

Fidele Ntie-Kang^{1,2,3}

Mechanistic role of plant-based bitter principles and bitterness prediction for natural product studies II: prediction tools and case studies

¹ Department of Chemistry, University of Buea, P. O. Box 63 Buea, Buea, Cameroon, E-mail: ntiekfidele@gmail.com² Department of Pharmaceutical Chemistry, Martin-Luther University Halle-Wittenberg, Wolfgang-Langenbeck Str. 4, Halle (Saale) 06120, Germany, E-mail: ntiekfidele@gmail.com³ Department of Informatics and Chemistry, University of Chemistry and Technology Prague, Technická 5 166 28 Prague 6, Dejvice, Czech Republic, E-mail: ntiekfidele@gmail.com

Abstract:

The first part of this chapter provides an overview of computer-based tools (algorithms, web servers, and software) for the prediction of bitterness in compounds. These tools all implement machine learning (ML) methods and are all freely accessible. For each tool, a brief description of the implemented method is provided, along with the training sets and the benchmarking results. In the second part, an attempt has been made to explain at the mechanistic level why some medicinal plants are bitter and how plants use bitter natural compounds, obtained through the biosynthetic process as important ingredients for adapting to the environment. A further exploration is made on the role of bitter natural products in the defense mechanism of plants against insect pest, herbivores, and other invaders. Case studies have focused on alkaloids, terpenoids, cyanogenic glucosides and phenolic derivatives.

Keywords: bitterness, computer-based tools, drug discovery, machine learning, medicinal plants, natural products

DOI: 10.1515/psr-2019-0007

1 Introduction

Many natural products (NPs) are bitter, including many drugs in the market [1]. In the previous chapter [2], an attempt was made to explain taste on a molecular basis in humans, based on molecules interacting with bitter taste receptors (hTas2Rs) [3]. NPs or secondary metabolites (SMs) are compounds of natural origin but which do not appear to have any direct functions in growth and development, i. e. they have no generally recognized roles in the process of photosynthesis, respiration, solute transport, translocation, nutrient assimilation, and differentiation. NPs and SMs play a significant role in the direct defense against herbivores by impairing performance by one of two general mechanisms:


- they may reduce the nutritional value of plant food, or
- they may act as feeding deterrents or toxins.

There has been considerable debate as to which of these two strategies is more important for host plant selection and herbivore resistance, particularly related to questions as to what extent variation in the levels of primary and secondary metabolites has evolved as a plant defense [4].

Medicinal plants are known to make several metabolites for the purpose of setting up their defense mechanism and to protect themselves against predators. Bitterness is an inherent property of many toxic chemicals, protecting humans and animals from self-poisoning. It must, however, be recalled that many toxins are not bitter, that many important bitter NPs have been used as rather curative agents (not toxins) [5], and that some known bitter chemicals do not deter herbivores [6]. Apart from the fact that several bitter principles are known to be therapeutic [5], a number of *in silico* models have been developed to predict the toxicity of chemicals based on chemical structure [7].

The focus was on structure-based, ligand-based, and machine learning (ML) methods. Bitterness is a deterrent factor for orally administered drugs. Due to the expensive and laborious experimental screening for

Fidele Ntie-Kang is the corresponding author.

 © 2019 Ntie-Kang, published by De Gruyter.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.

determining if a compound (in foods or drugs) is bitter, *in silico* models are urgently needed. Besides, bitter compounds are quite diverse in chemical structure and are currently known to bind 21 out of the 25 known hTas2Rs, making bitterness prediction quite a difficult task.

In this chapter, an attempt is made to provide a mechanistic view of the role of bitter principles in the defense mechanism of plants. The first part, however, provides an overview of computer-based tools for the prediction of bitterness. These are based on ML models, e. g. decision trees (DTs), random forest (RF), support vector machines (SVMs), etc., built from known datasets of bitter compounds (e. g. BitterDB [8–10]) *versus* known sweet compounds and/or known non-bitter compounds.

2 Available tools for predicting bitterness

Several tools are currently available for predicting bitterness in chemical compounds, all of which are based on ML methods (Table 1). The description is presented based on the methods implemented, the datasets used in the training and test sets, the main results obtained, etc. The tools in Table 1 are arranged according to the order of publication.

Table 1: Summary of software/tools for bitterness prediction.

Tools	Methods implemented/useful links	Advantages	References
BitterX	Support vector machines (SVMs) classifier. http://mdl.shsmu.edu.cn/BitterX	– Freely accessible web server. – Includes a user-friendly web interface.	[11]
BitterPredict	Decision trees-based tool for predicting bitterness. https://github.com/Niv-Lab/BitterPredict1	– The Matlab code is accessible.	[12]
BitterSweetForest	Random forest classifier for distinguishing bitter from sweet compounds. http://bioinformatics.charite.de/sweet/	– Open access tool. – Runs on a fully available KNIME workflow.	[13]
e-Bitter	Uses consensus votes from multiple machine learning methods (e. g. deep learning etc.). https://winpython.github.io/ https://www.dropbox.com/sh/3sebvza3qzmazda/AADgpCRXJtHAJzS8DK_P-q0ka?dl=0	– Adopts a consensus model for the prediction. – Has a free stand-alone software.	[14]
BitterSweet	Combines random forest and adaptive boosting (AB) to enhance classifier performance in order to predict both bitterness and sweetness. https://github.com/cosylabiiiit/bittersweet/ https://github.com/cosylabiiiit/bittersweet/data/	– The software is freely available for non-commercial use. – The prediction server is equipped with a user-friendly interface. – Data used in constructing the models are freely available for download. – Training sets cover a wider chemical space than previous models.	[15]

2.1 A note on machine learning methods

Full coverage of ML is beyond the scope of this chapter. Basically, ML methods are algorithms that are trained to find patterns within data and could be classified as supervised (e. g. deep neural networks, support vector machines, etc.) and unsupervised (e. g. random forests). Let us first define a couple of terms recurrent in ML:

2.1.1 Training set

The dataset of compounds used to build the model. Like the test set, this is often composed of compounds with a certain characteristic and those that do not have, e. g. bitter compounds *versus* non-bitter compounds.

2.1.2 Test set

The dataset of compounds used to prove or validate the model.

2.1.3 Decision trees

A DT uses a tree-like model of decisions and their possible consequences. They are used commonly used in operations research to arrive at decisions but are also popular in ML. Typically, a DT is a flowchart-like structure in which each internal node represents a “test” on an attribute (e. g. either an event occurs or not), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from the root to the leaf represent classification rules. DTs can easily become unstable, i. e. a small change in the data could lead to a large change in the structure of the optimal DT. As a result, predictors derived by other ML methods would perform better with similar data. One way to solve this problem could be by replacing a single DT with a random forest (RF) of DTs. However, RF is not often easy to interpret when compared with a single DT.

