinations of the original variables, with the first principal component having the largest variance, the second principal component having the second-largest variance, and so on. In this work, a PCA model was calculated on the training molecules and test molecules were projected in it [71]. Score and loading plots were used to evaluate the relationships between descriptors and modeled classes.

2.4.5 Applicability domain assessment

The applicability domain (AD) assessment of *k*NN-QSAR models has been already defined in the literature [72–76]. The approach consists in the calculation of the average distance of each test molecule from its *k*-nearest neighbors of the training set. Then, the average distance is compared to a user-defined threshold. If the average distance is lower than the threshold, then the test molecule is inside the AD and its prediction is considered reliable; otherwise, the molecule is outside the AD and its prediction is considered an extrapolation of the model. Thus, the evaluation of the applicability domain is implicitly defined on a similarity-based approach and this is supposed to better describe distribution of molecules due to the fact that *k*NN describes locally the covariance structure of the data, as well.

3 Software

HyperChem [59] was used for molecular design. Open Babel [57] was used to obtain the simplified molecular input line entry system (SMILES notations). Molecular descriptors were calculated by means of DRAGON version 6.0 [60]. V-WSP variable reduction toolbox [65], Classification toolbox [67] and PCA toolbox [77] for MATLAB have been used to perform descriptor reduction, model calibration and molecular descriptor analysis, respectively. Genetic algorithms were performed in MATLAB [78] by means of routines built by the authors.

4 Results and discussion

4.1 Discrimination of sweet and tasteless molecules

According to all known studies listed in Table 1, there are no QSAR studies for the discrimination between the sweetness and tastelessness. Thus, the QSAR model proposed here is the first model for such tastes. The understanding of the chemical structural features that influence these tastes is important due to the fact that the introduction or replacement of a functional group in a known scaffold or the change of its position generates the loss of sweetness. For example, 2-amino-4-nitro-propoxybenzene (sweet) become tasteless when the amino and nitro group swap positions (2-nitro-4-amino-propoxybenzene).

The initial dataset was split into a training set of 396 substances and a test set of 170 compounds. A total of 3763 molecular descriptors were first calculated for each molecule by the software, Dragon, but 2164 descriptors were retained after exclusion of those with constant (1394) and near-constant (123) values or those affected by missing values (82). Then, the V-WSP unsupervised variable reduction method was applied to further reduce the number of descriptors to minimize the potential of model over-fitting. In V-WSP, a correlation threshold of 0.95 was established as a limit and 1309 descriptors were excluded. As a result, only 855 molecular descriptors were retained for the subsequent modeling phase.

The selection of molecular descriptors by means of the GA-VSS coupled with the *k*NN classification methodology was performed in two steps in such a way as to handle the 855 descriptors and avoid potential overfitting of the final model. GA-VSS was initially carried out separately on each block of molecular descriptors; then, the descriptors selected from each block (141) were merged and GA-VSS was applied again to find the most useful subset of descriptors to calibrate the final QSAR model. The selection of the final model was done by taking into account the *NER*, as well as a balanced ratio between specificity and sensitivity of sweet and tasteless classes. Thus, a model based on nine conformation-independent descriptors was retained as the optimal one. The classification performance parameters of the sweet-tasteless QSAR model are listed in Table 3.

The selected QSAR model showed comparable performance in fitting and validation; *NER* in fitting and cross-validation were equal to 0.84 and 0.85, respectively, while the *NER* on the test set was lower (0.75). These results indicate the absence of potential overfitting in the model. On the other hand, the quality of the model should also be evaluated according to its ability to correctly predict sweetness: sensitivity and specificity slightly differ on the training set, while they are equal in the test set, showing an overall balanced discrimination between the modeled classes. A brief description of the nine descriptors included in the model is presented in Table 4.

PCA on model descriptors was performed in order to evaluate how these nine molecular descriptors are related to the discrimination of sweet-tasteless classes. The score plot of the first two principal components is shown in Fig. 1a, while the corresponding loading plot is shown in Fig. 1b. The combination of PC1 and PC2 explained 59 % of the variance and gave an acceptable separation between sweet and tasteless molecules, despite some overlaps.

The majority of the sweet molecules have positive scores on PC1, while tasteless molecules have negative scores. As can be seen in the loading plot (Fig. 1b), the descriptors placed in the positive zone on PC1 are *nRNH2*, *F03[N–O]*, and *BIC3*. These descriptors indicate that the sweetness of a molecule is potentially related to the number of amino radicals (NH_2) presented on an aliphatic skeleton, as well as to the number of nitrogen and oxygen atom pairs [79] at a topological distance of 3 *[F03(N–O)]* in the molecule. The Bond Information Content index (neighborhood symmetry of 3-order) gives information related to the

Table 3 Quality parameters for sweet-tasteless and sweet-bitter QSAR models; *d* is the number of descriptors; non-error rate (*NER*), sensitivity (*Sn*) and specificity (*Sp*) of the sweet class obtained on the full training set, in cross-validation and on the test set are reported

Model	*d*	Training set			Five fold cross-validation			Test set		
		NER	*Sn*	*Sp*	*NER*	*Sn*	*Sp*	*NER*	*Sn*	*Sp*
Sweet-tasteless	9	0.84	0.89	0.78	0.85	0.90	0.80	0.75	0.75	0.75
Sweet-bitter	4	0.86	0.96	0.77	0.86	0.95	0.77	0.79	0.95	0.63

