Flavonoid Production Predictions

Maggie J. Horton

Kettering University Department of Computer Science

# Author Note

This senior thesis is submitted as partial fulfillment of the graduation requirements for Kettering University for the purpose of obtaining a Bachelor of Science in Computer Science. Any conclusions drawn or opinions expressed within this thesis are my own and are not representative of the position and values held by Kettering University and any other individuals involved with this culminating undergraduate experience.

Though this thesis documents the outcomes of my personal work, I would like to extend my sincere appreciation to the following individuals for their valuable help: Dr. Veronica R. Moorman (Associate Professor of Biochemistry), Dr. Montserrat Rabago-Smith (Associate Professor of Chemistry), Dr. James I. Cohen (Former Professor of Biology), Abigail M. Hadden (Undergraduate Research Student), Jordan R. Wilson (Past Undergraduate Research Student), Jinny L. Shows (Past Undergraduate Research Student), Angela M. Ghannam (Past Undergraduate Research Student) and, Kiah S. Lowe (Past Undergraduate Research Student). Without their efforts this thesis would not be possible, and for that I am extremely grateful.

# Abstract

This project is focused on both theoretical predictions and experimental results of research concerning biosynthesis of select flavonoids: genistein, cyanidin, butein, isoliquiritigenin, catechin, epicatechin, epigallocatechin, gallocatechin, eriodictyol, naringenin, kaempferol, myricetin, quercetin, apigenin, and luteolin. The theoretical predictions are the result of code written in the Python programming language by utilizing data from the Kyoto Encyclopedia of Genes and Genomes (KEGG). The experimental data was collected by conducting a database and literature, involving different methods of data acquisition and processing. The results of the comparisons yielded somewhat mixed results, as there was a significant number of results from the literature and database search that both agreed and disagreed with the programmatic predictions. The comparisons for kaempferol and naringenin were very promising as the predictions for both had at least 40 experimental confirmations and zero experimental disagreements. This means that the species predicted for these compounds would be the best basis for designing tests for determining the phytochemical composition of selected species. The comparisons for cyanidin, eriodictyol, and quercetin were less promising. The predictions for eriodictyol had no experimental disagreements but only a small amount of experimental agreements, while the predictions for cyanidin and quercetin each had two experimental disagreements, which may require an additional literature search focused entirely on said compounds in order to better gauge the program accuracy.

Keywords: anthocyanidins, biosynthesis, chalcones, flavan-3-ols, flavonoids, flavonols, flavanones, flavones, isoflavonoids, python, scripting.

# Table of Contents

[Author Note](#_gjdgxs) 1

[Abstract](#_30j0zll) 2

[Table of Contents](#_1fob9te) 3

[Introduction](#_jvwsnynxj85h) 4

[Conclusions and Recommendations](#_j4xcm5nsppb8) 18

[Methods](#_nc9zsy2qppiy) 23

[Results](#_3l18frh) 37

[Discussion](#_5hys7x57leao) 57

[References](#_717g86qfuhxu) 76

[Appendix](#_9q2kdl80n0kw) 88

# 

# 

# 

# 

# Introduction

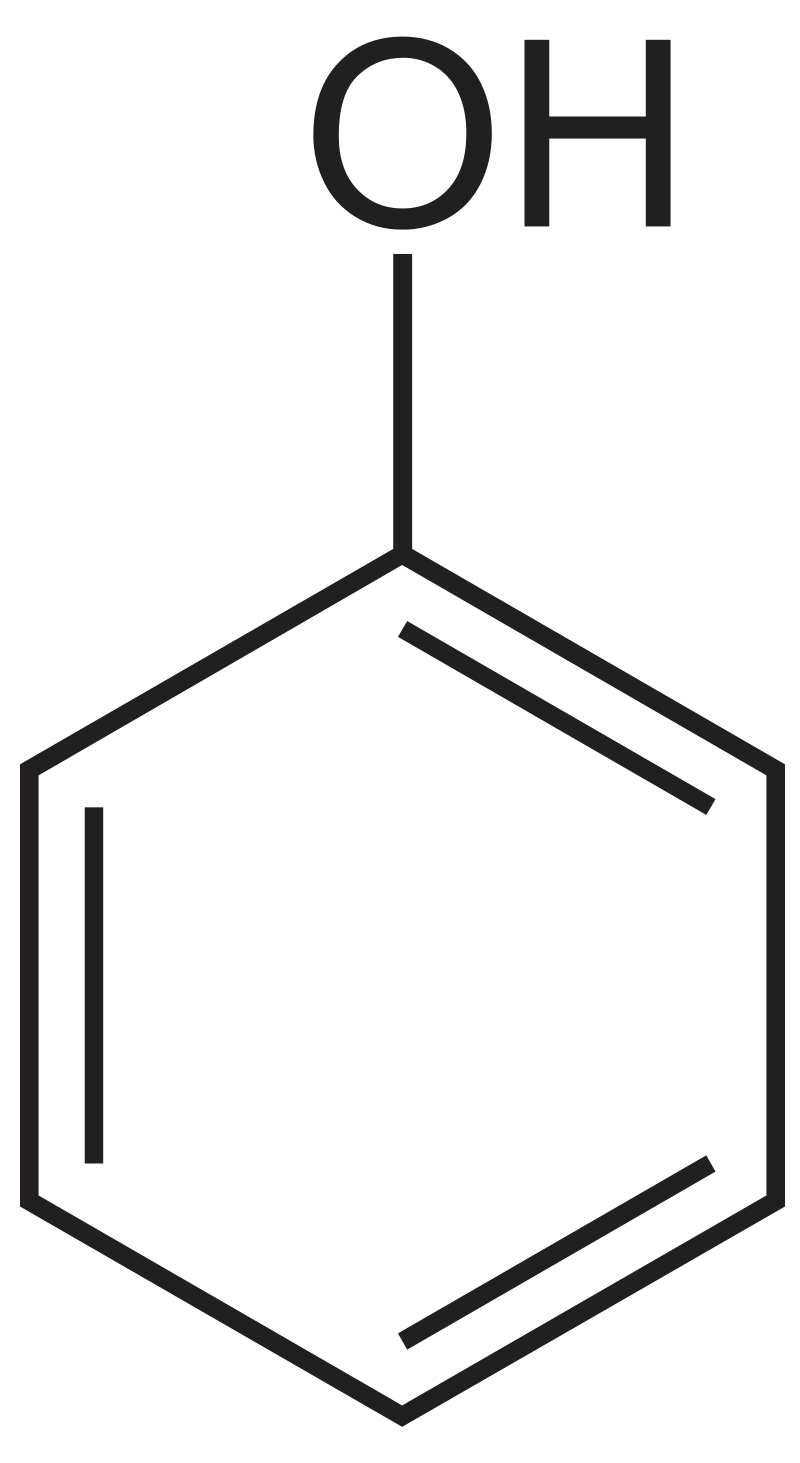
## Background

### Flavonoids

Flavonoids are a class of polyphenolic compounds, meaning that phenols are a defining part of their structure as well as a contributing factor of their beneficial activities. Phenol is a relatively simple compound as shown in Figures 1 and 2, yet is very useful due to its antimicrobial properties, making it a key ingredient in slimicides as well as a competent disinfectant and antiseptic (National Center for Biotechnology Information, 2021).

###### Figure 1

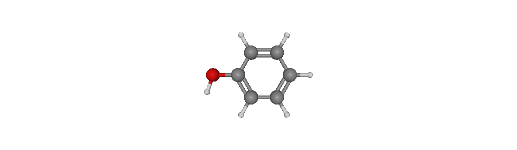
2D structure of phenol.

[](https://en.wikipedia.org/wiki/Phenol#/media/File:Phenol2.svg)

From the Wikipedia entry for phenol (NEUROtiker, 2007).

###### Figure 2

3D structure of phenol



From the PubChem entry for phenol (National Center for Biotechnology Information, 2021).

The defining characteristic that distinguishes flavonoids from other polyphenols is its characteristic 15-carbon skeleton. This simple skeleton consists of phenolic A and B rings, that typically have an accompanying OH group, and an additional C ring as shown in Figure 3.

###### Figure 3

Basic flavonoid structure.

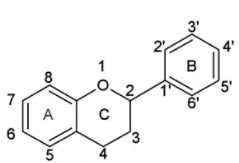
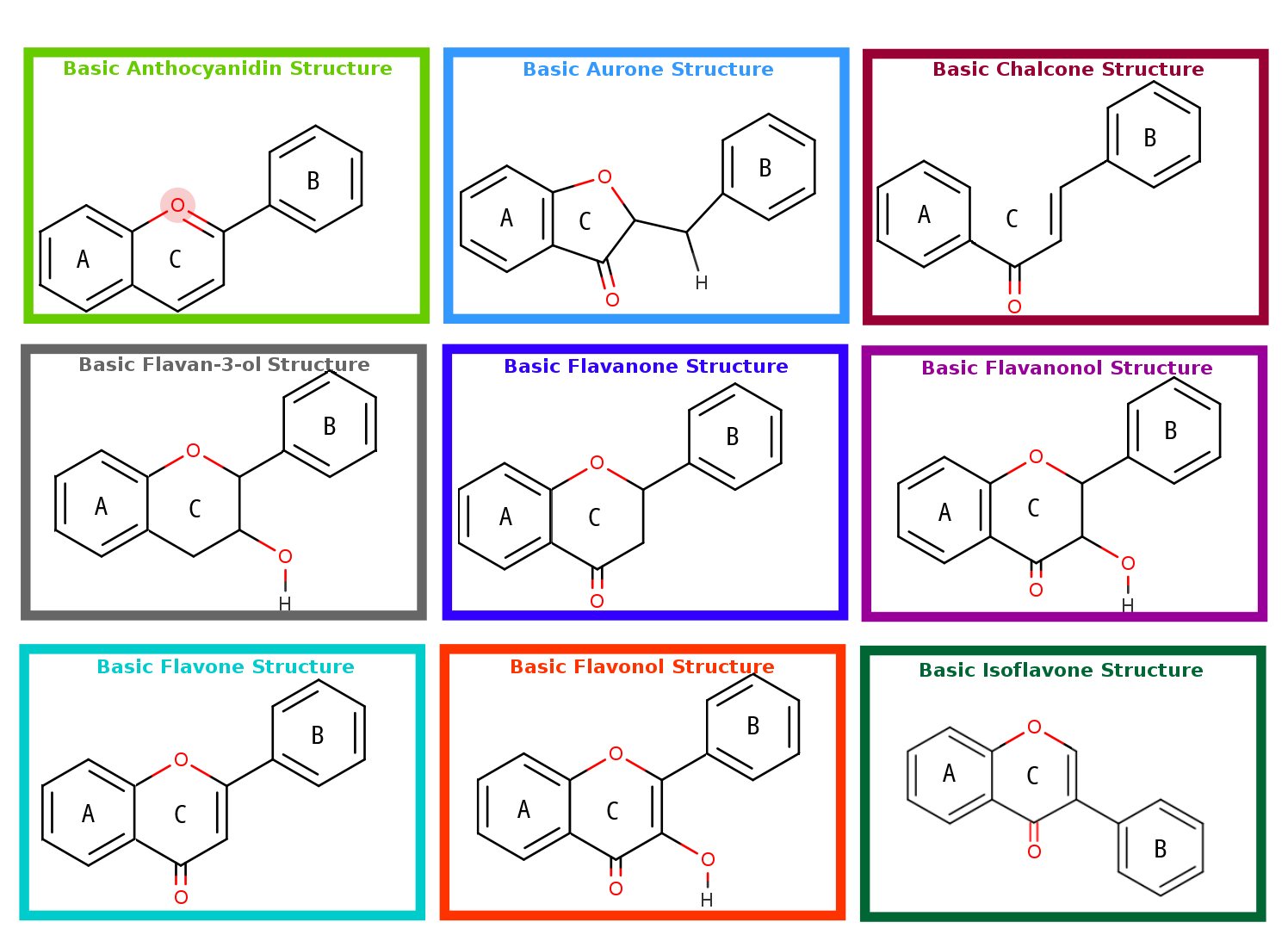


Image adapted from *Panche et al., 2016*

Since flavonoids are such a large class of compounds, it is helpful to break these compounds down into different subclasses. These subclasses are determined by the specific modifications of the C ring, which can be seen in the diagrams of each subclass' skeleton, shown in Figure 4.

###### Figure 4

Diagram of differing flavonoid structures.



Flavonoids are a highly diverse and extensive class of *secondary metabolites*, beneficial compounds that are not essential for reproduction and development (LibreTexts, 2021), and can be produced in most of the parts of a living plant (leaves, roots, flowers, etc.) (Dias et al., 2021). Since they are so diverse in their structures and distributions, it would then follow that their functional properties and availability would vary as well. For instance, anthocyanidins can provide red, blue, and purple pigmentation (Tuladhar et al., 2021) while flavones can provide white pigmentation or co-pigmentation (Hostetler et al., 2017), which means that a plant’s coloration could tie into the availability of select flavonoids. While pigments can be beneficial for attracting pollinators, flavonoids also can also play a part in tactical and protective advantages, including: UV protection, chemical signaling and antifungal activities (Robertson, 2021).

The helpful benefits of flavonoids are not exclusive to the plant world though. Humans greatly benefit from the antioxidant properties of the compounds as they are extremely common amongst the plants which play a part in the human diet (Panche, et al., 2016). The antioxidant properties of flavonoids are common across the large class of compounds, yet the medicinal potential does not stop there as several different beneficial properties have been recorded to date such as antibacterial (Xie et al., 2015), antiviral (Lalani & Poh, 2021) and anti-inflammatory (Panche, et al., 2016) activities. This potential is what makes flavonoids so interesting and promising, resulting in an abundance of ongoing research for the purpose of discovering new activities and medicinal applications.

### Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG (https://www.genome.jp.kegg/) is a freely accessible database resource, sponsored by the Kanehisa Laboratories of the Institute of Chemical Research of Kyoto University. It contains and has access to large amounts of information that is relevant for understanding the biological systems of animals, plants, fungi, protists, bacteria, and archaea, including: enzymes, proteins, pathways, genes and entire genomes. Since the database is a culmination of work that dates back to 1995 and is focused on connecting current knowledge and high throughput computing, the data that is available is extensive and is stored in an extremely interconnected fashion. That is, an entry in KEGG will have its own unique identifiers, along with the identifiers of associated entries such as genes, compounds or reactions, which makes the process of accessing specific information much easier.

### 

### Genes and Genomes

Genes are a part of living organisms and are the vehicle through which traits are passed down from parent to offspring (National Library of Medicine, 2021) and are responsible for providing the instructions for making essential macromolecules, like proteins. They are able to do this due to a special molecular building-block: *deoxyribonucleic acid* (DNA). DNA is a molecule which carries genetic information and instructions by using unique sequences of specific chemical bases (National Human Genome Research Institute) and is what allows for genes to code for specific proteins. A single gene only has so much influence over an organism however, and are but a smaller part of a collective referred to as a genome. An organism’s genome is the entirety of its genetic information which determines how an organism forms, grows and functions. Due to the advancements in technology, entire genomes of numerous organisms have been sequenced. In other words, their chemically-coded instructions have been determined, which has been an essential step in better understanding the underlying biological mechanisms of organisms, and has become incredibly valuable knowledge in the context of agriculture and medicine. In the context of this project though, genes and genomes are significant because they allow for a more comprehensive understanding of flavonoid biosynthesis, which relies on the analysis of the specific genes that code for the enzymes that are essential to this process.

### Proteins, Enzymes and EC Numbers

Proteins are another important class of organic macromolecules, which are large and complex molecules found in organisms. They are constructed by structures within a cell called ribosomes, which takes in a copied piece of DNA, in the form of a messenger ribonucleic acid (mRNA) molecule (Bartee et al., 2016), and using those instructions, then assembles a specific sequence of amino acids (National Human Genome Research Institute). Though they provide a wide range of purposes, the most relevant proteins in this context are ones that act as catalysts, or rather, chemicals that increase the rate of essential chemical reactions without being consumed in the process (National Human Genome Research Institute). These proteins are referred to as enzymes and are an essential group of molecules that regulates an organism’s biological functions. There are seven different types of enzymes commonly referred to that are grouped based on their functions and/or associated reactions: *oxidoreductases* catalyze oxidation-reduction reactions; *transferases* catalyze the transfer of molecules’ functional group(s); *hydrolases* break chemical bonds with water; *lyases* catalyze bond-breaking without oxidation or hydrolysis; *isomerases* rearrange the structure of molecules; *ligases* create chemical bonds between molecules (Chou & Elrod, 2003); and *translocases* aid in moving a molecule other than itself (Zhang & He, 2019).

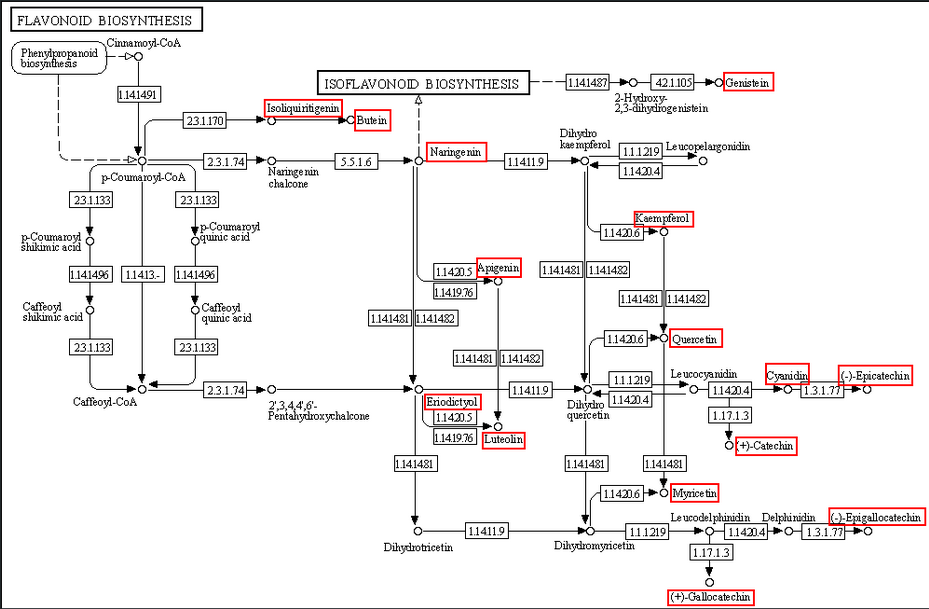
The classification of enzymes is more complex than sorting enzymes into seven -different groups, which is where the *enzyme commission* (EC) number becomes very helpful. EC numbers are used for classifying enzymes and take the format x.x.x.x, where the number in each position is associated with specific identifiers that are determined by the enzyme and its associated reaction(s) (Hu et al. 2012). The first digit corresponds with an enzyme’s class: (1) oxidoreductase, (2) transferase, (3) hydrolase, (4) lyase, (5) isomerase, (6) ligase, and (7) translocases (Tipton & McDonald). The second digit refers to the type of compound(s) involved, the third digit relays more information regarding the reaction type, and the final digit is a serial number used for identifying enzymes within their respective classifications. For example, as shown the enzyme database ExplorEnz’s webpage (https://www.enzyme-database.org/class.php), the EC number for hydroxycinnamoyl-CoA reductase is EC 1.3.1.117 because: it is an oxidoreductase (1.x.x.x) that acts on the CH-CH group of donors (1.3.x.x) with NAD+ or NADP+ as an acceptor (1.3.1.x) and has a serial number of 117 (1.3.1.117) (McDonald et al., 2009). EC numbers are an important portion of this project as they are included in the KEGG database entries for a gene, if the gene happens to code for one or more enzymes, and is extracted by parsing within the program.

### Biological Pathways and Flavonoid Biosynthesis

A *biological pathway* is a specific chain of molecular events within a cell that result in changes to the cell or lead to the production of one or more compounds (National Human Genome Research Institute, 2020). *Biosynthesis*, also referred to as *anabolism*, is a type of biological pathway that results in a product that is produced by enzyme-catalyzed reactions involving simpler precursor compounds (Biology Online, 2021). Numerous different pathways have been understood to date and are represented by diagrams referred to as pathway maps, which visualize the steps of a pathway. KEGG has a collection of pathway maps which use arrows to indicate the progression of reactions along with the pathways relevant precursors, products and enzymes and these maps have been an essential reference in better understanding flavonoid biosynthesis because their database entries also include the known associated genes and enzymes. The maps provided by KEGG are generalized, as in they are not representative of a specific organism and are made by utilizing information from numerous organisms in order to cover a wider range of reactions. Flavonoid biosynthesis requires the products of phenylpropanoid biosynthesis as precursor compounds and then ultimately ends up with the production of select compounds that feed into additional pathways: isoflavonoid biosynthesis, flavone biosynthesis, flavonol biosynthesis, and anthocyanin biosynthesis. Figure 5 shows a modified pathway map of KEGG’s flavonoid biosynthesis (Kanehisa Laboratories, 2019) where it has been connected to the isoflavonoid map (Kanehisa Laboratories, 2021) to include genistein. Additionally, the compounds that do not serve as precursors or intermediates to the flavonoids in the data set have been removed for the sake of simplicity.

###### Figure 5

Simplified combination of the flavonoid and isoflavonoid biosynthesis pathways. The flavonoids of interest are boxed in red. Figure was made using modified maps accessed from KEGG (Kanehisa Laboratories 2019 & 2021) .



### Python

Python is a very popular multi-purpose programming language that is utilized across a wide range of fields. The language has an emphasis on code readability by enforcing levels of indentation in order to create clearly structured and nested code. This emphasis on indentation along with the numerous handy built-in functions (that could take several lines to implement in other programming languages) makes Python easier to learn for those with little to no formal programming experience, making it a popular choice for writing programs relating to the natural sciences. As a result, a multitude of individuals have developed different libraries for the purpose of scientific programming, which only furthers the benefits of the language.

Python is also an ideal choice for this project because of the Bioservices (Cokelaer et al., 2013) library for Python. It is a free collection of code relating to biology and bioinformatics and is extremely diverse in its utility, which includes easy programmatic access of the data hosted by KEGG.

## Purpose and Nature

The data regarding the flavonoid composition of different plants can be considered lacking due to a bias or favoritism towards testing culinary or medicinal plants, as well as a bias which favors certain flavonoids. For instance, butein was first reportedly isolated in 1904 (Perkin & Hummel, 1904) and cyanidin does not appear to have any reference in literature until 1913 (Willstätter & Everest, 1913), yet when either compound is searched in Google scholar, as of December 2021 there are about 58,600 results for cyanidin and only around 13,900 results for butein. Many of the first results are regarding medical studies and applications though, yet one could anticipate that the usage in these applications could inspire more investigation into source organisms. The purpose of this project is to programmatically predict biosynthesis in specific species of plants. The predictions serve to inspire future testing of the chemical contents of said plants by using a method such as high-performance liquid chromatography (HPLC) in the efforts to find traces of select flavonoids. This would help to fill in the gap of missing data relating to the plants in the data set and compounds they are not yet experimentally known to synthesize. Additionally, this project consists of a literature and database search for the purpose of comparing the predictions to existing data and to analyze and comment on the available data in the databases.

### Flavonoids in the Data Set

The flavonoids of interest belong to seven different classes: isoflavonoids (genistein), anthocyanidins (cyanidin), chalcones (butein and isoliquiritigenin), flavan-3-ols ((+)-catechin, (-)-epicatechin, (-)-epigallocatechin and (+)-gallocatechin), flavanones ((-)-eriodictyol and (-)-naringenin), flavones (apigenin and luteolin), and flavonols (kaempferol , myricetin, and quercetin). The structures for each of the compounds can be seen in Figure 6. It should be noted that references to the flavan-3-ols and flavanones will omit the +/- as the project is not concerned with more than one of the compounds’ stereoisomers, which are compounds that are almost identical but differ in their orientation of atoms (Reusch, 2013).

###### Figure 6

Chemical structures of the flavonoids in the data set.

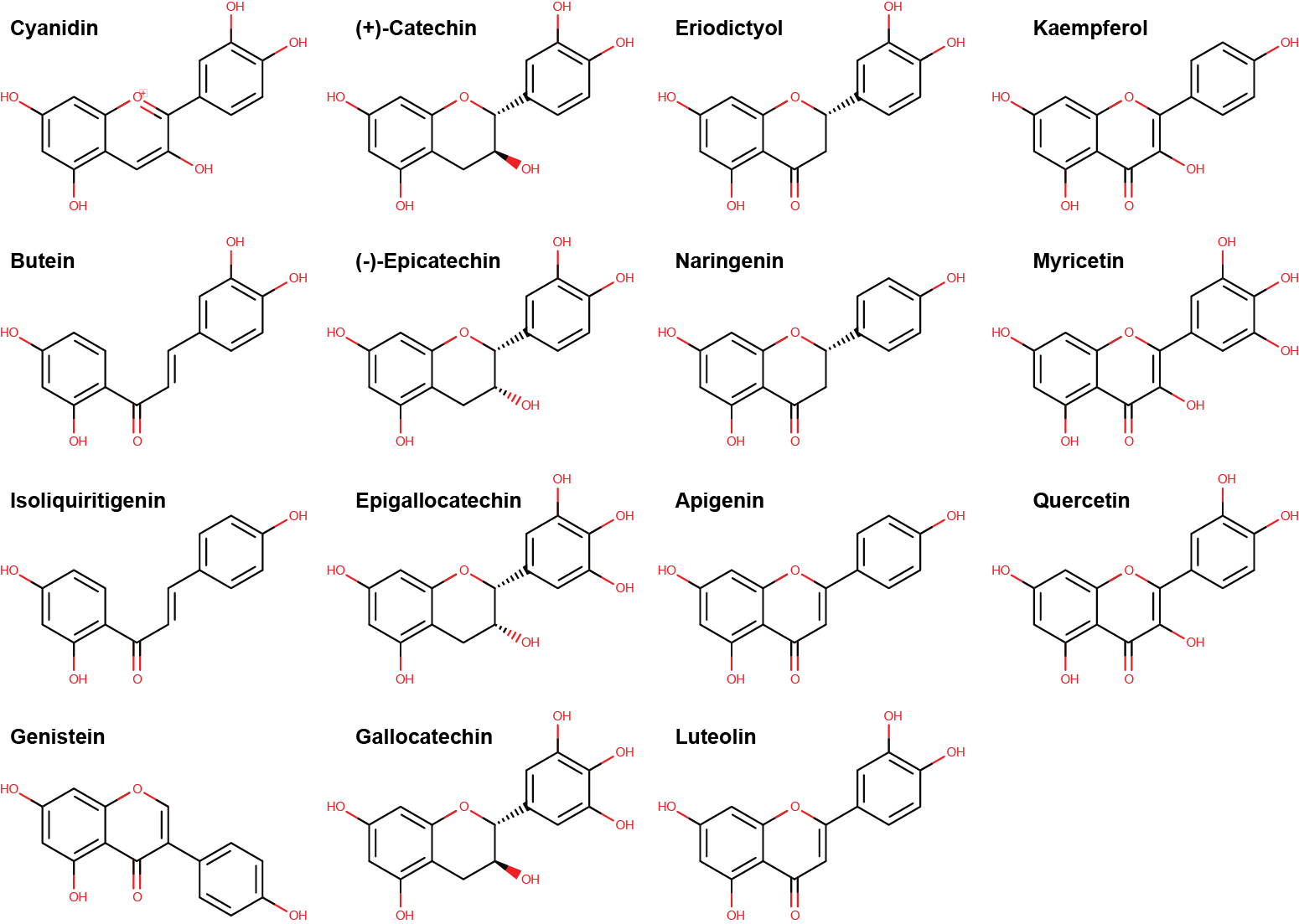


Image created by Dr. Veronica Moorman.

### Plants in the Data Set

There are 105 species within the project data set, which belong to 43 different plant families. The dataset had been determined by the available genomic data for plants provided by KEGG at the time of official project work (October 2020). The majority of these species (90) fall under the classification of flowering land plants. Green algae, red algae, mosses and club mosses make up the remainder of the data set. Table 1 lists the species of the data set, which have been sorted by select taxonomic classifications in order to serve as a visual aid of the species distribution. The taxonomic information was collected from four different online sources: UniProt (Velankar et al., 2013), the NCBI Taxonomy Browser (Schoch et al., 2020), Integrated Taxonomic Information System (ITIS, 2021), and AlgaeBase (Guiry & Guiry, 2022).

###### Table 1

List of the plants in the data set which includes the designated codes assigned by KEGG.

|  | | | | Family | *Species* | KEGG Code |
| --- | --- | --- | --- | --- | --- | --- |
| **Red Algae** | | | | Cyanidiaceae | *Cyanidioschyzon merolae* | cme |
| *Galdieria sulphuraria* | gsl |
| Gigartinaceae | *Chondrus crispus (carrageen moss)* | ccp |
| **Green Algae** | | | | Bathycoccaceae | *Bathycoccus prasinos* | bpg |
| *Ostreococcus lucimarinus* | olu |
| *Ostreococcus tauri* | ota |
| Chlamydomonadaceae | *Chlamydomonas reinhardtii* | cre |
| Chlorellaceae | *Auxenochlorella protothecoides* | apro |
| *Chlorella variabilis* | cvr |
| Coccomyxaceae | *Coccomyxa subellipsoidea* | csl |
| Mamiellaceae | *Micromonas commoda* | mis |
| *Micromonas pusilla* | mpp |
| Selenastraceae | *Monoraphidium neglectum* | mng |
| Volvocaceae | *Volvox carteri f. nagariensis* | vcn |
| **Land Plants** | **Non-vascular Plants** | | | Funariaceae | *Physcomitrium patens (spreading earth moss)* | ppp |
| **Vascular Plants** | **Seedless** | | Selaginellaceae | *Selaginella moellendorffii (spikemoss)* | smo |
| **Flowering Plants** | **Basal Angiosperms** | Amborellaceae | *Amborella trichopoda (primitive flowering shrub)* | atr |
| **Monocots** | Arecaceae | *Elaeis guineensis (African oil palm)* | egu |
| *Phoenix dactylifera (date palm)* | pda |
| Asparagaceae | *Asparagus officinalis (garden asparagus)* | aof |
| Musaceae | *Musa acuminata (wild Malaysian banana)* | mus |
| Orchidaceae | *Dendrobium catenatum (chained dendrobium)* | dct |
| *Phalaenopsis equestris (moth orchid)* | peq |
| Poaceae | *Aegilops tauschii (wheat D)* | ats |
| *Brachypodium distachyon (purple false brome)* | bdi |
| *Oryza brachyantha (malo sina)* | obr |
| *Oryza sativa subsp. japonica (Japanese rice)* | osa/dosa |
| *Setaria italica (foxtail millet)* | sita |
| *Sorghum bicolor (sorghum)* | sbi |
| *Zea mays (maize)* | zma |
| **Eudicots** | Amaranthaceae | *Beta vulgaris (sugar beet)* | bvg |
| *Spinacia oleracea (spinach)* | soe |
| Apiaceae | *Daucus carota (carrot)* | dcr |
| Asteraceae | *Cynara cardunculus var. scolymus (artichoke)* | ccav |
| *Helianthus annuus (common sunflower)* | han |
| *Lactuca sativa (garden lettuce)* | lsv |
| Bombacaceae | *Durio zibethinus (durian)* | dzi |
| Chenopodiaceae | *Chenopodium quinoa (quinoa)* | cqi |
| Convolvulaceae | *Ipomoea nil (Japanese morning glory)* | ini |
| Cucurbitaceae | *Cucumis melo (muskmelon)* | cmo |
| *Cucumis sativus (cucumber)* | csv |
| *Cucurbita maxima (winter squash)* | cmax |
| *Cucurbita moschata (crookneck pumpkin)* | cmos |
| *Cucurbita pepo subsp. pepo (vegetable marrow)* | cpep |
| *Momordica charantia (bitter melon)* | mcha |
| Euphorbiaceae | *Hevea brasiliensis (rubber tree)* | hbr |
| *Jatropha curcas (Barbados nut)* | jcu |
| *Manihot esculenta (cassava)* | mesc |
| *Ricinus communis (castor bean)* | rcu |
| Fabaceae | *Arachis duranensis (peanut diploid ancestor)* | adu |
| *Arachis ipaensis (peanut diploid ancestor)* | aip |
| *Cajanus cajan (pigeon pea)* | ccaj |
| *Cicer arietinum (chickpea)* | cam |
| *Glycine max (soybean)* | gmx |
| *Glycine soja (wild soybean)* | gsj |
| *Lotus japonicus (birdsfoot trefoil)* | lja |
| *Lupinus angustifolius (narrow-leaved blue lupine)* | lang |
| *Medicago truncatula (barrel medic)* | mtr |
| *Phaseolus vulgaris (common bean)* | pvu |
| *Vigna angularis (adzuki bean)* | var |
| *Vigna radiata (mung bean)* | vra |
| *Vigna unguiculata (cowpea)* | vun |
| *Quercus suber (cork oak)* | qsu |
| Juglandaceae | *Juglans regia (English walnut)* | jre |
| Malvaceae | *Gossypium arboreum (tree cotton)* | gab |
| *Gossypium hirsutum (upland cotton)* | ghi |
| *Gossypium raimondii (Peruvian cotton)* | gra |
| *Theobroma cacao (cacao)* | tcc |
| Myrtaceae | *Eucalyptus grandis (rose gum)* | egr |
| Oleaceae | *Olea europaea var. sylvestris (wild olive)* | oeu |
| Papaveraceae | *Papaver somniferum (opium poppy)* | psom |
| Pedaliaceae | *Sesamum indicum (sesame)* | sind |
| Rhamnaceae | *Ziziphus jujuba (Chinese jujube)* | zju |
| Rosaceae | *Fragaria vesca (woodland strawberry)* | fve |
| *Malus domestica (apple)* | mdm |
| *Prunus avium (sweet cherry)* | pavi |
| *Prunus mume (Japanese apricot)* | pmum |
| *Prunus persica (peach)* | pper |
| *Pyrus x bretschneideri (Chinese white pear)* | pxb |
| *Rosa chinensis (China rose)* | rcn |
| Rutaceae | *Citrus clementina (mandarin orange)* | cic |
| Rutaceae | *Citrus sinensis (Valencia orange)* | cit |
| Salicaceae | *Populus euphratica (Euphrates poplar)* | peu |
| *Populus trichocarpa (black cottonwood)* | pop |
| Solanaceae | *Capsicum annuum (cayenne pepper)* | cann |
| *Nicotiana attenuata (coyote tobacco)* | nau |
| *Nicotiana sylvestris (woodland tobacco)* | nsy |
| *Nicotiana tabacum (common tobacco)* | nta |
| *Nicotiana tomentosiformis (wild tobacco)* | nto |
| *Solanum lycopersicum (tomato)* | sly |
| *Solanum pennellii (wild tomato)* | spen |
| *Solanum tuberosum (potato)* | sot |
| Vitaceae | *Vitis vinifera (wine grape)* | vvi |
| Brassicaceae | *Arabidopsis lyrata (lyrate rockcress)* | aly |
| *Arabidopsis thaliana (thale cress)* | ath |
| *Brassica napus (rapeseed)* | bna |
| *Brassica oleracea (wild cabbage)* | boa |
| *Brassica rapa (field mustard)* | brp |
| *Camelina sativa (false flax)* | csat |
| *Capsella rubella (pink shepherd's-purse)* | crb |
| *Eutrema salsugineum (saltwater cress)* | eus |
| *Raphanus sativus (radish)* | rsz |
| Caricaceae | *Carica papaya (papaya)* | cpap |
| Cleomaceae | *Tarenaya hassleriana (spider flower)* | thj |
| Nelumbonaceae | *Nelumbo nucifera (sacred lotus)* | nnu |

It should be noted that for the sake of simplicity, there will be shortened names that will be used to refer to certain plants within relevant figures and tables since the data set does not contain more than one variety, cultivar, etc. for any given species. The shortened names are listed in Table 2. It also should be noted that there were two entries for *Oryza sativa subsp. japonica* due to the differences in the sources of the sequenced genomes, and for simplicity, the results from the predictions are condensed into one entry.