2.1.4 Artificial neural network

Artificial neural networks (ANN) are inspired by the neural networks in the biological system that form animal brains since the original goal of the ANN approach was to solve problems as the human brain would do. However, the neural network is not an algorithm in itself, but a framework where several different ML algorithms work together in order to treat the data inputs. As such, an ANN is based on a collection of connected units (nodes) called artificial neurons (AN), with each connection functioning like the synapses in the brain, processing and transmitting signals from one AN to another. At a connection, the signal between ANs is a real number, and the output is computed by some non-linear function of the sum of its inputs. The connections between ANs are called “edges”, ANs and edges typically having a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. ANs may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. Typically, ANs are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer (the input layer) to the last layer (the output layer), possibly after traversing the layers many times.

2.1.5 Deep neuron network

Deep neuron network or deep neural network (DNN) is a neural network with more than one hidden layer between the input and output layers. In DNN, thousands of neurons in each layer can be extensively applied to the dataset with thousands of features, and more advanced regularization technique such as the dropout can be used to prevent the overfitting problem. Nevertheless, DNN requires users to adjust a variety of parameters.

2.1.6 k-nearest neighbors

The k-nearest neighbors (k-NN) algorithm is among the simplest of all ML algorithms. Both for classification and regression, a weight is assigned to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of $1/d$, where d is the distance to the neighbor. This can be thought of as the training set for the algorithm, though no explicit training step is required.

2.1.7 Random forest

An RF algorithm is a type of ensemble learning method that constructs a large number of decisions trees (usually greater than 100), and outputs predictions based on a collection of the votes of the individual trees. A subset of the training dataset is chosen to grow individual trees, while the remaining samples are used to estimate the optimal fit. The constructed trees are grown by splitting the training set (subset) at each node according to the value of the random variable, which is sampled independently from a subset of variables.

2.1.8 Support vector machine

Support vector machine (SVM) or support vector network is a popular supervised ML technique that is used for classification and regression. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. The algorithm performs the classification by constructing the hyper-planes in the multi-dimensional space that separates the different classes.

2.1.9 Validation or performance assessment

The performance of statistical learning methods like ML is often measured by the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). In this scenario, TP , TN , FP , and FN would represent true bitterant, true non-bitterant, false bitterant, and false non-bitterant compounds, respectively.

Precision (P_{re}) is a measure of accuracy for a specific, predicted class.

$$P_{re} = \frac{TP}{TP + FP} \quad (1)$$

Accuracy (A_{cc}) is another frequently used index for the overall classification performance, but it may be misleading due to the highly unbalanced class distribution in the used datasets.

$$A_{cc} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Sensitivity (S_e) or *recall* and *specificity* (S_p) can assess a model's ability to correctly identify TP s and TN s, respectively. These two parameters are usually interpreted in combination with each other.

$$S_e = \frac{TP}{TP + FN} \quad (3)$$

$$S_p = \frac{TN}{TN + FP} \quad (4)$$

The indices in eqs. (1) to (4) are often used for model validation and comparison.

In addition, the *F1 measure* (or *F1-score*) and *non-error rate* (*NER*) are defined, respectively, as:

$$F1\ measure = \frac{2 \cdot P_{re} \cdot S_e}{P_{re} + S_e} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (5)$$

$$NER = \frac{S_e \cdot S_p}{2} \quad (6)$$

F1-score (cross-validation) is evaluated on the internal validation dataset during the cross-validation. Meanwhile, *F1-score* (test test) is when *F1-score* is assessed on the test set. $\Delta F1$ -score is the absolute value of the difference between *F1-score* (cross-validation) and *F1-score* (test set), i. e.:

$$\Delta F1 - score = |F1 - score (cross - validation) - F1 - score (test test)| \quad (7)$$

$\Delta F1$ -score is used to monitor the potential overfitting (or underfitting), i. e. if $\Delta F1$ -score is small, it means that the model performances are similar on the internal-validation dataset and test set.

2.1.10 Area under the curve (AUC)

This prediction metric is derived from a Receiver Operator Characteristics (ROC) plot. The ROC curve is a plot of the *FP* rate ($1 - S_p$) on the *y*-axis against the *TP* rate (S_e) on the *x*-axis while varying the decision threshold. The area under the curve (AUC) of the ROC plot provides a convenient way of comparing classifiers. An AUC value of 0.5 represents a random classifier, while an ideal classifier has an area of 1.0.

2.1.11 Matthews correlation coefficient

F1-score and Matthews correlation coefficient (MCC), eq. (8), are commonly used to measure the quality of binary classifications.

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

2.1.12 Y-randomization

This is a tool used in the validation of statistical models, whereby the performance of the original model in data description (r^2) is compared to that of models built for permuted (randomly shuffled) responses, based on the original descriptor pool and the original model building procedure [16].

2.1.13 Principal component analysis (PCA)

PCA is a mathematical technique that captures the linear interactions between the underlying attributes in a dataset. Every principal component can be expressed as a combination of one or more existing variables. All principal components are orthogonal to each other, and each one captures some amount of variance in the data.

2.2 BitterX – a support vector machines bitterness predictor

BitterX was the first web server tool that could be used to predict the human bitter taste receptors that bind certain small molecules [11]. It is available at <http://mdl.shsmu.edu.cn/BitterX>, with a web interface, Figure 1.

This tool serves two functions:

- identifying if a compound is a bitterant (or bitter taste receptor activator) and
- predicting its possible bitter taste receptors (Tas2Rs).

The SVMs model was built using a training set manually curated from the literature using PubMed and BitterDB [8]. This included 540 bitterants, i. e. 260 positive and 2379 negative bitterant-Tas2R interactions. Data on the bitterant and bitterant-Tas2R interactions were collected manually from the literature in order to be used for identifying bitterant-Tas2R interactions. The molecular structure file of each bitter compound obtained from PubChem [17] was input into a program Checker and ChemAxon's Standardizer for predicting the interactions with Tas2Rs. The benchmark evaluations showed that the models for bitterant determination and receptor recognition could accurately predict the activities of the test dataset [11]. Besides, BitterX could accurately predict the known Tas2Rs of several experimentally proven bitterants.



Figure 1: Web interface and output of BitterX [11]; (A) Query input in homepage (B) Bitterant-Tas2R interaction entries in “Browse” page. (C) An example of an output page after submitting a chemical molecule. A confidence score in probability is displayed along with the associated Tas2R in both the “Receptor List” and the Column Chart, which can be retrieved by clicking “Show Receptor Histogram”. Material reproduced from data originally published under a Creative Commons (CC BY) License.