Table 4 Brief description for the non-conformational Dragon descriptors included in the sweet-tasteless and sweet-bitter QSAR models

Name	Description	Block	Model
BIC3	Bond information content index (neighborhood symmetry of 3-order)	Information indices	Sweet-tasteless
CATS2D_05_AL	CATS2D acceptor-lipophilic at lag 05	CATS 2D	
nRNH2	Number of primary amines (aliphatic)	Functional group counts	
GATS6v	Geary autocorrelation of lag 6 weighted by van der Waals volume	2D autocorrelations	
GATS7v	Geary autocorrelation of lag 7 weighted by van der Waals volume		
AVS_B(e)	Average vertex sum from Burden matrix weighted by Sanderson electronegativity	2D matrix-based descriptors	
SpMax_B(i)	Leading eigenvalue from Burden matrix weighted by ionization potential		
B03[O–O]	Presence/absence of O–O at topological distance 3	2D atom pairs	
F03[N–O]	Frequency of N–O at topological distance 3		
SM4_B(s)	Spectral moment of order 4 from Burden matrix weighted by I-State	2D matrix-based descriptors	Sweet-bitter
C-026	R–CX–R	Atom-centered fragments	
F01[C-N]	Frequency of C–N at topological distance 1	2D Atom Pairs	
CATS2D_04_AL	CATS2D acceptor-lipophilic at lag 04	CATS 2D	

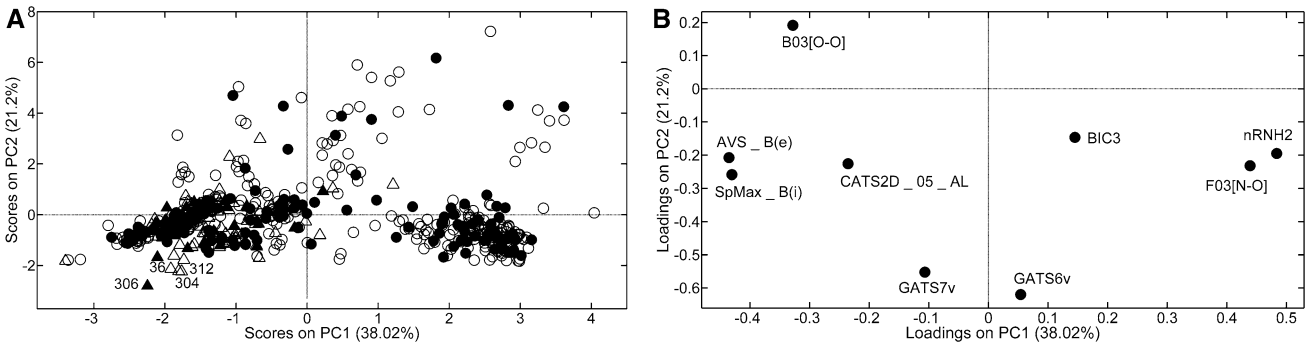

Fig. 1 PCA of the descriptors used in the *k*NN model for the sweet (*circles*), and tasteless (*triangles*) molecules. Score plot (**a**) and loading plot (**b**) of the first and second principal components (explained variance equal to 59.22 %). Training molecules are marked with *empty circles and triangles*, and test molecules are marker with *full circles and triangles*

molecular complexity of a compound. This descriptor takes on high values with increases in the number of equivalence classes of the substances; that is, the number of similar atoms of order 3.

On the other hand, the *AVS_B(e)*, *SpMax_B(i)*, and *CATS2D_05_AL* descriptors have high negative loading on PC1, characterizing, therefore, the tasteless compounds which have negative scores, as well as some sweet molecules. The Sanderson electronegativities (e) and the ionization potential (i) atomic properties were used to weight the molecular graph and obtain the corresponding weighted Burden matrices B(e) and B(i); from these matrices, the descriptor *AVS_B(e)* was calculated as the average row sum and the descriptor *SpMax_B(i)* as the leading eigenvalue. Figure 2a shows that these two descriptors are well correlated to each other. *CATS2D_05_AL* is among the 2D autocorrelation descriptors and counts the pairs of hydrogen bond acceptors (A) (i.e., all N or O with at least one available lone pair electron) and lipophilic atoms (L) (i.e., C bonded only to C or H atoms, as well as Cl, Br, I or S exhibiting a vertex degree of 2 and attached to two C atoms) separated by 5 bonds. In fact, Spillane et al. [36] claimed that the hydrophobicity is an important parameter for sweet taste determination, while Birch et al. [80] indicated that the presence of sweet taste may be attributed to the hydrophile-lipophile balance. Additionally, some CATS2D descriptors have been recently presented as useful descriptors to predict the relative sweetness (*RS*) of sweet compounds [81].