###### Table 2

Shortened plant names.

| **Scientific Name** | **Shortened Name** |
| --- | --- |
| Cucurbita pepo subsp. pepo | Cucurbita pepo |
| Cynara cardunculus var. scolymus | Cynara cardunculus |
| Olea europaea var. sylvestris | Olea europaea |
| Oryza sativa subsp. japonica | Oryza sativa |
| Pyrus x bretschneideri | Pyrus x |
| Volvox carteri f. nagariensis | Volvox carteri |

### Literature and Database Search

The literature and database search was a process of obtaining as much data as possible relating to the plants in the dataset and the synthesis of the flavonoids in the data set. Specifically, this search was focused primarily on finding plant-compound pairs, i.e. determining which of the plants in the data set are already known to synthesize the flavonoids in the data set. The motivation of conducting this search is based on the purely theoretical nature of the code. Any of the experimental data relating to the phytochemical contents of select plants that has been collected for this project is useful to compare against the programmatic predictions in order to provide a gauge of the accuracy of the program.

# 

# Conclusions and Recommendations

## Conclusions

Overall, there are undoubtedly further database and literature searches that will be necessary, as the data becomes available, in order to get a better grasp on the competency of the program as there are enough disagreements between the experimental and theoretical data to create doubt, yet there are also an abundance of agreements that do not completely dismiss the program as dysfunctional. Many of these disagreements come from either one single data source or are the results of studies that focus on culinary and other heavily cultivated (and thus potentially genetically modified) plants. This should not dismiss the validity of the data in its entirety however - it just means that these disagreements could be more likely to be resolved through further investigation of literature. Despite the number of experimental disagreements, there were enough experimental confirmations of the predictions which indicates that the program is at the very least directionally correct In terms of designing experiments in order to confirm or deny predictions made by the code, naringenin and kaempferol would serve as ideal candidates as there was a relatively large set of species that were both predicted and experimentally confirmed to synthesize the compounds. Experimenting with cyanidin, eriodictyol and quercetin would most likely be of some benefit for gauging the accuracy of the code as the experimental data agreed to a point where, at the very least, additional experimentation could address uncertainties in accuracy.

## Recommendations for Future Work

Since the data set of species has been solidified since October of 2020 and reflected the available genome sequences available through KEGG, the data set of plants used in this project is now limited as the number of species with accessible sequences has increased. The database now provides information for an additional 21 species (Table 3) of plants, which provides an opportunity for updating the data files for the code, and for conducting additional literature and database searches.

###### Table 3

The new species available from KEGG.

|  |  | Family | *Species* | KEGG Code |
| --- | --- | --- | --- | --- |
| **Flowering Plants** | **Basal Angiosperms** | Nymphaeaceae | *Nymphaea colorata* | [ncol](https://www.genome.jp/kegg-bin/show_organism?org=ncol) |
| **Monocots** | Poaceae | *Panicum hallii (Hall's panicgrass)* | phai |
| *Panicum virgatum (switchgrass)* | [pvir](https://www.genome.jp/kegg-bin/show_organism?org=pvir) |
| *Triticum aestivum (bread wheat)* | [taes](https://www.genome.jp/kegg-bin/show_organism?org=taes) |
| *Triticum dicoccoides (wild emmer wheat)* | [tdc](https://www.genome.jp/kegg-bin/show_organism?org=tdc) |
| **Eudicots** | Anacardiaceae | *Mangifera indica (mango)* | [minc](https://www.genome.jp/kegg-bin/show_organism?org=minc) |
| Anacardiaceae | *Pistacia vera (pistachio)* | [pvy](https://www.genome.jp/kegg-bin/show_organism?org=pvy) |
| Asteraceae | *Erigeron canadensis (horseweed)* | [ecad](https://www.genome.jp/kegg-bin/show_organism?org=ecad) |
| Celastraceae | *Tripterygium wilfordii (thunder duke vine)* | [twl](https://www.genome.jp/kegg-bin/show_organism?org=twl) |
| Convolvulaceae | *Ipomoea triloba (trilobed morning glory)* | [itr](https://www.genome.jp/kegg-bin/show_organism?org=itr) |
| Cucurbitaceae | *Benincasa hispida (wax gourd)* | [bhj](https://www.genome.jp/kegg-bin/show_organism?org=bhj) |
| Fabaceae | *Abrus precatorius (Indian licorice)* | [aprc](https://www.genome.jp/kegg-bin/show_organism?org=aprc) |
| Fabaceae | *Arachis hypogaea (peanut)* | [ahf](https://www.genome.jp/kegg-bin/show_organism?org=ahf) |
| Fagaceae | *Quercus lobata (valley oak)* | [qlo](https://www.genome.jp/kegg-bin/show_organism?org=qlo) |
| Lamiaceae | *Salvia splendens (scarlet sage)* | [sspl](https://www.genome.jp/kegg-bin/show_organism?org=sspl) |
| Moraceae | *Morus notabilis (mulberry species)* | [mnt](https://www.genome.jp/kegg-bin/show_organism?org=mnt) |
| Phrymaceae | *Erythranthe guttata (spotted monkey flower)* | [egt](https://www.genome.jp/kegg-bin/show_organism?org=egt) |
| Proteaceae | *Macadamia integrifolia (macadamia nut)* | [ming](https://www.genome.jp/kegg-bin/show_organism?org=ming) |
| Rosaceae | *Prunus dulcis (almond)* | [pdul](https://www.genome.jp/kegg-bin/show_organism?org=pdul) |
| Salicaceae | *Populus alba (white poplar)* | [palz](https://www.genome.jp/kegg-bin/show_organism?org=palz) |
| Vitaceae | *Vitis riparia (riverbank grape)* | [vri](https://www.genome.jp/kegg-bin/show_organism?org=vri) |

Additionally, it is recommended to extend the literature and database search to include the compounds from Table 4, in order to compare the new data to the predictions generated for these new compounds. The prediction code has been written for these compounds after the completion of the main project and is accessible through GitHub at *https://github.com/mbagg4152/pfpy* and can be found in the appendix. Since the basis of the program was already established it did not take long to update it to accommodate the new compounds. However, there was not enough time to conduct an updated literature and database search as the process is extremely time consuming and the new list of compounds is much larger than the working set from this project.

###### Table 4

Additional compounds, sorted by classification.

| **Classification** | **Compound** |
| --- | --- |
| Dihydrochalcones | Phloretin |
| Flavan 3-ols | Afzelechin |
| Epiafzelechin |
| Flavan 4-ols | Apiforol |
| Luteoforol |
| Flavanones | Butin |
| Dihydrotricetin |
| Hesperetin |
| Liquiritigenin |
| Pinocembrin |
| Flavones | Chrysin |
| Galangin |
| Tricetin |
| Flavononols | Ampelopsin (Dihydromyricetin) |
| Dihydrokaempferol |
| Fustin (2, 3-Dihydrofisetin) |
| Garbonzol |
| Pinobanksin |
| Taxifolin (Dihydroquercetin) |
| Isoflavanones | 2-Hydroxy-2, 3-dihydrogenistein |
| 2, 6, 7, 4'-Tetrahydroxyisoflavanone |
| 2, 7, 4'-Trihydroxyisoflavanone |
| 6, 7, 4'-Trihydroxyflavanone |
| Isoflavones | 6-Hydroxydaidzein |
| 7-4’-dihydroxyflavone |
| Daidzein |
| Leucoanthocyanidins (Flavan 3, 4-diols) | Leucocyanidin |
| Leucodelphinidin |
| Leucopelargonidin |

Another opportunity for future work lies in utilizing specific data that the code has already collected in order to analyze the genes and associated enzymes and flavonoids for the plants in the data set. This could be used to lead to a better understanding of the evolution of flavonoids, along with their distributions across different plant species.

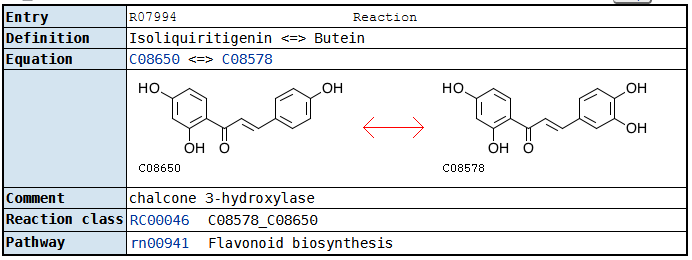
Additionally, there is an opportunity to conduct a literature and database search for compounds similar to the flavonoids in the data set. This could potentially include defining a strict definition for the similar compound and/or selecting specific similar compounds. Currently, the data that has been collected was collected while searching for data for the flavonoids in the data set and includes data from compounds that either derive from or have one of the flavonoids as a component. The data collected to date can be found at the beginning of the appendix in Tables 70-83.

The final recommendation is to test specific plants from the data sets for the purpose of comparing the predictions for isoliquiritigenin, butein, kaempferol, naringenin, cyanidin, eriodictyol and quercetin. It would be beneficial to test plants from Asteraceae and Fabaceaeas there is existing data which confirms butein synthesis in these families, and more importantly there are species from these families within the dataset used for this project. If the data set is expanded to include all available KEGG species, it may be beneficial to also work with *Pistacia vera*, as it belongs to Anacardiaceae, which also contains plants confirmed to synthesize butein (Nara Institute of Science and Technology). Additionally, it would be advantageous to test for isoliquiritigenin in these same species because (1) it is lacking experimental confirmations (albeit not to the same extent as butein), and (2) the KEGG flavonoid biosynthesis map appears to imply the reaction that transforms isoliquiritigenin to butein is bi-directional (Figure 7).

Testing kaempferol and naringenin would be greatly beneficial, and testing cyanidin, eriodictyol and quercetin could be of some benefit, given the predictions of the program and the comparisons made against experimentally known data, which will be covered later in the paper. If an experiment was conducted in order to focus on these five compounds, then it would be suggested that the initial subjects belong to *Fabaceae*. Each compound is predicted to be synthesized by most of the species of the dataset that belong to this family. Additionally, there is experimental data that shows that there are members of this family that have been known to synthesize each of these five compounds.

###### Figure 7

The KEGG entry for the reaction associated with isoliquiritigenin and butein.



From “KEGG REACTION: R07994” (Kanehisa Labs).

# 

# Methods

The bulk of the work for this project consisted of working on a program written in Python 3 and conducting literature and database searches. The most important goal of the program is to make the predictions regarding the flavonoid production in certain plant species.

The program was written using the standard distribution of Python, version 3.8.5. The program relies on additional library modules in order to function properly. Most of these modules are part of the Python standard library (The Python Software Foundation): datetime, json, os, re, sys, threading, and urllib. The final module needed for the program is Bioservices (version 1.7.11 was used for this project) and is the only external source which needs to be installed using the pip3 package manager. Bioservices is an extensive Python library and a crucial component of the program, meaning that it must be installed on a machine in order for the code to be able to execute and make its predictions. The program also requires JSON (JavaScript object notation) input files, which have been created specifically for this project. They serve as a better method of storing larger pieces of data compared to simply storing them as variables in the program, which takes out excess clutter from the code and makes data modification simpler. They look visually similar to the lists and dictionaries (key-value pairs) that are used in Python, which made the transition from hard-coded values (used in the previous project program versions) to JSON files extremely simple.

The literature and database searches were conducted as a way to gather as much information as possible in order to comment on the ranges of available data, but more importantly to compare experimentally known results with results predicted by the program. Much of the data referenced in the project comes from the databases, with additional information that has been collected from smaller sources and will be discussed further in the following subsections.

## Prediction Program Utilizing Data from KEGG

### Initialization

The initialization of the program is quite simple and focuses primarily on setting up the structure of directories for output and initializing key variables and constants. Many of the constants used are stored in JSON files in order to make the process of updating these values much simpler and to make the code more readable. Most of these values take the form of simple lists or dictionaries. The former is used for gathering data by using iteration while the latter is used for making the program output more human readable by using the keys to acquire the value that is to be printed. Some of the data used includes lists of the pathway and species codes, along with dictionaries that use these codes as the keys with the values being the actual name associated with the code. For example, *Arabidopsis thaliana* has a KEGG code of *“ath”* so the dictionary entry for this organism would be *“ath”: “Arabidopsis thaliana”*.

The most important values used within the program are contained within the list of organism-pathway codes, which are used to gather the core information that will be used throughout the program. As with any database, KEGG too has its own unique identifiers for all of the information that is stored and made publicly available. The codes for the organisms are a set of three to four alphabetic characters while the pathway codes are numeric strings. Conveniently, if an organism has genes that are involved in any given biological pathway, then the name and identifiers of these genes can be found by concatenating the organism and pathway codes.

The organism and pathway codes are held separately for the purpose of saving time and avoiding human error, as manually checking the names for each organism-pathway code would be extremely time consuming and tedious. The organism-pathway codes are generated iteratively by creating five new organism-pathway codes for each existing organism code, one for each of the biological pathways relating to this project as shown in Figures 8 and 9.

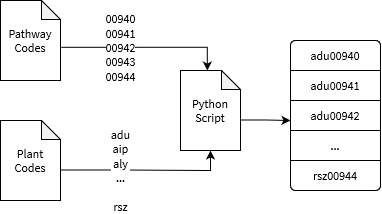
###### Figure 8

The code for generating organism-pathway pairs.



###### Figure 9

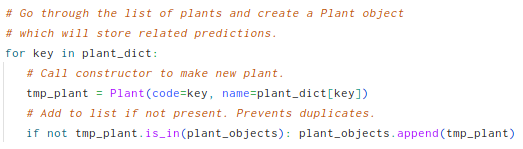
Visualization of organism-pathway code creation.



Once this is complete, a list of plant objects are created using the KEGG codes and their corresponding scientific names, which are python objects created specifically to store collections of information associated with each species as shown in Figure 10.

###### Figure 10

The process of creating the new plant objects.

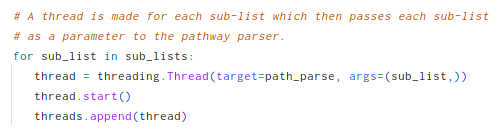


### Retrieving and Parsing Gene Data from KEGG

The list of organism-pathway codes created during initialization is first broken down into a collection of smaller lists. This is because the acquisition stage takes advantage of multithreading which greatly decreases the overall runtime of the program. Gathering the information for each organism-pathway code is a standalone process, which is what makes the usage of multithreading possible. Figure 11 highlights one of the implementations of multithreading in the program. Here, when each thread is created, a list will be passed as a parameter to the function that gathers the data.

###### Figure 11

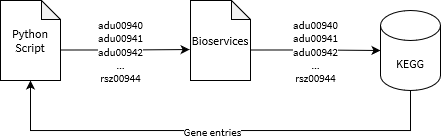
Portion of the code responsible for creating the threads.



The function path\_parse utilizes a very helpful library called bioservices which provides a multitude of functions for gathering and processing information from KEGG. When the function receives the list of organism-pathway codes, it iterates over the list, sending each code to KEGG. It then will return either an associated entry or an error message if the organism in question does not have any genes associated with the given pathway. Figure 12 visualizes this step of the program as the code consists of different sub-steps of tedious processing.

###### Figure 12

Visualization of data retrieval from KEGG.



Once a valid entry is gathered, it is then parsed to gather the gene’s KEGG ID, orthology, associated compound name, and most importantly the EC (enzyme commission) numbers. The EC numbers are added to a list belonging to the corresponding plant object that had been created earlier. This information alongside the organism’s code and name as well as the pathway code are written to output files for easier reference.

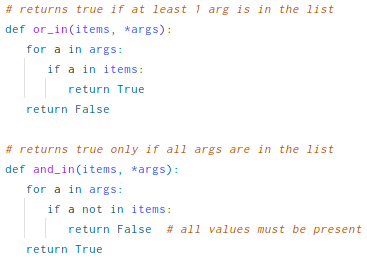
### Making Predictions

After the entries have been processed and each organisms' list of EC numbers has been appropriately updated, the program can move onto predicting the organism-compound pairs.This is done by first iterating through the list of plant objects and then for each individual plant object, the program then iterates through the list of flavonoid objects, which contains the chemical’s name and a shortened code. This code corresponds with the name of its respective function that is used for making its predictions and is used to trigger said function call.

The idea behind the functions used for making each compounds’ predictions are the same in nature: each flavonoid has a set of enzymes that are necessary for its production and are represented by EC numbers which means that the prediction can be handled by using a combination of logical ANDs and ORs. The logical operations were used because each step of synthesizing a compound may require one specific enzyme or may require one enzyme from a set of enzymes, then these steps must be chained together in order to get to a final product. These operations were set up by referencing KEGG’s pathway maps and are necessary to check if the enzymes required to synthesize a given flavonoid are present in a plant’s list, i.e., the plant has a specific set of genes that code for the required enzymes. For the sake of simplicity, two small utility functions have been created which are then used to run the logical operations when there are two or more elements that are to be checked as part of the overall logic flow, shown in Figure 13.

###### Figure 13

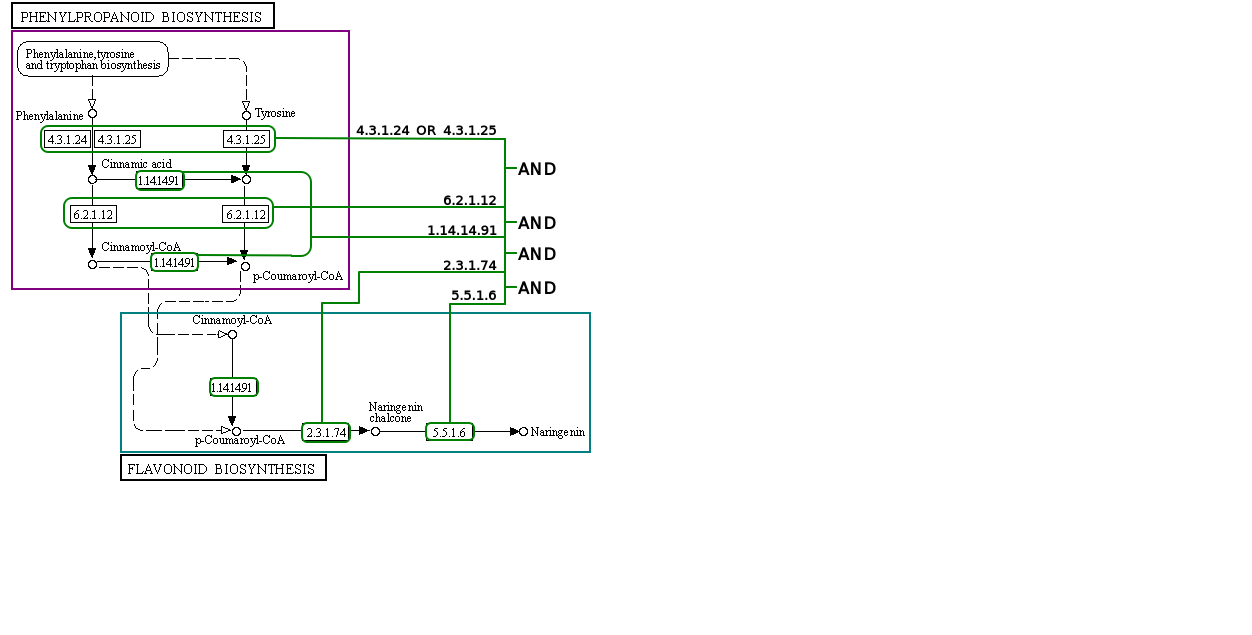
The logical functions that return true or false depending on the required enzymes and the input list.



In addition to the logical utility functions, some prediction functions can be simplified by using the output of another flavonoid's function due the structure of the biological pathways. For example, as shown in Figure 14, naringenin requires one more enzyme than its precursor naringenin chalcone which then in turn allows for a simpler logical function as seen in Figure 15. This means that the function that predicts naringenin only needs to call the function which predicts naringenin chalcone and then determine if a plant has genes that code for the additional enzyme, instead of checking the presence of numerous enzymes in one line of code.

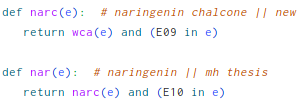
###### Figure 14

Logic for predicting naringenin, derived from KEGG pathway diagrams.



###### Figure 15

The prediction functions for naringenin chalcone and naringenin.



### Additional Code Functionality

Once the predictions have been made, the program will continue to complete two other tasks which have been deemed to be important in future research planned by Dr. Moorman. While this code does not affect the predictions in any way, it relies on different steps taken in the prediction process and is too significant of a part of the script to neglect entirely.

Right after the predictions have been completed, the program then moves on to look at the list of EC numbers for each plant that had been built by parsing the gene entries from KEGG. When building the list, duplicates were not removed so the program was able to count the occurrence of each EC number associated with each plant, which is helpful for determining the number of a plant’s genes that code for the same enzymes. Once this is done, the results are output into a tab delimited file where each row contains the plant name, the EC numbers and their counts.

After the EC number counts have been written to files, the program then shifts focus to grab the nucleotide sequences for each single gene that had been processed previously. The sequences are also hosted by KEGG but unlike gathering the gene entries, gathering the sequences had to be done differently due to bioservices not providing any functions relating to this specifically. Instead, the program uses DBGET URLs for accessing and downloading the sequences. DBGET is a database retrieval method implemented by KEGG which allows for easier data access since the URLs have a set format and are able to be built programmatically.

This portion of the code also utilizes multithreading as the number of genes from the selected list of organisms is roughly 24,000, which would take an extended amount of time to process if the program was written to only use a single thread. The function that handles the downloads and processing of these entries also saves each page as a local text file, which makes the execution time much faster if a user wishes to run the program multiple times. Each entry that is accessed is in the form of a simple HTML web page which is then parsed using regular expressions in order to extract the nucleotide sequence. Once a sequence has been gathered, the program then takes the key organism and gene identifiers and then uses these values to output the sequences in FASTA format. After every sequence is gathered and processed they are then sorted into collections based on associated EC numbers, which are then used to produce a set of FASTA files that, for each unique enzyme, contain the set of genes that code for them.

## Literature and Database Search

The literature and database search was centered around finding and collecting data associated with the plants from the working data set. The data used in this project primarily consists of data from various articles and journals and data gathered from large databases. Some sources referenced were collected by Jordan Wilson for his work on this project and his thesis *Computational Analysis of Flavonoid Pathways from Various Plant Species*, while others were found by utilizing different online resources such as Google Scholar and PubChem. The databases referenced for data collection are:

1. NPASS (Natural Product Activity and Species Source database) (Zeng et al., 2018), available at http://bidd.group/NPASS/search.php;
2. KNApSAcK (Afendi et al., 2012), available at http://www.knapsackfamily.com/knapsack\_core/top.php;
3. USDA Database for the Flavonoid Content of Selected Foods (Haytowitz et al., 2018), available at https://www.ars.usda.gov/ARSUserFiles/80400535/Data/Flav/Flav3.3.pdf;
4. Microbiome Analysis for Patients (Lassesen Consulting, LLC), available at https://www.microbiomeprescription.com/Library/Flavonoids;
5. IMPPAT (Indian Medicinal Plants Phytochemistry and Therapeutics) (Mohanraj et al., 2018), available at https://cb.imsc.res.in/imppat/home;
6. Dr. Duke’s Phytochemical and Ethnobotanical Databases at NAL (U.S. Department of Agriculture, Agricultural Research Service, 1992-2016), available at https://phytochem.nal.usda.gov/phytochem/search/list;
7. and NAPRALERT (NAtural PRoducts ALERT) (Loub et al., 1985), previously available at https://napralert.org/ but is unavailable as of January 2022.

Each database referenced in this project is available in different formats and accessible through different interfaces, meaning that data was collected both manually and programmatically. The data from NPASS and KNApSAcK were collected programmatically, while the remainder of the databases were accessed manually.

The data from NPASS was available to be exported as tab delimited text files, which then went through different levels of processing using regular expressions. This was done in order to create the appropriate JSON dictionaries consisting of the compound and organism IDs matched with their names. It should be noted that these values were collected on the basis of string matching, so there were additional codes collected as the names of some of the selected flavonoids were a substring in the names of additional flavonoids. In addition to creating the JSON data, a script was written for the purpose of processing the NPASS data. The script would use the collection of the specific NPASS codes that were collected previously in order to extract any pair from the NPASS data file that contained a selected organism and compound code. The selected pairs were then output and pasted into excel.

Accessing KNApSAcK was a simpler process given that the database is set up such that URLs can be generated in a formulaic fashion, which was extremely beneficial in creating the parsing script. For each plant in the data set, the program would create a KNApSAcK URL using its scientific name and then would download the corresponding HTML page. Once this was completed, the HTML files were then parsed one at a time using regular expressions, which would extract the compound names for each plant and would then complete a string search which would then determine if a compound and its corresponding plant could be saved as a matching pair. If the pair contains a flavonoid from the data set, then that specific pair is selected to be saved to an output file. This output file can then be pasted into excel and then filtered or searched to get a list of specific compounds. This could be done in the code, but it was preferable to finish this portion in excel so that the matches could be filtered by compound and pasted into their appropriate columns in the sheet used for storing the KNApSAcK data.

The remaining databases (USDA, Microbiome, IMPPAT, Dr. Duke’s and NAPRALERT) were presented in ways that did not make them ideal candidates for simple parsing scripts. All of the remaining databases, except for the USDA database, were searched by using the host website’s search functionality. The USDA database was accessed in PDF format so it was simply read through for finding matches. As each piece of data was collected, it had been added to a spreadsheet in order to keep track of the sources that stated whether a given plant is known to synthesize any of the compounds from the data set.

It should also be noted that the data from Microbiome is the result of theoretical predictions as well but should nonetheless still offer useful data for comparison and is grouped into the category of experimentally known data for the sake of simplicity. Almost none of the exact matches referenced can be attributed solely to Microbiome, as there is almost always one accompanying match from another database or literary source.

### Database Comparisons

Each database was created by different types of organizations (government agencies, academic groups, etc.) in order to serve different purposes, as described in the discussion. The difference in the purpose and nature of these databases means that the accessible data varies in their records on different organisms, compounds and their relationships. The data is also skewed towards specific organisms and compounds due to the differing levels of interest and perceived usefulness and as such, there were noticeable differences in the number and types of missing plants. This was found by keeping track of different database statistics in a spreadsheet. For each source, there is a list of compounds and organisms from the project lists that were missing. In addition to keeping track of the missing entries, the number of exact compound-organism matches were recorded to then determine how much each database was able to contribute. It should be noted that for each database, there was an attempt to search for each single plant and flavonoid from the dataset. For the programmatically searched databases, there was a manual search conducted to confirm the list of missing items as reported by the program.

### Supplemental Sources

Any other sources recorded and used for comparisons are simply a collection of smaller sources (i.e. academic papers, single studies, etc.).

### Comparisons Between the Literature and Program Output

This information relies entirely on data stored in and operations done on spreadsheets. In the main spreadsheet where these comparisons are made, there is a column which contains the name of each plant and additional columns which track if each plant is experimentally known or predicted to synthesize a compound.

For the experimentally known portion of data, the cell which corresponds with a specific plant and compound will be marked or blank, which is determined by attempting to find the plant name in specific named ranges which are named by compound and literature source. Then the appropriate letter code will be placed in the specific cell so long as a match is made. Keeping track of the predictions is far simpler. It is set up in a similar way to the experimentally known portion, but simply marks with an ‘x’ if there is a match for any given cell.

Using the aforementioned data, one can then use the values of those cells to then take count of the occurrences of a plant being experimentally known to synthesize a compound, a plant being predicted to be capable of synthesizing a compound, or a combination thereof. This can be used to put the outcome of the program into perspective.

# Results

## Predictions

The predictions listed in this section are the output of the project program which has been designed to, on a theoretical basis, predict if any of the 105 selected species could be capable of synthesizing any of the 15 selected flavonoids. A summary of the number of species predicted to synthesize each flavonoid can be seen in Table 5.

###### Table 5

Number of species predicted to synthesize each flavonoid.

| **Compound** | **Number of Species** |
| --- | --- |
| Genistein | 9 |
| Cyanidin | 76 |
| Butein | 90 |
| Isoliquiritigenin | 90 |
| Catechin | 58 |
| Epicatechin | 65 |
| Epigallocatechin | 43 |
| Gallocatechin | 42 |
| Eriodictyol | 89 |
| Naringenin | 90 |
| Kaempferol | 84 |
| Myricetin | 47 |
| Quercetin | 83 |
| Apigenin | 37 |
| Luteolin | 37 |

### Isoflavonoid

There are nine species that have been predicted to be capable of synthesizing genistein (Table 6), all of which belong to the legume family, Fabaceae.

###### Table 6

Plants predicted to synthesize genistein.

| *1. Cajanus cajan* | *4. Glycine soja* | *7. Vigna angularis* |
| --- | --- | --- |
| *2. Cicer arietinum* | *5. Lupinus angustifolius* | *8. Vigna radiata* |
| *3. Glycine max* | *6. Phaseolus vulgaris* | *9. Vigna unguiculata* |

### Anthocyanidin

There are 76 species which were predicted to synthesize cyanidin (Table 7), all of which are angiosperms of varying genera.

###### Table 7

Plants predicted to synthesize cyanidin.

| *1. Aegilops tauschii* | *27. Gossypium arboreum* | *52. Phoenix dactylifera* |
| --- | --- | --- |
| *2. Amborella trichopoda* | *28. Gossypium hirsutum* | *53. Populus euphratica* |
| *3. Arabidopsis lyrata* | *29. Gossypium raimondii* | *54. Populus trichocarpa* |
| *4. Arabidopsis thaliana* | *30. Helianthus annuus* | *55. Prunus avium* |
| *5. Arachis duranensis* | *31. Hevea brasiliensis* | *56. Prunus mume* |
| *6. Asparagus officinalis* | *32. Ipomoea nil* | *57. Prunus persica* |
| *7. Beta vulgaris* | *33. Jatropha curcas* | *58. Pyrus x* |
| *8. Brassica napus* | *34. Juglans regia* | *59. Quercus suber* |
| *9. Brassica oleracea* | *35. Lotus japonicus* | *60. Raphanus sativus* |
| *10. Brassica rapa* | *36. Lupinus angustifolius* | *61. Ricinus communis* |
| *11. Camelina sativa* | *37. Malus domestica* | *62. Rosa chinensis* |
| *12. Capsicum annuum* | *38. Manihot esculenta* | *63. Sesamum indicum* |
| *13. Carica papaya* | *39. Medicago truncatula* | *64. Setaria italica* |
| *14. Chenopodium quinoa* | *40. Musa acuminata* | *65. Solanum lycopersicum* |
| *15. Cicer arietinum* | *41. Nelumbo nucifera* | *66. Solanum pennellii* |
| *16. Citrus clementina* | *42. Nicotiana attenuata* | *67. Solanum tuberosum* |
| *17. Citrus sinensis* | *43. Nicotiana sylvestris* | *68. Spinacia oleracea* |
| *18. Cynara cardunculus* | *44. Nicotiana tabacum* | *69. Tarenaya hassleriana* |
| *19. Daucus carota* | *45. Nicotiana tomentosiformis* | *70. Theobroma cacao* |
| *20. Dendrobium catenatum* | *46. Olea europaea* | *71. Vigna angularis* |
| *21. Durio zibethinus* | *47. Oryza brachyantha* | *72. Vigna radiata* |
| *22. Elaeis guineensis* | *48. Oryza sativa* | *73. Vigna unguiculata* |
| *23. Eucalyptus grandis* | *49. Papaver somniferum* | *74. Vitis vinifera* |
| *24. Eutrema salsugineum* | *50. Phalaenopsis equestris* | *75. Zea mays* |
| *25. Glycine max* | *51. Phaseolus vulgaris* | *76. Ziziphus jujuba* |
| *26. Glycine soja* |  |  |

### Chalcones

There are 90 species which were predicted to synthesize both butein and isoliquiritigenin (Table 8), the majority of which are angiosperms. The only exceptions were the Byrophite *Physcomitrium patens* and the Lycophyte *Selaginella moellendorffii,* which are far removed from flowering plants as they both reproduce using spores instead of seeds (Rensing et al., 2020; Wallheimer, 2011).

###### Table 8

Plants predicted to synthesize the chalcones of the data set.

| *1. Aegilops tauschii* | *31. Eucalyptus grandis* | *61. Phalaenopsis equestris* |
| --- | --- | --- |
| *2. Amborella trichopoda* | *32. Eutrema salsugineum* | *62. Phaseolus vulgaris* |
| *3. Arabidopsis lyrata* | *33. Fragaria vesca* | *63. Phoenix dactylifera* |
| *4. Arabidopsis thaliana* | *34. Glycine max* | *64. Physcomitrium patens\** |
| *5. Arachis duranensis* | *35. Glycine soja* | *65. Populus euphratica* |
| *6. Arachis ipaensis* | *36. Gossypium arboreum* | *66. Populus trichocarpa* |
| *7. Asparagus officinalis* | *37. Gossypium hirsutum* | *67. Prunus avium* |
| *8. Beta vulgaris* | *38. Gossypium raimondii* | *68. Prunus mume* |
| *9. Brachypodium distachyon* | *39. Helianthus annuus* | *69. Prunus persica* |
| *10. Brassica napus* | *40. Hevea brasiliensis* | *70. Pyrus x* |
| *11. Brassica oleracea* | *41. Ipomoea nil* | *71. Quercus suber* |
| *12. Brassica rapa* | *42. Jatropha curcas* | *72. Raphanus sativus* |
| *13. Cajanus cajan* | *43. Juglans regia* | *73. Ricinus communis* |
| *14. Camelina sativa* | *44. Lactuca sativa* | *74. Rosa chinensis* |
| *15. Capsicum annuum* | *45. Lotus japonicus* | *75. Selaginella moellendorffii\** |
| *16. Carica papaya* | *46. Lupinus angustifolius* | *76. Sesamum indicum* |
| *17. Chenopodium quinoa* | *47. Malus domestica* | *77. Setaria italica* |
| *18. Cicer arietinum* | *48. Manihot esculenta* | *78. Solanum lycopersicum* |
| *19. Citrus clementina* | *49. Medicago truncatula* | *79. Solanum pennellii* |
| *20. Citrus sinensis* | *50. Momordica charantia* | *80. Solanum tuberosum* |
| *21. Cucumis melo* | *51. Musa acuminata* | *81. Sorghum bicolor* |
| *22. Cucumis sativus* | *52. Nelumbo nucifera* | *82. Spinacia oleracea* |
| *23. Cucurbita maxima* | *53. Nicotiana attenuata* | *83. Tarenaya hassleriana* |
| *24. Cucurbita moschata* | *54. Nicotiana sylvestris* | *84. Theobroma cacao* |
| *25. Cucurbita pepo* | *55. Nicotiana tabacum* | *85. Vigna angularis* |
| *26. Cynara cardunculus* | *56. Nicotiana tomentosiformis* | *86. Vigna radiata* |
| *27. Daucus carota* | *57. Olea europaea* | *87. Vigna unguiculata* |
| *28. Dendrobium catenatum* | *58. Oryza brachyantha* | *88. Vitis vinifera* |
| *29. Durio zibethinus* | *59. Oryza sativa* | *89. Zea mays* |
| *30. Elaeis guineensis* | *60. Papaver somniferum* | *90. Ziziphus jujuba* |

\* Bryophyte (non-flowering).

### Flavan-3-ols

The entirety of the species predicted to synthesize the flavan-3-ols were a diverse set of angiosperms. There are 58 predicted species for catechin (Table 9), 65 for epicatechin (Table 10), 43 for epigallocatechin (Table 11), and 42 for gallocatechin (Table 12). There were 40 species that were predicted to synthesize all four of these compounds, which can be seen in Table 13. The differences in the prediction sets can be attributed to the presence or absence of specific enzymes: flavonoid 3',5'-hydroxylase (EC:1.14.14.81) is needed to synthesize gallocatechin and epigallocatechin; leucoanthocyanidin reductase (EC:1.17.1.3) is needed to synthesize catechin and gallocatechin; and anthocyanidin reductase (EC:1.3.1.77) is needed to synthesize epicatechin and epigallocatechin.