2.3 BitterPredict – a decision trees-based tool for predicting taste from chemical structure

BitterPredict predicts whether a compound is bitter or not and is built on the adaptive boosting (AdaBoost) DTs classifier [12]. It implements an algorithm in which the DTs are built sequentially by learning from misclassified samples of the former DT. The positive training set includes 632 molecules from BitterDB [8], while about 2,000 non-bitter molecules were gathered from literature to create the negative set. The non-bitter set was composed into three subsets: non-bitter flavors, sweet molecules, and tasteless molecules. The classifier was based on physicochemical and ADME/Tox descriptors. BitterPredict was able to correctly classify >80 % of the compounds in the hold-out test set, and 70–90 % of the compounds in three independent external sets and in sensory test validation. This implies that BitterPredict is a quick and reliable tool for classifying large sets of compounds into bitter and non-bitter groups. In addition, the tool suggested ~ 40 % of random molecules, and a large portion (66 %) of clinical and experimental drugs, and of NPs (77 %) to be bitter. The Matlab code for BitterPredict is provided *via* BitterDB <http://bitterdb.agri.huji.ac.il/dbbitter.php#Bitter-Predict> and *via* GitHub repository <https://github.com/Niv-Lab/BitterPredict1>.

2.4 BitterSweetForest – a random forest open access tool

BitterSweetForest [13] uses a random forest (RF) classifier, based on molecular fingerprints that were used to discriminate between sweet- and bitter-tasting molecules. It is an open access model and is implemented on a KNIME workflow [18] that provides a platform for predicting if a compound would be bitter or sweet. A training set 1,202 of compounds, i. e. 517 artificial and natural sweeteners from the SuperSweet [19] against

685 bitter compounds from the BitterDB [8], was used to construct the model. The original model yielded an accuracy of 95 % and an area under the curve (AUC) of 0.98 in cross-validation. The model was validated using an independent test set with an accuracy of 96 % and an AUC of 0.98 for bitter and sweet taste prediction. This was then applied for the prediction of bitterness and sweetness in NPs from the Super natural II dataset [20], approved drugs from Drugbank [21], and known toxic compounds (with experimentally proven acute oral toxicity) from the Protox web server [22]. The BitterSweetForest tool predicted that up to 70 % and 10 % of the NPs from the Super natural II dataset, to be bitter and sweet, respectively, with a confidence score of 0.60 and above. In the same way, 77 % and 2 % of the approved drugs were predicted as bitter and as sweet, respectively, with a confidence score of 0.75 and above. Moreover, 75 % of the toxic compounds were predicted only as bitter with a minimum confidence score of 0.75. This model, thus, suggested that toxic compounds, NPs, and approved drugs are mostly bitter.

2.5 e-Bitter – a free software for bitterness prediction

The e-Bitter tool [13] is based on harnessed consensus votes from the multiple machine-learning methods (e. g. deep learning), combined with molecular fingerprints, to build classification models of compounds into either bitter or bitterless (non-bitter). The training set is composed of 707 experimentally proven bitterants (a majority from BitterDB [8]) and 592 non-bitterants (including sweet compounds downloaded from the Super-Sweet dataset [19] and SweetenersDB [23], along with 132 tasteless and 17 non-bitter compounds retrieved from the literature). The extended-connectivity fingerprint (ECFP) [24] was adopted as the molecular descriptor to build the bitter/bitterless classification models. Five algorithms – k-NN, SVM, RF, gradient boosting machine (GBM), and DNN – were used to train the models via the Scikit-learn, Keras, and TensorFlow python libraries. The model was validated with a five-fold cross-validation.

Through an exhaustive parameter exploration with the five-fold cross-validation, all the models are carefully scrutinized by the Y-randomization test to ensure their reliability, and subsequently nine consensus models are constructed based on the individual or average models, which differ in term of accuracy, speed, and diversity of models. One of the best consensus models showed that accuracy, precision, specificity, sensitivity, *F1-score*, and Matthews correlation coefficient (MCC) gave respective values of 0.929, 0.918, 0.898, 0.954, 0.936, and 0.856 on the test set. It was additionally demonstrated that e-Bitter outperforms BitterX on three test sets, while showing better results than BitterPredict for two test sets.

A graphical user interface (Figure 2) was developed for the convenience of users. The tool is unique in that it adopts a consensus model for bitterness prediction and was the first free stand-alone software for bitterness prediction. Another advantage is that the entire training dataset is publicly available from the e-Bitter program and users can view the 3D structure of each compound and its corresponding classification as bitter or bitterless (Y: bitterant or N: non-bitterant).

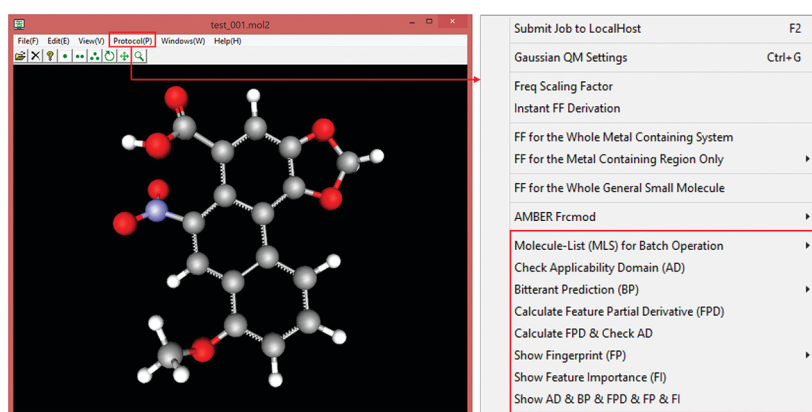


Figure 2: The basic functions in the e-Bitter program, which is highlighted by the red rectangle [14]. Material reproduced from data originally published under a Creative Commons (CC BY) License.

2.6 BitterSweet – a freely available state-of-the-art software for bitter versus sweet taste prediction

This is the most recently published tool for bitterness (and sweetness) prediction, which combines random forest and adaptive boosting to enhance classifier performance [15]. The dimensionality of the molecular descriptors was reduced using principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) [25]. The effort was motivated by the inconsistencies observed in the curation process that led

to the training datasets used to develop the models implemented in the previously described tools [11–14]. These include possible incorrect predictions that could result from including molecules with unverified taste information or incomplete representation of chemical space in the training set. For example, the training sets for constructing BitterX [11] and BitterPredict [12] included compounds with unverified non-bitterness (corresponding to 50 % and 55.6 % of non-bitter compounds, respectively), while BitterSweetForest [13] and e-Bitter [14] only used experimentally verified data. This considerably reduced the size of the training sets (and eventually) the possible bitter-sweet chemical space representation in the models.

BitterSweet is built on an exhaustive compilation of bitter, non-bitter, sweet, and non-sweet compounds from the literature, aimed at spanning the chemical space while not compromising the accuracy of taste information of the molecules. Its training set includes 918 bitter and 1510 non-bitter molecules as well as 1205 sweet and 1171 non-sweet molecules resulting from the careful curation of data from a wide variety of sources, ranging from scientific publications to books. Tasteless molecules were included as important controls for both bitter and sweet taste prediction. The datasets were separated into training and test sets, with the test set taken from the external validation/test sets obtained for the BitterPredict models [12].