In addition, PC2 provides information about other two autocorrelation indices (i.e., the Geary autocorrelations of lag 6 (*GATS6v*) and lag 7 (*GATS7v*), weighted by van der Waals volume). These descriptors are placed in the negative zone of PC2, indicating that the majority of tasteless compounds have atoms with large volume placed at a topological distance of 6 or 7, e.g., saccharine derivatives (36, 304,

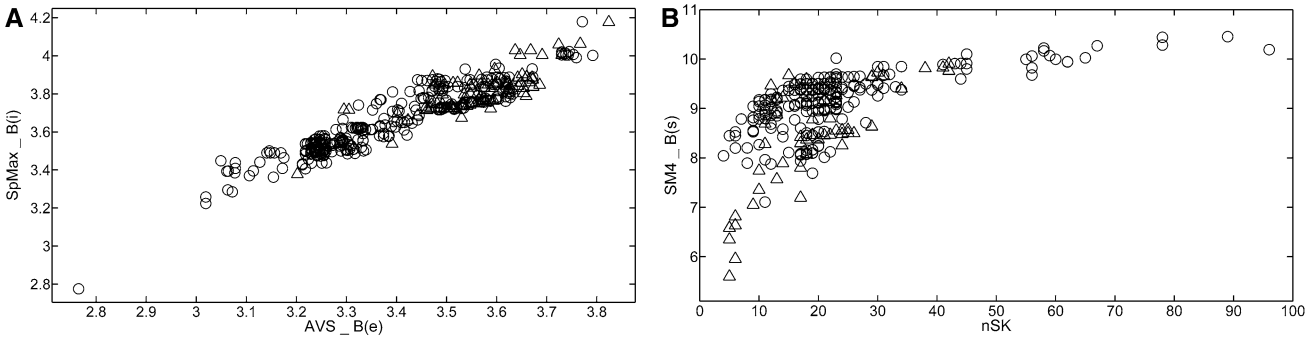

Fig. 2 **a** Plot between the average vertex sum from Burden matrix weighted by Sanderson electronegativity (AVS_B(e)) and the leading eigenvalue from Burden matrix weighted by ionization potential (SpMax_B(i)) for sweet (*circle*) and tasteless (*triangle*) molecules. **b** Plot between spectral moment of order 4 from Burden matrix weighted by I-State (SM4_B(s)) and the number of non-H atoms (nSK) for the sweet (*circle*) and bitter (*triangle*) molecules

306, 312). Moreover, a reasonable number of sweet molecules that are placed in the positive zone of PC2 exhibit low values of these two descriptors. Strong spatial autocorrelation produces small values of this index; moreover, positive autocorrelation translates to values between 0 and 1, whereas negative autocorrelation produces values larger than 1. Finally, in the positive zone of PC2, the presence/absence of a oxygen–oxygen bond at a topological distance of 3 (*B03[O–O]*) describes the sweetness of those sweet compounds characterized by positive scores in PC2. Since the sweet compounds are placed in the positive zone, it means that sweetness could be related to the presence of *B03[O–O]*, while the majority of tasteless compounds do not have this atom pair in their structures.

Table 3S shows a list of 117 sweeteners superimposing the tasteless molecules in the negative zone of PC1 (Fig. 1a). Fifteen sweeteners do not have values associated with the experimental values for the relative sweetness (RS) with respect to sucrose. Sweeteners placed in this zone exhibit low relative sweetness (RS lower than 2000) and only seven molecules present a RS higher than 2000 units (potent sweeteners). In particular, 76 substances have RS below 200 units, corresponding to the 65 % of the sweeteners located in the negative zone of PC1.

4.2 Discrimination of sweet and bitter molecules

From the 3763 calculated molecular descriptors, 2164 molecular descriptors were retained after exclusion of constant (1394), near-constant (123) or at least one missing value (82) descriptors. The same workflow as for the sweet-tasteless discrimination was used to develop the sweet-bitter QSAR model. Therefore, the dataset was split into a training set and a test set containing 356 and 152 compounds, respectively. Then, V-WSP method was used to obtain a reduced set of 855 molecular descriptors. After the application of GA-VSS, a QSAR model based on just four conformation-independent descriptors (refer to Table 4) with acceptable parameters in fitting ($NER_{train} = 0.86$), cross-validation ($NER_{cv} = 0.86$), and prediction of the test set of molecules ($NER_{test} = 0.79$) was obtained. Further classification indices are listed in Table 3. For this model, there is a relevant difference between sensitivity and specificity of the sweet class, indicating a superior capability of the model in discriminating sweet molecules with respect to bitter ones.

PCA analysis was applied to analyse the descriptor behavior in separating sweet and bitter molecules. The combination of PC1 and PC2 explained together 69 % of the variance. Figure 3a shows the scores plot, while the corresponding loading plot is shown in Fig. 3b.

The majority of the bitter compounds are placed in the positive zone of PC1 and the negative region of PC2, that is, these compounds are strongly characterized by the presence of C and N atom pairs in the molecule at topological distance of 1 (*F01[C-N]*) as well as the presence of a C atom linked to an electronegative atom (such as O, N, S, P, Se, halogens) and connected to any group through carbon separated by an aromatic bond as in benzene (*C-026*) [82]. On the other hand, the theobromine and caffeine alkaloid derivatives are isolated in the positive zone of both PC1 and PC2, indicating that these outliers are described by higher values of both *F01[C-N]* and *SM4_B(s)*. For the calculation of the *SM4_B(s)* descriptor, the spectral moment of order 4 operator is applied to the Burden matrix weighted by the intrinsic state (s). Figure 2b indicates that such a descriptor has a relation to the number of non-H atoms (*nSK*) in a molecule. Thus, the highest values of *SM4_B(s)* are particularly associated with large sweet molecules (e.g., mogroside V, rebaudioside D, mogroside IV). In contrast, low values for the *SM4_B(s)* descriptor describe better the small bitter compounds (e.g., pyrrolidine, piperazine,