###### Table 9

Plants predicted to synthesize catechin.

| *1. Aegilops tauschii* | *21. Hevea brasiliensis* | *40. Phoenix dactylifera* |
| --- | --- | --- |
| *2. Amborella trichopoda* | *22. Jatropha curcas* | *41. Populus euphratica* |
| *3. Arachis duranensis* | *23. Juglans regia* | *42. Populus trichocarpa* |
| *4. Asparagus officinalis* | *24. Lotus japonicus* | *43. Prunus avium* |
| *5. Beta vulgaris* | *25. Lupinus angustifolius* | *44. Prunus mume* |
| *6. Brachypodium distachyon* | *26. Malus domestica* | *45. Prunus persica* |
| *7. Carica papaya* | *27. Manihot esculenta* | *46. Pyrus x* |
| *8. Chenopodium quinoa* | *28. Medicago truncatula* | *47. Quercus suber* |
| *9. Cicer arietinum* | *29. Musa acuminata* | *48. Ricinus communis* |
| *10. Citrus clementina* | *30. Nelumbo nucifera* | *49. Rosa chinensis* |
| *11. Citrus sinensis* | *31. Nicotiana attenuata* | *50. Solanum pennellii* |
| *12. Durio zibethinus* | *32. Nicotiana sylvestris* | *51. Solanum tuberosum* |
| *13. Elaeis guineensis* | *33. Nicotiana tabacum* | *52. Spinacia oleracea* |
| *14. Eucalyptus grandis* | *34. Nicotiana tomentosiformis* | *53. Theobroma cacao* |
| *15. Glycine max* | *35. Olea europaea* | *54. Vigna angularis* |
| *16. Glycine soja* | *36. Oryza brachyantha* | *55. Vigna radiata* |
| *17. Gossypium arboreum* | *37. Oryza sativa* | *56. Vigna unguiculata* |
| *18. Gossypium hirsutum* | *38. Papaver somniferum* | *57. Vitis vinifera* |
| *19. Gossypium raimondii* | *39. Phaseolus vulgaris* | *58. Ziziphus jujuba* |
| *20. Helianthus annuus* |  |  |

##### 

###### Table 10

Plants predicted to synthesize epicatechin.

| *1. Aegilops tauschii* | *23. Gossypium arboreum* | *45. Populus euphratica* |
| --- | --- | --- |
| *2. Amborella trichopoda* | *24. Gossypium hirsutum* | *46. Populus trichocarpa* |
| *3. Arabidopsis lyrata* | *25. Gossypium raimondii* | *47. Prunus avium* |
| *4. Arabidopsis thaliana* | *26. Hevea brasiliensis* | *48. Prunus mume* |
| *5. Arachis duranensis* | *27. Jatropha curcas* | *49. Prunus persica* |
| *6. Asparagus officinalis* | *28. Juglans regia* | *50. Pyrus x* |
| *7. Beta vulgaris* | *29. Lotus japonicus* | *51. Quercus suber* |
| *8. Brassica napus* | *30. Lupinus angustifolius* | *52. Raphanus sativus* |
| *9. Brassica oleracea* | *31. Malus domestica* | *53. Ricinus communis* |
| *10. Brassica rapa* | *32. Manihot esculenta* | *54. Rosa chinensis* |
| *11. Camelina sativa* | *33. Medicago truncatula* | *55. Setaria italica* |
| *12. Carica papaya* | *34. Musa acuminata* | *56. Solanum lycopersicum* |
| *13. Chenopodium quinoa* | *35. Nelumbo nucifera* | *57. Solanum pennellii* |
| *14. Cicer arietinum* | *36. Nicotiana attenuata* | *58. Solanum tuberosum* |
| *15. Citrus clementina* | *37. Nicotiana sylvestris* | *59. Tarenaya hassleriana* |
| *16. Citrus sinensis* | *38. Nicotiana tabacum* | *60. Theobroma cacao* |
| *17. Durio zibethinus* | *39. Nicotiana tomentosiformis* | *61. Vigna angularis* |
| *18. Elaeis guineensis* | *40. Oryza brachyantha* | *62. Vigna radiata* |
| *19. Eucalyptus grandis* | *41. Oryza sativa* | *63. Vigna unguiculata* |
| *20. Eutrema salsugineum* | *42. Papaver somniferum* | *64. Zea mays* |
| *21. Glycine max* | *43. Phaseolus vulgaris* | *65. Ziziphus jujuba* |
| *22. Glycine soja* | *44. Phoenix dactylifera* |  |

###### Table 11

Plants predicted to synthesize epigallocatechin.

| *1. Aegilops tauschii* | *16. Juglans regia* | *30. Phoenix dactylifera* |
| --- | --- | --- |
| *2. Carica papaya* | *17. Lotus japonicus* | *31. Populus euphratica* |
| *3. Cicer arietinum* | *18. Lupinus angustifolius* | *32. Populus trichocarpa* |
| *4. Citrus clementina* | *19. Manihot esculenta* | *33. Quercus suber* |
| *5. Citrus sinensis* | *20. Medicago truncatula* | *34. Setaria italica* |
| *6. Durio zibethinus* | *21. Musa acuminata* | *35. Solanum lycopersicum* |
| *7. Elaeis guineensis* | *22. Nelumbo nucifera* | *36. Solanum pennellii* |
| *8. Eucalyptus grandis* | *23. Nicotiana attenuata* | *37. Solanum tuberosum* |
| *9. Glycine max* | *24. Nicotiana sylvestris* | *38. Theobroma cacao* |
| *10. Glycine soja* | *25. Nicotiana tabacum* | *39. Vigna angularis* |
| *11. Gossypium arboreum* | *26. Nicotiana tomentosiformis* | *40. Vigna radiata* |
| *12. Gossypium hirsutum* | *27. Oryza brachyantha* | *41. Vigna unguiculata* |
| *13. Gossypium raimondii* | *28. Oryza sativa* | *42. Zea mays* |
| *14. Hevea brasiliensis* | *29. Phaseolus vulgaris* | *43. Ziziphus jujuba* |
| *15. Jatropha curcas* |  |  |

###### Table 12

Plants predicted to synthesize gallocatechin.

| *1. Aegilops tauschii* | *15. Hevea brasiliensis* | *29. Oryza sativa* |
| --- | --- | --- |
| *2. Brachypodium distachyon* | *16. Jatropha curcas* | *30. Phaseolus vulgaris* |
| *3. Carica papaya* | *17. Juglans regia* | *31. Phoenix dactylifera* |
| *4. Cicer arietinum* | *18. Lotus japonicus* | *32. Populus euphratica* |
| *5. Citrus clementina* | *19. Lupinus angustifolius* | *33. Populus trichocarpa* |
| *6. Citrus sinensis* | *20. Manihot esculenta* | *34. Quercus suber* |
| *7. Durio zibethinus* | *21. Medicago truncatula* | *35. Solanum pennellii* |
| *8. Elaeis guineensis* | *22. Musa acuminata* | *36. Solanum tuberosum* |
| *9. Eucalyptus grandis* | *23. Nelumbo nucifera* | *37. Theobroma cacao* |
| *10. Glycine max* | *24. Nicotiana attenuata* | *38. Vigna angularis* |
| *11. Glycine soja* | *25. Nicotiana sylvestris* | *39. Vigna radiata* |
| *12. Gossypium arboreum* | *26. Nicotiana tabacum* | *40. Vigna unguiculata* |
| *13. Gossypium hirsutum* | *27. Nicotiana tomentosiformis* | *41. Vitis vinifera* |
| *14. Gossypium raimondii* | *28. Oryza brachyantha* | *42. Ziziphus jujuba* |

###### Table 13

Species predicted to synthesize the flavan-3-ols from the data set.

| *1. Aegilops tauschii* | *15. Jatropha curcas* | *28. Oryza sativa japonica* |
| --- | --- | --- |
| *2. Carica papaya* | *16. Juglans regia* | *29. Phaseolus vulgaris* |
| *3. Cicer arietinum* | *17. Lotus japonicus* | *30. Phoenix dactylifera* |
| *4. Citrus clementina* | *18. Lupinus angustifolius* | *31. Populus euphratica* |
| *5. Citrus sinensis* | *19. Manihot esculenta* | *32. Populus trichocarpa* |
| *6. Durio zibethinus* | *20. Medicago truncatula* | *33. Quercus suber* |
| *7. Elaeis guineensis* | *21. Musa acuminata* | *34. Solanum pennellii* |
| *8. Eucalyptus grandis* | *22. Nelumbo nucifera* | *35. Solanum tuberosum* |
| *9. Glycine max* | *23. Nicotiana attenuata* | *36. Theobroma cacao* |
| *10. Glycine soja* | *24. Nicotiana sylvestris* | *37. Vigna angularis* |
| *11. Gossypium arboreum* | *25. Nicotiana tabacum* | *38. Vigna radiata* |
| *12. Gossypium hirsutum* | *26. Nicotiana tomentosiformis* | *39. Vigna unguiculata* |
| *13. Gossypium raimondii* | *27. Oryza brachyantha* | *40. Ziziphus jujuba* |
| *14. Hevea brasiliensis* |  |  |

### Flavanones

There are 89 predicted species for eriodictyol and 90 for naringenin (Table 14). The predicted species for both compounds were exactly the same, with the exception of *Momordica charantia*, which was predicted to synthesize naringenin but not eriodictyol. This difference is caused by *M. charantia* not being able to produce specific enzymes (Figure 16), as per the data supplied by KEGG. All of the associated species are classified as angiosperms, with the exceptions of *P. patens* and *S. moellendorffii.*

###### Table 14

Plants predicted to synthesize the flavanones in the data set.

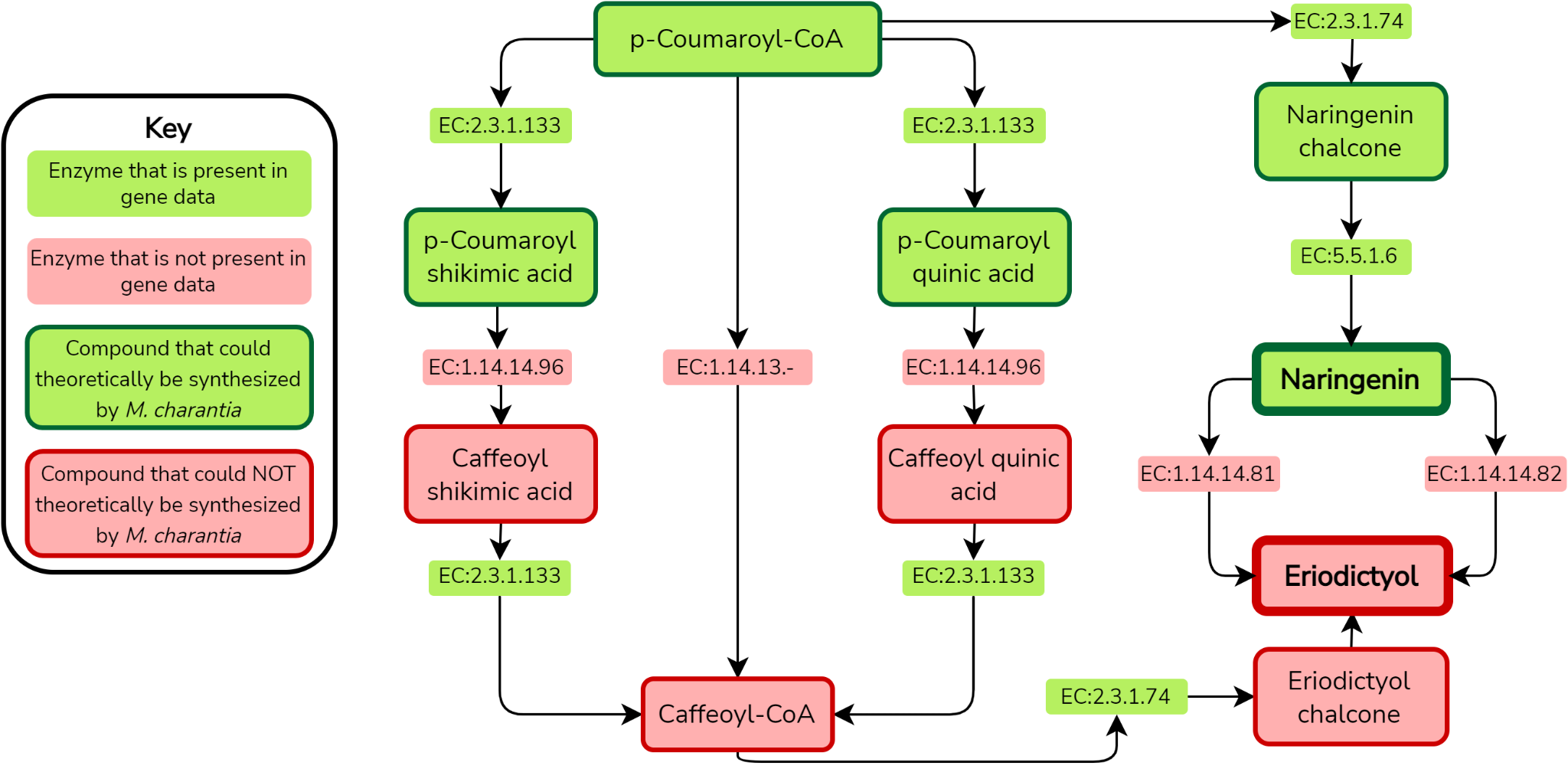
| *1. Aegilops tauschii* | *31. Eucalyptus grandis* | *61. Phalaenopsis equestris* |
| --- | --- | --- |
| *2. Amborella trichopoda* | *32. Eutrema salsugineum* | *62. Phaseolus vulgaris* |
| *3. Arabidopsis lyrata* | *33. Fragaria vesca* | *63. Phoenix dactylifera* |
| *4. Arabidopsis thaliana* | *34. Glycine max* | *64. Physcomitrium patens \** |
| *5. Arachis duranensis* | *35. Glycine soja* | *65. Populus euphratica* |
| *6. Arachis ipaensis* | *36. Gossypium arboreum* | *66. Populus trichocarpa* |
| *7. Asparagus officinalis* | *37. Gossypium hirsutum* | *67. Prunus avium* |
| *8. Beta vulgaris* | *38. Gossypium raimondii* | *68. Prunus mume* |
| *9. Brachypodium distachyon* | *39. Helianthus annuus* | *69. Prunus persica* |
| *10. Brassica napus* | *40. Hevea brasiliensis* | *70. Pyrus x* |
| *11. Brassica oleracea* | *41. Ipomoea nil* | *71. Quercus suber* |
| *12. Brassica rapa* | *42. Jatropha curcas* | *72. Raphanus sativus* |
| *13. Cajanus cajan* | *43. Juglans regia* | *73. Ricinus communis* |
| *14. Camelina sativa* | *44. Lactuca sativa* | *74. Rosa chinensis* |
| *15. Capsicum annuum* | *45. Lotus japonicus* | *75. Selaginella moellendorffii \** |
| *16. Carica papaya* | *46. Lupinus angustifolius* | *76. Sesamum indicum* |
| *17. Chenopodium quinoa* | *47. Malus domestica* | *77. Setaria italica* |
| *18. Cicer arietinum* | *48. Manihot esculenta* | *78. Solanum lycopersicum* |
| *19. Citrus clementina* | *49. Medicago truncatula* | *79. Solanum pennellii* |
| *20. Citrus sinensis* | *50. Momordica charantia \*\** | *80. Solanum tuberosum* |
| *21. Cucumis melo* | *51. Musa acuminata* | *81. Sorghum bicolor* |
| *22. Cucumis sativus* | *52. Nelumbo nucifera* | *82. Spinacia oleracea* |
| *23. Cucurbita maxima* | *53. Nicotiana attenuata* | *83. Tarenaya hassleriana* |
| *24. Cucurbita moschata* | *54. Nicotiana sylvestris* | *84. Theobroma cacao* |
| *25. Cucurbita pepo* | *55. Nicotiana tabacum* | *85. Vigna angularis* |
| *26. Cynara cardunculus* | *56. Nicotiana tomentosiformis* | *86. Vigna radiata* |
| *27. Daucus carota* | *57. Olea europaea* | *87. Vigna unguiculata* |
| *28. Dendrobium catenatum* | *58. Oryza brachyantha* | *88. Vitis vinifera* |
| *29. Durio zibethinus* | *59. Oryza sativa* | *89. Zea mays* |
| *30. Elaeis guineensis* | *60. Papaver somniferum* | *90. Ziziphus jujuba* |

\* Bryophyte (non-flowering).

\*\* Predicted to synthesize naringenin only.

###### Figure 16

Key enzymes that are produced by M. charantia in the context of eriodictyol and naringenin synthesis.



### Flavonols

There are 84 predicted angiosperm species for kaempferol (Table 15), 47 for myricetin (Table 16), and 83 for quercetin (Table 17). All of the species predicted to synthesize myricetin were also predicted to synthesize kaempferol and quercetin. There were 37 species predicted to synthesize kaempferol and quercetin (but not myricetin) because each plant was missing genes that code for flavonoid 3’, 5’-hydroxylase (EC:1.14.14.81) which is needed to synthesize myricetin. The set of species predicted to synthesize kaempferol and quercetin were almost identical, with the exception of *M. charantia* which was predicted to synthesize kaempferol but not quercetin. This is because *M. charantia* was missing the genes that code for flavonoid 3',5'-hydroxylase or flavonoid 3'-monooxygenase (EC:1.14.14.82), which are needed to synthesize quercetin from kaempferol or dihydrokaempferol. Additionally, *M. charantia* does have genes that code for naringenin 3-dioxygenase (EC:1.14.11.9) and flavonol synthase (EC:1.14.20.6), which are needed to go from eriodictyol to dihydroquercetin to quercetin, but would not work as *M. charantia* was not predicted to synthesize eriodictyol.

###### Table 15

Plants predicted to synthesize kaempferol .

| *1. Aegilops tauschii* | *29. Eucalyptus grandis* | *57. Papaver somniferum* |
| --- | --- | --- |
| *2. Amborella trichopoda* | *30. Eutrema salsugineum* | *58. Phalaenopsis equestris* |
| *3. Arabidopsis lyrata* | *31. Fragaria vesca* | *59. Phaseolus vulgaris* |
| *4. Arabidopsis thaliana* | *32. Glycine max* | *60. Phoenix dactylifera* |
| *5. Arachis duranensis* | *33. Glycine soja* | *61. Populus euphratica* |
| *6. Arachis ipaensis* | *34. Gossypium arboreum* | *62. Populus trichocarpa* |
| *7. Asparagus officinalis* | *35. Gossypium hirsutum* | *63. Prunus avium* |
| *8. Beta vulgaris* | *36. Gossypium raimondii* | *64. Prunus mume* |
| *9. Brachypodium distachyon* | *37. Helianthus annuus* | *65. Prunus persica* |
| *10. Brassica napus* | *38. Hevea brasiliensis* | *66. Pyrus x* |
| *11. Brassica oleracea* | *39. Ipomoea nil* | *67. Quercus suber* |
| *12. Brassica rapa* | *40. Jatropha curcas* | *68. Raphanus sativus* |
| *13. Camelina sativa* | *41. Juglans regia* | *69. Ricinus communis* |
| *14. Capsicum annuum* | *42. Lactuca sativa* | *70. Rosa chinensis* |
| *15. Carica papaya* | *43. Lotus japonicus* | *71. Sesamum indicum* |
| *16. Chenopodium quinoa* | *44. Lupinus angustifolius* | *72. Setaria italica* |
| *17. Cicer arietinum* | *45. Malus domestica* | *73. Solanum lycopersicum* |
| *18. Citrus clementina* | *46. Manihot esculenta* | *74. Solanum pennellii* |
| *19. Citrus sinensis* | *47. Medicago truncatula* | *75. Solanum tuberosum* |
| *20. Cucumis melo* | *48. Momordica charantia* | *76. Spinacia oleracea* |
| *21. Cucumis sativus* | *49. Musa acuminata* | *77. Tarenaya hassleriana* |
| *22. Cucurbita maxima* | *50. Nelumbo nucifera* | *78. Theobroma cacao* |
| *23. Cucurbita moschata* | *51. Nicotiana attenuata* | *79. Vigna angularis* |
| *24. Cucurbita pepo* | *52. Nicotiana sylvestris* | *80. Vigna radiata* |
| *25. Cynara cardunculus* | *53. Nicotiana tabacum* | *81. Vigna unguiculata* |
| *26. Daucus carota* | *54. Nicotiana tomentosiformis* | *82. Vitis vinifera* |
| *27. Durio zibethinus* | *55. Olea europaea* | *83. Zea mays* |
| *28. Elaeis guineensis* | *56. Oryza sativa* | *84. Ziziphus jujuba* |

###### Table 16

Plants predicted to synthesize myricetin.

| *1. Aegilops tauschii* | *17. Jatropha curcas* | *33. Phoenix dactylifera* |
| --- | --- | --- |
| *2. Brachypodium distachyon* | *18. Juglans regia* | *34. Populus euphratica* |
| *3. Capsicum annuum* | *19. Lotus japonicus* | *35. Populus trichocarpa* |
| *4. Carica papaya* | *20. Lupinus angustifolius* | *36. Quercus suber* |
| *5. Cicer arietinum* | *21. Manihot esculenta* | *37. Setaria italica* |
| *6. Citrus clementina* | *22. Medicago truncatula* | *38. Solanum lycopersicum* |
| *7. Citrus sinensis* | *23. Musa acuminata* | *39. Solanum pennellii* |
| *8. Durio zibethinus* | *24. Nelumbo nucifera* | *40. Solanum tuberosum* |
| *9. Elaeis guineensis* | *25. Nicotiana attenuata* | *41. Theobroma cacao* |
| *10. Eucalyptus grandis* | *26. Nicotiana sylvestris* | *42. Vigna angularis* |
| *11. Glycine max* | *27. Nicotiana tabacum* | *43. Vigna radiata* |
| *12. Glycine soja* | *28. Nicotiana tomentosiformis* | *44. Vigna unguiculata* |
| *13. Gossypium arboreum* | *29. Oryza sativa* | *45. Vitis vinifera* |
| *14. Gossypium hirsutum* | *30. Oryza brachyantha* | *46. Zea mays* |
| *15. Gossypium raimondii* | *31. Phalaenopsis equestris* | *47. Ziziphus jujuba* |
| *16. Hevea brasiliensis* | *32. Phaseolus vulgaris* |  |

###### Table 17

Plants predicted to synthesize quercetin.

| *1. Aegilops tauschii* | *29. Eucalyptus grandis* | *57. Phalaenopsis equestris* |
| --- | --- | --- |
| *2. Amborella trichopoda* | *30. Eutrema salsugineum* | *58. Phaseolus vulgaris* |
| *3. Arabidopsis lyrata* | *31. Fragaria vesca* | *59. Phoenix dactylifera* |
| *4. Arabidopsis thaliana* | *32. Glycine max* | *60. Populus euphratica* |
| *5. Arachis duranensis* | *33. Glycine soja* | *61. Populus trichocarpa* |
| *6. Arachis ipaensis* | *34. Gossypium arboreum* | *62. Prunus avium* |
| *7. Asparagus officinalis* | *35. Gossypium hirsutum* | *63. Prunus mume* |
| *8. Beta vulgaris* | *36. Gossypium raimondii* | *64. Prunus persica* |
| *9. Brachypodium distachyon* | *37. Helianthus annuus* | *65. Pyrus x* |
| *10. Brassica napus* | *38. Hevea brasiliensis* | *66. Quercus suber* |
| *11. Brassica oleracea* | *39. Ipomoea nil* | *67. Raphanus sativus* |
| *12. Brassica rapa* | *40. Jatropha curcas* | *68. Ricinus communis* |
| *13. Camelina sativa* | *41. Juglans regia* | *69. Rosa chinensis* |
| *14. Capsicum annuum* | *42. Lactuca sativa* | *70. Sesamum indicum* |
| *15. Carica papaya* | *43. Lotus japonicus* | *71. Setaria italica* |
| *16. Chenopodium quinoa* | *44. Lupinus angustifolius* | *72. Solanum lycopersicum* |
| *17. Cicer arietinum* | *45. Malus domestica* | *73. Solanum pennellii* |
| *18. Citrus clementina* | *46. Manihot esculenta* | *74. Solanum tuberosum* |
| *19. Citrus sinensis* | *47. Medicago truncatula* | *75. Spinacia oleracea* |
| *20. Cucumis melo* | *48. Musa acuminata* | *76. Tarenaya hassleriana* |
| *21. Cucumis sativus* | *49. Nelumbo nucifera* | *77. Theobroma cacao* |
| *22. Cucurbita maxima* | *50. Nicotiana attenuata* | *78. Vigna angularis* |
| *23. Cucurbita moschata* | *51. Nicotiana sylvestris* | *79. Vigna radiata* |
| *24. Cucurbita pepo* | *52. Nicotiana tabacum* | *80. Vigna unguiculata* |
| *25. Cynara cardunculus* | *53. Nicotiana tomentosiformis* | *81. Vitis vinifera* |
| *26. Daucus carota* | *54. Olea europaea* | *82. Zea mays* |
| *27. Durio zibethinus* | *55. Oryza sativa* | *83. Ziziphus jujuba* |
| *28. Elaeis guineensis* | *56. Papaver somniferum* |  |

### Flavones

There are 37 species that were predicted to synthesize both apigenin and luteolin (Table 18), with the entirety of the list of associated compounds consisting of angiosperms.

###### Table 18

Plants predicted to synthesize the flavones from the data set.

| *1. Aegilops tauschii* | *14. Glycine soja* | *26. Oryza sativa* |
| --- | --- | --- |
| *2. Arachis duranensis* | *15. Gossypium arboreum* | *27. Phaseolus vulgaris* |
| *3. Arachis ipaensis* | *16. Gossypium hirsutum* | *28. Populus euphratica* |
| *4. Brachypodium distachyon* | *17. Gossypium raimondii* | *29. Populus trichocarpa* |
| *5. Cajanus cajan* | *18. Helianthus annuus* | *30. Ricinus communis* |
| *6. Cicer arietinum* | *19. Hevea brasiliensis* | *31. Sesamum indicum* |
| *7. Citrus clementina* | *20. Lactuca sativa* | *32. Setaria italica* |
| *8. Citrus sinensis* | *21. Lupinus angustifolius* | *33. Sorghum bicolor* |
| *9. Cynara cardunculus* | *22. Manihot esculenta* | *34. Theobroma cacao* |
| *10. Daucus carota* | *23. Medicago truncatula* | *35. Vigna angularis* |
| *11. Durio zibethinus* | *24. Olea europaea* | *36. Vigna unguiculata* |
| *12. Eucalyptus grandis* | *25. Oryza brachyantha* | *37. Zea mays* |
| *13. Glycine max* |  |  |

## Literature and Database search

The literature and database search was conducted in order to find documentation that would provide exact matches for organism-compound pairs from the data set in order to be used later for comparisons against the predictions. A summary of these results can be found in Table 19.

###### Table 19

Number of species known to synthesize each flavonoid.

| **Compound** | **Unique Exact Matches** |
| --- | --- |
| Genistein | 17 |
| Cyanidin | 27 |
| Butein | 1 |
| Isoliquiritigenin | 5 |
| Catechin | 48 |
| Epicatechin | 38 |
| Epigallocatechin | 10 |
| Gallocatechin | 10 |
| Eriodictyol | 11 |
| Naringenin | 44 |
| Kaempferol | 49 |
| Myricetin | 19 |
| Quercetin | 59 |
| Apigenin | 30 |
| Luteolin | 36 |

### Isoflavonoid

There are 17 different species that are experimentally known to synthesize Genistein (Table 20).

###### Table 20

Plants known to produce Genistein.

| *1. Amborella trichopoda 3* | *7. Glycine max D, K, N, P, T* | *13. Spinacia oleracea 2* |
| --- | --- | --- |
| *2. Arabidopsis thaliana K* | *8. Gossypium hirsutum 1* | *14. Vigna angularis D, K* |
| *3. Beta vulgaris N* | *9. Lupinus angustifolius N* | *15. Vigna radiata D, K, N, T* |
| *4. Brassica oleracea K* | *10. Medicago truncatula K* | *16. Vigna unguiculata D, T* |
| *5. Cajanus cajan D, K, N* | *11. Phaseolus vulgaris K, D, A* | *17. Zea mays K* |
| *6. Cicer arietinum D, K, N* | *12. Prunus avium T* |  |

*Sources:*(D) Dr. Duke; *(K) KNApSAcK;* (N) NAPRALERT;  *(P) NPASS; (*T*) IMPPAT; (1)* (1) Nix et al., 2017; (2) Singh, 2016; (3) Wu et al., 2019.

### Anthocyanidin

There are 27 different species experimentally known to synthesize cyanidin (Table 21).

###### **Table 2**1

Plants known to produce cyanidin.

| *1. Arabidopsis thaliana K* | *10. Juglans regia U* | *19. Prunus avium U, K, M* |
| --- | --- | --- |
| *2. Brassica oleracea K, T* | *11. Lactuca sativa T* | *20. Prunus persica U* |
| *3. Capsicum annuum K* | *12. Lotus japonicus P* | *21. Raphanus sativus K* |
| *4. Carica papaya T* | *13. Malus domestica U, M, D* | *22. Solanum tuberosum K* |
| *5. Daucus carota K* | *14. Medicago truncatula K* | *23. Theobroma cacao D* |
| *6. Fragaria vesca P* | *15. Olea europaea N* | *24. Vigna angularis K* |
| *7. Glycine max P* | *16. Papaver somniferum N* | *25. Vigna unguiculata T* |
| *8. Gossypium hirsutum 1* | *17. Phaseolus vulgaris U* | *26. Vitis vinifera K, M, T, D, N* |
| *9. Ipomoea nil N* | *18. Phoenix dactylifera U, M* | *27. Zea mays K, T, D, A* |

Sources: (D) Dr. Duke; (K) KNApSAcK; (M) Microbiome; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (U) USDA; (1) Nix et al., 2017.

### Chalcones

There is only one species (*Glycine soja*) known to synthesize butein and this information was collected from NPASS. There are five species known to synthesize isoliquiritigenin (Table 22).

###### Table 22

Plants known to produce isoliquiritigenin.

| *1. Arabidopsis thaliana K* | *3. Glycine max D, K, P* | *5. Medicago truncatula K* |
| --- | --- | --- |
| *2. Cicer arietinum K, P* | *4. Helianthus annuus D* |  |

Sources: (D) Dr. Duke; (K) KNApSAcK; (P) NPASS.

### Flavan-3-ols

There are 48 species known to synthesize catechin (Table 23), 38 species known to synthesize Epcatechin (Table 24), 10 species known to synthesize epigallocatechin (Table 25), and 10 species known to synthesize gallocatechin (Table 26).

###### Table 23

Plants known to produce catechin.

| *1. Aegilops tauschii 3* | *17. Elaeis guineensis 11* | *33. Phaseolus vulgaris N, U* |
| --- | --- | --- |
| *2. Amborella trichopoda 23* | *18. Eucalyptus grandis 17* | *34. Phoenix dactylifera N* |
| *3. Arabidopsis thaliana K* | *19. Fragaria vesca N* | *35. Populus trichocarpa 22* |
| *4. Asparagus officinalis K* | *20. Glycine max D* | *36. Prunus avium K, M, N, U* |
| *5. Beta vulgaris 6* | *21. Gossypium hirsutum N* | *37. Prunus mume M, P* |
| *6. Brassica rapa 12* | *22. Jatropha curcas 10* | *38. Prunus persica K, M, U* |
| *7. Camelina sativa 20* | *23. Juglans regia N* | *39. Pyrus x 26* |
| *8. Capsicum annuum 7* | *24. Malus domestica D, U* | *40. Sesamum indicum 25* |
| *9. Carica papaya 4* | *25. Manihot esculenta 2* | *41. Solanum tuberosum 1* |
| *10. Chenopodium quinoa 8* | *26. Medicago truncatula K* | *42. Sorghum bicolor N* |
| *11. Cicer arietinum 16* | *27. Momordica charantia K* | *43. Spinacia oleracea 19* |
| *12. Citrus sinensis 15* | *28. Musa acuminata K* | *44. Theobroma cacao K, N, U* |
| *13. Cucumis melo 21* | *29. Nelumbo nucifera N, P* | *45. Vigna angularis K, P* |
| *14. Cucumis sativus 24* | *30. Nicotiana tabacum 18* | *46. Vigna radiata 5* |
| *15. Cucurbita pepo 14* | *31. Olea europaea D* | *47. Vitis vinifera D, K, M, N, P, U* |
| *16. Durio zibethinus 9* | *32. Oryza sativa 12* | *48. Ziziphus jujuba K, M, N, P, U* |

Sources: (D) Dr. Duke; (K) KNApSAcK; (M) Microbiome; (N) NAPRALERT; (P) NPASS; (U) USDA; (1) Akyol et al., 2016; (2) Buschmann et al., 2000; (3) Dong-Dong et al., 2015; (4) Duru & Duru, 2019; (5) Gan et al., 2016; (6) Georgiev et al., 2010; (7) Ghasemnezhad et al., 2010; (8) Hemalatha et al., 2016; (9) Ho & Bhat, 2015; (10) Igbinosa et al., 2011; (11) Jaffri et al., 2011; (12) Kim et al., 2015; (13) Kim et al., 2017; (14) Kostecka-Gugała et al., 2020; (15) Liew et al., 2018; (16) Quintero-Soto et al., 2018; (17) Santos et al., 2013; (18) Shimoda et al., 2007; (19) Singh, 2016; (20) Terpinc et al., 2012; (21) Tsanova-Savova et al., 2015; (22) Wang et al., 2013; (23) Wu et al., 2019; (24) Xu et al., 2019; (25) Zeb et al., 2017; (26) Zhai et al., 2014.

###### Table 24

Plants known to produce epicatechin.

| *1. Amborella trichopoda 14* | *14. Juglans regia 15* | *27. Prunus persica U* |
| --- | --- | --- |
| *2. Arabidopsis thaliana K* | *15. Malus domestica D, M, U* | *28. Pyrus x 17* |
| *3. Beta vulgaris 6* | *16. Medicago truncatula K* | *29. Ricinus communis N* |
| *4. Brassica napus 2* | *17. Momordica charantia K* | *30. Sesamum indicum 16* |
| *5. Capsicum annuum 4* | *18. Musa acuminata K* | *31. Solanum tuberosum K, T* |
| *6. Carica papaya 11* | *19. Nelumbo nucifera T* | *32. Sorghum bicolor 8* |
| *7. Cucumis melo M, U* | *20. Nicotiana sylvestris P* | *33. Spinacia oleracea 9* |
| *8. Durio zibethinus 7* | *21. Nicotiana tabacum 3* | *34. Theobroma cacao D, K, M, N, P, T, U* |
| *9. Elaeis guineensis 1* | *22. Phaseolus vulgaris U* | *35. Vigna angularis P* |
| *10. Eucalyptus grandis 12* | *23. Phoenix dactylifera N* | *36. Vigna radiata 5* |
| *11. Fragaria vesca N* | *24. Populus trichocarpa 13* | *37. Vitis vinifera D, K, M, N, P, T, U* |
| *12. Glycine max N, T* | *25. Prunus avium K, M, U* | *38. Ziziphus jujuba N, P, U* |
| *13. Gossypium hirsutum 10* | *26. Prunus mume P* |  |

Sources: (D) Dr. Duke; (K) KNApSAcK; (M) Microbiome; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (U) USDA; (1) Ahmad et al., 2018; (2) Auger et al., 2010; (3) Chen et al., 2013; (4) di Sotto et al., 2018; (5) Ganesan & Xu, 2018 ; (6) Georgiev et al., 2010; (7) Liu et al., 2013; (8) Luo et al., 2018; (9) Singh, 2016; (10) Nix et al., 2017; (11) Oboh et al., 2013; (12) Santos et al., 2012; (13) Wang et al., 2013; (14) Wu et al., 2019; (15) Yan et al., 2019; (16) Zeb et al., 2017; (17) Zhai et al., 2014.

###### Table 25

Plants known to produce epigallocatechin.

| *1. Citrus sinensis 2* | *5. Phaseolus vulgaris 1* | *8. Theobroma cacao D, N* |
| --- | --- | --- |
| *2. Cucumis melo M, U* | *6. Prunus avium M, U* | *9. Vitis vinifera D, N* |
| *3. Gossypium hirsutum N* | *7. Prunus persica M, U* | *10. Ziziphus jujuba N* |
| *4. Malus domestica U* |  |  |

Sources: (D) Dr. Duke; (M) Microbiome; (N) NAPRALERT; (U) USDA; (1) Ganesan & Xu, 2017; (2) Liew et al., 2018.