The bitter-sweet taste prediction models were trained and evaluated using a wide spectrum of molecular descriptors, e. g. Dragon 2D/3D quantitative structure-activity relationships (QSAR) descriptors [26], ECFPs, physicochemical as well as ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties from Canvas [27], as well as structural and physicochemical descriptors from ChemoPy [28]. Thus, BitterSweet implements state-of-the-art ML models for bitter-sweet taste prediction, whose performance has been proven on large specialized chemical sets, e. g. FlavorDB [29], FoodDB (<http://foodb.ca>), SuperSweet [19], Super Natural II [20], DSSTox [30], and DrugBank [21]. All datasets for building the BitterSweet models have been made publicly available (<https://github.com/cosylabiiit/bittersweet/>). In addition, the BitterSweet predictor is implementable in a freely available software for bitter- and sweet-taste prediction.

3 Bitter natural products from plants

In general, plants make several bitter principles, including polyphenols (e. g. flavonoids, isoflavonoids, tannins, etc.) and alkaloids. Tannins, for example, are particularly useful in repelling unwanted insect predators, while flavonoids are cytotoxic to the herbivores by interacting with different enzymes through complexation [31]. The defense mechanism of plants against predators and harsh environmental conditions is complex, involving direct defense (e. g. by forming thorns, spines, prickles, hard waxy leaves, etc.), induced resistance (Figure 3) and indirect plant defense (Figure 4). Bitter principles only play a minor role, since some herbivores are known to tolerate bitter principles [5].

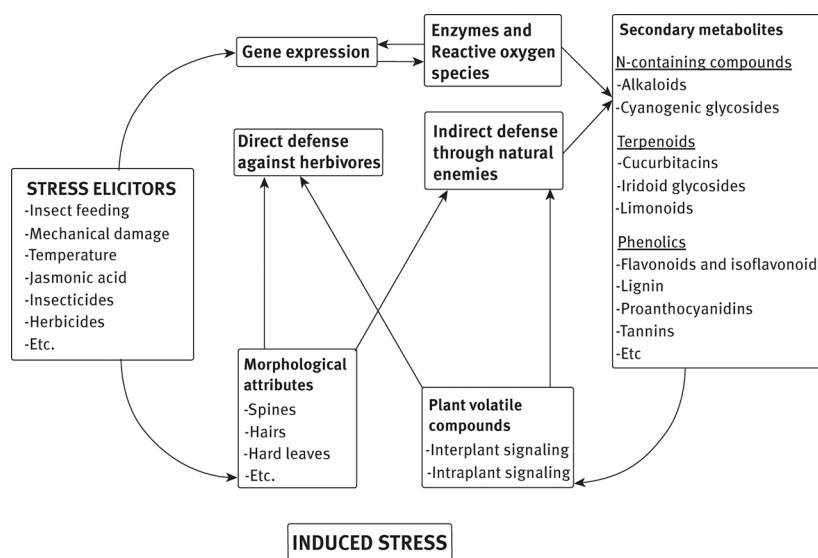


Figure 3: Induced resistance in plants. Figure adapted from reference [32].

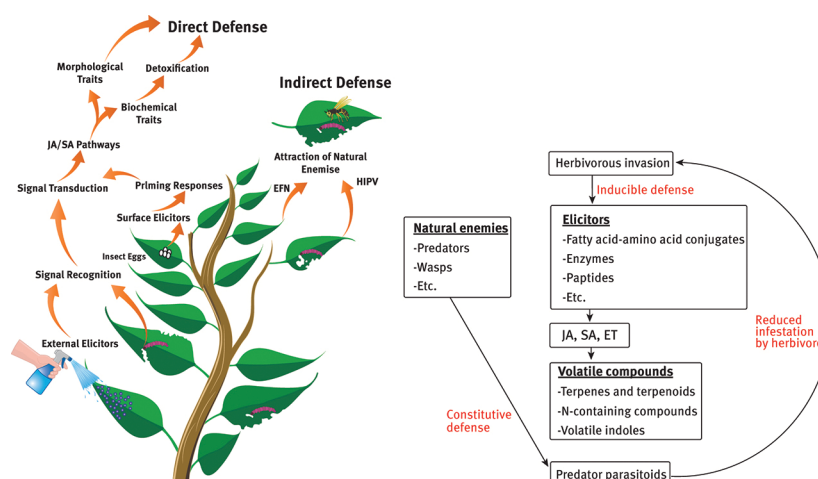


Figure 4: (Left) Plant defense against insect pests (EPF = extra floral nectar; HIPV = herbivore induced plant volatiles; JA = jasmonic acid; SA = salicylic acid). (Right) Major components and pathways involved in indirect plant defense (ET = Ethylene). Figure adapted from references [32, 33].

Induced resistance, against insect attack, for example, involves many signal transduction pathways mediated by a network of phytohormones (or plant hormones) [32, 33]. Phytohormones play a critical role in regulating plant growth, development, and defense mechanisms. The signal transduction pathways are mediated by jasmonic acid (JA, an important phytohormone), salicylic acid, and ethylene (Figure 4). Specific sets of defense-related genes are activated by these pathways upon wounding the plant or by insect feeding. JA is derived in octadecanoid pathway from linolenic acid. Rising levels of JA in response to herbivore attack often triggers the production of several proteins involved in plant defenses, e. g. proteins that inhibit digestion in the herbivore. Phytohormones may act individually, synergistically, or antagonistically, depending upon the attacker. The phytohormone accumulates upon wounding and or destruction of plant tissue by herbivores. Chewing of plant parts by insects, for example, causes the dioxygenation of linoleic and linolenic acids.

Ethylene is another important phytohormone, which plays an active role in plant defense against many insects, the ethylene signaling pathway playing an important role in induced plant defense against insects and pathogens both directly and indirectly. Ethylene signaling pathway works either synergistically or antagonistically, with JA in expression of plant defense responses against pathogens and herbivorous insects. It has been reported that ethylene and JA work together in tomato in proteinase inhibitors expression.

3.1 Why are some bitter plants medicinal?

It has been mentioned from the previous paragraphs that many NPs, including those contained in medicinal plants, are bitter [12]. The bitter taste and hence the therapeutic potential of these plants could be attributed to the presence of the former. The fact that plants do contain bioactive principles does not mean that plant had humans in mind when biosynthesizing the bioactive metabolite. The plants actually biosynthesized the bitter principle to defend itself against predators (including insects, mammal herbivores, etc.) and disease-causing pathogens. Humans only later discovered the curative properties of the plants in the quest to find treatment for their own diseases. The identification of the bioactive principles in the plants responsible for their therapeutic uses only followed later. SMs can either be stored in the inactive form or induced in response to the herbivore or microbe attack.