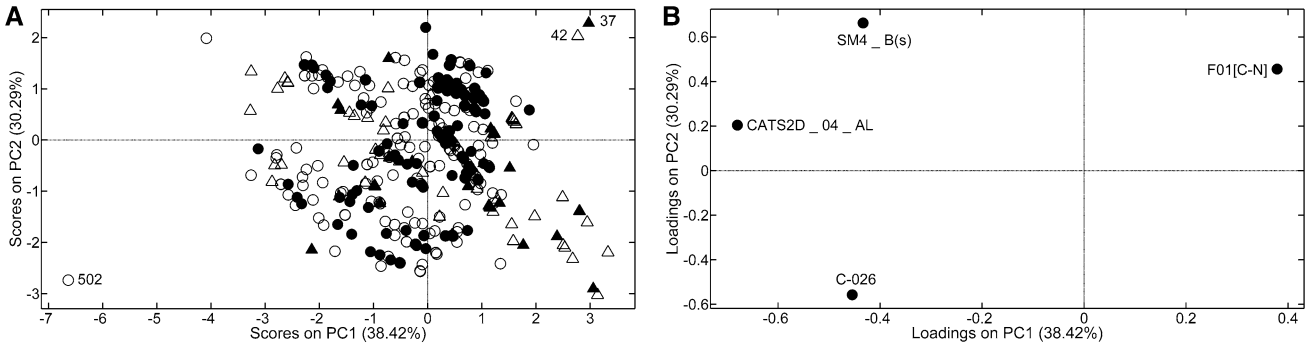

Fig. 3 PCA of the descriptors used in the *k*NN model for the sweet (*circles*), and bitter (*triangles*) molecules. Score plot (**a**) and loading plot (**b**) of the first and second principal components (explained variance equal to 68.71 %). Training molecules are marked with *empty circles and triangles*, and test molecules are marker with *full circles and triangles*

pyrrole). Additionally, almost all these compounds located in the positive zone of PC1 and PC2 are compounds exhibiting nitrogen atoms. For example, theobromine (42) and caffeine (37) contain each 4 *N* atoms inside their scaffolds.

The majority of the sweet molecules are placed in the negative zone of PC1 indicating that they are related to the CATS2D Acceptor-Lipophilic (AL) at lag 04, R–CX–R, and spectral moment of the order 4 from the Burden matrix weighted by intrinsic state. As explained for the sweet-tasteless model, the sweetness is described by the presence of linear fragments of acceptor and lipophilic atoms; that is, this activity is strongly related to the presence of N, O, and S with at least one available electron lone pair, as well as lipophilic atoms separated by 4 units of topological distance. In addition, a Burden matrix-based descriptor is selected to describe the sweetness activity (*SM4_B(s)*). It is interesting to observe the link between *C-026* and *SM4_B(s)* descriptors. Since these two descriptors are placed in the negative and the positive region of PC2, respectively, they explain opposite information; i.e., the higher the value of *SM4_B(s)* the lower the presence of carbon atoms linked to an electronegative atom. On the other hand, the sweetener Selligueain A (502), a trimeric proanthocyanidin extracted from the *Selligrrea fei* plant [83] is also an outlier in the negative zone of both principal components. This compound exhibits high values of the *C-026* and *CATS2D_04_AL*. In fact, this sweetener is characterized by the presence of 6 aromatic rings in which 12 carbon atoms are bonded to an oxygen atom.

Natural sugar substances have been regarded to taste better than artificial sweeteners; however, chemical modification of any of these compounds is sufficient to convert them to a bitter taste substance [4, 84]. In fact, some sweeteners exist that exhibit both sweetness and bitterness as intrinsic features [85], e.g., acesulfame potassium, sodium saccharin, hernandulcin, stevioside, isocoumarin derivatives and some sugar derivatives [84]. For substances exhibiting the sweet and bitter tastes, one side of the molecule binds to the sweet receptor, while the other side binds to the bitter receptor, that is, these molecules appeared to be polarized on taste receptors. These data suggest that these two receptors are very close to each other [85].

As shown in Table 1, the only *k*NN model related to the discrimination of sweet and bitter molecules was published by Takahashi et al. [18]. Its fitting performance ($NER_{train} = 0.909$) is comparable to the model proposed in this study; however, it included only 22 substances, while our QSAR model was calibrated on a significantly bigger set of 508 compounds. Consider that QSAR models presented in Table 1 were mainly calibrated by using homog eneous families of molecules, which do not allow their gene ralization to other families of substances. Moreover, external validation of proposed QSAR models was not performed in the majority of the published studies [15, 18, 21, 24–26] and therefore a comparison of predictive capabilities with respect to the model proposed in this study is not feasible.

4.3 Definition applicability domain

As previously described, the applicability domain estimation of both models consists of the comparison of the average distance between each test molecule and its *k*-nearest neighbors with a defined threshold value. Figure 4 shows the distribution of average distances of test molecules for the sweet-tasteless and sweet-bitter QSAR models. Because of these distributions, two thresholds equal to 2 (Fig. 4a) and 0.5 (Fig. 4b) were defined for the applicability domain definition of the sweet-tasteless and sweet-bitter QSAR models, respectively.