###### Table 26

Plants known to produce gallocatechin.

| *1. Gossypium hirsutum 2* | *5. Phaseolus vulgaris 1* | *8. Vigna radiata K* |
| --- | --- | --- |
| *2. Juglans regia T* | *6. Theobroma cacao N* | *9. Vitis vinifera D, K, N, P, T* |
| *3. Medicago truncatula K* | *7. Vigna angularis K* | *10. Ziziphus jujuba N* |
| *4. Nelumbo nucifera N, T* |  |  |

Sources: (D) Dr. Duke; (K) KNApSAcK; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (1) Ganesan & Xu, 2017; (2) Nix et al., 2017.

### Flavanones

There are 11 species known to synthesize eriodictyol (Table 27) and 44 species known to synthesize naringenin (Table 28).

###### Table 27

Plants known to produce eriodictyol.

| *1. Amborella trichopoda 6* | *5. Malus domestica 7* | *9. Theobroma cacao 1* |
| --- | --- | --- |
| *2. Eucalyptus grandis 3* | *6. Prunus avium 4* | *10. Vigna radiata 5* |
| *3. Glycine soja P* | *7. Prunus persica N* | *11. Ziziphus jujuba N* |
| *4. Gossypium hirsutum 2* | *8. Sorghum bicolor N* |  |

Sources: (N) NAPRALERT; (P) NPASS; (1) Gallego et al., 2019; (2) Nix et al., 2017; (3) Santos et al., 2013; (4) Sanz et al., 2010; (5) Tang et al., 2014 (6) Wu et al., 2019; (7) Yıldırım et al., 2015.

###### Table 28

Plants known to produce naringenin.

| *1. Amborella trichopoda 18* | *16. Daucus carota K* | *31. Prunus avium D, K, N* |
| --- | --- | --- |
| *2. Arabidopsis thaliana K* | *17. Dendrobium catenatum P* | *32. Prunus mume K, N, P* |
| *3. Asparagus officinalis K* | *18. Eucalyptus grandis 12* | *33. Prunus persica K, N, T* |
| *4. Brachypodium distachyon 7* | *19. Fragaria vesca 3* | *34. Raphanus sativus K* |
| *5. Brassica napus 10* | *20. Glycine max D, K, N, T* | *35. Selaginella moellendorffii 20* |
| *6. Brassica oleracea K, T* | *21. Gossypium hirsutum 8* | *36. Solanum lycopersicum D* |
| *7. Brassica rapa 16* | *22. Juglans regia 19* | *37. Solanum tuberosum 17* |
| *8. Cajanus cajan 8* | *23. Medicago truncatula K* | *38. Sorghum bicolor N* |
| *9. Capsicum annuum K* | *24. Momordica charantia K* | *39. Spinacia oleracea 13* |
| *10. Carica papaya K* | *25. Musa acuminata K* | *40. Theobroma cacao K* |
| *11. Chenopodium quinoa 4* | *26. Nelumbo nucifera 15* | *41. Vigna angularis N* |
| *12. Cicer arietinum N* | *27. Olea europaea 14* | *42. Vigna radiata D* |
| *13. Citrus sinensis D, N, P, U* | *28. Oryza sativa 5* | *43. Vitis vinifera 11* |
| *14. Cucumis melo 6* | *29. Phaseolus vulgaris 2* | *44. Zea mays K, T* |
| *15. Cynara cardunculus D, U* | *30. Phoenix dactylifera 1* |  |

Sources: (D) Dr. Duke; (K) KNApSAcK; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (U) USDA; (1) El-Kholy et al., 2019; (2) Ganesan & Xu, 2017; (3) Gasperotti et al., 2015; (4) Hemalatha et al., 2016; (5) Kim et al., 2017; (6) Mallek-Ayadi et al., 2017; (7) Napoleão et al., 2017; (8) Nix et al., 2015; (9) Nix et al., 2017; (10) Pei et al., 2012; (11) Perestrelo et al., 2012; (12) Santos et al., 2017; (13) Singh, 2016; (14) Tamasi et al., 2019; (15) Tang et al., 2017; (16) Thiruvengadam & Chung, 2015; (17) Valiñas et al., 2017; (18) Wu et al., 2019; (19) Yan et al., 2019; (20) Yobi et al., 2012.

### Flavonols

There are 49 different species known to synthesize kaempferol (Table 29), 19 species known to synthesize myricetin (Table 30), and there are 59 species known to synthesize quercetin (Table 31).

###### Table 29

Plants known to produce kaempferol .

| *1. Arabidopsis thaliana K, N, T* | *18. Fragaria vesca 1* | *34. Papaver somniferum N, T* |
| --- | --- | --- |
| *2. Asparagus officinalis D, K, M, T, U* | *19. Glycine max D, K, M, N, T, U* | *35. Phaseolus vulgaris N, U* |
| *3. Beta vulgaris N, P* | *20. Gossypium hirsutum 3* | *36. Prunus avium D, M, N, U* |
| *4. Brassica napus N* | *21. Hevea brasiliensis N* | *37. Prunus persica K, M, N, T, U* |
| *5. Brassica oleracea D, K, M, N, T, U* | *22. Juglans regia D* | *38. Raphanus sativus K, M, N, U* |
| *6. Brassica rapa U* | *23. Lactuca sativa D, N* | *39. Ricinus communis D* |
| *7. Capsicum annuum K, M, U* | *24. Lotus japonicus P* | *40. Rosa chinensis N, P* |
| *8. Carica papaya K, M, U* | *25. Lupinus angustifolius N* | *41. Solanum lycopersicum D, K* |
| *9. Chenopodium quinoa  N* | *26. Malus domestica D, M, U* | *42. Solanum tuberosum D, M, N, U* |
| *10. Cicer arietinum N, T* | *27. Medicago truncatula K, N* | *43. Spinacia oleracea D, M, N, U* |
| *11. Citrus sinensis M, N, U* | *28. Momordica charantia K* | *44. Theobroma cacao D, N* |
| *12. Cucumis melo M, U* | *29. Musa acuminata K, N* | *45. Vigna radiata D, M, N, U* |
| *13. Cucumis sativus M, N, T, U* | *30. Nelumbo nucifera D, M, N, P, T, U* | *46. Vigna unguiculata T* |
| *14. Cucurbita maxima N* | *31. Nicotiana tabacum D, N* | *47. Vitis vinifera D, M, N, U* |
| *15. Cucurbita pepo D* | *32. Olea europaea N* | *48. Zea mays K, T* |
| *16. Daucus carota D, M, N, T, U* | *33. Oryza sativa 2* | *49. Ziziphus jujuba N* |
| *17. Elaeis guineensis 4* |  |  |

Sources: (D) Dr. Duke; (K) KNApSAcK; (M) Microbiome; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (U) USDA; (1) Gasperotti et al., 2015; (2) Kim et al., 2017; (3) Nix et al., 2017; (4) Zhou et al., 2019.

###### Table 30

Plants known to produce myricetin.

| *1. Asparagus officinalis K* | *8. Juglans regia 4* | *14. Prunus avium M, U* |
| --- | --- | --- |
| *2. Brassica rapa U* | *9. Lotus japonicus P* | *15. Solanum tuberosum D, N* |
| *3. Capsicum annuum D, N, U* | *10. Malus domestica 3* | *16. Spinacia oleracea M, U* |
| *4. Carica papaya M, U* | *11. Medicago truncatula K, N* | *17. Vigna radiata 1* |
| *5. Citrus sinensis D, M, N, U* | *12. Nelumbo nucifera M, P, T, U* | *18. Vigna unguiculata T* |
| *6. Daucus carota D, M, N, U* | *13. Phaseolus vulgaris U* | *19. Vitis vinifera D, M, N, P, U* |
| *7. Gossypium hirsutum 2* |  |  |

Sources: (D) Dr. Duke; (K) KNApSAcK; (M) Microbiome; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (U) USDA; (1) Ganesan & Xu, 2018; (2) Nix et al., 2017; (3) Petkovska et al., 2017; (4) Yan et al., 2019.

###### Table 31

Plants known to produce quercetin.

| *1. Amborella trichopoda 7* | *21. Fragaria vesca 2* | *41. Phaseolus vulgaris N, T, U* |
| --- | --- | --- |
| *2. Arabidopsis thaliana K, T* | *22. Glycine max D, M, N, U* | *42. Phoenix dactylifera D, N, U* |
| *3. Asparagus officinalis D, K, M, T, U* | *23. Gossypium arboreum N* | *43. Prunus avium D, K, M, N, U* |
| *4. Beta vulgaris D, M, N, P, U* | *24. Gossypium hirsutum 5* | *44. Prunus mume P* |
| *5. Brassica napus N* | *25. Helianthus annuus D, N* | *45. Prunus persica M, N, U* |
| *6. Brassica oleracea D, K, M, N, U* | *26. Hevea brasiliensis N* | *46. Quercus suber N, P* |
| *7. Brassica rapa U* | *27. Jatropha curcas 3* | *47. Raphanus sativus K, N* |
| *8. Cajanus cajan 4* | *28. Juglans regia D, N* | *48. Ricinus communis D, N* |
| *9. Camelina sativa 6* | *29. Lactuca sativa D, N, T* | *49. Rosa chinensis N, P* |
| *10. Capsicum annuum D, K, M, P, U* | *30. Lotus japonicus P* | *50. Solanum lycopersicum 1* |
| *11. Carica papaya K, N* | *31. Lupinus angustifolius N* | *51. Solanum tuberosum D, M, N, U* |
| *12. Chenopodium quinoa N* | *32. Malus domestica D, M, U* | *52. Spinacia oleracea D, M, N, U* |
| *13. Cicer arietinum N, T* | *33. Manihot esculenta N* | *53. Theobroma cacao D, N, U* |
| *14. Citrus sinensis D, M, N, P, U* | *34. Medicago truncatula K, N* | *54. Vigna angularis P* |
| *15. Cucumis melo M, U* | *35. Momordica charantia K* | *55. Vigna radiata D, M, N, U* |
| *16. Cucumis sativus M, N, U* | *36. Musa acuminata K, M, N* | *56. Vigna unguiculata M, N, U* |
| *17. Cucurbita maxima N* | *37. Nelumbo nucifera D, M, P, U* | *57. Vitis vinifera D, M, N, P, U* |
| *18. Cucurbita pepo D* | *38. Nicotiana tabacum D, N* | *58. Zea mays D, K* |
| *19. Cynara cardunculus D* | *39. Olea europaea D, N* | *59. Ziziphus jujuba M, P, U* |
| *20. Daucus carota D, M, N, T, U* | *40. Papaver somniferum N, T* |  |

Sources: (D) Dr. Duke; (K) KNApSAcK; (M) Microbiome; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (U) USDA; (1) Bhandari & Lee, 2016; (2) Gasperotti et al., 2015; (3) Igbinosa et al., 2011; (4) Nix et al., 2015; (5) Nix et al., 2017; (6) Terpinc et al., 2012; (7) Wu et al., 2019.

### Flavones

There are 30 species known to synthesize apigenin (Table 32), and 36 species known to synthesize luteolin (Table 33).

###### Table 32

Plants known to produce apigenin.

| *1. Asparagus officinalis K* | *11. Cucurbita pepo 1* | *21. Olea europaea D, N, T* |
| --- | --- | --- |
| *2. Beta vulgaris 4* | *12. Cynara cardunculus D, U* | *22. Phaseolus vulgaris D* |
| *3. Brassica oleracea K, M* | *13. Daucus carota D, K, N, T* | *23. Phoenix dactylifera D, N* |
| *4. Brassica rapa  U* | *14. Fragaria vesca P* | *24. Populus trichocarpa N* |
| *5. Cajanus cajan  N* | *15. Glycine max N, P* | *25. Prunus avium K* |
| *6. Capsicum annuum T* | *16. Jatropha curcas D, N* | *26. Raphanus sativus K* |
| *7. Carica papaya M, U* | *17. Juglans regia D, M* | *27. Ricinus communis T* |
| *8. Chenopodium quinoa 2* | *18. Lactuca sativa N* | *28. Sorghum bicolor N* |
| *9. Citrus sinensis 3* | *19. Lupinus angustifolius N* | *29. Theobroma cacao K* |
| *10. Cucurbita moschata P* | *20. Medicago truncatula K* | *30. Zea mays D, K, N* |

Sources: (D) Dr. Duke; (K) KNApSAcK; (M) Microbiome; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (U) USDA; (1) Akomolafe et al., 2016; (2) Hemalatha et al., 2016; (3) Liew et al., 2018; (4) Silva et al., 2020.

###### Table 33

Plants known to produce luteolin.

| *1. Amborella trichopoda 7* | *13. Cucurbita moschata U* | *25. Olea europaea D, M,, N, U* |
| --- | --- | --- |
| *2. Asparagus officinalis K* | *14. Cucurbita pepo 1* | *26. Phaseolus vulgaris  D, U* |
| *3. Beta vulgaris M, U* | *15. Cynara cardunculus D, U* | *27. Phoenix dactylifera 4* |
| *4. Brassica napus 5* | *16. Daucus carota D, K, M, N, T, U* | *28. Raphanus sativus K* |
| *5. Brassica oleracea K, M, T, U* | *17. Fragaria vesca P* | *29. Setaria italica N* |
| *6. Brassica rapa U* | *18. Helianthus annuus D, N* | *30. Solanum lycopersicum 2* |
| *7. Cajanus cajan N* | *19. Lactuca sativa D, N* | *31. Solanum tuberosum N* |
| *8. Capsicum annuum D, K, M, N, P, T, U* | *20. Lupinus angustifolius N* | *32. Sorghum bicolor 6* |
| *9. Carica papaya M, U* | *21. Malus domestica D, M, U* | *33. Spinacia oleracea M, U* |
| *10. Chenopodium quinoa 4* | *22. Medicago truncatula K* | *34. Theobroma cacao D, K* |
| *11. Citrus sinensis M, N, U* | *23. Momordica charantia T* | *35. Vitis vinifera D, M, N* |
| *12. Cucumis melo M, U* | *24. Nelumbo nucifera D, M, P, T, U* | *36. Zea mays K, N* |

Sources: (D) Dr. Duke; (K) KNApSAcK; (M) Microbiome; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (U) USDA; (1) Akomolafe et al., 2016; (2) Bhandari & Lee, 2016; (3) Hemalatha et al., 2016; (4) Khallouki et al., 2018; (5) Pei et al., 2012; (6) Svensson et al., 2010; (7) Wu et al., 2019.

## Database Comparisons

NAPRALERT contained the greatest number of exact plant-compound matches, and only the USDA, IMPPAT, and Microbiome were missing data for flavonoids from the data set, as shown in Table 34. After looking at the species which were present in each database, it was then possible to create Figures 17 and 18 by editing figures generated by UpSet (Lex et al., 2014). As can be seen in Figure 17, KNApSAck, NPASS, and NAPRALERT, in this order, had the most species from the dataset. Additionally, there are 10 that have no records in any of the databases, and 22 species that had records in all of the databases. This information can be seen in Figure 18, which shows the number of overlaps in species data amongst the databases. The darkened dots indicates species’ presence in a database: the first row indicates that there are 22 species in all of the databases, the second row shows there are 10 missing from all databases, and the third row indicates there are 10 species with associated records in KNApSAcK, NPASS and NAPRALERT only.

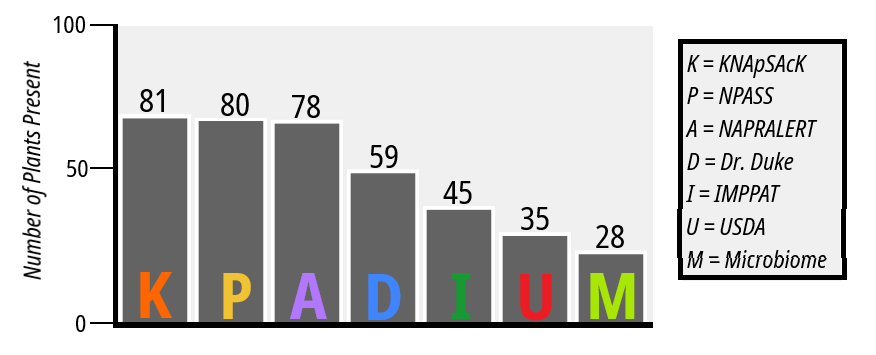
###### Table 34

Notable information about the databases.

| **Database** | **Accessed** | **Exact Matches** | **Missing Plants** | **Missing Compounds** |
| --- | --- | --- | --- | --- |
| NPASS | April 2021 | 45 | 25 | N/A |
| KNApSAcK | June 2021 | 107 | 24 | N/A |
| NAPRALERT | May 2021 | 147 | 27 | N/A |
| Dr. Duke | June 2021 | 98 | 46 | N/A |
| IMPPAT | June 2021 | 52 | 60 | Epigallocatechin  Isoliquiritigenin |
| USDA | June 2021 | 98 | 70 | Butein  Genistein  Isoliquiritigenin |
| Microbiome | June 2021 | 76 | 77 | Isoliquiritigenin |

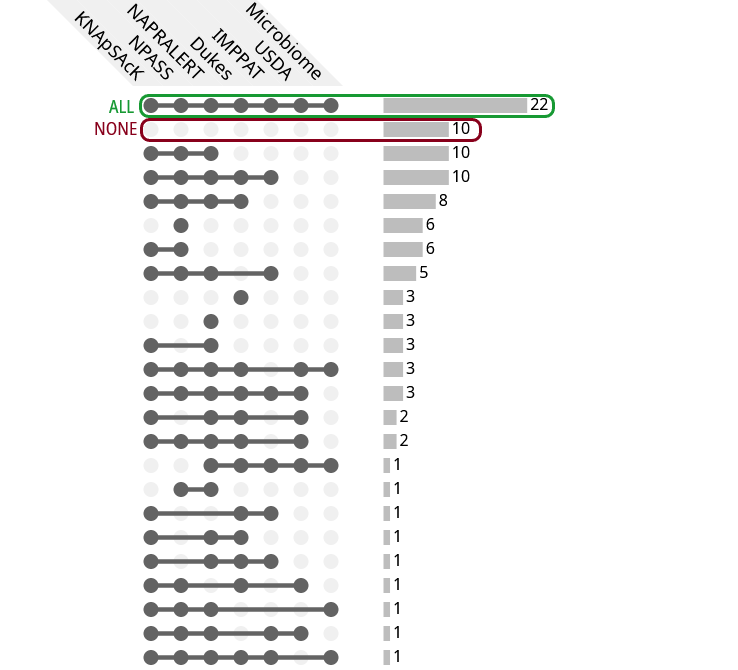
###### Figure 17

Number of plants from the dataset that are present in each database.



###### Figure 18

Grouping of the databases by overlapping the presence of each specific species.



# 

# 

# Discussion

## Databases

As mentioned earlier, KNApSAck, NPASS, and NAPRALERT were the databases with the fewest missing species, with NAPRALERT and KNApSAcK having the most and second most exact pairs, respectively. This follows accordingly with important claims from each of these databases, as they seemingly aim to be more generalized and extensive. NAPRALERT reports that data is pulled from research and reports that cover a range of topics which include: medically relevant activities, activities that affect plant growth, ethnomedical plants, and natural insecticides (Loub, et al., 1985). KNApSAcK states that data is pulled from numerous sources, especially those with data relating to biological pathways, and on their search engine (NAIST Comparative Genomics Laboratory, 2021), it is stated to have 57,906 metabolite entries and 137,333 metabolite-species pairs as of their most recent update in August 2021. NPASS states that their 25,041 source organisms include plants, microbes and marine species from differing Kingdoms, and that there are records for 35,032 different natural products (Zeng et al., 2018).

The list of species found in all of the databases referenced in this project (Table 35) are influenced by the nature of Dr. Duke’s database, the USDA database, IMPPAT and Microbiome, which are mostly concerned with data relating to culinary and/or traditional medicinal applications. The species consist almost entirely of primarily culinary plants, except for Nelumbo nucifera, which has a long history of both culinary and traditional medicinal applications (Board of Trustees of the Royal Botanic Gardens, Kew).

###### Table 35

List of species present in all the databases, as of the first half of 2021.

| *1. Asparagus officinalis* | *7. Citrus sinensis* | *13. Nelumbo nucifera* | *19. Spinacia oleracea* |
| --- | --- | --- | --- |
| *2. Beta vulgaris* | *8. Cucumis melo* | *14. Phaseolus vulgaris* | *20. Vigna radiata* |
| *3. Brassica oleracea* | *9. Cucumis sativus* | *15. Prunus avium* | *21. Vigna unguiculata* |
| *4. Capsicum annuum* | *10. Daucus carota* | *16. Prunus persica* | *22. Vitis vinifera* |
| *5. Carica papaya* | *11. Glycine max* | *17. Raphanus sativus* |  |
| *6. Cicer arietinum* | *12. Juglans regia* | *18. Solanum tuberosum* |  |

The list of species missing from each database (Table 36) makes sense in the context of the information provided by the databases, diverse and specialized alike. While some provide more diverse data than others, there is only so much available information that is available for acquisition, which could provide an explanation as to why half of these species (*M. commoda, M. neglectum, O. lucimarinus, O. tauri, and V. carteri f. nagariensis*) are green algae. The remaining species do not appear to have any culinary or medicinal applications, which also offers a better understanding of their absence. Though it should be noted that each of the non-algae species listed have been subjects of interest relating to genomes sequencing: *A. tauschii* and the evolution of hexaploid wheat (Dvorak et al., 1998); Genome analysis of diploid the ancestors of *A. hypogea* (peanut), *A. duranensis* and *A. ipaensis* (Bertioli et al., 2016); Studies of self-compatibility and incompatibility by analyzing *C. rubella* (Slotte et al., 2013); Comparing genomes of *Brassicaceae* to sister family *Cleomaceae* by sequencing the genome of *T. hassleriana* (Cheng et al., 2013).

###### Table 36

List of species missing from all databases, as of the first half of 2021.

| *1. Aegilops tauschii* | *5. Micromonas commoda* | *8. Ostreococcus tauri* |
| --- | --- | --- |
| *2. Arachis duranensis* | *6. Monoraphidium neglectum* | *9. Tarenaya hassleriana* |
| *3. Arachis ipaensis* | *7. Ostreococcus lucimarinus* | *10. Volvox carteri* |
| *4. Capsella rubella* |  |  |

Overall, the databases seemed to be most effective at providing data relating to culinary plants, while also providing information on a decent number of traditional medicinal species, with the exception of the USDA database as it was specifically concerned with the flavonoid content in food items. The focus on these types of plants is in no way surprising as medicine and the contents of the human diet are heavily intertwined with the collective interest of extending and improving human life.

While the more diverse and extensive databases provided a wider range of plant species, the information on species of algae were lacking as the data collected is somewhat telling of the focus of research from which is reported. Research regarding phytochemicals has been a long, ongoing endeavor which was seemingly more focused on land plants, especially earlier in the time span which could have been affected by something as simple as the accessibility of subjects or perhaps believing that there was nothing noteworthy about pond scum.

## Comparisons of Experimental Data and Program Output

The experimentation and research relating to flavonoid biosynthesis has biases in terms of the flavonoids and plants chosen for the subject of study, which leads to an uneven distribution of experimentally known data. As a result, there are programmatically predicted organism-compound pairs that do not have any sort of representations within the data gathered during the literature and database search, which is kept in mind when comparing data. In addition to the predicted pairs without matching experimental data, it is also important to look at the pairs which are experimentally reported but have no predicted counterparts.

At the time of writing, there are some potential reasons that could offer an explanation for the data that is experimentally known but not predicted:

1. There is some issue in the code and/or input files which has remained consistent and undetected to date and thus has negatively impacted the predictions.
2. There could be some missing context in designing the logic for the prediction functions, which could potentially have a significant impact on the outcome.
3. There could be issues that could arise from the assembled pathways from KEGG, especially if there is a popularized or standardized usage of a specific strain, such as the “Grandsen” strain of *P. patens* (Rensing et al., 2020).
4. Some sources used dried plant matter for testing, which is important to note as the drying process can impact the chemical composition of a plant (di Cesare et al., 2003).
5. Some plants, especially commonly edible plants, can have numerous cultivars, varieties, and subspecies, such as *Brassica oleracea*. There could be compositions unique to different sub-groups and as a result, if these sub-groups are not considered or reported then there could be some unanticipated errors.
6. KEGG’s flavonoid biosynthesis map involves specific stereoisomers, while the sources have had mentions of both stereoisomers of the flavan-3-ols.
7. There could be some type of unknown or undocumented convergence of biosynthetic capabilities between different lineages.
8. As the disclaimer from Dr. Duke's database states, "To err is human", which sets the expectation of some degree of error in the collected data.

### Isoflavonoid

Of the nine species of plants predicted to synthesize genistein, eight of them are experimentally known to do so (Table 37). *Glycine soja* is the only species predicted to synthesize genistein without an experimental confirmation. Additionally, there are nine species which are only experimentally known to synthesize genistein (Table 38), though there is only one data source for each of these exact pairs. Therefore, at this point, there could be issues with the program or the data that has been collected. Additionally, some of the disagreements between the experimentally known data and the predictions could be attributed to the mass cultivation and *Beta vulgaris*, *Brassica oleracea*, *Prunus avium*, *Spinacia oleracea*, and *Zea mays*.

###### Table 37

Plants predicted and experimentally known to synthesize genistein

| *1. Cajanus cajan* | *4. Lupinus angustifolius* | *7. Vigna radiata* |
| --- | --- | --- |
| *2. Cicer arietinum* | *5. Phaseolus vulgaris* | *8. Vigna unguiculata* |
| *3. Glycine max* | *6. Vigna angularis* |  |

###### Table 38

Plants experimentally known to synthesize genistein but were not predicted.

| *1. Amborella trichopoda* | *4. Brassica oleracea* | *7. Prunus avium* |
| --- | --- | --- |
| *2. Arabidopsis thaliana* | *5. Gossypium hirsutum* | *8. Spinacia oleracea* |
| *3. Beta vulgaris* | *6. Medicago truncatula* | *9. Zea mays* |

### Anthocyanidin

Of the 76 species predicted to synthesize cyanidin, 25 of them were experimentally known to synthesize cyanidin (Table 39), while the remaining 51 species currently have no corresponding data (Table 40). There were only two species that were experimentally known but not predicted to synthesize cyanidin: *Fragaria vesca* and *Lactuca sativa*. Given that there is a small number of disagreements and that these disagreements could be attributed to the heavy cultivation and selective breeding of these species, the predictions for cyanidin could be useful in designing experiments for testing phytochemical contents in the selected species.

###### Table 39

Plants predicted and experimentally known to synthesize cyanidin.

| *1. Arabidopsis thaliana* | *10. Lotus japonicus* | *18. Prunus persica* |
| --- | --- | --- |
| *2. Brassica oleracea* | *11. Malus domestica* | *19. Raphanus sativus* |
| *3. Capsicum annuum* | *12. Medicago truncatula* | *20. Solanum tuberosum* |
| *4. Carica papaya* | *13. Olea europaea* | *21. Theobroma cacao* |
| *5. Daucus carota* | *14. Papaver somniferum* | *22. Vigna angularis* |
| *6. Glycine max* | *15. Phaseolus vulgaris* | *23. Vigna unguiculata* |
| *7. Gossypium hirsutum* | *16. Phoenix dactylifera* | *24. Vitis vinifera* |
| *8. Ipomoea nil* | *17. Prunus avium* | *25. Zea mays* |
| *9. Juglans regia* |  |  |

###### Table 40

Plants predicted to synthesize cyanidin but not yet found experimentally.

| *1. Aegilops tauschii* | *18. Eucalyptus grandis* | *35. Oryza sativa* |
| --- | --- | --- |
| *2. Amborella trichopoda* | *19. Eutrema salsugineum* | *36. Phalaenopsis equestris* |
| *3. Arabidopsis lyrata* | *20. Glycine soja* | *37. Populus euphratica* |
| *4. Arachis duranensis* | *21. Gossypium arboreum* | *38. Populus trichocarpa* |
| *5. Asparagus officinalis* | *22. Gossypium raimondii* | *39. Prunus mume* |
| *6. Beta vulgaris* | *23. Helianthus annuus* | *40. Pyrus x* |
| *7. Brassica napus* | *24. Hevea brasiliensis* | *41. Quercus suber* |
| *8. Brassica rapa* | *25. Jatropha curcas* | *42. Ricinus communis* |
| *9. Camelina sativa* | *26. Lupinus angustifolius* | *43. Rosa chinensis* |
| *10. Chenopodium quinoa* | *27. Manihot esculenta* | *44. Sesamum indicum* |
| *11. Cicer arietinum* | *28. Musa acuminata* | *45. Setaria italica* |
| *12. Citrus clementina* | *29. Nelumbo nucifera* | *46. Solanum lycopersicum* |
| *13. Citrus sinensis* | *30. Nicotiana attenuata* | *47. Solanum pennellii* |
| *14. Cynara cardunculus* | *31. Nicotiana sylvestris* | *48. Spinacia oleracea* |
| *15. Dendrobium catenatum* | *32. Nicotiana tabacum* | *49. Tarenaya hassleriana* |
| *16. Durio zibethinus* | *33. Nicotiana tomentosiformis* | *50. Vigna radiata* |
| *17. Elaeis guineensis* | *34. Oryza brachyantha* | *51. Ziziphus jujuba* |

### 

### Chalcones

*G. soja* is the only species that was experimentally known and predicted to synthesize butein. While it is good that the only experimental confirmation from the dataset is also predicted, it is difficult to determine the accuracy of these predictions from such a small data set.

The entirety of the plants that were experimentally known to synthesize isoliquiritigenin were also predicted in the program (Table 41). While it is slightly promising to find data that confirms a number of these predictions, it would be more advantageous to find more experimental data before coming to a solid conclusion regarding the accuracy of the predictions for isoliquiritigenin.

###### Table 41

Plants predicted and experimentally known to synthesize isoliquiritigenin.

| *1. Arabidopsis thaliana* | *3. Glycine max* | *5. Medicago truncatula* |
| --- | --- | --- |
| *2. Cicer arietinum* | *4. Helianthus annuus* |  |

### Flavan-3-ols

Of the 58 species predicted to synthesize catechin, there were 37 of these which were also experimentally known to synthesize catechin (Table 42). The species that were predicted but have no corresponding experimental data can be found in Table 43. However there were 11 species which have been experimentally known to synthesize catechin but were not predicted by the program. The number of predicted results with experimental confirmations does look promising in terms of program accuracy, though it may require further investigation of the species in Table 44 to better understand the discrepancies before finding more confidence with the predictions for catechin. Each experimental disagreement had only one corresponding data source, which could potentially be indicative of isolated incidents of catechin synthesis in a specific variety or cultivar which may have not been specified in the documented experimental results.

###### Table 42

Plants predicted and experimentally known to synthesize catechin.

| *1. Aegilops tauschii* | *14. Jatropha curcas* | *26. Populus trichocarpa* |
| --- | --- | --- |
| *2. Amborella trichopoda* | *15. Juglans regia* | *27. Prunus avium* |
| *3. Asparagus officinalis* | *16. Malus domestica* | *28. Prunus mume* |
| *4. Beta vulgaris* | *17. Manihot esculenta* | *29. Prunus persica* |
| *5. Carica papaya* | *18. Medicago truncatula* | *30. Pyrus x* |
| *6. Chenopodium quinoa* | *19. Musa acuminata* | *31. Solanum tuberosum* |
| *7. Cicer arietinum* | *20. Nelumbo nucifera* | *32. Spinacia oleracea* |
| *8. Citrus sinensis* | *21. Nicotiana tabacum* | *33. Theobroma cacao* |
| *9. Durio zibethinus* | *22. Olea europaea* | *34. Vigna angularis* |
| *10. Elaeis guineensis* | *23. Oryza sativa* | *35. Vigna radiata* |
| *11. Eucalyptus grandis* | *24. Phaseolus vulgaris* | *36. Vitis vinifera* |
| *12. Glycine max* | *25. Phoenix dactylifera* | *37. Ziziphus jujuba* |
| *13. Gossypium hirsutum* |  |  |

###### Table 43

Plants predicted to synthesize catechin but not yet found experimentally.

| *1. Arachis duranensis* | *8. Hevea brasiliensis* | *15. Papaver somniferum* |
| --- | --- | --- |
| *2. Brachypodium distachyon* | *9. Lotus japonicus* | *16. Populus euphratica* |
| *3. Citrus clementina* | *10. Lupinus angustifolius* | *17. Quercus suber* |
| *4. Glycine soja* | *11. Nicotiana attenuata* | *18. Ricinus communis* |
| *5. Gossypium arboreum* | *12. Nicotiana sylvestris* | *19. Rosa chinensis* |
| *6. Gossypium raimondii* | *13. Nicotiana tomentosiformis* | *20. Solanum pennellii* |
| *7. Helianthus annuus* | *14. Oryza brachyantha* | *21. Vigna unguiculata* |

###### Table 44

Plants experimentally known to synthesize catechin but were not predicted.

| *1. Arabidopsis thaliana* | *5. Cucumis melo* | *9. Momordica charantia* |
| --- | --- | --- |
| *2. Brassica rapa* | *6. Cucumis sativus* | *10. Sesamum indicum* |
| *3. Camelina sativa* | *7. Cucurbita pepo* | *11. Sorghum bicolor* |
| *4. Capsicum annuum* | *8. Fragaria vesca* |  |

Of the species that were predicted to synthesize epicatechin, there are 30 species which provide experimental confirmation (Table 45) and 35 which do not (Table 46). There were also eight species which were experimentally known to synthesize epicatechin but were not predicted (Table 47). There too should be investigation into these species before drawing any further conclusions, as the entirety of these species are culinary and thus have been subject to selective breeding and genetic variation.

###### Table 45

Plants predicted and experimentally known to synthesize epicatechin.

| *1. Amborella trichopoda* | *11. Juglans regia* | *21. Prunus avium* |
| --- | --- | --- |
| *2. Arabidopsis thaliana* | *12. Malus domestica* | *22. Prunus mume* |
| *3. Beta vulgaris* | *13. Medicago truncatula* | *23. Prunus persica* |
| *4. Brassica napus* | *14. Musa acuminata* | *24. Pyrus x* |
| *5. Carica papaya* | *15. Nelumbo nucifera* | *25. Ricinus communis* |
| *6. Durio zibethinus* | *16. Nicotiana sylvestris* | *26. Solanum tuberosum* |
| *7. Elaeis guineensis* | *17. Nicotiana tabacum* | *27. Theobroma cacao* |
| *8. Eucalyptus grandis* | *18. Phaseolus vulgaris* | *28. Vigna angularis* |
| *9. Glycine max* | *19. Phoenix dactylifera* | *29. Vigna radiata* |
| *10. Gossypium hirsutum* | *20. Populus trichocarpa* | *30. Ziziphus jujuba* |

###### Table 46

Plants predicted to synthesize epicatechin but not yet found experimentally.

| *1. Aegilops tauschii* | *13. Glycine soja* | *25. Papaver somniferum* |
| --- | --- | --- |
| *2. Arabidopsis lyrata* | *14. Gossypium arboreum* | *26. Populus euphratica* |
| *3. Arachis duranensis* | *15. Gossypium raimondii* | *27. Quercus suber* |
| *4. Asparagus officinalis* | *16. Hevea brasiliensis* | *28. Raphanus sativus* |
| *5. Brassica oleracea* | *17. Jatropha curcas* | *29. Rosa chinensis* |
| *6. Brassica rapa* | *18. Lotus japonicus* | *30. Setaria italica* |
| *7. Camelina sativa* | *19. Lupinus angustifolius* | *31. Solanum lycopersicum* |
| *8. Chenopodium quinoa* | *20. Manihot esculenta* | *32. Solanum pennellii* |
| *9. Cicer arietinum* | *21. Nicotiana attenuata* | *33. Tarenaya hassleriana* |
| *10. Citrus clementina* | *22. Nicotiana tomentosiformis* | *34. Vigna unguiculata* |
| *11. Citrus sinensis* | *23. Oryza brachyantha* | *35. Zea mays* |
| *12. Eutrema salsugineum* | *24. Oryza sativa* |  |

###### Table 47

Plants experimentally known to synthesize epicatechin but were not predicted.

| *1. Capsicum annuum* | *4. Momordica charantia* | *7. Spinacia oleracea* |
| --- | --- | --- |
| *2. Cucumis melo* | *5. Sesamum indicum* | *8. Vitis vinifera* |
| *3. Fragaria vesca* | *6. Sorghum bicolor* |  |

Of the species predicted to synthesize epigallocatechin, there are only five experimentally known to synthesize the compound (Table 48), while the remaining 38 species are not yet known to do so (Table 49). There were five species of plants that were experimentally known to synthesize epigallocatechin but were not predicted (Table 50), which is concerning when compared to the number of experimental confirmations. Most likely these disagreements could be attributed to the usage of different cultivars in testing as these species are all culinary plants, but more research will need to be done in order to confirm or deny these disagreements before making any further judgments of the predictions for epigallocatechin.