An example where a plant produces bitter bioactive metabolites as a result of pathogenic attack is seen in the process of cyanogenesis (i. e. the ability of living organisms to liberate a toxic compound hydrogen cyanide from stored nontoxic cyanogenic glycosides) [34–37]. This process involves the conversion of phytoanticipins (e. g. cyanogenic glycosides) to phytoalexins (e. g. antimicrobial compounds synthesized by plants that accumulate rapidly at areas of pathogen infection). The phytoanticipins are mainly activated by β -glucosidase when the plant is consumed by herbivores (Figure 5). This in turn mediates the release of various biocidal aglycone metabolites. Phytoalexins include isoflavonoids, terpenoids, alkaloids, etc., which, for example, influence the performance and survival of insects. Apart of the fact that the aforementioned compounds classes have a broad range of known biological activities useful for treatment of several human ailments, such SMs both defend the plants from different stresses and also increase the fitness of the plants [38].

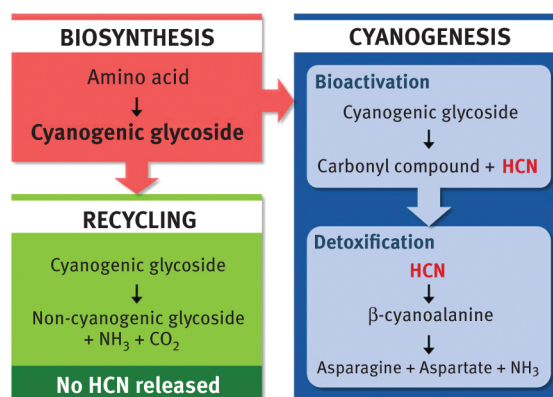


Figure 5: Summary of biosynthesis of cyanogenic glucosides and mechanism of cyanogenesis [34]. Figure reproduced by permission.

3.2 Understanding the role of natural bitter compounds in plant defense mechanisms

3.2.1 Defensive plant-based bitter principles

SMs involved in plant defense (collectively known as antiherbivory compounds) are often classified into three sub-groups: nitrogen-containing compounds (including alkaloids, cyanogenic glycosides, glucosinolates, and benzoxazinoids), terpenoids, and phenolics. The chemical structures of some plant-based bitter principles which are known to play a defensive role are shown in Figure 6. Among them, plant phenols are the most common and widespread group of defensive compounds. Their major role is in resistance against insects, microorganisms, and competing plants. Cucurbitacins, for example, are bitter compounds whose hostility to a wide range of herbivores, including lepidopteran larvae, beetles, mites, and vertebrate grazers, can be explained by their taste [32]. The known defensive functions of some bitter SMs have been summarized in Table 2.

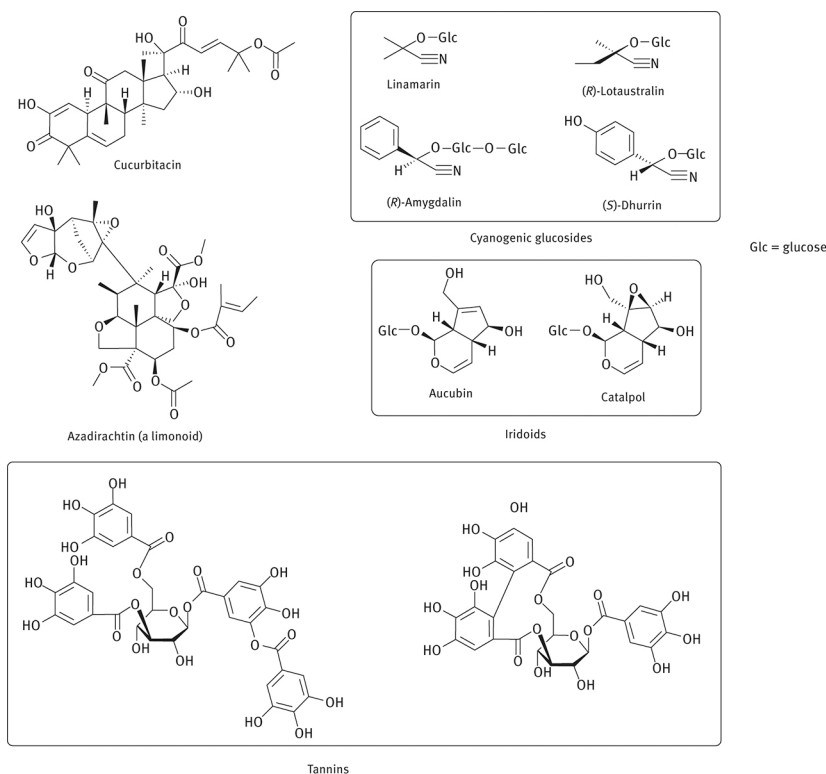


Figure 6: Chemical structures of some bitter principles involved in diverse defensive roles in plants.

Table 2: Defensive role of plant-based bitter compounds.

Compound or compound class	Description	Function/mode of action	References
Alkaloids	NPs containing an N-containing heterocycle and are biosynthesized from various amino acids.	– Can inhibit or activate enzymes. – Alter carbohydrate and fat storage by inhibiting the formation of phosphodiester bonds involved in their breakdown. – Bind to nucleic acids and can inhibit the synthesis of proteins and affect DNA repair mechanisms. – Affect cell membrane and cytoskeletal structure causing the cells to weaken, collapse, or leak, and can affect nerve transmission.	[39, 40]
Cucurbitacins	A bitter class of triterpenes that makes plants hostile to a wide range of herbivores.	– Directly affect insect growth and development. – Act indirectly by acting as oviposition deterrents. – Act as phagostimulants to insect pests. – Cucurbitacins B and D have been reported as major phagostimulants for leaf beetles.	[41–43]
Cyanogenic glycosides	Bitter and toxic principles. Found in more than 2650 different plant species.	– These are bitter compounds that deter herbivores. – They are degraded by enzymes (e. g. β -glucosidases and hydroxynitrile lyases) to liberate (toxic) hydrogen cyanide to deter herbivores. – Formation of their breakdown products is initiated by microbial attack or consumption of plant material by herbivores.	[34–37]
Flavonoids and isoflavonoids	Bitter polyphenols common in several plant families.	– Cytotoxic and interact with different enzymes through complexation. – Both classes of compounds protect the plant against insect pests by influencing the behavior, and growth and development of insects. – They are very important in plant resistance against pathogenic bacteria and fungi. – They may also be directly involved in the inhibition of the pathogen's enzymes, especially those digesting the plant cell wall, by chelating metals required for their activity.	[42–44]
Iridoid glycosides	A class of bitter cyclopentanoid monoterpene-derived compounds.	– In plants, they act primarily as a defense against herbivores or against infection by microorganisms. – Directly toxic to insect pests. – Reduce the nutritional quality of plant tissues, thereby rendering them less digestible to insects. – Denature amino acids, proteins, and nucleic acids by binding covalently to nucleophilic side chains via imine formation. – Inhibit the activity of enzymes involved in the formation of prostaglandins and leukotrienes.	[45]
Lignin	A phenolic heteropolymer whose synthesis has been found to be induced by insect or pathogen attack.	– Limits the entry of pathogens by physically blocking the leaves. – Increases the leaf toughness, hence reducing the number of leaves consumed by insects. – Decreases the nutritional content of the leaf. – Its rapid deposition further reduces the growth of the pathogen or reduces insect fecundity.	[46, 47]
Limonoids	Phytochemicals of the triterpenoid class.	– Azadirachtin, for example, is a naturally occurring insecticide from this class of compounds. – Active as a feeding inhibitor towards the desert locust (<i>Schistocerca gregaria</i>). – Acts mainly as an antifeedant and growth disruptor.	[48, 49]