In fact, for the sweet-tasteless QSAR model, the majority of the test molecules have low average distances with respect to their neighbors, as expected. Therefore, according to the AD definition, a test molecule is predicted only if its average distance is lower than 2; otherwise, it is

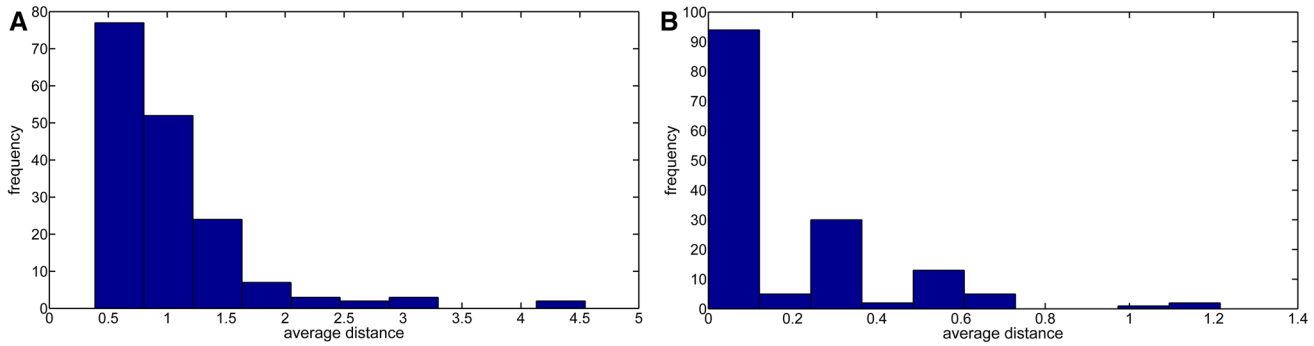

Fig. 4 Histograms of the average distances of test molecules with respect to their neighbors for **a** the sweet-tasteless model and **b** sweet-bitter model

considered outside the AD of the sweet-tasteless model. A total of 24 test molecules (19 sweet and 4 tasteless) exhibit an average distance higher than 2 and their prediction can be considered unreliable. Thus, recalculating the test set classification parameters by excluding these molecules, an improvement of the classification performance was achieved since the non-error rate increased from 0.75 (Table 3) to 0.76, the specificity for the sweet class increased from 0.75 to 0.78, while the sensitivity remains the same (0.75). This result confirms the appropriateness of the proposed strategy for the AD definition, since molecules with distant neighbors (therefore excluded from the model AD) are mostly related to misclassifications.

Similar results were obtained for the sweet-bitter QSAR model, for which a distance threshold equal to 0.5 was selected. Nine molecules had average distances from their neighbors higher than the threshold. Interestingly, caffeine, which was already discussed as an outlier, is effectively outside the model AD. As for the sweet-tasteless model, even for the sweet-bitter QSAR model, the exclusion of test molecules outside the AD allowed the classification enhancement, the non-error rate being improved from 0.79 to 0.83, as well as the sweet class sensitivity (from 0.95 to 0.97) and the sweet class specificity (from 0.63 to 0.68).

4.4 Discussion of the classification performance

The classification performance of both of the proposed QSAR models can be considered adequate, taking into account the simplicity of the classification technique (*k*NN), the small number of molecular descriptors included in the model and the fact that several experimental factors can affect the model calibration. First of all, the experimental values of sweetness can be considered noisy, due to the fact that these values are assessed by trained panelists who measure both the types of tastes and their intensities. Since not all the sweeteners exhibit a purely sweet taste, humans are unlikely to discern differences when a substance exhibits two or more tastes (multisapophoric or potential multisapophoric molecules), thus introducing a potential experimental error in the responses modeled by the proposed QSAR models. This may be due to receptor saturation on the taste buds of the tongue or the polarization of the taste receptors [85].

Moreover, the study of the structure–activity relationships in sweeteners is difficult due to the flexibility of the molecules and complications in establishing their conformation. In addition, having sweeteners with more than one AH-B site, it is difficult to indicate which one interacts with the receptor to elicit the sweetness [9, 25, 86, 87]. Also, the molecules considered in the datasets belong to a wide variety of families of compounds, which did not allow these mathematical models to completely predict the sweetness of new potential sweeteners exhibiting such diverse scaffolds.

5 Conclusions

In this study, conformation-independent QSAR models were performed to discriminate sweet molecules from tasteless and bitter compounds. Experimental values were collected from several sources, and the resulting datasets were accurately verified. Each dataset was randomly split into training and test sets. The use of the V-WSP variable reduction and the Genetic algorithms coupled with the *k*NN classification approach allowed the selection of an optimal subset of Dragon descriptors for each model. The results showed that the two models demonstrated good statistics in fitting and cross-validation, as well as an acceptable accuracy in predicting the test molecules. Moreover, the similar performance in both models for the training set, cross-validation and test set indicates the absence of overfitting. PCA was used to highlight the potential relationships between modeled classes and the selected molecular descriptors. It is interesting to highlight that Dragon molecular descriptors as

well as the multivariate methodologies employed here were used for the first time to perform classification models for the discrimination between sweet and non-sweet tastes. These models are based on a simple classification technique as well as a reduced number of molecular descriptors and could be useful for scientist working to develop superior low-calorie sweeteners in order to design, in a rational way, new potent sweetener candidates. Finally, the conformation-independent QSAR methodology represents an efficient alternative approach to develop models based on topological and constitutional molecular aspects of chemical compounds.