###### Table 48

Plants predicted and experimentally known to synthesize epigallocatechin.

| *1. Citrus sinensis* | *3. Phaseolus vulgaris* | *5. Ziziphus jujuba* |
| --- | --- | --- |
| *2. Gossypium hirsutum* | *4. Theobroma cacao* |  |

###### Table 49

Plants predicted to synthesize epigallocatechin but not yet found experimentally.

| *1. Aegilops tauschii* | *14. Juglans regia* | *27. Phoenix dactylifera* |
| --- | --- | --- |
| *2. Carica papaya* | *15. Lotus japonicus* | *28. Populus euphratica* |
| *3. Cicer arietinum* | *16. Lupinus angustifolius* | *29. Populus trichocarpa* |
| *4. Citrus clementina* | *17. Manihot esculenta* | *30. Quercus suber* |
| *5. Durio zibethinus* | *18. Medicago truncatula* | *31. Setaria italica* |
| *6. Elaeis guineensis* | *19. Musa acuminata* | *32. Solanum lycopersicum* |
| *7. Eucalyptus grandis* | *20. Nelumbo nucifera* | *33. Solanum pennellii* |
| *8. Glycine max* | *21. Nicotiana attenuata* | *34. Solanum tuberosum* |
| *9. Glycine soja* | *22. Nicotiana sylvestris* | *35. Vigna angularis* |
| *10. Gossypium arboreum* | *23. Nicotiana tabacum* | *36. Vigna radiata* |
| *11. Gossypium raimondii* | *24. Nicotiana tomentosiformis* | *37. Vigna unguiculata* |
| *12. Hevea brasiliensis* | *25. Oryza brachyantha* | *38. Zea mays* |
| *13. Jatropha curcas* | *26. Oryza sativa* |  |

###### Table 50

Plants experimentally known to synthesize epigallocatechin but were not predicted.

| *1. Cucumis melo* | *3. Prunus avium* | *5. Vitis vinifera* |
| --- | --- | --- |
| *2. Malus domestica* | *4. Prunus persica* |  |

Of the species predicted to synthesize gallocatechin, 10 species are known to synthesize the compound (Table 51), while the remaining 32 species have not yet been recorded to do so (Table 52). The lack of only experimentally known results is promising in terms of analyzing the accuracy of the predictions, though it is suggested that more experimental data should be collected as available before designing experiments based on the predictions for gallocatechin.

###### Table 51

Plants predicted and experimentally known to synthesize gallocatechin.

| *1. Gossypium hirsutum* | *5. Phaseolus vulgaris* | *8. Vigna radiata* |
| --- | --- | --- |
| *2. Juglans regia* | *6. Theobroma cacao* | *9. Vitis vinifera* |
| *3. Medicago truncatula* | *7. Vigna angularis* | *10. Ziziphus jujuba* |
| *4. Nelumbo nucifera* |  |  |

###### Table 52

Plants predicted to synthesize gallocatechin but not yet found experimentally.

| *1. Aegilops tauschii* | *12. Gossypium arboreum* | *23. Nicotiana tomentosiformis* |
| --- | --- | --- |
| *2. Brachypodium distachyon* | *13. Gossypium raimondii* | *24. Oryza brachyantha* |
| *3. Carica papaya* | *14. Hevea brasiliensis* | *25. Oryza sativa* |
| *4. Cicer arietinum* | *15. Jatropha curcas* | *26. Phoenix dactylifera* |
| *5. Citrus clementina* | *16. Lotus japonicus* | *27. Populus euphratica* |
| *6. Citrus sinensis* | *17. Lupinus angustifolius* | *28. Populus trichocarpa* |
| *7. Durio zibethinus* | *18. Manihot esculenta* | *29. Quercus suber* |
| *8. Elaeis guineensis* | *19. Musa acuminata* | *30. Solanum pennellii* |
| *9. Eucalyptus grandis* | *20. Nicotiana attenuata* | *31. Solanum tuberosum* |
| *10. Glycine max* | *21. Nicotiana sylvestris* | *32. Vigna unguiculata* |
| *11. Glycine soja* | *22. Nicotiana tabacum* |  |

### Flavanones

Of the species predicted to synthesize eriodictyol, 11 species are known to synthesize the compound (Table 53), while the remaining 78 species have not yet been recorded to do so (Table 54).The lack of experimental disagreements looks promising in the context of the predictions, though, the number of confirmations are still small compared to the overall set of predicted species meaning that more literature is needed before labeling the predictions for eriodictyol as accurate.

###### Table 53

Plants predicted and experimentally known to synthesize eriodictyol.

| *1. Amborella trichopoda* | *5. Malus domestica* | *9. Theobroma cacao* |
| --- | --- | --- |
| *2. Eucalyptus grandis* | *6. Prunus avium* | *10. Vigna radiata* |
| *3. Glycine soja* | *7. Prunus persica* | *11. Ziziphus jujuba* |
| *4. Gossypium hirsutum* | *8. Sorghum bicolor* |  |

###### 

###### Table 54

Plants predicted to synthesize eriodictyol but not yet found experimentally.

| *1. Aegilops tauschii* | *27. Dendrobium catenatum* | *53. Oryza sativa* |
| --- | --- | --- |
| *2. Arabidopsis lyrata* | *28. Durio zibethinus* | *54. Papaver somniferum* |
| *3. Arabidopsis thaliana* | *29. Elaeis guineensis* | *55. Phalaenopsis equestris* |
| *4. Arachis duranensis* | *30. Eutrema salsugineum* | *56. Phaseolus vulgaris* |
| *5. Arachis ipaensis* | *31. Fragaria vesca* | *57. Phoenix dactylifera* |
| *6. Asparagus officinalis* | *32. Glycine max* | *58. Physcomitrium patens* |
| *7. Beta vulgaris* | *33. Gossypium arboreum* | *59. Populus euphratica* |
| *8. Brachypodium distachyon* | *34. Gossypium raimondii* | *60. Populus trichocarpa* |
| *9. Brassica napus* | *35. Helianthus annuus* | *61. Prunus mume* |
| *10. Brassica oleracea* | *36. Hevea brasiliensis* | *62. Pyrus x* |
| *11. Brassica rapa* | *37. Ipomoea nil* | *63. Quercus suber* |
| *12. Cajanus cajan* | *38. Jatropha curcas* | *64. Raphanus sativus* |
| *13. Camelina sativa* | *39. Juglans regia* | *65. Ricinus communis* |
| *14. Capsicum annuum* | *40. Lactuca sativa* | *66. Rosa chinensis* |
| *15. Carica papaya* | *41. Lotus japonicus* | *67. Selaginella moellendorffii* |
| *16. Chenopodium quinoa* | *42. Lupinus angustifolius* | *68. Sesamum indicum* |
| *17. Cicer arietinum* | *43. Manihot esculenta* | *69. Setaria italica* |
| *18. Citrus clementina* | *44. Medicago truncatula* | *70. Solanum lycopersicum* |
| *19. Citrus sinensis* | *45. Musa acuminata* | *71. Solanum pennellii* |
| *20. Cucumis melo* | *46. Nelumbo nucifera* | *72. Solanum tuberosum* |
| *21. Cucumis sativus* | *47. Nicotiana attenuata* | *73. Spinacia oleracea* |
| *22. Cucurbita maxima* | *48. Nicotiana sylvestris* | *74. Tarenaya hassleriana* |
| *23. Cucurbita moschata* | *49. Nicotiana tabacum* | *75. Vigna angularis* |
| *24. Cucurbita pepo* | *50. Nicotiana tomentosiformis* | *76. Vigna unguiculata* |
| *25. Cynara cardunculus* | *51. Olea europaea* | *77. Vitis vinifera* |
| *26. Daucus carota* | *52. Oryza brachyantha* | *78. Zea mays* |

Of the 90 species predicted to synthesize naringenin, 44 species are known to synthesize the compound (Table 55), while the remaining 46 species have not yet been recorded to do so (Table 56). Due to the absence of species that were experimentally known but not predicted and the comparably larger amount of experimental confirmations for the predicted species, these predictions could provide a better basis for research in comparison to the other flavonoids.

###### Table 55

Plants predicted and experimentally known to synthesize naringenin.

| *1. Amborella trichopoda* | *16. Daucus carota* | *31. Prunus avium* |
| --- | --- | --- |
| *2. Arabidopsis thaliana* | *17. Dendrobium catenatum* | *32. Prunus mume* |
| *3. Asparagus officinalis* | *18. Eucalyptus grandis* | *33. Prunus persica* |
| *4. Brachypodium distachyon* | *19. Fragaria vesca* | *34. Raphanus sativus* |
| *5. Brassica napus* | *20. Glycine max* | *35. Selaginella moellendorffii* |
| *6. Brassica oleracea* | *21. Gossypium hirsutum* | *36. Solanum lycopersicum* |
| *7. Brassica rapa* | *22. Juglans regia* | *37. Solanum tuberosum* |
| *8. Cajanus cajan* | *23. Medicago truncatula* | *38. Sorghum bicolor* |
| *9. Capsicum annuum* | *24. Momordica charantia* | *39. Spinacia oleracea* |
| *10. Carica papaya* | *25. Musa acuminata* | *40. Theobroma cacao* |
| *11. Chenopodium quinoa* | *26. Nelumbo nucifera* | *41. Vigna angularis* |
| *12. Cicer arietinum* | *27. Olea europaea* | *42. Vigna radiata* |
| *13. Citrus sinensis* | *28. Oryza sativa* | *43. Vitis vinifera* |
| *14. Cucumis melo* | *29. Phaseolus vulgaris* | *44. Zea mays* |
| *15. Cynara cardunculus* | *30. Phoenix dactylifera* |  |

###### 

###### Table 56

Plants predicted to synthesize naringenin but not yet found experimentally.

| *1. Aegilops tauschii* | *17. Gossypium raimondii* | *32. Papaver somniferum* |
| --- | --- | --- |
| *2. Arabidopsis lyrata* | *18. Helianthus annuus* | *33. Phalaenopsis equestris* |
| *3. Arachis duranensis* | *19. Hevea brasiliensis* | *34. Physcomitrium patens* |
| *4. Arachis ipaensis* | *20. Ipomoea nil* | *35. Populus euphratica* |
| *5. Beta vulgaris* | *21. Jatropha curcas* | *36. Populus trichocarpa* |
| *6. Camelina sativa* | *22. Lactuca sativa* | *37. Pyrus x* |
| *7. Citrus clementina* | *23. Lotus japonicus* | *38. Quercus suber* |
| *8. Cucumis sativus* | *24. Lupinus angustifolius* | *39. Ricinus communis* |
| *9. Cucurbita maxima* | *25. Malus domestica* | *40. Rosa chinensis* |
| *10. Cucurbita moschata* | *26. Manihot esculenta* | *41. Sesamum indicum* |
| *11. Cucurbita pepo* | *27. Nicotiana attenuata* | *42. Setaria italica* |
| *12. Durio zibethinus* | *28. Nicotiana sylvestris* | *43. Solanum pennellii* |
| *13. Elaeis guineensis* | *29. Nicotiana tabacum* | *44. Tarenaya hassleriana* |
| *14. Eutrema salsugineum* | *30. Nicotiana tomentosiformis* | *45. Vigna unguiculata* |
| *15. Glycine soja* | *31. Oryza brachyantha* | *46. Ziziphus jujuba* |
| *16. Gossypium arboreum* |  |  |

### Flavonols

Of the 84 species predicted to synthesize kaempferol, 49 species are known to synthesize the compound (Table 57), while the remaining 35 species have not yet been recorded to do so (Table 58). Similar to naringenin, the large number of experimental confirmations and the lack of disagreements between the experimental and theoretical data makes these predictions ideal to use as a basis for experimentation.

###### Table 57

Plants predicted and experimentally known to synthesize kaempferol .

| *1. Arabidopsis thaliana* | *18. Fragaria vesca* | *34. Papaver somniferum* |
| --- | --- | --- |
| *2. Asparagus officinalis* | *19. Glycine max* | *35. Phaseolus vulgaris* |
| *3. Beta vulgaris* | *20. Gossypium hirsutum* | *36. Prunus avium* |
| *4. Brassica napus* | *21. Hevea brasiliensis* | *37. Prunus persica* |
| *5. Brassica oleracea* | *22. Juglans regia* | *38. Raphanus sativus* |
| *6. Brassica rapa* | *23. Lactuca sativa* | *39. Ricinus communis* |
| *7. Capsicum annuum* | *24. Lotus japonicus* | *40. Rosa chinensis* |
| *8. Carica papaya* | *25. Lupinus angustifolius* | *41. Solanum lycopersicum* |
| *9. Chenopodium quinoa* | *26. Malus domestica* | *42. Solanum tuberosum* |
| *10. Cicer arietinum* | *27. Medicago truncatula* | *43. Spinacia oleracea* |
| *11. Citrus sinensis* | *28. Momordica charantia* | *44. Theobroma cacao* |
| *12. Cucumis melo* | *29. Musa acuminata* | *45. Vigna radiata* |
| *13. Cucumis sativus* | *30. Nelumbo nucifera* | *46. Vigna unguiculata* |
| *14. Cucurbita maxima* | *31. Nicotiana tabacum* | *47. Vitis vinifera* |
| *15. Cucurbita pepo* | *32. Olea europaea* | *48. Zea mays* |
| *16. Daucus carota* | *33. Oryza sativa* | *49. Ziziphus jujuba* |
| *17. Elaeis guineensis* |  |  |

###### Table 58

Plants predicted to synthesize kaempferol but not yet found experimentally.

| *1. Aegilops tauschii* | *13. Eutrema salsugineum* | *25. Phoenix dactylifera* |
| --- | --- | --- |
| *2. Amborella trichopoda* | *14. Glycine soja* | *26. Populus euphratica* |
| *3. Arabidopsis lyrata* | *15. Gossypium arboreum* | *27. Populus trichocarpa* |
| *4. Arachis duranensis* | *16. Gossypium raimondii* | *28. Prunus mume* |
| *5. Arachis ipaensis* | *17. Helianthus annuus* | *29. Pyrus x* |
| *6. Brachypodium distachyon* | *18. Ipomoea nil* | *30. Quercus suber* |
| *7. Camelina sativa* | *19. Jatropha curcas* | *31. Sesamum indicum* |
| *8. Citrus clementina* | *20. Manihot esculenta* | *32. Setaria italica* |
| *9. Cucurbita moschata* | *21. Nicotiana attenuata* | *33. Solanum pennellii* |
| *10. Cynara cardunculus* | *22. Nicotiana sylvestris* | *34. Tarenaya hassleriana* |
| *11. Durio zibethinus* | *23. Nicotiana tomentosiformis* | *35. Vigna angularis* |
| *12. Eucalyptus grandis* | *24. Phalaenopsis equestris* |  |

Of the 47 species predicted to synthesize myricetin, 13 species are known to synthesize the compound (Table 59), while the remaining 34 species have not yet been recorded to do so (Table 60). However, there are six species experimentally known to synthesize myricetin but were not predicted by the program (Table 61). It is promising that the list of disagreements consists entirely of culinary plants, but further investigation is necessary before completely dismissing the data as inaccurately reported data from specific cultivars. Currently, the limited and conflicting nature of the data do not indicate that the predictions for myricetin are correct, so the predictions should not be used as a basis for experimentation.

###### Table 59

Plants predicted and experimentally known to synthesize myricetin.

| *1. Capsicum annuum* | *6. Lotus japonicus* | *10. Solanum tuberosum* |
| --- | --- | --- |
| *2. Carica papaya* | *7. Medicago truncatula* | *11. Vigna radiata* |
| *3. Citrus sinensis* | *8. Nelumbo nucifera* | *12. Vigna unguiculata* |
| *4. Gossypium hirsutum* | *9. Phaseolus vulgaris* | *13. Vitis vinifera* |
| *5. Juglans regia* |  |  |

###### 

###### Table 60

Plants predicted to synthesize myricetin but not yet found experimentally.

| *1. Aegilops tauschii* | *13. Jatropha curcas* | *24. Phoenix dactylifera* |
| --- | --- | --- |
| *2. Brachypodium distachyon* | *14. Lupinus angustifolius* | *25. Populus euphratica* |
| *3. Cicer arietinum* | *15. Manihot esculenta* | *26. Populus trichocarpa* |
| *4. Citrus clementina* | *16. Musa acuminata* | *27. Quercus suber* |
| *5. Durio zibethinus* | *17. Nicotiana attenuata* | *28. Setaria italica* |
| *6. Elaeis guineensis* | *18. Nicotiana sylvestris* | *29. Solanum lycopersicum* |
| *7. Eucalyptus grandis* | *19. Nicotiana tabacum* | *30. Solanum pennellii* |
| *8. Glycine max* | *20. Nicotiana tomentosiformis* | *31. Theobroma cacao* |
| *9. Glycine soja* | *21. Oryza brachyantha* | *32. Vigna angularis* |
| *10. Gossypium arboreum* | *22. Oryza sativa* | *33. Zea mays* |
| *11. Gossypium raimondii* | *23. Phalaenopsis equestris* | *34. Ziziphus jujuba* |
| *12. Hevea brasiliensis* |  |  |

###### 

###### Table 61

Plants experimentally known to synthesize myricetin but were not predicted.

| *1. Asparagus officinalis* | *3. Daucus carota* | *5. Prunus avium* |
| --- | --- | --- |
| *2. Brassica rapa* | *4. Malus domestica* | *6. Spinacia oleracea* |

Of the 83 species predicted to synthesize Quercetin, 57 species are known to synthesize the compound (Table 62), while the remaining 26 species have not yet been recorded to do so (Table 63). There were two species *C. cajan* and *M. charantia* that were experimentally known to synthesize quercetin but were not predicted. Although there are two species which offer some disagreement, they are culinary plants and thus have different cultivars that can affect phytochemical composition. The sheer amount of available literature combined with the largest set of experimental confirmations for the predictions makes quercetin, at the very least, worth the investigation in order to better address the disagreements.

###### Table 62

Plants predicted and experimentally known to synthesize quercetin.

| *1. Amborella trichopoda* | *20. Fragaria vesca* | *39. Phaseolus vulgaris* |
| --- | --- | --- |
| *2. Arabidopsis thaliana* | *21. Glycine max* | *40. Phoenix dactylifera* |
| *3. Asparagus officinalis* | *22. Gossypium arboreum* | *41. Prunus avium* |
| *4. Beta vulgaris* | *23. Gossypium hirsutum* | *42. Prunus mume* |
| *5. Brassica napus* | *24. Helianthus annuus* | *43. Prunus persica* |
| *6. Brassica oleracea* | *25. Hevea brasiliensis* | *44. Quercus suber* |
| *7. Brassica rapa* | *26. Jatropha curcas* | *45. Raphanus sativus* |
| *8. Camelina sativa* | *27. Juglans regia* | *46. Ricinus communis* |
| *9. Capsicum annuum* | *28. Lactuca sativa* | *47. Rosa chinensis* |
| *10. Carica papaya* | *29. Lotus japonicus* | *48. Solanum lycopersicum* |
| *11. Chenopodium quinoa* | *30. Lupinus angustifolius* | *49. Solanum tuberosum* |
| *12. Cicer arietinum* | *31. Malus domestica* | *50. Spinacia oleracea* |
| *13. Citrus sinensis* | *32. Manihot esculenta* | *51. Theobroma cacao* |
| *14. Cucumis melo* | *33. Medicago truncatula* | *52. Vigna angularis* |
| *15. Cucumis sativus* | *34. Musa acuminata* | *53. Vigna radiata* |
| *16. Cucurbita maxima* | *35. Nelumbo nucifera* | *54. Vigna unguiculata* |
| *17. Cucurbita pepo* | *36. Nicotiana tabacum* | *55. Vitis vinifera* |
| *18. Cynara cardunculus* | *37. Olea europaea* | *56. Zea mays* |
| *19. Daucus carota* | *38. Papaver somniferum* | *57. Ziziphus jujuba* |

###### Table 63

Plants predicted to synthesize quercetin but not yet found experimentally.

| *1. Aegilops tauschii* | *10. Eucalyptus grandis* | *19. Phalaenopsis equestris* |
| --- | --- | --- |
| *2. Arabidopsis lyrata* | *11. Eutrema salsugineum* | *20. Populus euphratica* |
| *3. Arachis duranensis* | *12. Glycine soja* | *21. Populus trichocarpa* |
| *4. Arachis ipaensis* | *13. Gossypium raimondii* | *22. Pyrus x* |
| *5. Brachypodium distachyon* | *14. Ipomoea nil* | *23. Sesamum indicum* |
| *6. Citrus clementina* | *15. Nicotiana attenuata* | *24. Setaria italica* |
| *7. Cucurbita moschata* | *16. Nicotiana sylvestris* | *25. Solanum pennellii* |
| *8. Durio zibethinus* | *17. Nicotiana tomentosiformis* | *26. Tarenaya hassleriana* |
| *9. Elaeis guineensis* | *18. Oryza sativa* |  |

### Flavones

Of the 37 species predicted to synthesize apigenin, 15 species are known to synthesize the compound (Table 64), while the remaining 22 species have not yet been recorded to do so (Table 65). Additionally there were 15 species that were only experimentally known to synthesize apigenin (Table 66), which does not instill much confidence in the accuracy of these predictions. All of the disagreements, besides Jatropha curcas, are commonly known culinary plants which then means an additional investigation should be conducted regarding the cultivars utilized in the experiments before investigating the logic of the program.

###### Table 64

Plants predicted and experimentally known to synthesize apigenin.

| *1. Cajanus cajan* | *6. Lactuca sativa* | *11. Populus trichocarpa* |
| --- | --- | --- |
| *2. Citrus sinensis* | *7. Lupinus angustifolius* | *12. Ricinus communis* |
| *3. Cynara cardunculus* | *8. Medicago truncatula* | *13. Sorghum bicolor* |
| *4. Daucus carota* | *9. Olea europaea* | *14. Theobroma cacao* |
| *5. Glycine max* | *10. Phaseolus vulgaris* | *15. Zea mays* |

###### Table 65

Plants predicted to synthesize apigenin but not yet found experimentally.

| *1. Aegilops tauschii* | *9. Glycine soja* | *16. Oryza brachyantha* |
| --- | --- | --- |
| *2. Arachis duranensis* | *10. Gossypium arboreum* | *17. Oryza sativa* |
| *3. Arachis ipaensis* | *11. Gossypium hirsutum* | *18. Populus euphratica* |
| *4. Brachypodium distachyon* | *12. Gossypium raimondii* | *19. Sesamum indicum* |
| *5. Cicer arietinum* | *13. Helianthus annuus* | *20. Setaria italica* |
| *6. Citrus clementina* | *14. Hevea brasiliensis* | *21. Vigna angularis* |
| *7. Durio zibethinus* | *15. Manihot esculenta* | *22. Vigna unguiculata* |
| *8. Eucalyptus grandis* |  |  |

###### Table 66

Plants experimentally known to synthesize apigenin but were not predicted.

| *1. Asparagus officinalis* | *6. Carica papaya* | *11. Jatropha curcas* |
| --- | --- | --- |
| *2. Beta vulgaris* | *7. Chenopodium quinoa* | *12. Juglans regia* |
| *3. Brassica oleracea* | *8. Cucurbita moschata* | *13. Phoenix dactylifera* |
| *4. Brassica rapa* | *9. Cucurbita pepo* | *14. Prunus avium* |
| *5. Capsicum annuum* | *10. Fragaria vesca* | *15. Raphanus sativus* |

Of the 37 species predicted to synthesize luteolin, 14 species are known to synthesize the compound (Table 67), while the remaining 23 species have not yet been recorded to do so (Table 68). Unfortunately, there were 22 species that were experimentally known to synthesize luteolin but were not predicted (Table 69), which needs to be further investigated due to most of the plants having a long history of cultivation.

###### Table 67

Plants predicted and experimentally known to synthesize luteolin.

| *1. Cajanus cajan* | *6. Lactuca sativa* | *11. Setaria italica* |
| --- | --- | --- |
| *2. Citrus sinensis* | *7. Lupinus angustifolius* | *12. Sorghum bicolor* |
| *3. Cynara cardunculus* | *8. Medicago truncatula* | *13. Theobroma cacao* |
| *4. Daucus carota* | *9. Olea europaea* | *14. Zea mays* |
| *5. Helianthus annuus* | *10. Phaseolus vulgaris* |  |

###### Table 68

Plants predicted to synthesize luteolin but not yet found experimentally.

| *1. Aegilops tauschii* | *9. Glycine max* | *17. Oryza sativa* |
| --- | --- | --- |
| *2. Arachis duranensis* | *10. Glycine soja* | *18. Populus euphratica* |
| *3. Arachis ipaensis* | *11. Gossypium arboreum* | *19. Populus trichocarpa* |
| *4. Brachypodium distachyon* | *12. Gossypium hirsutum* | *20. Ricinus communis* |
| *5. Cicer arietinum* | *13. Gossypium raimondii* | *21. Sesamum indicum* |
| *6. Citrus clementina* | *14. Hevea brasiliensis* | *22. Vigna angularis* |
| *7. Durio zibethinus* | *15. Manihot esculenta* | *23. Vigna unguiculata* |
| *8. Eucalyptus grandis* | *16. Oryza brachyantha* |  |

###### Table 69

Plants experimentally known to synthesize luteolin but were not predicted.

| *1. Amborella trichopoda* | *9. Chenopodium quinoa* | *16. Nelumbo nucifera* |
| --- | --- | --- |
| *2. Asparagus officinalis* | *10. Cucumis melo* | *17. Phoenix dactylifera* |
| *3. Beta vulgaris* | *11. Cucurbita moschata* | *18. Raphanus sativus* |
| *4. Brassica napus* | *12. Cucurbita pepo* | *19. Solanum lycopersicum* |
| *5. Brassica oleracea* | *13. Fragaria vesca* | *20. Solanum tuberosum* |
| *6. Brassica rapa* | *14. Malus domestica* | *21. Spinacia oleracea* |
| *7. Capsicum annuum* | *15. Momordica charantia* | *22. Vitis vinifera* |
| *8. Carica papaya* |  |  |

# 

# 

# References

Abe, N., Sato, H., & Sakamura, S. (1987). Antifungal Stress Compounds from Adzuki Bean, Vigna angularis, Treated withCephalosporium gregatumType B. Agricultural and Biological Chemistry, 51(2), 349–353. doi:10.1080/00021369.1987.10868021

Afendi, F. M., Okada, T., Yamazaki, M., Aki-Hirai-Morita, Nakamura, Y., Nakamura, K., Ikeda, S., Takahashi, H., Altaf-Ul-Amin, Latifah, Darusman, Saito, K., & Kanaya, S. (2012).“KNApSAcK Family Databases: Integrated Metabolite-Plant Species Databases for Multifaceted Plant Research, ” Plant Cell Physiol., 53, e1(1-12). doi:10.1093/pcp/pcr165

Ahmad, T., Bustam, M. A., Irfan, M., Moniruzzaman, M., Anwaar Asghar, H. M., & Bhattacharjee, S. (2018). Green synthesis of stabilized spherical shaped gold nanoparticles using novel aqueous Elaeis guineensis (oil palm) leaves extract. Journal of Molecular Structure, 1159, 167–173. doi:10.1016/j.molstruc.2017.11.095

Akomolafe, S. F., Oboh, G., Oyeleye, S. I., Molehin, O. R., & Ogunsuyi, O. B. (2016). Phenolic Composition and Inhibitory Ability of Methanolic Extract from Pumpkin (Cucurbita pepo L) Seeds on Fe-induced Thiobarbituric acid reactive species in Albino Rat’s Testicular Tissue In-Vitro. Journal of Applied Pharmaceutical Science, 6(9), 115–120. https://doi.org/10.7324/JAPS.2016.60917

Akyol, H., Riciputi, Y., Capanoglu, E., Caboni, M., & Verardo, V. (2016). Phenolic Compounds in the Potato and Its Byproducts: An Overview. International Journal of Molecular Sciences, 17(6), 835. doi:10.3390/ijms17060835

Arts, I. C. W., van de Putte, B., & Hollman, P. C. H. (2000). Catechin Contents of Foods Commonly Consumed in The Netherlands. 1. Fruits, Vegetables, Staple Foods, and Processed Foods. Journal of Agricultural and Food Chemistry, 48(5), 1746–1751. doi:10.1021/jf000025h

Auger, B., Marnet, N., Gautier, V., Maia-Grondard, A., Leprince, F., Renard, M., Guyot, S., Nesi, N., & Routaboul, J. M. (2010). A Detailed Survey of Seed Coat Flavonoids in Developing Seeds ofBrassica napusL. Journal of Agricultural and Food Chemistry, 58(10), 6246–6256. https://doi.org/10.1021/jf903619v

Bartee, L., Shriner, W., & Creech, C. (2016). The Production of a Protein – Principles of Biology. Open Oregon Educational Resources. https://openoregon.pressbooks.pub/mhccmajorsbio/chapter/production-of-a-protein/

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K. S., Liu, X., Gao, D., Clevenger, J., Dash, S., Ren, L., Moretzsohn, M. C., Shirasawa, K., Huang, W., Vidigal, B., Abernathy, B., Chu, Y., Niederhuth, C. E., Umale, P., . . . Ozias-Akins, P. (2016). The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. Nature Genetics, 48(4), 438–446. https://doi.org/10.1038/ng.3517

Bhandari, S. R., & Lee, J. G. (2016). Ripening-Dependent Changes in Antioxidants, Color Attributes, and Antioxidant Activity of Seven Tomato (Solanum lycopersicumL.) Cultivars. Journal of Analytical Methods in Chemistry, 2016, 1–13. doi:10.1155/2016/5498618

Biology Online. (2021). Biosynthesis Definition and Examples. Biology Articles, Tutorials & Dictionary Online. https://www.biologyonline.com/dictionary/biosynthesis

Board of Trustees of the Royal Botanic Gardens, Kew. Nelumbo nucifera Gaertn. | Plants of the World Online | Kew Science. Plants of the World Online. https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:605422–1

Buschmann, H., Reilly, K., Rodriguez, M. X., Tohme, J., & Beeching, J. R. (2000). Hydrogen Peroxide and Flavan-3-ols in Storage Roots of Cassava (Manihot esculentaCrantz) during Postharvest Deterioration. Journal of Agricultural and Food Chemistry, 48(11), 5522–5529. doi:10.1021/jf000513p

Chen, J., Leng, H., Duan, Y., Zhao, W., Yang, G., Guo, Y., Hu, Q. (2013). Three new flavonoids from the leaves of oriental tobacco and their cytotoxicity. Phytochemistry Letters, 6(1), 144–147. doi:10.1016/j.phytol.2012.12.001

Cheng, S., van den Bergh, E., Zeng, P., Zhong, X., Xu, J., Liu, X., Hofberger, J., de Bruijn, S., Bhide, A. S., Kuelahoglu, C., Bian, C., Chen, J., Fan, G., Kaufmann, K., Hall, J. C., Becker, A., Bräutigam, A., Weber, A. P., Shi, C., . . . Schranz, M. E. (2013). The Tarenaya hassleriana Genome Provides Insight into Reproductive Trait and Genome Evolution of Crucifers. The Plant Cell, 25(8), 2813–2830. https://doi.org/10.1105/tpc.113.113480

Chou, K. C., & Elrod, D. W. (2003). Prediction of Enzyme Family Classes. Journal of Proteome Research, 2(2), 183–190. https://doi.org/10.1021/pr0255710

Cokelaer, T., Pultz, D., Harder, L. M., Serra-Musach, J., & Saez-Rodriguez, J. (2013). BioServices: a common Python package to access biological Web Services programmatically. Bioinformatics, 29(24), 3241–3242. https://doi.org/10.1093/bioinformatics/btt547

Cuong, D., Kwon, S.-J., Jeon, J., Park, Y., Park, J., & Park, S. (2018). Identification and Characterization of Phenylpropanoid Biosynthetic Genes and Their Accumulation in Bitter Melon (Momordica charantia). Molecules, 23(2), 469. doi:10.3390/molecules23020469

De Pascual-Teresa, S., Santos-Buelga, C., & Rivas-Gonzalo, J. C. (2000). Quantitative Analysis of Flavan-3-ols in Spanish Foodstuffs and Beverages. Journal of Agricultural and Food Chemistry, 48(11), 5331–5337. doi:10.1021/jf000549h

Di Cesare, L. F., Forni, E., Viscardi, D., Nani, R. C. (2003). Changes in the Chemical Composition of Basil Caused by Different Drying Procedures. Journal of Agricultural and Food Chemistry, 51(12), 3575–3581. doi:10.1021/jf021080o

di Sotto, A., Vecchiato, M., Abete, L., Toniolo, C., Giusti, A. M., Mannina, L., Locatelli, M., Nicoletti, M., & di Giacomo, S. (2018). Capsicum annuum L. var. Cornetto di Pontecorvo PDO: Polyphenolic profile and in vitro biological activities. Journal of Functional Foods, 40, 679–691. https://doi.org/10.1016/j.jff.2017.11.041

Dias, M. C., Pinto, D. C. G. A., & Silva, A. M. S. (2021). Plant Flavonoids: Chemical Characteristics and Biological Activity. Molecules, 26(17), 5377. https://doi.org/10.3390/molecules26175377

DONG DONG, Z., CHEN, J., LI, T., CHEN, F., & QUN CUI, D. (2015). Molecular survey of Tamyb10-1 genes and their association with grain colour and germinability in Chinese wheat and Aegilops tauschii. Journal of Genetics, 94(3), 453–459. doi:10.1007/s12041-015-0559-0

Dueñas, M., Hernández, T., Estrella, I., & Fernández, D. (2009). Germination as a process to increase the polyphenol content and antioxidant activity of lupin seeds (Lupinus angustifolius L.). Food Chemistry, 117(4), 599–607. doi:10.1016/j.foodchem.2009.04.051

Duru, I. A., & Duru, C. E. (2019). IDENTIFICATION AND QUANTIFICATION OF PHYTOCHEMICALS FROM Carica papaya Linn (Caricaceae) ROOT EXTRACT USING GC-FID. Journal of Chemical Society of Nigeria, 44(7). Retrieved from http://journals.chemsociety.org.ng/index.php/jcsn/article/view/406

Dvorak, J., Luo, M. C., Yang, Z. L., & Zhang, H. B. (1998). The structure of the Aegilops tauschii genepool and the evolution of hexaploid wheat. Theoretical and Applied Genetics, 97(4), 657-670.