Proanthocyanidins	Oligomeric or polymeric flavonoids, also known as condensed tannins.	They serve among other chemical and induced defense mechanisms against plant pathogens and predators.	[50–53]
Tannins	Bitter, toxic, and poor tasting polyphenols that deter insects from feeding on plants containing them.	– Known to influence insect growth and development by binding to the predator vital proteins. – Reduce nutrient absorption efficiency and cause midgut lesions. – Precipitate proteins nonspecifically (including the digestive enzymes of insects), by hydrogen bonding or covalent bonding of protein NH ₂ groups. – Chelate the metal ions, thereby reducing their bioavailability to insects. – When ingested, they reduce the digestibility of the proteins thereby decreasing the nutritive value of plants and plant parts to insects.	[31, 54]

Popular plant-based alkaloids include nicotine, caffeine, morphine, cocaine, colchicine, ergolines, strychnine, and quinine. Their (mostly) aversively bitter taste is a natural deterrent to herbivores. While alkaloids mostly act on receptors of neurotransmitters, others (such as phenolics and terpenoids) are less specific and attack a multitude of proteins by building hydrogen, hydrophobic, and ionic bonds, thus modulating their 3D structures and in consequence their bioactivities.

Cyanogenic glycosides (often stored in inactive forms in plant vacuoles) become toxic when consumed by herbivores. This is because the rupture of the plant cell membranes releases the glycosides which bring them into contact with HCN-releasing enzymes in the cytoplasm (Figure 5). HCN is known to be highly toxic by blocking cellular respiration [55]. Glucosinolates are activated in a much similar manner, but the products of their breakdown rather cause milder effects like gastroenteritis, salivation, diarrhea, and mouth irritation [40]. Benzoxazinoids are also stored as inactive glucosides in plant vacuoles. Upon plant tissue disruption when herbivores feed, these antiherbivory compounds get into contact with β -glucosidases from the chloroplasts and lead to the enzymatic release of toxic aglucones [56]. Some benzoxazinoids are only synthesized when herbivores start feeding. Such SMs are considered to act by an induced plant defense mechanism [57].

Among the terpenoids, diterpenoids are widely distributed in latex and resins and can be quite toxic. The high toxicity of *Rhododendron* leaves can be attributed to the presence of diterpenoids [58]. Saponins are complex triterpenoids which are known to break down the red blood cells of herbivores [59]. Among the limonoids (a sub-class of terpenoids), azadirachtin is a well-known naturally occurring insecticide, which is active as a feeding inhibitor towards the desert locust (*Schistocerca gregaria*), acting mainly as an antifeedant and growth disruptor [48, 49]. Iridoid glycosides (another sub-class terpenoids), e. g. aucubin and catalpol, prevent the invasion of plants by insects and microorganisms [45].

Phenolics are known to exhibit antiseptic properties, while others disrupt endocrine activity. From simple tannins to the more complex flavonoids which confer on plants much of their colorful pigments, polyphenols are known for several activities, e. g. antioxidant activity. Some polyphenols are involved in plant defense mechanisms, e. g. lignin, silymarin, and cannabinoids [46, 47]. Condensed tannins (e. g. proanthocyanidins), including 2 to >50 flavonoid molecules, inhibit herbivore digestion by binding to the consumed plant proteins, rendering digestion difficult or almost impossible. This is done by the SMs interfering with protein absorption and digestive enzymes [50–53, 60].

3.2.2 Case study: investigating human taste receptors of phenolic compounds

Soares et al. undertook an investigation of 6 polyphenols against several human taste receptors, with the view of identifying which receptors are activated by these compounds [61]. The compounds included; the hydrolyzable tannin pentagalloylglucose (PGG), the precursor of condensed tannins (-)-epicatechin, two procyanidin oligomers or condensed tannins (the dimer B3 and the trimer C2), and the anthocyanins malvidin-3-glucoside and cyanidin-3-glucoside, which are commonly found in plant-based foods and drinks, e. g. red wine, beer, tea, and chocolate. The chemical structures of the investigated compounds have been shown on Figure 7.

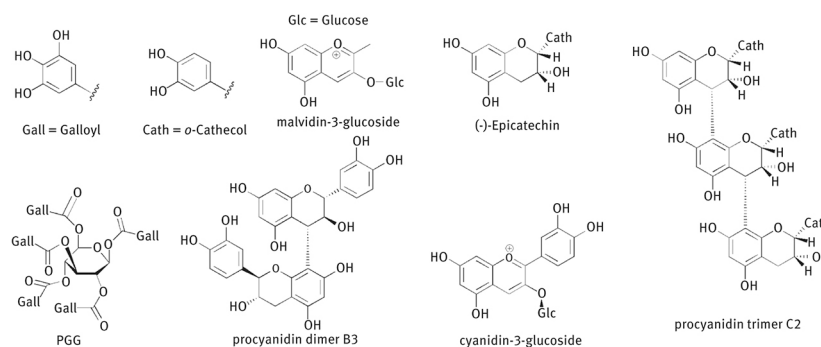


Figure 7: Chemical structures of some plant-based bitter principles investigated in the study [61].

The observed EC_{50} values for the different compounds vary 100-fold, the lowest values being for PGG and malvidin-3-glucoside. The compounds were shown to activate different combinations of the 25 hTas2Rs, e. g. (-)-epicatechin activated 3 of the receptors (i. e. hTas2R4, hTas2R5, and hTas2R39), while PGG activated only two receptors (i. e. hTas2R5 and hTas2R39). Meanwhile, malvidin-3-glucoside and procyanidin trimer only stimulated one receptor each (i. e. hTas2R7 and hTas2R5, respectively). The authors remarkably discovered tannins to be the first selective natural agonists for the receptor hTas2R5, with a high potency only toward this receptor. The authors also suggested the catechol and/or galloyl groups to be important structural determinants for mediating the interaction of these polyphenolic compounds with this receptor. This hypothesis could be verified by docking the compounds against the receptor site. The conclusions of this study would lead to the suggestion that the presence of these polyphenols in the food items could explain the bitterness of fruits, vegetables, and derived products even when in low concentrations.