Acknowledgments Cristian Rojas is grateful for his PhD Fellowship from the National Secretary of Higher Education, Science, Technology and Innovation (SENESCYT) from the Republic of Ecuador, as well as for the financial support provided by the Ministry of Foreign Affairs and International Cooperation (FARNESINA) from the Italian Government for the PhD research conducted at the University of Milano-Bicocca.

References

1. Shallenberger RS (1993) Taste chemistry. Springer Science & Business Media, Berlin
2. Hugot E, Jenkins GH (1972) Handbook of cane sugar engineering, vol 114. Elsevier, Philadelphia
3. Asadi M (2006) Beet-sugar handbook. Wiley, NewYork
4. Birch GG (1999) Modulation of sweet taste. BioFactors 9(1):73–80
5. deMan JM (1999) Principles of food chemistry, 3rd edn. Berlin, Springer
6. Oertly E, Myers RG (1919) A new theory relating constitution to taste. Simple relations between the constitution of aliphatic compounds and their sweet taste. J Am Chem Soc 41(6):855–867
7. Shallenberger RS, Acree TE (1967) Molecular theory of sweet taste. Nature 216:480–482
8. Kier LB (1972) A molecular theory of sweet taste. J Pharm Sci 61(9):1394–1397
9. Nofre C, Tinti J-M (1996) Sweetness reception in man: the multipoint attachment theory. Food Chem 56(3):263–274
10. Ellis JW (1995) Overview of sweeteners. J Chem Educ 72(8):671
11. Katritzky AR, Lobanov VS, Karelson M (1995) QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. Chem Soc Rev 24:279–287
12. Trinajstic N (1992) Chemical graph theory. CRC Press, Boca Raton
13. Diudea MV (2001) QSPR/QSAR studies by molecular descriptors. Nova Science Publishers, New York
14. Todeschini R, Consonni V (2009) Molecular descriptors for chemoinformatics, vol 2. Wiley-VCH, Weinheim
15. Iwamura H (1980) Structure-taste relationship of perillartine and nitro-and cyanoaniline derivatives. J Med Chem 23(3):308–312
16. van der Wel H, van der Heijden A, Peer H (1987) Sweeteners. Food Rev Int 3(3):193–268
17. Kier LB (1980) Molecular structure influencing either a sweet or bitter taste among aldoximes. J Pharm Sci 69(4):416–419
18. Takahashi Y, Miyashita Y, Tanaka Y, Abe H, Sasaki S (1982) A consideration for structure-taste correlations of perillartines using pattern-recognition techniques. J Med Chem 25(10):1245–1248
19. Takahashi Y, Abe H, Miyashita Y, Tanaka Y, Hayasaka H, Sasaki SI (1984) Discriminative structural analysis using pattern recognition techniques in the structure-taste problem of perillartines. J Pharm Sci 73(6):737–741
20. Miyashita Y, Takahashi Y, Takayama C, Ohkubo T, Funatsu K, Sasaki S-I (1986) Computer-assisted structure/taste studies on sulfamates by pattern recognition methods. Anal Chim Acta 184:143–149
21. Miyashita Y, Takahashi Y, Takayama C, Sumi K, Nakatsuka K, Ohkubo T, Abe H, Sasaki S (1986) Structure-taste correlation of L-aspartyl dipeptides using the SIMCA method. J Med Chem 29(6):906–912
22. Okuyama T, Miyashita Y, Kanaya S, Katsumi H, S-i Sasaki, Randić M (1988) Computer assisted structure-taste studies on sulfamates by pattern recognition method using graph theoretical invariants. J Comput Chem 9(6):636–646
23. Spillane WJ, McGlinchey G (1981) Structure-activity studies on sulfamate sweeteners II: semiquantitative structure-taste relationship for sulfamate ($RNHSO_3^-$) sweeteners-the role of R. J Pharm Sci 70(8):933–935
24. Spillane WJ, McGlinchey G, Muircheartaigh IÓ, Benson GA (1983) Structure-activity studies on sulfamate sweetners III: structure-taste relationships for heterosulfamates. J Pharm Sci 72(8):852–856
25. Spillane WJ, Sheahan MB (1989) Semi-quantitative and quantitative structure-taste relationships for carboand hetero-sulphamate ($RNHSO_3^-$) sweeteners. J Chem Soc, Perkin Trans 2(7):741–746
26. Spillane WJ, Sheahan M (1991) Structure-taste relationships for sulfamate sweeteners ($RNHSO_3^-$). Phosphorus Sulfur Silicon Relat Elem 59(1–4):255–258
27. Spillane WJ, Sheahan MB, Ryder CA (1993) Synthesis and taste properties of sodium disubstituted phenylsulfamates. Structure-taste relationships for sweet and bitter/sweet sulfamates. Food Chem 47(4):363–369
28. Drew MGB, Wilden GRH, Spillane WJ, Walsh RM, Ryder CA, Simmie JM (1998) Quantitative structure-activity relationship studies of sulfamates $RNHSO_3Na$: distinction between sweet, sweet-bitter, and bitter molecules. J Agric Food Chem 46(8):3016–3026
29. Spillane WJ, Ryder CA, Curran PJ, Wall SN, Kelly LM, Feeney BG, Newell J (2000) Development of structure-taste relationships for sweet and non-sweet heterosulfamates. J Chem Soc Perkin Trans 2(7):1369–1374
30. Spillane WJ, Feeney BG, Coyle CM (2002) Further studies on the synthesis and tastes of monosubstituted benzenesulfamates. A semi-quantitative structure-taste relationship for the meta-compounds. Food Chem 79(1):15–22
31. Spillane WJ, Kelly LM, Feeney BG, Drew MG, Hattotuwagama CK (2003) Synthesis of heterosulfamates. Search for structure-taste relationships. Arkivoc 7:297–309
32. Kelly DP, Spillane WJ, Newell J (2005) Development of structure-taste relationships for monosubstituted phenylsulfamate sweeteners using classification and regression tree (CART) analysis. J Agric Food Chem 53(17):6750–6758
33. Spillane WJ, Kelly DP, Curran PJ, Feeney BG (2006) Structure-taste relationships for disubstituted phenylsulfamate tastants using classification and regression tree (CART) Analysis. J Agric Food Chem 54(16):5996–6004
34. Spillane WJ, Coyle CM, Feeney BG, Thompson EF (2009) Development of structure-taste relationships for thiazolyl-, benzothiazolyl-, and thiadiazolylsulfamates. J Agric Food Chem 57(12):5486–5493
35. Spillane WJ (1993) Structure taste studies of sulphamates. In: Mathlouthi M, Kanters JA, Birch GG (eds) Sweet-taste chemoreception. Elsevier Science Publishers, Philadelphia, p 283