El-Kholy, W. M., Soliman, T. N., & Darwish, A. M. G. (2019). Evaluation of date palm pollen (Phoenix dactylifera L.) encapsulation, impact on the nutritional and functional properties of fortified yoghurt. PLOS ONE, 14(10), e0222789. doi:10.1371/journal.pone.0222789

Gallego, A. M., Rojas, L. F., Rodriguez, H. A., Mora, C., Atehortúa, L., Urrea, A. I., … Pabón-Mora, N. (2019). Metabolomic profile of cacao cell suspensions growing in blue light/dark conditions with potential in food biotechnology. Plant Cell, Tissue and Organ Culture (PCTOC). doi:10.1007/s11240-019-01679-3

Gan, R.-Y., Wang, M.-F., Lui, W.-Y., Wu, K., & Corke, H. (2016). Dynamic changes in phytochemical composition and antioxidant capacity in green and black mung bean (Vigna radiata) sprouts. International Journal of Food Science & Technology, 51(9), 2090–2098. doi:10.1111/ijfs.13185

Ganesan, K., & Xu, B. (2017). Polyphenol-Rich Dry Common Beans (Phaseolus vulgaris L.) and Their Health Benefits. International Journal of Molecular Sciences, 18(11), 2331. doi:10.3390/ijms18112331

Ganesan, K., & Xu, B. (2018). A critical review on phytochemical profile and health promoting effects of mung bean ( Vigna radiata ). Food Science and Human Wellness, 7(1), 11–33. doi:10.1016/j.fshw.2017.11.002

Gao, Q.-H., Wu, C.-S., Wang, M., Xu, B.-N., & Du, L.-J. (2012). Effect of Drying of Jujubes (Ziziphus jujuba Mill.) on the Contents of Sugars, Organic Acids, α-Tocopherol, β-Carotene, and Phenolic Compounds. Journal of Agricultural and Food Chemistry, 60(38), 9642–9648. doi:10.1021/jf3026524

Gardner, R. L., Kerst, A. F., Wilson, D. M., & Payne, M. G. (1967). Beta vulgaris L.: The characterization of three polyphenols isolated from the leaves. Phytochemistry, 6(3), 417–422. doi:10.1016/s0031-9422(00)86299-8

Gasperotti, M., Masuero, D., Mattivi, F., & Vrhovsek, U. (2015). Overall dietary polyphenol intake in a bowl of strawberries: The influence of Fragaria spp. in nutritional studies. Journal of Functional Foods, 18, 1057–1069. doi:10.1016/j.jff.2014.08.013

Georgiev, V. G., Weber, J., Kneschke, E.-M., Denev, P. N., Bley, T., & Pavlov, A. I. (2010). Antioxidant Activity and Phenolic Content of Betalain Extracts from Intact Plants and Hairy Root Cultures of the Red Beetroot Beta vulgaris cv. Detroit Dark Red. Plant Foods for Human Nutrition, 65(2), 105–111. doi:10.1007/s11130-010-0156-6

Ghasemnezhad, M., Sherafati, M., & Payvast, G. A. (2011). Variation in phenolic compounds, ascorbic acid and antioxidant activity of five coloured bell pepper (Capsicum annum) fruits at two different harvest times. Journal of Functional Foods, 3(1), 44–49. doi:10.1016/j.jff.2011.02.002

Guiry, M.D. & Guiry, G.M. (2022). AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. https://www.algaebase.org

Hammouda, H., Chérif, J. K., Trabelsi-Ayadi, M., Baron, A., & Guyot, S. (2013). Detailed Polyphenol and Tannin Composition and Its Variability in Tunisian Dates (Phoenix dactylifera L.) at Different Maturity Stages. Journal of Agricultural and Food Chemistry, 61(13), 3252–3263. doi:10.1021/jf304614j

Hashmi, M. A., Khan, A., Hanif, M., Farooq, U., & Perveen, S. (2015). Traditional Uses, Phytochemistry, and Pharmacology ofOlea europaea(Olive). Evidence-Based Complementary and Alternative Medicine, 2015, 1–29. doi:10.1155/2015/541591

Haytowitz, D.B., Wu, X., Bhagwat, S. (2018). USDA Database for the Flavonoid Content of Selected Foods, Release 3.3. U.S. Department of Agriculture, Agricultural Research Service. Nutrient Data Laboratory Home Page:http://www.ars.usda.gov/nutrientdata/flavv

Hemalatha, P., Bomzan, D. P., Sathyendra Rao, B. V., & Sreerama, Y. N. (2016). Distribution of phenolic antioxidants in whole and milled fractions of quinoa and their inhibitory effects on α-amylase and α-glucosidase activities. Food Chemistry, 199, 330–338. doi:10.1016/j.foodchem.2015.12.025

Ho, L.-H., & Bhat, R. (2015). Exploring the potential nutraceutical values of durian (Durio zibethinus L.) – An exotic tropical fruit. Food Chemistry, 168, 80–89. doi:10.1016/j.foodchem.2014.07.020

Hostetler, G. L., Ralston, R. A., & Schwartz, S. J. (2017). Flavones: Food Sources, Bioavailability, Metabolism, and Bioactivity. Advances in Nutrition, 8(3), 432–435. https://doi.org/10.3945/an.116.012948

Hou, Z., Luo, J., & Kong, L. (2009). Medium-Pressure Liquid Chromatography Coupled to Electrospray Ionization Mass Spectrometry for Separation and On-Line Characterization of Flavonoids from Asparagus officinalis. Chromatographia, 70(9-10), 1447–1450. doi:10.1365/s10337-009-1329-z

Hu, Q. N., Zhu, H., Li, X., Zhang, M., Deng, Z., Yang, X., & Deng, Z. (2012). Assignment of EC Numbers to Enzymatic Reactions with Reaction Difference Fingerprints. PLoS ONE, 7(12). https://doi.org/10.1371/journal.pone.0052901

Huang, C. F., Chen, Y. W., Yang, C. Y., Lin, H. Y., Way, T. D., Chiang, W., & Liu, S. H. (2011). Extract of Lotus Leaf (Nelumbo nucifera) and Its Active Constituent Catechin with Insulin Secretagogue Activity. Journal of Agricultural and Food Chemistry, 59(4), 1087–1094. doi:10.1021/jf103382h

Igbinosa, O. O., Igbinosa, I. H., Chigor, V. N., Uzunuigbe, O. E., Oyedemi, S. O., Odjadjare, E. E., Okoh, A. I., & Igbinosa, E. O. (2011). Polyphenolic Contents and Antioxidant Potential of Stem Bark Extracts from Jatropha curcas (Linn). International Journal of Molecular Sciences, 12(5), 2958–2971. https://doi.org/10.3390/ijms12052958

ITIS. (2021). Integrated Taxonomic Information System. Integrated Taxonomic Information System. https://www.itis.gov/

Jaffri, J. M., Mohamed, S., Ahmad, I. N., Mustapha, N. M., Manap, Y. A., & Rohimi, N. (2011). Effects of catechin-rich oil palm leaf extract on normal and hypertensive rats’ kidney and liver. Food Chemistry, 128(2), 433–441. doi:10.1016/j.foodchem.2011.03.050

Kanehisa Laboratories. (1995). KEGG: Kyoto Encyclopedia of Genes and Genomes. KEGG: Kyoto Encyclopedia of Genes and Genomes. https://www.genome.jp/kegg/

Kanehisa Laboratories. (2019). KEGG PATHWAY: Flavonoid biosynthesis - Reference pathway. Kyoto Encyclopedia of Genes and Genomes. https://www.genome.jp/kegg-bin/show\_pathway?map00941

Kanehisa Laboratories. (2021). KEGG PATHWAY: Isoflavonoid biosynthesis - Reference pathway. Kyoto Encyclopedia of Genes and Genomes. https://www.genome.jp/pathway/map00943

Kanehisa Labs. KEGG REACTION: R07994. Kyoto Encyclopedia of Genes and Genomes. https://www.genome.jp/entry/R07994

Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. Protein Science, 28(11), 1947–1951. https://doi.org/10.1002/pro.3715

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 28(1), 27–30. https://doi.org/10.1093/nar/28.1.27

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2020). KEGG: integrating viruses and cellular organisms. Nucleic Acids Research, 49(D1), D545–D551. https://doi.org/10.1093/nar/gkaa970

Khallouki, F., Ricarte, I., Breuer, A., & Owen, R. W. (2018). Characterization of phenolic compounds in mature Moroccan Medjool date palm fruits ( Phoenix dactylifera ) by HPLC-DAD-ESI-MS. Journal of Food Composition and Analysis, 70, 63–71. doi:10.1016/j.jfca.2018.03.005

Kim, M. Y., Lee, S. H., Jang, G. Y., Li, M., Lee, Y. R., Lee, J., & Jeong, H. S. (2017). Changes of phenolic-acids and vitamin E profiles on germinated rough rice ( Oryza sativa L.) treated by high hydrostatic pressure. Food Chemistry, 217, 106–111. doi:10.1016/j.foodchem.2016.08.069

Kim, Y. J., Kim, Y. B., Li, X., Choi, S. R., Park, S., Park, J. S., Lim, Y. P., & Park, S. U. (2015). Accumulation of Phenylpropanoids by White, Blue, and Red Light Irradiation and Their Organ-Specific Distribution in Chinese Cabbage (Brassica rapa ssp. pekinensis). Journal of Agricultural and Food Chemistry, 63(30), 6772–6778. https://doi.org/10.1021/acs.jafc.5b02086

Klejdus, B., Lojková, L., Plaza, M., Šnóblová, M., & Štěrbová, D. (2010). Hyphenated technique for the extraction and determination of isoflavones in algae: Ultrasound-assisted supercritical fluid extraction followed by fast chromatography with tandem mass spectrometry. Journal of Chromatography A, 1217(51), 7956–7965. doi:10.1016/j.chroma.2010.07.020

Koopman, F., Beekwilder, J., Crimi, B., van Houwelingen, A., Hall, R. D., Bosch, D., van Maris, A. J., Pronk, J. T., & Daran, J. M. (2012). De novo production of the flavonoid naringenin in engineered Saccharomyces cerevisiae. Microbial Cell Factories, 11(1). https://doi.org/10.1186/1475-2859-11-155

Koprivica, M. R., Trifković, J. Đ., Dramićanin, A. M., Gašić, U. M., Akšić, M. M. F., & Milojković-Opsenica, D. M. (2018). Determination of the phenolic profile of peach (Prunus persica L.) kernels using UHPLC–LTQ OrbiTrap MS/MS technique. European Food Research and Technology. doi:10.1007/s00217-018-3116-2

Kostecka-Gugała, A., Kruczek, M., Ledwożyw-Smoleń, I., & Kaszycki, P. (2020). Antioxidants and Health-Beneficial Nutrients in Fruits of Eighteen Cucurbita Cultivars: Analysis of Diversity and Dietary Implications. Molecules, 25(8), 1792. doi:10.3390/molecules25081792

Lalani, S., & Poh, C. L. (2020). Flavonoids as Antiviral Agents for Enterovirus A71 (EV-A71). Viruses, 12(2), 184. https://doi.org/10.3390/v12020184

Lassesen Consulting, LLC. MicrobiomePrescription : Flavonoids. Microbiome Analysis for Patients. https://www.microbiomeprescription.com/Library/Flavonoids

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics, 20(12), 1983–1992. https://doi.org/10.1109/tvcg.2014.2346248

LibreTexts. (2021). 17.1C: Primary and secondary metabolites. Biology LibreTexts. Retrieved from https://bio.libretexts.org/@go/page/12413

Liew, S. S., Ho, W. Y., Yeap, S. K., & Sharifudin, S. A. B. (2018). Phytochemical composition and in vitro antioxidant activities of Citrus sinensis peel extracts. PeerJ, 6, e5331. doi:10.7717/peerj.5331

Liu, Y., Feng, S., Song, L., He, G., Chen, M., & Huang, D. (2013). Secondary Metabolites in Durian Seeds: Oligomeric Proanthocyanidins. Molecules, 18(11), 14172–14185. doi:10.3390/molecules181114172

Loub W. D., Farnsworth N. R., Soejarto D. D., Quinn M. L. (1985) NAPRALERT: computer handling of natural product research data. J Chem Inf Model 25:99–103. https://doi.org/10.1021/ci00046a009

Luo, X., Cui, J., Zhang, H., & Duan, Y. (2018). Subcritical water extraction of polyphenolic compounds from sorghum ( Sorghum bicolor L.) bran and their biological activities. Food Chemistry, 262, 14–20. doi:10.1016/j.foodchem.2018.04.073

Malenčić, D., Popović, M., & Miladinović, J. (2007). Phenolic Content and Antioxidant Properties of Soybean (Glycine max (L.) Merr.) Seeds. Molecules, 12(3), 576–581. doi:10.3390/12030576

Mallek-Ayadi, S., Bahloul, N., & Kechaou, N. (2017). Characterization, phenolic compounds and functional properties of Cucumis melo L. peels. Food Chemistry, 221, 1691–1697. doi:10.1016/j.foodchem.2016.10.117

McDonald, A.G., Boyce, S. & Tipton, K.F. (2009). ExplorEnz: the primary source of the IUBMB enzyme list. Nucleic Acids Res. 37, D593–D597. DOI: 10.1093/nar/gkn582

Mohanraj, K., Karthikeyan, B. S., Vivek-Ananth, R.P., Chand, R. P. B., Aparna, S. R., Mangalapandi, P., and Samal, A. (2018). IMPPAT: A curated database of Indian Medicinal Plants, Phytochemistry And Therapeutics. Scientific Reports 8:4329.

NAIST Comparative Genomics Laboratory. (2021). KNApSAcK: A Comprehensive Species-Metabolite Relationship Database. KNApSAcK Family. http://www.knapsackfamily.com/KNApSAcK/

Napoleão, T. A., Soares, G., Vital, C. E., Bastos, C., Castro, R., Loureiro, M. E., & Giordano, A. (2017). Methyl jasmonate and salicylic acid are able to modify cell wall but only salicylic acid alters biomass digestibility in the model grass Brachypodium distachyon. Plant Science, 263, 46–54. doi:10.1016/j.plantsci.2017.06.014

Nara Institute of Science and Technology. KNApSAcK Metabolite Information - C00006941. KNApSAcK. http://www.knapsackfamily.com/knapsack\_core/information.php?word=C00006941

National Center for Biotechnology Information (2021). PubChem Compound Summary for CID 5280441, Vitexin. Retrieved from https://pubchem.ncbi.nlm.nih.gov/compound/Vitexin.

National Center for Biotechnology Information (2021). PubChem Compound Summary for CID 996, Phenol. Retrieved December 8, 2021 from https://pubchem.ncbi.nlm.nih.gov/compound/Phenol.

National Human Genome Research Institute. (2020). Biological Pathways Fact Sheet. Genome.Gov. https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet

National Human Genome Research Institute. Deoxyribonucleic Acid (DNA). Genome.Gov. https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid

National Human Genome Research Institute. Enzyme. Genome.Gov. https://www.genome.gov/genetics-glossary/Enzyme

National Human Genome Research Institute. Protein. Genome.Gov. https://www.genome.gov/genetics-glossary/Protein

National Library of Medicine. (2021). What is a gene?: MedlinePlus Genetics. MedlinePlus. https://medlineplus.gov/genetics/understanding/basics/gene/

NEUROtiker. (2007). Structure of Phenol [Illustration]. Wikipedia Media Viewer. https://en.wikipedia.org/wiki/Phenol#/media/File:Phenol2.svg

Nix, A., Paull, C. A., & Colgrave, M. (2015). The flavonoid profile of pigeonpea, Cajanus cajan: a review. SpringerPlus, 4(1). doi:10.1186/s40064-015-0906-x

Nix, A., Paull, C., & Colgrave, M. (2017). Flavonoid Profile of the Cotton Plant, Gossypium hirsutum: A Review. Plants, 6(4), 43. doi:10.3390/plants6040043

Oboh, G., Olabiyi, A. A. & Akinyemi, A. J. (2013). Inhibitory effect of aqueous extract of different parts of unripe pawpaw (Carica papaya) fruit on Fe2+-induced oxidative stress in rat pancreasin vitro. Pharmaceutical Biology, 51(9), 1165–1174. doi:10.3109/13880209.2013.782321

Panche, A. N., Diwan, A. D., & Chandra, S. R. (2016). Flavonoids: an overview. Journal of Nutritional Science, 5(e47). https://doi.org/10.1017/jns.2016.41

Pei, D., Liu, D., Liu, J.-X., & Di, D.-L. (2012). Chemical investigation of the pollen of Brassica napus. Chemistry of Natural Compounds, 48(2), 310–312. doi:10.1007/s10600-012-0232-2

Perestrelo, R., Lu, Y., Santos, S. A. O., Silvestre, A. J. D., Neto, C. P., Câmara, J. S., & Rocha, S. M. (2012). Phenolic profile of Sercial and Tinta Negra Vitis vinifera L. grape skins by HPLC–DAD–ESI-MSn. Food Chemistry, 135(1), 94–104. doi:10.1016/j.foodchem.2012.04.102

Perkin, A. G., & Hummel, J. J. (1904). CXLVII.—The colouring principle of the flowers of the Butea frondosa. J. Chem. Soc., Trans., 85(0), 1459–1472. doi:10.1039/ct9048501459

Petkovska, A., Gjamovski, V., Stanoeva, J. P., Stefova, M. (2017). Characterization of the Polyphenolic Profiles of Peel, Flesh and Leaves of Malus domestica Cultivars Using UHPLC-DAD-HESI-MS n. Natural Product Communications, 12(1), 1934578X1701200–. doi:10.1177/1934578x1701200111

Preuß, A., Stracke, R., Weisshaar, B., Hillebrecht, A., Matern, U., & Martens, S. (2009). Arabidopsis thaliana expresses a second functional flavonol synthase. FEBS Letters, 583(12), 1981–1986. doi:10.1016/j.febslet.2009.05.006

Quintero-Soto, M. F., Saracho-Peña, A. G., Chavez-Ontiveros, J., Garzon-Tiznado, J. A., Pineda-Hidalgo, K. V., Delgado-Vargas, F., & Lopez-Valenzuela, J. A. (2018). Phenolic profiles and their contribution to the antioxidant activity of selected chickpea genotypes from Mexico and ICRISAT collections. Plant Foods for Human Nutrition, 73(2), 122–129. doi:10.1007/s11130-018-0661-6

Rensing, S. A., Goffinet, B., Meyberg, R., Wu, S. Z., & Bezanilla, M. (2020). The Moss Physcomitrium (Physcomitrella) patens: A Model Organism for Non-Seed Plants. The Plant Cell, 32(5), 1361–1376. https://doi.org/10.1105/tpc.19.00828

Repo-Carrasco-Valencia, R., Hellström, J. K., Pihlava, J.-M., & Mattila, P. H. (2010). Flavonoids and other phenolic compounds in Andean indigenous grains: Quinoa (Chenopodium quinoa), kañiwa (Chenopodium pallidicaule) and kiwicha (Amaranthus caudatus). Food Chemistry, 120(1), 128–133. doi:10.1016/j.foodchem.2009.09.087

Reusch, W. (2013). Stereoisomers. Chemistry - MSU. https://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/sterisom.htm

Robertson, S. B. (2021). What are Flavonoids? News-Medical.Net. https://www.news-medical.net/health/What-are-Flavonoids.aspx

San, B., & Yildirim, A. N. (2010). Phenolic, alpha-tocopherol, beta-carotene and fatty acid composition of four promising jujube (Ziziphus jujuba Miller) selections. Journal of Food Composition and Analysis, 23(7), 706–710. doi:10.1016/j.jfca.2010.02.008

Santos, S. A. O., Vilela, C., Freire, C. S. R., Neto, C. P., & Silvestre, A. J. D. (2013). Ultra-high performance liquid chromatography coupled to mass spectrometry applied to the identification of valuable phenolic compounds from Eucalyptus wood. Journal of Chromatography B, 938, 65–74. doi:10.1016/j.jchromb.2013.08.034

Santos, S. A. O., Villaverde, J. J., Freire, C. S. R., Domingues, M. R. M., Neto, C. P., & Silvestre, A. J. D. (2012). Phenolic composition and antioxidant activity of Eucalyptus grandis, E. urograndis (E. grandis×E. urophylla) and E. maidenii bark extracts. Industrial Crops and Products, 39, 120–127. doi:10.1016/j.indcrop.2012.02.003

Santos, S. A., Vilela, C., Domingues, R. M., Oliveira, C. S., Villaverde, J. J., Freire, C. S., Neto, C. P., & Silvestre, A. J. (2017). Secondary metabolites from Eucalyptus grandis wood cultivated in Portugal, Brazil and South Africa. Industrial Crops and Products, 95, 357–364. https://doi.org/10.1016/j.indcrop.2016.10.044

Sanz, M., Cadahía, E., Esteruelas, E., Muñoz, A. M., Fernández De Simón, B., Hernández, T., & Estrella, I. (2010). Phenolic Compounds in Cherry (Prunus avium) Heartwood with a View to Their Use in Cooperage. Journal of Agricultural and Food Chemistry, 58(8), 4907–4914. doi:10.1021/jf100236v

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O’Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database, 2020. https://doi.org/10.1093/database/baaa062

Shimoda, K., Otsuka, T., Morimoto, Y., Hamada, H., & Hamada, H. (2007). Glycosylation and Malonylation of Quercetin, Epicatechin, and Catechin by Cultured Plant Cells. Chemistry Letters, 36(10), 1292–1293. doi:10.1246/cl.2007.1292

Silva, L. G. S., Morelli, A. P., Pavan, I. C. B., Tavares, M. R., Pestana, N. F., Rostagno, M. A., Simabuco, F. M., Bezerra, R. M. N. (2020). Protective effects of beet ( Beta vulgaris ) leaves extract against oxidative stress in endothelial cells in vitro. Phytotherapy Research. doi:10.1002/ptr.6612

Singh, S. (2016). Enhancing phytochemical levels, enzymatic and antioxidant activity of spinach leaves by chitosan treatment and an insight into the metabolic pathway using DART-MS technique. Food Chemistry, 199, 176–184. doi:10.1016/j.foodchem.2015.11.127

Slotte, T., Hazzouri, K. M., ÅGren, J. A., Koenig, D., Maumus, F., Guo, Y. L., Steige, K., Platts, A. E., Escobar, J. S., Newman, L. K., Wang, W., Mandáková, T., Vello, E., Smith, L. M., Henz, S. R., Steffen, J., Takuno, S., Brandvain, Y., Coop, G., . . . Wright, S. I. (2013). The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nature Genetics, 45(7), 831–835. https://doi.org/10.1038/ng.2669

Solar, A., Colarič, M., Usenik, V., & Stampar, F. (2006). Seasonal variations of selected flavonoids, phenolic acids and quinones in annual shoots of common walnut (Juglans regia L.). Plant Science, 170(3), 453–461. doi:10.1016/j.plantsci.2005.09.012

Svensson, L., Sekwati-Monang, B., Lutz, D. L., Schieber, A., & Gänzle, M. G. (2010). Phenolic Acids and Flavonoids in Nonfermented and Fermented Red Sorghum (Sorghum bicolor (L.) Moench). Journal of Agricultural and Food Chemistry, 58(16), 9214–9220. doi:10.1021/jf101504v

Tamasi, G., Baratto, M. C., Bonechi, C., Byelyakova, A., Pardini, A., Donati, A., Leone, G., Consumi, M., Lamponi, S., Magnani, A., & Rossi, C. (2019). Chemical characterization and antioxidant properties of products and by‐products from Olea europaea L. Food Science & Nutrition, 7(9), 2907–2920. https://doi.org/10.1002/fsn3.1142

Tang, D., Dong, Y., Ren, H., Li, L., & He, C. (2014). A review of phytochemistry, metabolite changes, and medicinal uses of the common food mung bean and its sprouts (Vigna radiata). Chemistry Central Journal, 8(1), 4. doi:10.1186/1752-153x-8-4

Tang, X., Tang, P., & Liu, L. (2017). Molecular Structure–Affinity Relationship of Flavonoids in Lotus Leaf (Nelumbo nucifera Gaertn.) on Binding to Human Serum Albumin and Bovine Serum Albumin by Spectroscopic Method. Molecules, 22(7), 1036. doi:10.3390/molecules22071036

Terpinc, P., Polak, T., Makuc, D., Ulrih, N. P., & Abramovič, H. (2012). The occurrence and characterisation of phenolic compounds in Camelina sativa seed, cake and oil. Food Chemistry, 131(2), 580–589. doi:10.1016/j.foodchem.2011.09.033

The Python Software Foundation. The Python Standard Library — Python 3.10.1 documentation. Python. https://docs.python.org/3/library/index.html

Thiruvengadam, M., & Chung, I.-M. (2015). Selenium, putrescine, and cadmium influence health-promoting phytochemicals and molecular-level effects on turnip (Brassica rapa ssp. rapa). Food Chemistry, 173, 185–193. doi:10.1016/j.foodchem.2014.10.012

Tipton, K. & McDonald, A. A Brief Guide to Enzyme Nomenclature and Classification. Joint Commission on Biochemical Nomenclature Newsletter. Retrieved from https://iubmb.org/wp-content/uploads/sites/10116/2018/11/A-Brief-Guide-to-Enzyme-Classification-and-Nomenclature-rev.pdf

Tsanova-Savova, S., Ribarova, F., & Gerova, M. (2005). (+)-Catechin and (−)-epicatechin in Bulgarian fruits. Journal of Food Composition and Analysis, 18(7), 691–698. doi:10.1016/j.jfca.2004.06.008

Tuladhar, P., Sasidharan, S., & Saudagar, P. (2021). Role of phenols and polyphenols in plant defense response to biotic and abiotic stresses. In Biocontrol Agents and Secondary Metabolites: Applications and Immunization for Plant Growth and Protection (1st ed., pp. 419–441). Woodhead Publishing. https://doi.org/10.1016/C2019-0-03577-7

U.S. Department of Agriculture, Agricultural Research Service. (1992-2016). Dr. Duke's Phytochemical and Ethnobotanical Databases. Home Page, http://phytochem.nal.usda.gov/ http://dx.doi.org/10.15482/USDA.ADC/1239279

Valiñas, M. A., Lanteri, M. L., ten Have, A., & Andreu, A. B. (2017). Chlorogenic acid, anthocyanin and flavan-3-ol biosynthesis in flesh and skin of Andean potato tubers ( Solanum tuberosum subsp. andigena ). Food Chemistry, 229, 837–846. doi:10.1016/j.foodchem.2017.02.150

Velankar, S., Dana, J.M., Jacobsen, J., van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C., Martin, M.J., Kleywegt, G.J. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res. 2013 Jan;41(Database issue):D483-9. doi: 10.1093/nar/gks1258. Epub 2012 Nov 29. PMID: 23203869; PMCID: PMC3531078.

Vissers, A., Kiskini, A., Hilgers, R., Marinea, M., Wierenga, P. A., Gruppen, H., & Vincken, J. P. (2017). Enzymatic Browning in Sugar Beet Leaves (Beta vulgaris L.): Influence of Caffeic Acid Derivatives, Oxidative Coupling, and Coupled Oxidation. Journal of Agricultural and Food Chemistry, 65(24), 4911–4920. https://doi.org/10.1021/acs.jafc.7b01897

Wallheimer, B. (2011). Selaginella genome adds piece to plant evolutionary puzzle. Purdue University News Service. https://www.purdue.edu/newsroom/research/2011/110505BanksSelaginella.html

Wang, L., Jiang, Y., Yuan, L., Lu, W., Yang, L., Karim, A., & Luo, K. (2013). Isolation and Characterization of cDNAs Encoding Leucoanthocyanidin Reductase and Anthocyanidin Reductase from Populus trichocarpa. PLoS ONE, 8(5), e64664. doi:10.1371/journal.pone.0064664

Willstätter, R., & Everest, A. E. (1913). Untersuchungen über die Anthocyane. I. Über den Farbstoff der Kornblume. Justus Liebigs Annalen der Chemie, 401(2), 189-232.

Wilson, Jordan R. “Computational Analysis of Flavonoid Pathways from Various Plant Species.” Kettering University.

Wu, S., Wilson, A. E., Chang, L., & Tian, L. (2019). Exploring the Phytochemical Landscape of the Early-Diverging Flowering Plant Amborella trichopoda Baill. Molecules, 24(21), 3814. doi:10.3390/molecules24213814

Xie, Y., Yang, W., Tang, F., Chen, X., & Ren, L. (2014). Antibacterial Activities of Flavonoids: Structure-Activity Relationship and Mechanism. Current Medicinal Chemistry, 22(1), 132–149. https://doi.org/10.2174/0929867321666140916113443

Xu, X., Tian, H., He, M., Gebretsadik, K., Qi, X., Xu, Q., & Chen, X. (2019). Changes in catechin contents and expression of catechin biosynthesis-associated genes during early cucumber fruit development. Acta Physiologiae Plantarum, 41(8). doi:10.1007/s11738-019-2925-7

Yan, M., Chen, M., Zhou, F., Cai, D., Bai, H., Wang, P., Lei, H., & Ma, Q. (2019). Separation and analysis of flavonoid chemical constituents in flowers of Juglans regia L. by ultra-high-performance liquid chromatography-hybrid quadrupole time-of-flight mass spectrometry. Journal of Pharmaceutical and Biomedical Analysis, 164, 734–741. https://doi.org/10.1016/j.jpba.2018.11.029

Yıldırım, F., Yıldırım, A. N., San, B., & Ercişli, S. (2015). The Relationship Between Growth Vigour of Rootstock and Phenolic Contents in Apple (Malus × domestica). Erwerbs-Obstbau, 58(1), 25–29. doi:10.1007/s10341-015-0253-7

Yobi, A., Wone, B. W. M., Xu, W., Alexander, D. C., Guo, L., Ryals, J. A., Oliver, M. J., Cushman, J. C. (2012). Comparative metabolic profiling between desiccation-sensitive and desiccation-tolerant species ofSelaginellareveals insights into the resurrection trait. The Plant Journal, 72(6), 983–999. doi:10.1111/tpj.12008

Zahir, A. A., Rahuman, A. A., Bagavan, A., Geetha, K., Kamaraj, C., & Elango, G. (2011). Evaluation of medicinal plant extracts and isolated compound epicatechin from Ricinus communis against Paramphistomum cervi. Parasitology Research, 111(4), 1629–1635. doi:10.1007/s00436-011-2589-8

Zeb, A., Muhammad, B., & Ullah, F. (2017). Characterization of sesame (Sesamum indicum L.) seed oil from Pakistan for phenolic composition, quality characteristics and potential beneficial properties. Journal of Food Measurement and Characterization, 11(3), 1362–1369. doi:10.1007/s11694-017-9514-5

Zeng, X., Zhang, P., He, W., Qin, C., Chen, S., Tao, L., Wang, Y., et al. (2018). NPASS: natural product activity and species source database for natural product research, discovery and tool developmentNucleic Acids Research, 46(D1):D1217-D1222. PMID:29106619

Zhai, R., Liu, X. T., Feng, W. T., Chen, S. S., Xu, L. F., Wang, Z. G., Zhang, J. L., Li, P. M., & Ma, F. W. (2014). Different Biosynthesis Patterns among Flavonoid 3-glycosides with Distinct Effects on Accumulation of Other Flavonoid Metabolites in Pears (Pyrus bretschneideri Rehd.). PLoS ONE, 9(3), e91945. https://doi.org/10.1371/journal.pone.0091945

Zhang, P. & He, Y. Translocation Reactions (TRs) Family Classification. Database for Polyphenol Utilization Proteins from Microbiome. https://bcb.unl.edu/dbpup/classes/TRs

Zhang, X., Lin, Z., Fang, J., Liu, M., Niu, Y., Chen, S., & Wang, H. (2015). An on-line high-performance liquid chromatography–diode-array detector–electrospray ionization–ion-trap–time-of-flight–mass spectrometry–total antioxidant capacity detection system applying two antioxidant methods for activity evaluation of the edible flowers from Prunus mume. Journal of Chromatography A, 1414, 88–102. doi:10.1016/j.chroma.2015.08.033

Zhou, J., Ma, Y., Jia, Y., Pang, M., Cheng, G., & Cai, S. (2019). Phenolic profiles, antioxidant activities and cytoprotective effects of different phenolic fractions from oil palm (Elaeis guineensis Jacq.) fruits treated by ultra-high pressure. Food Chemistry, 288, 68–77. doi:10.1016/j.foodchem.2019.03.002

# Appendix

## Similar Compounds

### Isoflavonoids

###### Table 70

Similar and derivative compounds of genistein synthesized by plants in the data set.

| **Compound Name(s)** | | **Species** | **Source(s)** |
| --- | --- | --- | --- |
| 1. | 2'-Hydroxygenistein | Cajanus cajan  Lupinus angustifolius  Phaseolus vulgaris  Vigna angularis  Vigna radiata  Vigna unguiculata | K, T  K, N  K, N, T  K, N  T, N  T |
| 2. | 4'-Methylgenistein  Biochanin A | Cicer arietinum  Glycine max  Phaseolus vulgaris | K, P  K  2 |
| 3. | Genistein 7-O-β-D-glucoside  Genistin | Glycine max  Lupinus angustifolius  Medicago truncatula  Raphanus sativus | K, P  1  K  K |
| 4. | Genistein 7-O-β-D-glucoside 6''-O-malonate | Glycine max  Medicago truncatula | K  K |
| 5. | Genistein 7-O-glucosyl-glucoside | Lupinus angustifolius | D |
| 6. | Genistein apiofuranosyl diglycoside | Lupinus angustifolius | 1 |
| 7. | Genistein apiofuranosyl glycoside | Lupinus angustifolius | 1 |
| 8. | Genistin 6''-O-acetate | Glycine max | K |
| 9. | Genistin 6''-O-malonate  Malonylgenistin | Glycine max | N |
| 10. | Hidroxygenistein 7-glucoside | Lupinus angustifolius | 1 |
| 11. | Hidroxygenistein 7-acetylglucoside | Lupinus angustifolius | 1 |
| 12. | Hydroxygenistein | Vigna angularis  Vigna radiata | D  D |
| 13. | Isogenistein 7-O-glucoside | Cajanus cajan | D |
| 14. | Methylgenistein | Glycine max | D |

Sources: (D) Dr. Duke; (K) KNApSAcK; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (1) Dueñas et al., 2009; (2) Ganesan & Xu, 2017

### Anthocyanidins

###### Table 71

Similar and derivative compounds of cyanidin synthesized by plants in the data set.