4 Conclusions

Predicting bitterness is a very costly and challenging task for both the pharmaceutical and the food industry. In this work, a brief presentation of ML methods and recent ML tools (algorithms, web servers, and software) for taste prediction has been provided, with an emphasis on bitter/sweet compounds. Some of these methods are freely implemented in software packages and in some cases the tools for implementing them, e. g. curated datasets, KNIME workflows are available for download. Such tools/datasets would be quite useful for predicting if a SM is bitter or not and which receptor it is most likely to activate. SMs in general and bitter principles are known to be important in defending plants against predators and harsh environmental conditions. The usefulness of bitter principles from plants in defense against pathogens and herbivores has been highlighted in the second part of this work. Emphasis has been laid on nitrogen compounds (including alkaloids, cyanogenic glycosides, glucosinolates, and benzoxazinoids), terpenoids (iridoids, limonoids, and cucurbitacins), and phenolic compounds (flavonoids, isoflavonoids, lignin, proanthocyanidins, and tannins), which are involved in the defense of plants against herbivores and pathogens.

Acknowledgements

The author acknowledges a return fellowship and an equipment subsidy from the Alexander von Humboldt Foundation, Germany. Financial support for this work is acknowledged from the Ministry of Education, Youth and Sports of the Czech Republic.

List of Abbreviations

Abbreviation Full meaning

AB adaptive boosting

AN artificial neurons

ANN artificial neural networks

AUC area under the curve

BitterDB database of bitter compounds and receptors
 DNN deep neuron network
 DT decision tree
 ECFP extended-connectivity fingerprints
 FN false negatives
 FP false positives
 hTas2Rs human bitter taste receptors
 JA jasmonic acid or jasmonate
 k-NN k-nearest neighbors
 MCC Matthews correlation coefficient
 ML machine learning
 NPs natural products
 PCA principal component analysis
 PGG pentagalloylglucose
 RF random forest
 ROC Receiver Operator Characteristics
 SVMs support vector machines
 TN true negatives
 TP true positives
 t-SNE t-distributed stochastic neighbor embedding

References

- [1] Drewnowski A, Gomez-Carneros C. Bitter taste, phytonutrients and the consumer: a review. Am J Clin Nutr. 2000;72:1424–35.
- [2] Ntie-Kang F. Mechanistic role of plant-based bitter principles and bitterness prediction for natural product studies I: database and methods. Phys Sci Rev 2019. DOI: 10.1515/psr-2018-0117.
- [3] Shaik FA, Singh N, Arakawa M, Duan K, Bhullar RP, Chelikani P. Bitter taste receptors: extraoral roles in pathophysiology. Int J Biochem Cell Biol. 2016;77:197–204.
- [4] Berenbaum MR. Turnabout is fair play: secondary roles for primary compounds. J Chem Ecol. 1995;21:925–40.
- [5] Nolte DL, Russell Mason J, Lewis SL. Tolerance of bitter compounds by an herbivore, *Cavia porcellus*. J Chem Ecol. 1994;20:303–8.
- [6] Barratt-Fornell A, Drewnowski A. The taste of health: nature's bitter gifts. Nutr Today. 2002;37:144–50.
- [7] Bahia MS, Nissim I, Niv MY. Bitterness prediction *in-silico*: a step towards better drugs. Int J Pharm. 2017;536:526–9.
- [8] Wiener A, Shudler M, Levit A, Niv MY. BitterDB: a database of bitter compounds. Nucleic Acids Res. 2012;40:D413–9.
- [9] Dagan-Wiener A, Di Pizio A, Nissim I, Bahia MS, Dubovski N, Margulis E, et al. Bitter DB: taste ligands and receptors database in 2019. Nucleic Acids Res. 2019;47:D1179–85.
- [10] Bitter DB. Institute of Biochemistry, Food Science and Nutrition, Faculty of Agriculture, The Hebrew University of Jerusalem. Available at: <http://bitterdb.agri.huji.ac.il/dbbitter.php>. Accessed: 15 Jan 2019.
- [11] Huang W, Shen Q, Su X, Ji M, Liu X, Chen Y, et al. BitterX: a tool for understanding bitter taste in humans. Sci Rep. 2016;6:23450.
- [12] Wiener AD, Nissim I, Abu NB, Borgonovo G, Bassoli A, Niv MY. Bitter or not? BitterPredict, a tool for predicting taste from chemical structure. Sci Rep. 2017;7:12074.
- [13] Banerjee P, Preissner R. BitterSweetForest: a random forest based binary classifier to predict bitterness and sweetness of chemical compounds. Front Chem. 2018;6:93.
- [14] Zheng S, Jiang M, Zhao C, Zhu R, Hu Z, Xu Y, et al. e-Bitter: bitterant prediction by the consensus voting from the machine-learning methods. Front Chem. 2018;6:82.
- [15] Tuwani R, Wadhwa S, Bagler G. BitterSweet: building machine learning models for predicting the bitter and sweet taste of small molecules. bioRxiv. 2018. doi:10.1101/426692.
- [16] Rücker C, Rücker G, Meringer M. y-Randomization and its variants in QSPR/QSAR. J Chem Inf Model. 2007;47:2345–57.
- [17] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. Nucleic Acids Res. 2016;44:D1202–13.
- [18] Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME: the Konstanz information Miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, editors. Data analysis, machine learning and applications SE - 38, Studies in Classification, Data Analysis, and Knowledge Organization. Berlin Heidelberg: Springer, 2008:319–26.
- [19] Ahmed J, Preissner S, Dunkel M, Worth CL, Eckert A, Preissner R. SuperSweet – a resource on natural and artificial sweetening agents. Nucleic Acids Res. 2011;39:D377–82.