36. Spillane WJ, Ryder CA, Walsh MR, Curran PJ, Concagh DG, Wall SN (1996) Sulfamate sweeteners. Food Chem 56(3):255–261
37. Walters DE (2006) Analysing and predicting properties of sweet-tasting compounds. In: Spillane WJ (ed) Optimising sweet taste in foods. pp 283–291
38. Rojas C, Duchowicz PR, Pis Diez R, Tripaldi P (2016) Applications of quantitative structure-relative sweetness relationships in food chemistry. In: Mercader AG, Duchowicz PR, Sivakumar PM (eds) Chemometrics applications and research: QSAR in medicinal chemistry. CRC Press, Taylor & Francis Group, pp 317–339
39. van der Heijden A (1997) Historical overview on structure-activity relationships among sweeteners. Pure Appl Chem 69(4):667–674
40. Spillane W, Malaubier J-B (2014) Sulfamic acid and its N-and O-substituted derivatives. Chem Rev 114(4):2507–2586
41. Organisation for Economic Co-operation and Development (2007) Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models. OECD Publishing, Paris
42. Arnoldi A, Bassoli A, Borgonovo G, Drew MG, Merlini L, Morini G (1998) Sweet isovanillyl derivatives: synthesis and structure-taste relationships of conformationally restricted analogues. J Agric Food Chem 46(10):4002–4010
43. Arnoldi A, Bassoli A, Borgonovo G, Merlini L (1995) Synthesis and sweet taste of optically active (−)-haematoxylin and of some (±)-haematoxylin derivatives. J Chem Soc Perkin Trans 1(19):2447–2453
44. Arnoldi A, Bassoli A, Borgonovo G, Merlini L, Morini G (1997) Synthesis and structure-activity relationships of sweet 2-benzoylbenzoic acid derivatives. J Agric Food Chem 45(6):2047–2054
45. Arnoldi A, Bassoli A, Merlini L (1996) Progress in isovanillyl sweet compounds. Food Chem 56(3):247–253
46. Arnoldi A, Bassoli A, Merlini L, Ragg E (1991) Isovanillyl sweeteners. Synthesis, conformational analysis, and structure-activity relationship of some sweet oxygen heterocycles. J Chem Soc Perkin Trans 2(9):1399–1406
47. Arnoldi A, Bassoli A, Merlini L, Ragg E (1993) Isovanillyl sweeteners. Synthesis and sweet taste of sulfur heterocycles. J Chem Soc Perkin Trans 1(12):1359–1366
48. Bassoli A, Borgonovo G, Drew MG, Merlini L (2000) Enantiodifferentiation in taste perception of isovanillic derivatives. Tetrahedron Asymmetry 11(15):3177–3186
49. Bassoli A, Drew MGB, Hattotuwagama CK, Merlini L, Morini G, Wilden GRH (2001) Quantitative structure-activity relationships of sweet isovanillyl derivatives. Quant Struct-Act Relat 20(1):3–16
50. Belitz H-D, Grosch W, Schieberle P (2009) Food chemistry, 4th edn. Springer-Verlag, Heidelberg
51. Nanayakkara NPD, Hussain RA, Pezzuto JM, Soejarto DD, Kinghorn AD (1988) An intensely sweet dihydroflavonol derivative based on a natural product lead compound. J Med Chem 31(6):1250–1253
52. O'Brien-Nabors L (2001) Alternative sweeteners, 3rd edn. New York, Marcel Dekker Inc
53. Yamato M, Hashigaki K (1979) Chemical structure and sweet taste of isocoumarins and related compounds. Chem Senses 4(1):35–47
54. Yang X, Chong Y, Yan A, Chen J (2011) In-silico prediction of sweetness of sugars and sweeteners. Food Chem 128(3):653–658
55. Zhong M, Chong Y, Nie X, Yan A, Yuan Q (2013) Prediction of sweetness by multilinear regression analysis and support vector machine. J Food Sci 78(9):S1445–S1450
56. Paulus K, Reisch AM (1980) The influence of temperature on the threshold values of primary tastes. Chem Senses 5(1):11–21
57. Open Babel, Open Babel: The Open Source Chemistry Toolbox. http://openbabel.org/
58. Berthold M, Cebron N, Dill F, Gabriel T, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2008) KNIME: the konstanz information miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R (eds) Data analysis, machine learning and applications. Studies in classification, data analysis, and knowledge organization. Springer, Berlin Heidelberg, pp 319–326
59. Hypercube, Inc., HyperChem. http://www.hyper.com
60. TALETE, srl., Dragon (version 6) (2015). Software for Molecular Descriptor Calculation, http://www.talete.mi.it/
61. Pearlman RS (1993) 3D molecular structures: generation and use in 3D searching. In: Kubinyi H (ed) 3D QSAR in drug design. Theory and applications, Springer Science & Business Media, pp 41–79
62. Doweyko AM (2004) 3D-QSAR illusions. J Comput Aided Mol Des 18(7–9):587–596
63. Hechinger M, Leonhard K, Marquardt W (2012) What is wrong with quantitative structure-property relations models based on three-dimensional descriptors? J Chem Inf Model 52(8):1984–1993
64. Kowalski B, Bender C (1972) k-Nearest neighbor classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. Anal Chem 44(8):1405–1411
65. Ballabio D, Consonni V, Mauri A, Claeys-Bruno M, Sergent M, Todeschini R (2014) A novel variable reduction method adapted from space-filling designs. Chemometr Intell Lab Syst 136:147–154
66. Leardi R, Gonzalez AL (1998) Genetic algorithms applied to feature selection in PLS regression: how and when to use them. Chemometr Intell Lab Syst 41(2):195–207
67. Ballabio D, Consonni V (2013) Classification tools in chemistry. Part 1: linear models. PLS-DA. Anal Methods 5(16):3790–3798
68. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemometr Intell Lab Syst 2(1):37–52
69. Jolliffe IT (1986) Principal component analysis. Springer Science + Business Media, Berlin
70. Krzanowski W (1988) Principles of multivariate analysis: a user's perspective. Oxford University Press, Oxford
71. Bro R, Smilde AK (2014) Principal component analysis. Anal Methods 6(9):2812–2831
72. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R (2012) Comparison of different approaches to define the applicability domain of QSAR models. Molecules 17(5):4791–4810
73. Sahigara F, Ballabio D, Todeschini R, Consonni V (2013) Defining a novel *k*-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. J Cheminfor 5:27
74. Cassotti M, Consonni V, Mauri A, Ballabio D (2014) Validation and extension of a similarity-based approach for prediction of acute aquatic toxicity towards Daphnia magna. SAR QSAR Environ Res 25(12):1013–1036
75. Cassotti M, Ballabio D, Consonni V, Mauri A, Tetko IV, Todeschini R (2014) Prediction of acute aquatic toxicity toward Daphnia magna by using the GA-kNN method. Altern Lab Anim 42:31–41
76. Cassotti M, Ballabio D, Todeschini R, Consonni V (2015) A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas). SAR QSAR Environ Res 26(3):217–243
77. Ballabio D (2015) A MATLAB toolbox for principal component analysis and unsupervised exploration of data structure. Chemometr Intell Lab Syst 149:1–9
78. MathWorks: Natick, MatLab (version 7.13.0.564) (2011). http://www.mathworks.com