| **Compound Name(s)** | | **Species** | **Source(s)** |
| --- | --- | --- | --- |
| **1.** | Cyanidin 3-(3'',6''-dimalonylglucoside) | Zea mays | K, N |
| **2.** | Cyanidin 3-(4''-acetylglucoside) | Citrus sinensis | K |
| **3.** | Cyanidin 3-(6''-acetylglucoside) | Vitis vinifera | K |
| **4.** | Cyanidin 3-(6''-caffeyl-6'''-feruly-lsophoroside)-5-glucoside | Raphanus sativus | K |
| **5.** | Cyanidin 3-(6''-malonylglucoside) | Citrus sinensis  Lactuca sativa  Zea mays | K  K  K, N |
| **6.** | Cyanidin 3-(6'',6'''-disinapylsophoroside)-5-glucoside | Brassica oleracea | K |
| **7.** | Cyanidin 3-(sinapoyl-xylosyl-glucosyl)-galactoside | Daucus carota | N |
| **8.** | Cyanidin 3-[6-(3-glucosylcaffeyl)glucoside]-5-glucoside | Ipomoea nil | K |
| **9.** | Cyanidin 3-[6''-(4-glucosyl-p-coumaryl)sophoroside]-5-glucoside | Brassica oleracea | K |
| **10.** | Cyanidin 3-[6''-(4-glucosylcaffeyl)sophoroside]-5-glucoside | Ipomoea nil | K |
| **11.** | Cyanidin 3-acetyl-glucoside  Cyanidin 3-O-acetyl-glucoside | Vitis vinifera | N |
| **12.** | Cyanidin 3-arabinoside | Malus domestica  Theobroma cacao | D  K |
| **13.** | Cyanidin 3-coyumaroyl-glucoside | Vitis vinifera | N |
| **14.** | Cyanidin 3-dimalonyl-glucoside | Zea mays | D, N |
| **15.** | Cyanidin 3-galactoside  Cyanidin 3-O-galactoside  Cyanidin 3-O-β-D-galactopyranoside  Idaein | Daucus carota  Fragaria vesca  Glycine max  Theobroma cacao  Vitis vinifera  Zea mays | K, N  2  P  D, K  D  D, N |
| **16.** | Cyanidin 3-galactoside p-coumaric acid ester | Zea mays | D |
| **17.** | Cyanidin 3-glucogalactoside | Daucus carota | N |
| **18.** | Cyanidin 3-glucoside  Cyanidin 3-O-glucoside  Cyanidin 3-O-β-D-glucoside  Cyanidin 3-monoglucoside  Chrysanthemin  Kuromanin | Amborella trichopoda  Asparagus officinalis  Citrus sinensis  Fragaria vesca  Olea europaea  Phaseolus vulgaris  Vigna radiata  Vitis vinifera | 3  D  K  2  D  1  D  D, P |
| **19.** | Cyanidin 3-glucosyl-rutinoside  Cyanidin 3-O-β-D-glucosyl-rutinoside | Olea europaea  Prunus avium | N  N |
| **20.** | Cyanidin 3-lathyroside | Daucus carota | K |
| **21.** | Cyanidin 3-malonyl-glucoside | Helianthus annuus | N |
| **22.** | Cyanidin 3-malonyl-xyloside | Helianthus annuus | N |
| **23.** | Cyanidin 3-O-(6-O-para-coumaroyl)-glucoside | Vitis vinifera | N |
| **24.** | Cyanidin 3-O-[2-O-(2-O-(sinapoyl)-bata-D-  xylopyranosyl) 6-O-(4-O-(β-D-glucopyranosyl)-p-  coumaroyl-bata-D-glucopyranoside] 5-O-[6-O-  (malonyl) bata-D-glucopyranoside] | Arabidopsis thaliana | K |
| **25.** | Cyanidin 3-O-[2''-O-(2'''-O-(sinapoyl) xylosyl) 6''-O-  (p-coumaroyl) glucoside] 5-O-glucoside | Arabidopsis thaliana | K |
| **26.** | Cyanidin 3-O-[2''-O-(2'''-O-(sinapoyl) xylosyl) 6''-O-  (p-O-(glucosyl) p-coumaroyl) glucoside] 5-O-glucoside | Arabidopsis thaliana | K |
| **27.** | Cyanidin 3-O-[2''-O-(xylosyl) 6''-O-(p-O-(glucosyl) p-  coumaroyl) glucoside] 5-O-[6'''-O-(malonyl) glucoside] | Arabidopsis thaliana | K |
| **28.** | Cyanidin 3-O-[2''-O-(xylosyl) glucoside] 5-O-  (6'''-O-malonyl) glucoside | Arabidopsis thaliana | K |
| **29.** | Cyanidin 3-O-[2''-O-(xylosyl)-6''-O-(p-O-(glucosyl)-p-  coumaroyl) glucoside] 5-O-glucoside | Arabidopsis thaliana | K |
| **30.** | Cyanidin 3-O-[6-O-(malonyl)-β-D-glucopyranoside]-7,3'-  di-O-[6-O-(sinapyl)-β-D-glucopyranoside] | Phalaenopsis equestris | K |
| **31.** | Cyanidin 3-O-[β-D-glucopyranoside]-7,3'-di-O-  [6-O-(sinapyl)-β-D-glucopyranoside] | Phalaenopsis equestris | K |
| **32.** | Cyanidin 3-rhamnosyl-glucosyl-glucoside | Asparagus officinalis | N |
| **33.** | Cyanidin 3-rhamnosyl-glucoside | Asparagus officinalis | D |
| **34.** | Cyanidin 3-rutinoside  Cyanidin 3-O-β-D-rutinoside | Olea europaea  Prunus avium | D, N  N |
| **35.** | Cyanidin 3-sambubioside-5-glucoside | Arabidopsis thaliana | K |
| **36.** | Cyanidin 3-sophoroside-5-glucoside  Cyanidin 3-O-sophoroside-5-O-glucoside  Cyanidin 3-sophorobiose-5-O-β-D-glucoside | Brassica oleracea  Raphanus sativus | K, N  N |
| **37.** | Cyanidin 3-xyloside | Helianthus annuus | K, N |
| **38.** | Cyanidin 3,5-digalactoside | Daucus carota | N |
| **39.** | Cyanidin 3,5-diglucoside  Cyanidin 3,5-O-diglucoside  Cyanin | Asparagus officinalis  Brassica oleracea  Fragaria vesca  Malus domestica | D  N  P  D |
| **40.** | Cyanidin 7-arabinoside | Malus domestica | D |
| **41.** | Cyanidin chloride | Glycine max | P |
| **42.** | Leucocyanidin | Amborella trichopoda  Arabidopsis thaliana  Brassica napus  Capsicum annuum  Gossypium hirsutum  Medicago truncatula  Momordica charantia  Nelumbo nucifera  Phaseolus vulgaris  Prunus avium  Ziziphus jujuba | 3  N  K  K  N  K  K  K  K  K  N |

Sources: (D) Dr. Duke; (K) KNApSAcK; (N) NAPRALERT; (P) NPASS; (1) Ganesan & Xu, 2017; (2) Gasperotti et al., 2015; (3) Wu et al., 2019

### Chalcones

###### Table 72

Similar and derivative compounds of isoliquiritigenin synthesized by plants in the data set.

| **Compound Name** | | **Species** | **Source(s)** |
| --- | --- | --- | --- |
| 1. | Isoliquiritin | Cicer arietinum | P |
| 2. | Liquirtigenin | Glycine soja | P |
| 3. | Neoisoliquiritigenin | Glycine max | K |

Sources: (K) KNApSAcK; (P) NPASS

### Flavan-3-ols

###### Table 73

Similar and derivative compounds of catechin synthesized by plants in the data set.

| **Compound Name(s)** | | **Species** | **Source(s)** |
| --- | --- | --- | --- |
| **1.** | (+)-Gallocatechin-(4α->8)-(+)-catechin | Nelumbo nucifera | N |
| **2.** | Catechin (4α->6)-catechin  Procyanidin B6 | Nelumbo nucifera  Vitis vinifera | P  N, P |
| **3.** | Catechin (4α->6)-epicatechin  Procyanidin B8 | Nelumbo nucifera  Vitis vinifera | P  N, P |
| **4.** | Catechin (4α->8)-catechin  Procyanidin B3 | Fragaria vesca  Nelumbo nucifera  Vitis vinifera | 1  P  N, P |
| **5.** | Catechin (4α->8)-catechin-(4α->8)-catechin  Procyanidin C2 | Vitis vinifera | N |
| **6.** | Catechin (4α->8)-epicatechin  Procyanidin B4 | Nelumbo nucifera  Theobroma cacao  Vitis vinifera | P  N  P |
| **7.** | Catechin (4α->8)-epicatechin-3-O-gallate | Vitis vinifera | K |
| **8.** | Catechin 3'-O-gallate | Ziziphus jujuba | N |
| **9.** | Catechin 7-O-β-D-glucopyranoside | Vigna angularis | K, N |
| **10.** | Catechin gallate | Lotus japonicus  Vigna angularis  Vitis vinifera | P  N  P |
| **11.** | Epicatechin (4β->6)-catechin  Procyanidin B7 | Theobroma cacao  Vitis vinifera | N  P |
| **12.** | Epicatechin (4β->8)-catechin  Procyanidin B1 | Amborella trichopoda  Fragaria vesca  Nelumbo nucifera  Theobroma cacao  Vigna angularis  Vitis vinifera | 2  N, 1  P  N, P  P  N, P |
| **13.** | Epicatechin (4β->8)-epicatechin-(4β->8)-catechin  Procyanidin T2 | Vitis vinifera | N |

Sources: (K) KNApSAcK; (N) NAPRALERT; (P) NPASS; (1) Gasperotti et al., 2015; (2) Wu et al., 2019

###### Table 74

Similar and derivative compounds of epicatechin synthesized by plants in the data set.

| **Compound Name(s)** | | **Species** | **Source(s)** |
| --- | --- | --- | --- |
| **1.** | [Epicatechin (4β->8)]2-epicatechin  Procyanidin C1 | Theobroma cacao  Vitis vinifera | N, P  P |
| **2.** | 9,10-Dihydro-8-hydroxy-10-methyl-8H-  pyrano[2,3-h]epicatechin | Lupinus angustifolius | K |
| **3.** | Catechin (4α->8)-catechin-(4α->8)-catechin  Procyanidin C2 | Vitis vinifera | N |
| **4.** | Catechin (4α->8)-epicatechin  Procyanidin B4 | Theobroma cacao  Vitis vinifera | N  P |
| **5.** | Catechin-(4α->6)-epicatechin  Procyanidin B8 | Vitis vinifera | N |
| **6.** | Epicatechin (4β->6)-epicatechin  Procyanidin B5 | Fragaria vesca  Nicotiana sylvestris  Theobroma cacao  Vitis vinifera | N  P  N, P  N, P |
| **7.** | Epicatechin (4β->6)catechin  Procyanidin B7 | Theobroma cacao  Vitis vinifera | P  N |
| **8.** | Epicatechin (4β->8)-catechin  Procyanidin B1 | Fragaria vesca  Nelumbo nucifera  Theobroma cacao  Vigna angularis  Vitis vinifera | N  P  N  P  N, P |
| **9.** | Epicatechin (4β->8)-epicatechin  Procyanidin B2 | Fragaria vesca  Nelumbo nucifera  Nicotiana sylvestris  Theobroma cacao  Vitis vinifera  Ziziphus jujuba | N  P  P  N, P  N, P  P |
| **10.** | Epicatechin (4β->8)-epicatechin-(4β->6)-epicatechin | Vitis vinifera | P |
| **11.** | Epicatechin (4β->8)-epicatechin-(4β->8)-catechin  Procyanidin T2 | Vitis vinifera | N |
| **12.** | Epicatechin (6'->8)-epicatechin | Vitis vinifera | N |
| **13.** | Epicatechin 3-gallate | Citrus sinensis  Malus domestica  Prunus avium  Prunus persica  Vitis vinifera | M  M, U  M, U  U  D, M, U |
| **14.** | Epicatechin 3-O-gallate | Prunus persica  Theobroma cacao  Vitis vinifera  Ziziphus jujuba | N  N  K  N |
| **15.** | Epicatechin 8-C-β-D-galactopyranoside | Theobroma cacao | N |
| **16.** | Epicatechin 8-C-β-D-galactoside | Theobroma cacao | K |
| **17.** | Epicatechin gallate | Lotus japonicus  Vitis vinifera  Zea mays | P  P, T  T |
| 18. | Epicatechin pentaacetate | Prunus persica | P |

Sources: (D) Dr. Duke; (K) KNApSAcK; (M) Microbiome; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (U) USDA

###### Table 75

Similar and derivative compounds of epigallocatechin synthesized by plants in the data set.

| **Compound Name** | | **Species** | **Source(s)** |
| --- | --- | --- | --- |
| **1.** | Epigallocatechin 3-gallate  Epigallocatechin 3-O-gallate | Citrus sinensis  Malus domestica  Phoenix dactylifera  Prunus persica  Vitis vinifera | M  M, U  M  M, U  M, N |
| **2.** | Epigallocatechin gallate | Lotus japonicus  Vitis vinifera | P  K |

Sources: (K) KNApSAcK; (M) Microbiome; (N) NAPRALERT; (P) NPASS; (U) USDA

###### Table 76

Similar and derivative compounds of gallocatechin synthesized by plants in the data set.

| **Compound Name** | | **Species** | **Source(s)** |
| --- | --- | --- | --- |
| **1.** | Gallocatechin 3-O-gallate | Vitis vinifera | K, N |
| **2.** | (+)-Gallocatechin-(4α->8)-(+)-catechin | Nelumbo nucifera | N |

Sources: (K) KNApSAcK; (N) NAPRALERT

### Flavanones

###### Table 77

Similar and derivative compounds of eriodictyol synthesized by plants in the data set.

| **Compound Name(s)** | | **Species** | **Source** |
| --- | --- | --- | --- |
| **1.** | Eriodictyol 5-O-β-glucoside | Sorghum bicolor | N |
| **2.** | Eriodictyol 7-O-glucoside  Eriodictyol 7-O-β-D-glucoside | Amborella trichopoda  Prunus avium | Wu  N |
| **3.** | Eriodictyol-7-O-rutinoside  Eriocitrin | Citrus sinensis | P |

Sources: (N) NAPRALERT; (P) NPASS; (1) Wu et al., 2019

###### Table 78

Similar and derivative compounds of naringenin synthesized by plants in the data set.

| **Compound Name(s)** | | **Species** | **Source(s)** |
| --- | --- | --- | --- |
| **1.** | (2R) Naringenin 5-O-β-D-glucopyranoside | Prunus persica | P |
| **2.** | 2-hydroxynaringenin | Zea mays | K |
| **3.** | 6-prenylnaringenin | Raphanus sativus | K |
| **4.** | 6,8-di-C-glucopyranosylnaringenin | Ziziphus jujuba | K |
| **5.** | 6,8-diprenylnaringenin  Lonchocarpol A  Senegalensein | Glycine max | K |
| **6.** | Naringenin 4',7-dimethyl ether | Cajanus cajan | N |
| **7.** | Naringenin 7-(2-p-coumaroylglucoside) | Ricinus communis | K |
| **8.** | Naringenin 7-methyl ether  Naringenin 7-O-methyl ether  7-O-Methylnaringenin  Sakuranetin | Daucus carota  Glycine max  Phaseolus vulgaris  Prunus avium | D  P  2  K |
| **9.** | Naringenin 7-O-glucoside  Naringenin 7-O-β-D-glucoside  Prunin | Amborella trichopoda  Cynara cardunculus  Phaseolus vulgaris  Prunus persica  Solanum lycopersicum  Theobroma cacao  Vitis vinifera | 4  D  2  K  K  K  P |
| **10.** | Naringenin 7-O-neohesperidoside  Naringin | Amborella trichopoda  Citrus sinensis  Phaseolus vulgaris  Prunus mume  Solanum tuberosum  Theobroma cacao | 4  D  2  P  K  1 |
| **11.** | Naringenin 7-rutinoside  Naringenin 7-O-rutinoside  Narirutin | Citrus sinensis  Cynara cardunculus  Phaseolus vulgaris | D, K, P, T  D, K  2 |
| **12.** | Naringenin chalcone | Arabidopsis thaliana  Asparagus officinalis  Capsicum annuum  Daucus carota  Ipomoea nil  Medicago truncatula  Momordica charantia  Solanum lycopersicum  Vitis vinifera | K  K  K  K  N  K  K  K  K |
| **13.** | Naringenin hexoside | Sorghum bicolor  Theobroma cacao | 3  1 |

Sources: (D) Dr. Duke; (K) KNApSAcK; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (1) Gallego et al., 2019; (2) Ganesan & Xu, 2017; (3) Luo et al., 2018;(4) Wu et al., 2019

### Flavonols

###### Table 79

Similar and derivative compounds of kaempferol synthesized by plants in the data set.

| **Compound Name(s)** | | **Species** | **Source(s)** |
| --- | --- | --- | --- |
| **1.** | 3,6-Dimethoxy Kaempferol | Beta vulgaris | P |
| **2.** | Dihydrokaempferol  Aromadendrin | Fragaria vesca  Glycine soja  Momordica charantia | 4  P  K |
| **3.** | Dihydrokaempferol 7-O-β-D-glucoside | Ipomoea nil | N |
| **4.** | Kaempferol 3-(2''-p-coumarylglucoside)  2''-p-coumarylastragalin | Quercus suber | K |
| **5.** | Kaempferol 3-(2''-sinapylglucoside)-7-sophoroside | Brassica napus | K |
| **6.** | Kaempferol 3-(2'',3''-diacetyl-4''-(Z)-p-coumaryl-6''-(E)-p-  coumarylglucoside) | Quercus suber | K |
| **7.** | Kaempferol 3-(2'''-(E)-caffeoylglucosyl)-(1->2)-glucoside-  7-cellobioside | Brassica oleracea | K |
| **8.** | Kaempferol 3-(2'''-(E)-caffeylsophoroside)-7-glucoside | Brassica oleracea | K |
| **9.** | Kaempferol 3-(2'''-(E)-p-coumarylsophoroside)-7-glucoside | Brassica oleracea | K |
| **10.** | Kaempferol 3-(2G-glucosylrutinoside)  Kaempferol 3-(6'''-rhamnosyl-2'''-glucosyl-glucoside) | Prunus avium | K |
| **11.** | Kaempferol 3-(3'',4''-diacetyl-2'',6''-di-(E)-p-  coumarylglucoside) | Quercus suber | K |
| **12.** | Kaempferol 3-(4''-acetyl-6''-p-coumarylglucoside) | Quercus suber | K |
| **13.** | Kaempferol 3-(6''-acetylglucosyl)-(1->3)-galactoside  Ricinitin | Ricinus communis | K |
| **14.** | Kaempferol 3-[2'''-(E)-ferulylsophoroside]-7-glucoside | Brassica oleracea | K |
| **15.** | Kaempferol 3-apiosyl-(1->2)-glucoside | Cicer arietinum | K |
| **16.** | Kaempferol 3-feruloyl-sophoroside | Brassica napus  Brassica oleracea  Brassica rapa | N  N  N |
| **17.** | Kaempferol 3-galactoside-7-rhamnoside | Arabidopsis thaliana | K |
| **18.** | Kaempferol 3-galactosyl-7-diglucoside | Prunus persica | N |
| **19.** | Kaempferol 3-gentiobioside  Kaempferol 3-O-gentiobioside  Kaempferol 3-O-β-D-gentiobioside | Glycine max | K, N |
| **20.** | Kaempferol 3-gentiobioside-7-rhamnoside  Kaempferol 3-glucoside-(1->6)-glucoside-7-α-L-rhamnoside | Arabidopsis thaliana | K |
| **21.** | Kaempferol 3-glucoside  Kaempferol 3-O-β-D-glucoside  Kaempferol 3-O-β-D-glucopyranoside  Astragalin | Cicer arietinum  Daucus carota  Fragaria vesca  Glycine max  Lotus japonicus  Nelumbo nucifera  Nicotiana tabacum  Phaseolus vulgaris  Rosa chinensis  Spinacia oleracea  Vitis vinifera | T  D  4  D  K  P, T  D  3  P  D  D, P |
| **22.** | Kaempferol 3-glucosyl-7-arabinoside | Raphanus sativus | N |
| **23.** | Kaempferol 3-glucosyl-7-diglucoside | Prunus persica | N |
| **24.** | Kaempferol 3-glucuronide  Kaempferol-3-β-D-glucuronide  Kaempferol 3-O-β-D-glucuronide | Fragaria vesca  Nelumbo nucifera  Phaseolus vulgaris  Vigna radiata | 4  D, N, T  K, T  K, N |
| **25.** | Kaempferol 3-neohesperidoside  Kaempferol 3-O-β-D-neohesperidoside | Glycine max | N |
| **26.** | Kaempferol 3-O-(2,6-di-O-α-L-rhamnopyranosyl)-β-D-  galactopyranoside  Mauritianin | Chenopodium quinoa | K |
| **27.** | Kaempferol 3-O-(2'',6''-di-(E)-p-coumaroyl)-β-glucopyranoside | Quercus suber | P |
| **28.** | Kaempferol 3-O-(3'',4"-diacetyl-2",6"-di-(E)-p-coumaroyl)-  glucoside | Quercus suber | P |
| **29.** | Kaempferol 3-O-[α-L-rhamnopyranosyl(1->2)-β-D-  glucopyranosyl]-7-O-α-L-rhamnopyranoside | Arabidopsis thaliana | N |
| **30.** | Kaempferol 3-O-D-galactoside  Kaempferol 3-O-α-L-galactoside  Kaempferol 3-O-β-D-galactoside  Kaempferol 3-O-β-D-galactopyranoside  Trifolin | Nelumbo nucifera  Nicotiana tabacum  Prunus persica  Zea mays | P, N  2  N  N |
| **31.** | Kaempferol 3-O-digalactopyranoside | Glycine max | N |
| **32.** | Kaempferol 3-O-diglucopyranoside | Glycine max | N |
| **33.** | Kaempferol 3-O-glucosyl-gentiobioside | Glycine max | N |
| **34.** | Kaempferol 3-O-glucosyl-rutinoside | Glycine max | N |
| **35.** | Kaempferol 3-O-neohesperidoside | Glycine max  Gossypium hirsutum | N  N |
| **36.** | Kaempferol 3-O-rhamnosyl-gentiobioside | Glycine max | N |
| **37.** | Kaempferol 3-O-rhamnosyl-rutinoside | Glycine max | N |
| **38.** | Kaempferol 3-O-robinobioside | Nelumbo nucifera | P |
| **39.** | Kaempferol 3-O-robinoside-7-O-rhamnoside  Robinin | Amborella trichopoda  Nelumbo nucifera  Vigna radiata | 6  D  D |
| **40.** | Kaempferol 3-O-rutinoside  Kaempferol 3-O-β-rutinoside  Kaempferol 3-O-(α-L-rhamnopyranosyl(1->6)-β-D-  glucopyranoside  Nicotiflorin  Nicotifloroside | Nelumbo nucifera  Phaseolus vulgaris  Prunus mume  Ricinus communis  Ziziphus jujuba | P  3  7  D, K  P |
| **41.** | Kaempferol 3-O-sophoroside  Kaempferol 3-O-β-sophoroside  Kaempferol 3-O-β-D-sophoroside  Sophoraflavonoloside | Brassica napus  Brassica oleracea  Brassica rapa  Glycine max  Solanum tuberosum  Vigna angularis  Vigna radiata | N  N  N  K, N  D  N  N |
| **42.** | Kaempferol 3-O-sophoroside-7-O-β-D-glucopyranoside | Brassica rapa | K |
| **43.** | Kaempferol 3-O-α-L-rhamnopyranosyl (1->6)-β-D-  glucopyranoside | Nelumbo nucifera | N |
| **44.** | Kaempferol 3-O-β-D-diglucoside | Prunus persica | N |
| **45.** | Kaempferol 3-O-β-D-glucopyranoside-7-O-α-L-  rhamnopyranoside | Arabidopsis thaliana | N |
| **46.** | Kaempferol 3-O-β-D-glucopyranosyl-7-O-α-L-  rhamnopyranoside | Arabidopsis thaliana | N |
| **47.** | Kaempferol 3-O-β-D-glucuronopyranosyl methyl ester | Nelumbo nucifera | N |
| **48.** | Kaempferol 3-O-β-D-glycoside | Nicotiana tabacum | 2 |
| **49.** | Kaempferol 3-O-β-D-xylopyranoside | Ricinus communis | D, N |
| **50.** | Kaempferol 3-rhamnoglucoside  Kaempferol 3-O-rhamnoglucoside | Nicotiana tabacum  Phoenix dactylifera | N  5 |
| **51.** | Kaempferol 3-rhamnoside  Kaempferol 3-O-rhamnoside  Kaempferol 3-O-α-rhamnoside  Kaempferol 3-O-α-L-rhamnopyranoside  Afzelin | Raphanus sativus  Rosa chinensis | K  P |
| **52.** | Kaempferol 3-rutinosyl-4'-diglucoside | Prunus avium | N |
| **53.** | Kaempferol 3-sinapoyl-sophoroside | Brassica napus  Brassica oleracea  Brassica rapa | N  N  N |
| **54.** | Kaempferol 3-sophoroside-7-cellobioside | Brassica oleracea | K |
| **55.** | Kaempferol 3-sophoroside-7-glucoside | Brassica napus  Brassica oleracea  Brassica rapa | N  N  N |
| **56.** | Kaempferol 3-sophoroside-7-rhamnoside | Solanum tuberosum | K |
| **57.** | Kaempferol 3-sophoroside-rhamnoside | Solanum tuberosum | D |
| **58.** | Kaempferol 3-sophorotrioside-7-rhamnoside | Solanum tuberosum | K |
| **59.** | Kaempferol 3-triglucoside-7-rhamnoside | Solanum tuberosum | D |
| **60.** | Kaempferol 3-xyloside | Ricinus communis | K |
| **61.** | Kaempferol 3-β-D-glucuronopyranoside | Phaseolus vulgaris | N |
| **62.** | Kaempferol 3,4'-diglucoside  Kaempferol 3,4'-di-O-β-D-glucopyranoside | Prunus avium | K |
| **63.** | Kaempferol 3,4'-dimethyl ether | Arabidopsis thaliana | N |
| **64.** | Kaempferol 3,7-diglucoside  Kaempferol 3,7-di-O-β-D-glucoside  Kaempferol 3,7-O-β-D-diglucopyranoside  Kaempferol 3,7-di-O-β-D-glucopyranoside | Arabidopsis thaliana  Brassica napus  Brassica rapa | P  N  N |
| **65.** | Kaempferol 3,7-dirhamnoside  Kaempferitrin | Arabidopsis thaliana  Lotus japonicus  Raphanus sativus | D, P  K  K |
| **66.** | Kaempferol 4'-methyl ether  Kaempferide | Populus trichocarpa  Raphanus sativus | N  N |
| **67.** | Kaempferol 7-glucoside  Kaempferol 7-O-glucoside  Populnin | Cicer arietinum  Vitis vinifera | T  D |
| **68.** | Kaempferol 7-O-rutinoside | Prunus mume | 7 |
| **69.** | Kaempferol 7-O-β-d-glucopyranoside | Nelumbo nucifera | P |
| **70.** | Kaempferol 7-rhamnoside  Kaempferol 7-O-rhamnoside  Kaempferol 7-O-α-L-rhamnopyranoside | Phaseolus vulgaris  Vigna angularis  Vigna radiata | K  N  N |
| **71.** | Kaempferol 7-α-L-arabinoside | Raphanus sativus | K |
| **72.** | Kaempferol 7,4'-dimethoxy ether | Raphanus sativus | N |
| **73.** | Kaempferol sinapoyl-trihexoside | Brassica napus | 1 |

Sources: (D) Dr. Duke; (K) KNApSAcK; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (1) Auger et al., 2010; (2) Chen et al., 2013; (3) Ganesan & Xu, 2017; (4) Gasperotti et al., 2015; (5) Khallouki et al., 2018; (6) Wu et al., 2019 ; (7) Zhang et al., 2015.

###### 

###### Table 80

Similar and derivative compounds of myricetin synthesized by plants in the data set.

| **Compound Name(s)** | | **Species** | **Source(s)** |
| --- | --- | --- | --- |
| **1.** | 3'-O-Methylmyricetin  Laricitrin | Medicago truncatula | K |
| **2.** | Dihydromyricetin  Ampelopsin | Capsicum annuum | K |
| **3.** | Myricetin 3-glucoside  Myricetin 3-O-glucoside  Myricetin 3-O-β-D-glucoside  Myricetin 3-O-β-D-glucopyranoside  Isomyricitrin | Nelumbo nucifera  Phaseolus vulgaris  Vitis vinifera | P  K, 1  N |
| **4.** | Myricetin 3-glucuronide  Myricetin 3-O-β-D-glucuronide | Nelumbo nucifera  Vitis vinifera | P  N |
| **5.** | Myricetin 3-O-rhamnoside | Phaseolus vulgaris | 1 |
| **6.** | Myricetin 3-O-β-D-galactopyranoside | Nelumbo nucifera  Vitis vinifera | P  P |
| **7.** | Myricetin 3',5'-dimethyl ether  Syringetin | Nelumbo nucifera | P |
| **8.** | Syringetin 3-O-glucoside  Syringetin 3-O-β-D-glucopyranoside | Nelumbo nucifera | N, P |

Sources: (K) KNApSAcK; (N) NAPRALERT; (P) NPASS; (1) Ganesan & Xu, 2017

###### Table 81

Similar and derivative compounds of quercetin synthesized by plants in the data set.

| **Compound Name(s)** | | **Species** | **Source(s)** |
| --- | --- | --- | --- |
| 1. | Dihydroquercetin  Taxifolin  Distylin | Asparagus officinalis  Capsicum annuum  Daucus carota  Fragaria vesca  Glycine soja  Medicago truncatula  Momordica charantia  Nelumbo nucifera  Nicotiana tabacum  Olea europaea  Prunus avium  Zea mays | K  K  K  5  P  K  K  P  2  6  K  K |
| 2. | Dihydroquercetin 3-O-rhamnoside  Taxifolin 3-O-rhamnoside | Phoenix dactylifera | 7 |
| 3. | Isorhamnetin 3-galactoside  Isorhamnetin 3-O-galactoside | Beta vulgaris  Phaseolus vulgaris  Prunus mume | P  K  10 |
| 4. | Isorhamnetin 3-glucoside  Isorhamnetin 3-O-glucoside | Amborella trichopoda  Brassica napus  Fragaria vesca  Nelumbo nucifera  Phoenix dactylifera  Quercus suber | 9  1  5  P  7  P |
| 5. | Isorhamnetin 3-O-neohesperidoside | Prunus mume | 10 |
| 6. | Isorhamnetin 3-O-rhamnoglucoside | Phoenix dactylifera | 7 |
| 7. | Isorhamnetin 3-O-rhamnoside | Prunus mume | K, P |
| 8. | Isorhamnetin 3-O-rutinoside | Amborella trichopoda  Prunus mume | 9  10 |
| 9. | Isorhamnetin 7-O-rutinoside | Prunus mume | 10 |
| 10. | Isorhamnetin dihexoside | Brassica napus | 1 |
| 11. | Isorhamnetin hexoside-sulfate | Brassica napus | 1 |
| 12. | Isorhamnetin sinapoyl-trihexoside | Brassica napus | 1 |
| 13. | Quercetin 3-(2'''-(E)-caffeylsophoroside]-7-glucoside | Brassica oleracea | K |
| 14. | Quercetin 3-(6''-malonylglucoside)-7-glucoside | Lactuca sativa | K |
| 15. | Quercetin 3-[2'''-(E)-feruloylsophoroside]-7-glucoside | Brassica oleracea | K |
| 16. | Quercetin 3-2G-rhamnosylrutinoside  Manghaslin | Glycine max | K |
| 17. | Quercetin 3-acetylglucoside | Fragaria vesca | 5 |
| 18. | Quercetin 3-arabinoside | Juglans regia | D |
| 19. | Quercetin 3-D-xyloside  Quercetin 3-O-xyloside | Malus domestica | D |
| 20. | Quercetin 3-diglucoside  Quercetin 3-O-diglucoside  Quercetin 3-O-β-D-diglucoside | Nicotiana tabacum  Prunus persica  Ziziphus jujuba | D  N  N |
| 21. | Quercetin 3-galactoside  Quercetin 3-O-galactoside  Quercetin 3-β-galactoside  Quercetin 3-O-β-D-galactoside  Quercetin 3-O-galactopyranoside  Quercetin 3-β-galactopyranoside  Quercetin 3-O-β-D-galactopyranoside  Hyperin  Hyperoside | Juglans regia  Malus domestica  Musa acuminata  Nelumbo nucifera  Nicotiana tabacum  Phaseolus vulgaris  Phoenix dactylifera  Rosa chinensis  Theobroma cacao  Ziziphus jujuba | D  D  K  K, P  2  4  7  P  K  P |
| 22. | Quercetin 3-gentiobioside | Glycine max  Papaver somniferum | K  D, K, N |
| 23. | Quercetin 3-glucoside  Quercetin 3-O-glucoside  Quercetin 3-O-β-glucoside  Quercetin 3-O-β-D-glucoside  Quercetin 3-O-β-D-glucopyranoside  Isoquercetin  Isoquercitrin  Hirsutrin | Asparagus officinalis  Brassica napus  Cicer arietinum  Daucus carota  Fragaria vesca  Glycine max  Gossypium hirsutum  Juglans regia  Lactuca sativa  Lotus japonicus  Momordica charantia  Musa acuminata  Nelumbo nucifera  Nicotiana tabacum  Papaver somniferum  Phaseolus vulgaris  Phoenix dactylifera  Prunus mume  Quercus suber  Ricinus communis  Rosa chinensis  Solanum tuberosum  Theobroma cacao  Vigna radiata  Vitis vinifera  Ziziphus jujuba | P  1  T  D  5  D  K, N  T  T  K  K  K  K, P, T  D  D  T, 4  7  P, 10  P  D, K, T  P  D, K  T  D  P  P |
| 24. | Quercetin 3-glucosilarabinoside | Fragaria vesca | 5 |
| 25. | Quercetin 3-glucuronide  Quercetin 3-O-glucuronide  Miquelianin | Fragaria vesca  Nelumbo nucifera  Phaseolus vulgaris  Vitis vinifera | 5  D  D  N |
| 26. | Quercetin 3-methyl ether  3-O-methylquercetin | Beta vulgaris  Carica papaya | P  P |
| 27. | Quercetin 3-O-(2,6-di-α-L-rhamnopyranosyl)-β-D-  galactopyranoside | Chenopodium quinoa | N |
| 28. | Quercetin 3-O-[6''-O-rhamnosyl-glucoside] 7-  O-rhamnoside | Arabidopsis thaliana | K |
| 29. | Quercetin 3-O-glucosyl-galactoside | Juglans regia  Ziziphus jujuba | N  N |
| 30. | Quercetin 3-O-methyl-ether | Phaseolus vulgaris | D |
| 31. | Quercetin 3-O-neohesperidoside | Glycine max  Gossypium hirsutum  Prunus mume  Zea mays | K  N  K, N, P  K |
| 32. | Quercetin 3-O-rhamnosyl(1->6)-galactoside | Prunus mume | N |
| 33. | Quercetin 3-O-rutinoside  Rutin | Asparagus officinalis  Carica papaya  Citrus sinensis  Cucumis melo  Cynara cardunculus  Glycine max  Lotus japonicus  Malus domestica  Momordica charantia  Musa acuminata  Nelumbo nucifera  Nicotiana tabacum  Olea europaea  Phaseolus vulgaris  Phoenix dactylifera  Prunus mume  Ricinus communis  Solanum lycopersicum  Solanum tuberosum  Spinacia oleracea  Theobroma cacao  Vigna angularis  Vigna radiata  Vitis vinifera  Ziziphus jujuba | D, K, P  K  D  D, P  D  D  P  D  K  K  D  D, P  D, 6  4  D, 7  K  D  K  D, K  D  D  P  D  D  D, K, P |
| 34. | Quercetin 3-O-α-L-arabinopyranoside  Guaijaverin  Guajavarin | Theobroma cacao | N |
| 35. | Quercetin 3-O-α-L-rhamnopyranosyl-7-O-β-D-glucopyranoside | Capsicum annuum | N |
| 36. | Quercetin 3-O-α-L-rhamnosyl-glucoside | Manihot esculenta | N |
| 37. | Quercetin 3-O-β-D-2-glucosyl-rutinoside | Glycine max | D |
| 38. | Quercetin 3-O-β-D-glucopyranosyl-7-O-α-L-  rhamnopyranoside | Arabidopsis thaliana | K |
| 39. | Quercetin 3-O-β-D-xylopyranoside  Reynoutrin | Malus domestica  Ricinus communis | D  D, K |
| 40. | Quercetin 3-O-β-robinoside 7-O-α-L-rhamnopyranoside | Phaseolus vulgaris | D |
| 41. | Quercetin 3-rhamnoglucoside | Malus domestica | D |
| 42. | Quercetin 3-rhamnoside  Quercetin 3-O-rhamnoside  Quercetin 3-O-rhamnopyranoside  Quercitrin | Arabidopsis thaliana  Capsicum annuum  Daucus carota  Glycine max  Juglans regia  Malus domestica  Nicotiana tabacum  Olea europaea  Raphanus sativus  Ricinus communis  Rosa chinensis  Solanum lycopersicum  Solanum tuberosum  Spinacia oleracea | T  K, P  D  D  D  D  D  D, 6  K  D  P  K  D  D |
| 43. | Quercetin 3-rhamnoside-7-glucoside | Capsicum annuum | K |
| 44. | Quercetin 3-rhamnosyl-(1->6)-(2''-acetylglucoside) | Prunus avium  Prunus mume | K  K |
| 45. | Quercetin 3-rutinoside-4'-glucoside | Prunus avium | K |
| 46. | Quercetin 3-rutinoside-7-glucoside | Nicotiana tabacum | K |
| 47. | Quercetin 3-sophoroside  Quercetin 3-O-sophoroside  Quercetin 3-O-β-D-sophoroside  Baimaside | Brassica oleracea  Glycine max  Vigna radiata | N  K, N  K, N |
| 48. | Quercetin 3-sophoroside-7-glucoside | Brassica napus  Brassica oleracea | K  K |
| 49. | Quercetin 3,3'-diglucoside | Zea mays | K |
| 50. | Quercetin 3,3'-dimethyl ether | Nicotiana tabacum | D |
| 51. | Quercetin 3,4-diglucoside | Fragaria vesca | 5 |
| 52. | Quercetin 3,4'-diglucoside | Solanum lycopersicum | P |
| 53. | Quercetin 3,4'-dimethyl ether  3',4'-Dimethoxyquercetin  Dillenetin | Vigna unguiculata | N |
| 54. | Quercetin 3,5,3',4'-tetramethyl ether | Gossypium hirsutum | K |
| 55. | Quercetin 3,5,7,4'-tetramethyl ether | Gossypium hirsutum | K |
| 56. | Quercetin 3,7-di-O-α-L-rhamnopyranoside | Arabidopsis thaliana | K |
| 57. | Quercetin 3,7-diglucoside  Quercetin 3,7-di-O-β-D-glucoside  Quercetin 3,7-O-β-diglucopyranoside  Quercetin 3,7-di-O-β-D-glucopyranoside  7-O-β-D-glucopyranosyl-quercetin-3-O-β-D-  glucopyranoside | Brassica napus  Brassica rapa  Solanum lycopersicum  Zea mays | N  N, P  P  D, K, P |
| 58. | Quercetin 3,7-dimethyl ether 5-glucoside | Zea mays | K |
| 59. | Quercetin 3'-glucoside | Gossypium arboreum  Gossypium hirsutum  Helianthus annuus | K  K  D |
| 60. | Quercetin 3'-methyl ether  3'-O-Methylquercetin  Isorhamnetin | Amborella trichopoda  Brassica oleracea  Cicer arietinum  Lotus japonicus  Nelumbo nucifera  Nicotiana tabacum  Prunus avium  Prunus mume  Raphanus sativus | 9  K  K, T  P  P  D  M  10  K |
| 61. | Quercetin 4-O-β-D-glucoside | Ipomoea nil | N |
| 62. | Quercetin 4'-methyl ether  4'-Methoxyquercetin  Tamarixetin | Raphanus sativus | K |
| 63. | Quercetin 4'-O-β-D-glucopyranoside  Spiraeoside |  |  |
| 64. | Quercetin 7-methyl ether  7-Methoxyquercetin  7-O-Methxyl quercetin  Rhamnose  Rhamnetin | Sesamum indicum | P |
| 65. | Quercetin 7-O-glucoside | Olea europaea | 6 |
| 66. | Quercetin 7-O-β-D-glucoside  Quercimeritrin | Helianthus annuus | D |
| 67. | Quercetin 7,4'-di-O-β-D-glucopyranoside | Solanum lycopersicum | P |
| 68. | Quercetin dihexoside | Brassica napus | 1 |
| 69. | Quercetin glucuronide | Phaseolus vulgaris  Rosa chinensis  Vitis vinifera | P  P  P |
| 70. | Quercetin hexoside | Beta vulgaris  Theobroma cacao | 8  3 |