- [20] Banerjee P, Erehman J, Gohlke BO, Wilhelm T, Preissner R, Dunkel M. Super natural II – a database of natural products. *Nucleic Acids Res.* 2015;43:D935–9.
- [21] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res.* 2018;46:D1074–82.
- [22] Drwal MN, Banerjee P, Dunkel M, Wettig MR, Preissner R. ProTox: a web server for the *in silico* prediction of rodent oral toxicity. *Nucleic Acids Res.* 2014;42:W53–8.
- [23] Chéron JB, Casciuc I, Golebiowski J, Antonczak S, Fiorucci S. [Sweetness prediction of natural compounds](#). *Food Chem.* 2017;221:1421–5.
- [24] Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model.* 2010;50:742–54.
- [25] van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res.* 2008;9:2579–605.
- [26] Mauri A, Consonni V, Pavan M, Todeschini R. Dragon software: an easy approach to molecular descriptor calculations. *Match Commun Math Comput Chem.* 2006;56:237–48.
- [27] Duan J, Dixon SL, Lowrie JF, Sherman W. Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J Mol Graph Model.* 2010;29:157–70.
- [28] Cao DS, Xu QS, Hu QN, Liang YZ. ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics.* 2013;29:1092–4.
- [29] Garg N, Garg N, Sethupathy A, Tuwani R, Nk R, Dokania S, et al. [FlavorDB: a database of flavor molecules](#). *Nucleic Acids Res.* 2018;46:D1210–6.
- [30] Richard AM, Williams CR. [Distributed structure-searchable toxicity \(DSSTox\) public database network: a proposal](#). *Mutat Res.* 2002;499:27–52.
- [31] Belete T. Defense mechanisms of plants to insect pests: from morphological to biochemical approach. *Trends Tech Sci Res.* 2018;2:555584.
- [32] War AR, Taggar GK, Hussain B, Taggar MS, Nair RM, Hari C, et al. Plant defence against herbivory and insect adaptations. *AoB Plants.* 2018;10:ply037.
- [33] War AR, Paulraj MC, Ahmad T, Buhroo AA, Hussain B, Ignacimuthu S, et al. Mechanisms of plant defense against insect herbivores. *Plant Signal Behav.* 2012;7:1306–20.
- [34] Zagrobelny M, Bak S, Møller BL. [Cyanogenesis in plants and arthropods](#). *Phytochemistry.* 2008;69:1457–68.
- [35] Fürstenberg-Hägg J, Zagrobelny M, Bak S. [Plant defense against insect herbivores](#). *Int J Mol Sci.* 2013;14:10242–97.
- [36] Sánchez-Pérez R, Jørgensen K, Olsen CE, Dicenta F, Lindberg Møller B. [Bitterness in almonds](#). *Plant Physiol.* 2008;146:1040–52.
- [37] Malagón J, Garrido A. Relación entre el contenido de glicósidos cianogénicos y la resistencia a *Capnodis tenebrionis* L. en frutales de hueso. *Bol Sanid Veg Plagas.* 1990;16:499–503.
- [38] Agrawal AA. Induced responses to herbivory in wild radish: effects on several herbivores and plant fitness. *Ecology.* 1999;80:1713–23.
- [39] Roberts MF, Wink M. Alkaloids: biochemistry, ecology, and medicinal applications. New York: Plenum Press, 1998. ISBN 978-0-306-45465-3.
- [40] Rhoades DF. Evolution of plant chemical defense against herbivores. In: Rosenthal GA, Janzen DH, editors. *Herbivores: their interaction with secondary plant metabolites*. New York: Academic Press, 1979:3–54. ISBN 978-0-12-597180-5.
- [41] Chen JC, Chiu MH, Nie RL, Cordell GA, Qiu SX. [Cucurbitacins and cucurbitane glycosides: structures and biological activities](#). *Nat Prod Rep.* 2005;22:386–99.
- [42] Treutter D. Significance of flavonoids in plant resistance and enhancement of their biosynthesis. *Plant Biol.* 2005;7:581–91.
- [43] Blount JW, Dixon RA, Paiva NL. Stress responses in alfalfa (*Medicago sativa* L.) XVI. Antifungal activity of medicarpin and its biosynthetic precursors; implications for the genetic manipulation of stress metabolites. *Physiol Mol Plant Pathol.* 1992;41:333–49.
- [44] Dai GH, Nicole M, Andary C, Martinez C, Bresson E, Boher B, et al. Flavonoids accumulate in cell walls, middle lamellae and callose-rich papillae during an incompatible interaction between *Xanthomonas campestris* pv. *malvacearum* and cotton. *Physiol Mol Plant Pathol.* 1996;49:285–306.
- [45] Stephenson AG. Iridoid glycosides in the nectar of *Catalpa speciosa* are unpalatable to nectar thieves. *J Chem Ecol.* 1982;8:1025–34.
- [46] Bagniewska-Zadworna A, Barakat A, Łakomy P, Smoliński DJ, Zadworny M. Lignin and lignans in plant defence: insight from expression profiling of cinnamyl alcohol dehydrogenase genes during development and following fungal infection in *Populus*. *Plant Sci.* 2014;229:111–21.
- [47] Liu Q, Luo L, Zheng L. [Lignins: biosynthesis and biological functions in plants](#). *Int J Mol Sci.* 2018;19:335.
- [48] Mordue AJ, Blackwell A. [Azadirachtin: an update](#). *J Insect Physiol.* 1993;39:903–24.
- [49] Lee JW, Jin CL, Jang KC, Choi GH, Lee HD, Kim JH. Investigation on the insecticidal limonoid content of commercial biopesticides and neem extract using solid phase extraction. *J Agric Chem Env.* 2013;2:81–5.
- [50] Laaksonen OA, Salminen JP, Mäkilä L, Kallio HP, Yang B. [Proanthocyanidins and their contribution to sensory attributes of black currant juices](#). *J Agric Food Chem.* 2015;63:5373–80.
- [51] Ferreira D, Marais JP, Coleman CM, Slade D. Comprehensive natural products II 6.18 - proanthocyanidins: chemistry and biology. In: Liu HW, Mander L, editor(s). *Comprehensive natural products II chemistry and biology* Vol. 6. Oxford, UK: Elsevier, 2010:605–61.
- [52] Ma X, Yang W, Laaksonen O, Nylander M, Kallio H, Yang B. Role of flavonols and proanthocyanidins in the sensory quality of sea buckthorn (*Hippophaë rhamnoides* L.) berries. *J Agric Food Chem.* 2017;65:9871–9.
- [53] Amil-Ruiz F, Blanco-Portales R, Munoz-Blanco J, Caballero JL. [The strawberry plant defense mechanism: a molecular review](#). *Plant Cell Physiol.* 2011;52:1873–903.
- [54] Wink M. [Modes of action of herbal medicines and plant secondary metabolites](#). *Medicines.* 2015;2:251–86.
- [55] Niemeyer HM. [Plant cyanogenic glycosides](#). *Toxicon.* 2000;38:11–36.
- [56] Niemeyer HM. Hydroxamic acids derived from 2-hydroxy-2H-1,4-benzoxazin-3(4H)-one: key defense chemicals of cereals. *J Agric Food Chem.* 2009;57:1677–96.
- [57] Glauser G, Marti G, Villard N, Doyen GA, Wolfender JL, Turlings TC, et al. Induction and detoxification of maize 1,4-benzoxazin-3-ones by insect herbivores. *Plant J.* 2011;68:901–11.
- [58] Wink M, van Wyk BE. *Mind-altering and poisonous plants of the world*. Portland, OR, USA: Timber Press; 2010.

- [59] Herrmann F, Wink M. Synergistic interactions of saponins and monoterpenes in HeLa and Cos7 cells and in erythrocytes. *Phytomedicine*. 2011;18:1191–6.
- [60] van Soest PJ. Nutritional ecology of the ruminant: ruminant metabolism, nutritional strategies, the cellulolytic fermentation, and the chemistry of forages and plant fibers. Corvallis, Oregon: O & B Books, 1982. ISBN 978-0-9601586-0-7.
- [61] Soares S, Kohl S, Thalmann S, Mateus N, Meyerhof W, De Freitas V. Different phenolic compounds activate distinct human bitter taste receptors. *J Agric Food Chem*. 2013;61:1525–33.