79. Carhart RE, Smith DH, Venkataraghavan R (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. J Chem Inf Comput Sci 25(2):64–73
80. Birch GG, Karim R, Lopez A (1994) Novel aspects of structure-activity relationships in sweet taste chemoreception. Food Qual Prefer 5(1):87–93
81. Rojas C, Tripaldi P, Duchowicz PR (2016) A new QSPR study on relative sweetness. Int J Quant Struct Prop Relatsh 1(1):76–90
82. Ghose AK, Viswanadhan VN, Wendoloski JJ (1998) Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. J Phys Chem A 102(21):3762–3772
83. Baek N-I, Chung M-S, Shamon L, Kardono LBS, Tsauri S, Padmawinata K, Pezzuto JM, Soejarto DD, Kinghorn AD (1993) Selligueain A, a novel highly sweet proanthocyanidin from the rhizomes of *Selliguea feei*. J Nat Prod 56(9):1532–1538
84. Birch GG (1987) Sweetness and sweeteners. Endeavour 11(1):21–24
85. Birch G, Mylvaganam A (1976) Evidence for the proximity of sweet and bitter receptor sites. Nature 260:632–634
86. van der Heijden A, van der Wel H, Peer HG (1985) Structure-activity relationships in sweeteners. I. Nitroanilines, sulphamates, oximes, isocoumarins and dipeptides. Chem Senses 10(1):57–72
87. Katritzky AR, Petrukhin R, Perumal S, Karelson M, Prakash I, Desai N (2002) A QSPR study of sweetness potency using the CODESSA program. Croat Chem Acta 75(2):475–502