Sources: (D) Dr. Duke; (K) KNApSAcK; (M) Microbiome; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (1) Auger et al., 2010; (2) Chen et al., 2013; (3) Gallego et al., 2019; (4) Ganesan & Xu, 2017; (5) Gasperotti et al., 2013; (6) Hashmi et al., 2015; (7) Khallouki et al., 2018; (8) Vissers et al., 2017; (9) Wu et al., 2019; (10) Zhang et al., 2015

### Flavones

###### Table 82

Similar and derivative compounds of apigenin synthesized by plants in the data set.

| **Compound Name(s)** | | **Species** | **Source(s)** |
| --- | --- | --- | --- |
| **1.** | 6-Hydroxyapigenin  Scutellarin | Fragaria vesca | P |
| **2.** | Apigenin 4'-methyl ether  Acacetin | Populus trichocarpa | N |
| **3.** | Apigenin 4'-O-rhamnosyl-glucoside | Olea europaea | N |
| **4.** | Apigenin 4'-O-β-D-glucoside | Daucus carota | D, N |
| **5.** | Apigenin 4',7-diglucoside | Cynara cardunculus | D |
| **6.** | Apigenin 5-O-β-D-glucoside | Daucus carota | N |
| **7.** | Apigenin 6-C-glucoside  Isovitexin | Cucumis sativus  Jatropha curcas  Nelumbo nucifera  Theobroma cacao  Vigna angularis  Vigna radiata  Zea mays  Ziziphus jujuba | K  D  P  K  P  K  K  P |
| **8.** | Apigenin 6-C-α-L-arabinopyranoside-  8-C-β-D-glucopyranoside  Isoschaftoside | Capsicum annuum  Nelumbo nucifera | K  P |
| **9.** | Apigenin 6-C-β-D-glucopyranosyl-  8-C-α-L-arabinopyranoside  Schaftoside | Capsicum annuum | N |
| **10.** | Apigenin 6,8-di-C-glucoside  Vicenin II | Cucumis sativus  Fragaria vesca  Medicago truncatula  Olea europaea  Ziziphus jujuba | K  P  K  4  P |
| **11.** | Apigenin 7-apioglucoside  Apiin | Capsicum annuum  Lupinus angustifolius | D  1 |
| **12.** | Apigenin 7-di-O-xyloside  Apigenin 7-di-O-β-D-xyloside | Olea europaea | D, N |
| **13.** | Apigenin 7-galactomannoside | Daucus carota | N |
| **14.** | Apigenin 7-glucoside  Apigenin 7-O-glucoside  Apigenin 7-O-β-D-glucoside  Apigenin 7-O-β-D-glucopyranoside  Apigetrin  Cosmosiin | Asparagus officinalis  Capsicum annuum  Cynara cardunculus  Daucus carota  Fragaria vesca  Olea europaea  Phaseolus vulgaris  Theobroma cacao | K  K  D  D, K  P  D, 4  3  D, K |
| **15.** | Apigenin 7-glycoside | Cucumis melo | 6 |
| **16.** | Apigenin 7-methyl-glucuronide | Cynara cardunculus | N |
| **17.** | Apigenin 7-neohesperidoside | Lupinus angustifolius | 1 |
| **18.** | Apigenin 7-O-(6''-O-para-coumaroyl-  glucoside) | Cucumis sativus | N |
| **19.** | Apigenin 7-O-apiofuranosyl(1->2)-  β-D-glucoside | Lactuca sativa | N |
| **20.** | Apigenin 7-O-α-L-rhamno-glucoside | Citrus sinensis | N |
| **21.** | Apigenin 7-O-β-D-galactomannoside | Daucus carota | D |
| **22.** | Apigenin 7-O-β-D-glucosyl-  6-C-glucoside | Cucurbita pepo | N |
| **23.** | Apigenin 7-rutinoside  Apigenin 7-O-rutinoside  Apigenin 7-O-β-D-rutinoside  Isorhoifolin | Cynara cardunculus  Daucus carota  Olea europaea | D  D, K  4 |
| **24.** | Apigenin 7,4'-dimethyl ether | Populus trichocarpa | N |
| **25.** | Apigenin 8-C-glucoside  Vitexin | Beta vulgaris  Cucumis sativus  Glycine max  Jatropha curcas  Nelumbo nucifera  Theobroma cacao  Vigna angularis  Vigna radiata  Ziziphus jujuba | P  K  D, K  D  P  D, K  P  K  P |
| **26.** | Apigenin 8-C-glucoside-2'-rhamnoside  Vitexin 2'-O-rhamnoside | Glycine max | D |
| **27.** | Apigenin C-diglucoside | Phoenix dactylifera | 5 |
| **28.** | Apigenin di-C-glycoside | Lupinus angustifolius | N |
| **29.** | Apigenin hexoside | Theobroma cacao | 2 |
| **30.** | Apigenin trimethyl ether | Citrus sinensis | N |
| **31.** | Apigeninidin | Sorghum bicolor  Zea mays | K  N |
| **32.** | Trimethylapigenin | Citrus sinensis | P |

Sources: (D) Dr. Duke; (K) KNApSAcK; (N) NAPRALERT; (P) NPASS; (1) Dueñas et al., 2009; (2) Gallego et al., 2019; (3) Ganesan & Xu, 2017; (4) Hashmi et al., 2015; (5) Khallouki et al., 2018; (6) Mallek-Ayadi et al., 2017.

###### Table 83

Similar and derivative compounds of luteolin synthesized by plants in the data set.

| **Compound Name(s)** | | **Species** | **Source(s)** |
| --- | --- | --- | --- |
| **1.** | 2''-O-α-L-rhamnosyl-6-C-fucosyl-3'-  methoxyluteolin | Zea mays | N |
| **2.** | 2''-O-α-L-rhamnosyl-6-C-fucosyl-luteolin | Zea mays | N |
| **3.** | 2"-O-α-L-rhamnosyl-6-C-quinovosyl-luteolin | Zea mays | N |
| **4.** | Luteolin 3'-methyl ether  Chrysoeriol | Amborella trichopoda  Fragaria vesca  Olea europaea  Phoenix dactylifera  Zea mays  Ziziphus jujuba | 6  P  3  4  K  P |
| **5.** | Luteolin 3'-methyl ether 7-O-glucoside  Chrysoeriol 7-O-glucoside | Olea europaea  Phoenix dactylifera | 3  4 |
| **6.** | Luteolin 3'-methyl ether 7-O-  rhamnoglucoside  Chrysoeriol 7-O-rhamnoglucoside | Phoenix dactylifera | 4 |
| **7.** | Luteolin 3'-O-β-D-glucoside | Cynara cardunculus | D |
| **8.** | Luteolin 4-β-D-glucoside | Cynara cardunculus | D |
| **9.** | Luteolin 4'-glucoside  Luteolin 4'-O-glucoside  Luteolin 4'-β-D-glucoside  Luteolin 4'-O-β-D-glucoside  Luteolin 4'-O-β-D-glucopyranoside  Juncein | Cynara cardunculus  Daucus carota  Olea europaea | D  D  D, N |
| **10.** | Luteolin 4'-methyl ether  4'-Methylluteolin  Diosmetin | Amborella trichopoda  Nelumbo nucifera  Olea europaea | 6  P  3 |
| **11.** | Luteolin 4'-sulfate | Daucus carota | K, T |
| **12.** | Luteolin 4',7-diglucoside | Cynara cardunculus | D |
| **13.** | Luteolin 5-glucoside | Olea europaea | D |
| **14.** | Luteolin 6-C-glucoside  Luteolin 6-C-β-D-glucopyranoside  Homoorientin  Isoorientin | Nelumbo nucifera | P |
| **15.** | Luteolin 7-gentiobioside | Cynara cardunculus | D |
| **16.** | Luteolin 7-glucoside  Luteolin 7-O-glucoside  Luteolin 7-O-β-D-glucoside  Cynaroside  Glucoluteolin  Luteoloside | Cajanus cajan  Capsicum annuum  Cynara cardunculus  Fragaria vesca  Lupinus angustifolius  Nelumbo nucifera  Olea europaea  Prunus mume  Theobroma cacao | D  D, P  D  P  1  D, P  D, 3  P  D |
| **17.** | Luteolin 7-glycoside | Cucumis melo | 5 |
| **18.** | Luteolin 7-O-(2-apiofuranosyl-4-  glucopyranosyl-6-malonyl) glucopyranoside | Capsicum annuum | K |
| **19.** | Luteolin 7-O-(6''-malonylglucoside) | Daucus carota | K |
| **20.** | Luteolin 7-O-glucuronic acid | Fragaria vesca | P |
| **21.** | Luteolin 7-O-rhamnoglucoside | Phoenix dactylifera | 4 |
| **22.** | Luteolin 7-O-β-glucuronide  Luteolin 7-O-β-D-glucuronide | Daucus carota | D, K, N |
| **23.** | Luteolin 7-rutinoside  Luteolin 7-O-rutinoside  Luteolin 7-O-β-rutinoside  Luteolin 7-O-β-D-rutinoside | Cynara cardunculus  Daucus carota  Olea europaea  Phoenix dactylifera | D  D, K  3  D, N |
| **24.** | Luteolin 7-β-rutinoside | Cynara cardunculus | D |
| **25.** | Luteolin 7,4-O-diglucoside | Olea europaea | 3 |
| **26.** | Luteolin 8-C-glucoside  Luteolin 8-C-β-D-glucopyranoside  Lutexin  Orientin | Beta vulgaris  Capsicum annuum  Cucumis melo  Nelumbo nucifera  Theobroma cacao | P  P  K  P  K |
| **27.** | Luteolin glucopyranosyl acetate | Lupinus angustifolius | 1 |
| **28.** | Luteolin hexoside | Theobroma cacao | 2 |

Sources: (D) Dr. Duke; (K) KNApSAcK; (N) NAPRALERT; (P) NPASS; (T) IMPPAT; (1) Dueñas et al., 2009; (2) Gallego et al., 2019; (3) Hashmi et al., 2015; (4) Khallouki et al., 2018; (5) Mallek-Ayadi et al., 2017; (6) Wu et al., 2019

## GitHub

All of the files are hosted on GitHub at https://github.com/mbagg4152/pfpy/tree/master/projects/mh/. The README for the repository can be found here: https://github.com/mbagg4152/pfpy#readme.

## Script Files

### Main file: keggv3.py

*# custom project library imports*

from flib.fconstants import \*

from flib.data\_types import \*

import flib.prediction\_logic as predict

*# other library imports*

from urllib import request

from urllib.error import HTTPError, URLError

import datetime

import sys

import threading

sys.path.append(os.getcwd().replace(os.sep + 'flavonoid', ''))

from sharedlib.regexlines import \*

try:

import bioservices

from bioservices.kegg import KEGG

except *ImportError* as e:

*print*('Program needs bioservices in order to work. In the terminal, try:\n'

'`pip3 install bioservices` or `pip install bioservices`')

*print*(e)

*exit*(1)

*####################*

*# Global Variables #*

*####################*

init\_time = datetime.datetime.now() *# get time of program execution*

kegg = KEGG() *# used to access KEGG's functions from bioservices*

*# The lists are used for processing/accessing data as well as making output files.*

all\_genes = []

ec\_collections = []

pathgenes = []

plant\_matrix = []

plant\_objects = []

plant\_pathways = []

*# The locks are to protect the values of the global variables when using multithreading.*

lock\_ec = threading.Lock()

lock\_gene = threading.Lock()

lock\_kegg = threading.Lock()

lock\_plant = threading.Lock()

lock\_dbget = threading.Lock()

*# These will hold the file path values for the programs outputs and current*

*# and project directory.*

path\_chem = ''

path\_cwd = ''

path\_fasta = ''

path\_gene = ''

path\_main = ''

path\_raw\_gene = ''

path\_raw\_fasta = ''

thread\_lim = 5 *# The max number of processor threads to be used by program.*

def main():

*"""*

*This is the main function of the file which calls specific functions in order when*

*running and displays the total run time at the end of the code execution.*

*"""*

setup()

get\_parse\_pathway\_genes()

flavonoid\_predictions()

make\_plant\_ec\_counts()

build\_nt\_fasta\_by\_ec()

runtime = datetime.datetime.now() - init\_time

*print*('\nRun time: ' + *str*(runtime))

def setup():

*"""*

*This is the initial setup function for the program. If the user supplies a directory*

*name in the command line args, then that name will be used when outputting the data.*

*If no name is supplied, then the data is outputted to the directory named data. This*

*function also creates the list of pathway and plant codes based on the KEGG codes*

*that can be found in the JSON files. After making the list, it creates a list of plant*

*objects that will be used throughout the program.*

*"""*

*# Make sure global values can be used and updated*

global path\_chem

global path\_cwd

global path\_fasta

global path\_gene

global path\_main

global path\_raw\_gene

global plant\_objects

global plant\_pathways

global path\_raw\_fasta

*# Check if user supplied directory name as a commandline argument.*

*# if not, default to 'data'*

opt\_dir\_name = ''

if *len*(sys.argv) > 1: *# Alternate directory name supplied by user*

opt\_dir\_name = sys.argv[1] *# Get directory name*

else:

*print*('No directory name supplied, defaulting to `data`. '

'Supply name using `python3 keggv2.py dir\_name`')

opt\_dir\_name = 'data'

*# Update the global path values based on user decision*

path\_cwd = os.getcwd() + SEP

path\_main = path\_cwd + opt\_dir\_name

path\_chem = path\_main + DIR\_CHEM

path\_fasta = path\_main + DIR\_FASTA

path\_gene = path\_main + DIR\_GENE

path\_raw\_gene = path\_main + DIR\_RGENE

path\_raw\_fasta = path\_main + DIR\_RFASTA

*# Initialize the output directories if they don't exist*

init\_dir(path\_main)

init\_dir(path\_gene)

init\_dir(path\_fasta)

init\_dir(path\_chem)

init\_dir(path\_raw\_gene)

init\_dir(path\_raw\_fasta)

*# Combines plant and pathway codes.*

plant\_pathways = [i + j for i in plant\_list for j in path\_map\_list]

*# Go through the list of plants and create a Plant object which will store*

*# related predictions.*

for key in plant\_dict:

*# Call constructor to make new plant.*

tmp\_plant = Plant(code=key, name=plant\_dict[key])

*# Add to list if not present. Prevents duplicates.*

if not tmp\_plant.is\_in(plant\_objects):

plant\_objects.append(tmp\_plant)

def get\_parse\_pathway\_genes():

*"""*

*This function breaks the list of plant pathways into different lists in order for*

*different data to be processed at the same time using multithreading. Once all of*

*the threads have finished, then the list of genes by path will be looped through*

*in order to create both the gene data output files for each pathway and for the*

*master file that contains all of the gene information.*

*"""*

global path\_gene

global plant\_objects

global plant\_pathways

global thread\_lim

*# Chunk up the list of plant-pathway codes.*

sub\_lists = list\_partition(plant\_pathways, thread\_lim)

threads = [] *# Will be used to keep track of and kill off threads.*

*# A thread is made for each sub-list which then passes each sub-list as a parameter*

*# to the pathway parser.*

for sub\_list in sub\_lists:

thread = threading.Thread(target=path\_parse, args=(sub\_list,))

thread.start()

threads.append(thread)

for thread in threads: thread.join() *# Wait for each thread to die before continuing.*

master\_output = '' *# Data for the master file which holds all gathered gene data.*

for item in pathgenes:

tmp\_file = path\_gene + SEP + item.path + '.csv'

tmp\_output = '' *# Data for current gene's output file.*

for gene in item.genes:

*# Get the formatted string of gene information for the current and master files.*

tmp\_output += gene.simple() + '\n'

master\_output += gene.simple() + '\n'

write\_append(tmp\_file, tmp\_output, write\_over=True)

write\_append(path\_main + SEP + 'MasterList.csv', master\_output, write\_over=True)

def path\_parse(paths):

*"""*

*Given a pathway for a specific plant, the program then passes it to KEGG where it*

*retrieves the appropriate entry. The EC number, KO number and each GENE entry are*

*parsed from the data sent back by KEGG and then are appropriately saved by updating*

*the plant objects.*

*"""*

for path in paths:

global pathgenes

global plant\_objects

*print*(path) *# Not necessary, but is nice for a gauge of progress.*

with lock\_kegg:

raw = ''

*# Look for local raw gene file to save time by preventing the call to kegg.get().*

try:

with *open*(path\_raw\_gene + SEP + path + '.csv', 'r') as tmp\_read:

raw = tmp\_read.read()

tmp\_read.close()

*# No local raw gene data file found so the data must be downloaded.*

except *FileNotFoundError*:

raw = kegg.get(path)

with *open*(path\_raw\_gene + SEP + path + '.csv', 'w') as tmp\_get:

try: tmp\_get.write(raw)

*# Error should only ever occur for plant-pathway codes that don't exist.*

except *TypeError*: pass

tmp\_get.close()

gene\_entry = {}

if *isinstance*(raw, *str*):

if *len*(raw) > 0: *# Don't bother parsing the empty entries.*

kegg\_entry = kegg.parse(raw) *# Parses kegg entry into dictionary.*

gene\_entry = kegg\_entry.get(GKY) *# Get data from dictionary w/ key `GENE`.*

*# If gene\_entry is not empty, the statement will succeed. Otherwise ignore blanks.*

if gene\_entry:

plant\_code = ''.join(re.split(RE\_ALPH, path)) *# Plant code is alpha-only.*

plant\_name = plant\_dict.get(plant\_code)

with lock\_gene:

pathgenes.append(PathGene(path=path)) *# Make new PathGene & add to list.*

for key in gene\_entry:

tmp\_entry = gene\_entry[key]

try:

*# Get all EC nums from entry using regular expressions*

ec\_nums = re.findall(RE\_EC, tmp\_entry)

*# Remove unwanted characters and format each EC num.*

for i in *range*(0, *len*(ec\_nums)):

item = multi\_replace(ec\_nums[i], [('[', ''), (']', ''),

(':', ''), (' ', '')])

ec\_nums[i] = 'EC:' + item

*# Find and process the orthology ID.*

orthology = multi\_replace(quick\_fetch(RE\_KO, tmp\_entry),

[('[', ''), (']', '')])

*# Remove EC & KO in order to get compound name.*

name = re.sub(RE\_KO, '', (re.sub(RE\_EC, '', tmp\_entry)))

*# Call Gene constructor and pass in the parsed values.*

tmp\_gene = Gene(gene\_id=key, plant=plant\_name, ec\_nums=ec\_nums,

path=path, ortho=orthology, compound=name,

plant\_code=plant\_code)

with lock\_gene:

for pathgene in pathgenes: *# Update list of genes for current path.*

if pathgene.path == path: pathgene.genes.append(tmp\_gene)

with lock\_plant:

for index, plant in *enumerate*(plant\_objects):

if plant.name == tmp\_gene.plant:

tmp\_plant = plant

*# Add to list of plant genes if not present.*

if not tmp\_gene.is\_in(tmp\_plant.genes):

tmp\_plant.genes.append(tmp\_gene)

*# Add to list of all genes if not present.*

if not tmp\_gene.is\_in(all\_genes):

all\_genes.append(tmp\_gene)

*# Update the plant's EC nums. Dupes preferred*

*# (for the master count matrix).*

tmp\_plant.ec\_nums.extend(ec\_nums)

*# Update the plant list with the new/additional info*

*# for current plant.*

plant\_objects[index] = plant

except *IndexError*:

pass *# There was nothing found using regular expressions.*

def flavonoid\_predictions():

*"""*

*This is the function that goes through each plant, looks at the list of EC numbers*

*then applies a function in order to determine whether or not the plant has the*

*required EC numbers needed to synthesize each compound.*

*"""*

global plant\_objects

output\_list = '' *# Master list of all flavonoids and predicted plants.*

output\_yn = '' *# Master list of all flavonoids & a Y/N based on predicted status*

plant\_ec\_output = '' *# For outputting each plant & EC numbers from their gene entries.*

plant\_names = 'Name\n' *# Each plant's name, used in the first line of the Y/N output.*

for plant in plant\_objects:

unique\_nums = []

plant\_ec\_output += '\n' + plant.name + '\t'

plant\_names = plant\_names + plant.name + '\n'

*# Go through the plant's EC nums and add each EC number to the output once.*

for num in plant.ec\_nums:

if num not in unique\_nums:

plant\_ec\_output += num + '\t' *# Add EC number to the output*

unique\_nums.append(num) *# Ensures EC num isn't added to output again.*

for chem\_data in flav\_data\_lists: *# Make the call to the prediction functions.*

if predict.flav\_check(*getattr*(predict, chem\_data.code.lower()), unique\_nums):

*# Add the name of the plant, if predicted.*

chem\_data.plants.append(plant.name)

*# Create the formatted strings for the output prediction file.*

output\_yn = plant\_names.replace('\n', '\t')

for key in flav\_data\_lists:

save\_file([key.plants], key.file\_name, path\_chem, sep='\n')

output\_list = output\_list + '\n' + key.code

for plant in key.plants:

output\_list = output\_list + '\t' + plant

output\_yn = output\_yn + '\n' + key.code

for item in plant\_objects:

if item.name in key.plants:

output\_yn = output\_yn + '\tY'

else:

output\_yn = output\_yn + '\tN'

item\_count = *len*(key.plants)

*print*(key.label + ' predicted in ' + *str*(item\_count) + ' organisms.')

write\_append(path\_chem+SEP+'\_plant-ec-nums.tsv', plant\_ec\_output, write\_over=True)

write\_append(path\_chem+SEP+'\_predictions\_list.tsv', output\_list, write\_over=True)

write\_append(path\_chem+SEP+'\_predictions\_yn.tsv', output\_yn, write\_over=True)

def make\_plant\_ec\_counts():

*"""*

*This function creates and outputs a 'matrix' relating to each species and EC number*

*by running the fill\_matrix function on multiple threads. For each gene entry*

*containing a specific EC number, the program will increase the counter and display*

*it at the end next to the appropriate EC number.*

*"""*

global plant\_objects

for plant in plant\_objects:

fill\_count\_matrix(plant) *# Update master matrix using the info from each plant.*

out = '' *# Output string for the master EC count matrix.*

for plant in plant\_matrix:

out += plant.name + ':\t'

for count in plant.ec\_counts:

*# Update with EC num & its occurrence.*

out += *str*(count.number) + ' (' + *str*(count.count) + ')\t'

out += '\n'

write\_append(path\_main + SEP + 'MasterECCountMatrix.tsv', out, write\_over=True)

def fill\_count\_matrix(plant):

*"""*

*This function builds the ec counts for each list.*

*"""*

global plant\_objects

tmp\_plant = plant

for num in tmp\_plant.ec\_nums:

if tmp\_plant.has\_ec\_count(num):

tmp\_plant.incr\_ec\_count(num) *# Increment EC number's count.*

else: *# Need to create count object for current EC number.*

tmp\_count = EcCounts(number=num, count=1)

plant.ec\_counts.append(tmp\_count)

with lock\_plant: *# Update the count matrix.*

plant\_matrix.append(tmp\_plant)

def build\_nt\_fasta\_by\_ec():

*"""*

*This function uses multithreading and the information gathered from running*

*run\_path\_parse in order to get the FASTA/DNA sequence for each of the gene entries*

*that were found. As before, the program parses the list after all threads are done*

*and then created a FASTA file for each EC number and created the Master FASTA file.*

*"""*

*# Chunk up the list of all plant genes.*

sub\_lists = list\_partition(all\_genes, thread\_lim)

threads = [] *# For keeping track of each thread.*

*print*('getting data for ' + *str*(*len*(all\_genes)) + ' genes')

for sub\_list in sub\_lists: *# Make new thread for each sublist & pass to build\_fasta().*

thread = threading.Thread(target=build\_fasta, args=(sub\_list,))

thread.start()

threads.append(thread)

for thread in threads: thread.join() *# Don't continue until all threads are done.*

*print*('Starting to gather data for FASTA files...')

master\_fasta = path\_main + SEP + 'MasterFASTA.fasta'

master\_output = ''

for item in ec\_collections: *# just to make sure file is cleared before populating*

item.ec\_name = item.ec\_name.replace('.', '-').replace(':', '')

tmp\_file\_path = path\_fasta + SEP + item.ec\_name + '.fasta'

write\_append(tmp\_file\_path, '', write\_over=True)

for item in ec\_collections:

item.ec\_name = item.ec\_name.replace('.', '-').replace(':', '')

tmp\_file\_path = path\_fasta + SEP + item.ec\_name + '.fasta'

tmp\_output = ''

for entry in item.ec\_entries:

*# add to string to be printed into specific EC file*

tmp\_output += entry.simple() + '\n'

*# add to string to be printed into master FASTA*

master\_output += entry.simple() + '\n'

*# write the file for current EC number*

write\_append(tmp\_file\_path, tmp\_output, write\_over=False)

*# write the master FASTA file*

write\_append(master\_fasta, master\_output, write\_over=True)

*print*('Done making the FASTA files.')

def build\_fasta(genes):

*"""*

*This function uses the plant code and gene id in order to find the matching FASTA*

*sequence using the appropriate dbget url. The pages are saved into memory as HTML*

*and are parsed in order to extract the important information from the web page.*

*After parsing, the FASTA sequences are added to EcFastaCollection objects in order*

*to maintain proper association when writing all of the sequences out to files.*

*"""*

global ec\_collections

for gene in genes:

*# using the plant code and gene id create a string formatted as code:gene*

combined = gene.plant\_code.strip() + ':' + \

gene.gene\_id.replace('(RAP-DB) ', '').strip()

*# append code-gene string to the end of the dbget incomplete URL*

db\_url = URL\_DBGET + combined

url\_data = ''

with lock\_dbget:

tmp\_code = gene.plant\_code

tmp\_id = gene.gene\_id

try: *# Look for local raw fasta file to save time by not having to download.*

with *open*(path\_raw\_fasta + SEP + tmp\_code + tmp\_id + TF, 'r') as tmp\_read:

url\_data = tmp\_read.read()

tmp\_read.close()

*# No local raw fasta data file found so the data must be downloaded.*

except *FileNotFoundError*:

try:

*# read the html from the dbget url*

with request.urlopen(db\_url) as db\_site:

url\_data = db\_site.read().decode('utf-8')

with *open*(path\_raw\_fasta+SEP+tmp\_code+tmp\_id+TF, 'w') as tmp\_get:

tmp\_get.write(url\_data)

tmp\_get.close()

except (HTTPError, URLError) as url\_http\_err:

*print*('Something went wrong with error ' + url\_http\_err)

continue

*# get the header of the FASTA entry using regular expressions,*

*# &gt; is the HTML representation of >*

fasta\_header = ''.join(re.findall(RE\_NT\_HEAD, url\_data)).replace('&gt; ', '>')+\

' {' + plant\_dict[gene.plant\_code.strip()] + '}'

*# get the DNA sequence body using regular expressions*

fasta\_body = ''.join(re.findall(RE\_NT\_SEQ, url\_data))

full\_fasta\_entry = fasta\_header + '\n' + fasta\_body *# create FASTA entry string*

tmp\_entry = FastaEcEntry(plant=plant\_dict.get(gene.plant\_code),

gene=gene.gene\_id, dna=full\_fasta\_entry)

with lock\_ec:

for g in gene.ec\_nums:

tmp\_ec = EcFastaCollection(ec\_num=g, ec\_entries=[tmp\_entry])

ec\_collections.append(tmp\_ec)

if \_\_name\_\_ == '\_\_main\_\_':

main()

### Library file: data\_types.py

*# custom project library import*

from flib.fconstants import \*

class Flav:

*"""*

*This class holds the data for each flavonoid. The objects are initialized with their*

*file name and label and only later in the program, their empty list of plants will*

*be filled.*

*-------------------------------------------------------------------------------------*

*ATTRIBUTES*

*self.label: string that contains the flavonoids name*

*self.plants: list of plants predicted to produce the flavonoid*

*self.file\_name: string that holds the flavonoids output file name*

*-------------------------------------------------------------------------------------*

*FUNCTIONS*

*\_\_init\_\_: constructor for the object*

*\_\_eq\_\_: defines equality of the object*

*is\_in: determines if an identical or nearly identical object is already in the list*

*"""*

def *\_\_init\_\_*(*self*, label: *str*, plants: [str], file\_name: *str*, code=''):

*self*.plants = plants

*self*.label = label

*self*.file\_name = file\_name

*self*.code = code

def *\_\_eq\_\_*(*self*, other):

return *self*.plants == other.plants and \

*self*.label == other.label and \

*self*.file\_name == other.file\_name

def is\_in(*self*, items):

for item in items:

if *self* == item: return True

return False

class Plant:

*"""*

*This object holds information about each plant used in the program. The plant objects*

*are initialized with their scientific name and their code and then have different*

*information added later.*

*---------------------------------------------------------------------------------------*

*ATTRIBUTES*

*self.name: scientific name of the plant*

*self.code: KEGG code for the plant*

*self.genes: the gene entries for the plant*

*self.ec\_nums: the EC numbers parsed from the plants gene entries*

*self.flavonoids: the list of flavonoids that the plant could potentially produce*

*self.ec\_counts: list of objects that hold the number of times each EC number appears*

*---------------------------------------------------------------------------------------*

*FUNCTIONS*

*\_\_init\_\_: constructor for the object*

*\_\_eq\_\_: defines equality of the object*

*is\_in: determines if an identical or nearly identical object is already in the list*

*has\_ec\_count: used to determine whether or not a specific EC number is already in the*

*list of EC counts*

*incr\_ec\_count: used to increase the count for the EC count objects.*

*"""*

def *\_\_init\_\_*(*self*, name=None, code=None, genes=None, ec\_nums=None, flavonoids=None,

ec\_counts=None):

*self*.name = name if name is not None else ' '

*self*.code = code if code is not None else ' '

*self*.genes = genes if genes is not None else []

*self*.ec\_nums = ec\_nums if ec\_nums is not None else []

*self*.flavonoids = flavonoids if flavonoids is not None else []

*self*.ec\_counts = ec\_counts if ec\_counts is not None else []

def *\_\_eq\_\_*(*self*, other):

return *self*.name == other.name and \

*self*.code == other.code and \

*self*.genes == other.genes and \

*self*.ec\_nums == other.ec\_nums and \

*self*.flavonoids == other.flavonoids

def is\_in(*self*, items):

for item in items:

if *self* == item: return True

return False

def simple(*self*):

gstr = ''

for gene in *self*.genes: gstr += gene.no\_plant() + ' || '

return *self*.name + ', ' + *self*.code + ', ' + gstr + ', ' + \

*str*(*self*.ec\_nums) + ', ' + *str*(*self*.flavonoids)

def has\_ec\_count(*self*, ec\_number):

for ec in *self*.ec\_counts:

if ec.number == ec\_number: return True

return False

def incr\_ec\_count(*self*, ec\_number):

for ec in *self*.ec\_counts:

if ec.number == ec\_number:

ec.count += 1

class PathGene:

*"""*

*This object is used to hold Gene objects in a way such that they are sorted by the*

*pathway from which they were found.*

*--------------------------------------------------------------------------------------*

*ATTRIBUTES*

*self.path: the pathway that resulted in the gene entry*

*self.genes: the list of gene entries from this pathway*

*--------------------------------------------------------------------------------------*

*FUNCTIONS*

*\_\_init\_\_: constructor for the object*

*\_\_eq\_\_: defines equality of the object*

*is\_in: determines if an identical or nearly identical object is already in the list*

*"""*

def *\_\_init\_\_*(*self*, path=None, genes=None):

*self*.path = path if path is not None else ' '

*self*.genes = genes if genes is not None else []

def *\_\_eq\_\_*(*self*, other):

return *self*.path == other.path and \

*self*.genes == other.genes

def is\_in(*self*, items):

for item in items:

if *self* == item: return True

return False

class Gene:

*"""*

*This object holds data gathered from KEGG for each plant's pathway (like aip00491).*

*--------------------------------------------------------------------------------------*

*ATTRIBUTES*

*self.gene\_id: the ID of the gene from a plant*

*self.plant: the scientific name of the plant that has this gene*

*self.plant\_code: the KEGG code for the plant*

*self.compound: the compound name listed in the entry*

*self.ec\_nums: the list of EC numbers found in the entry*

*self.ortho: the KEGG orthology code for the compound*

*self.path: the pathway where the gene was found*

*--------------------------------------------------------------------------------------*

*FUNCTIONS*

*\_\_init\_\_: constructor for the object*

*\_\_eq\_\_: defines equality of the object*

*is\_in: determines if an identical or nearly identical object is already in the list*

*simple: returns a formatted string that contains information from the object*

*no\_plant: same as simple, but without including the plant name*

*"""*

def *\_\_init\_\_*(*self*, gene\_id=None, plant=None, compound=None, ec\_nums=None,

ortho=None, path=None, plant\_code=None):

*self*.gene\_id = gene\_id if gene\_id is not None else ' '

*self*.plant = plant if plant is not None else ' '

*self*.plant\_code = plant\_code if plant\_code is not None else ' '

*self*.compound = compound if compound is not None else ' '

*self*.ec\_nums = ec\_nums if ec\_nums is not None else []

*self*.ortho = ortho if ortho is not None else ' '

*self*.path = path if path is not None else ' '

def *\_\_eq\_\_*(*self*, other):

return *self*.gene\_id == other.gene\_id and \

*self*.plant\_code == other.plant\_code and \

*self*.ec\_nums == other.ec\_nums

def is\_in(*self*, items):

for item in items:

if *self* == item: return True

return False

def simple(*self*):

return *self*.plant + ', ' + *self*.gene\_id + ', ' + *self*.compound + ', ' + \

*str*(*self*.ec\_nums) + ', ' + *self*.ortho

def no\_plant(*self*):

return *self*.gene\_id + ', ' + *self*.compound + ', ' + *self*.ec\_nums + ', ' + *self*.ortho

class EcFastaCollection:

*"""*

*This object is used to hold the associated FASTA entries for any given EC number.*

*--------------------------------------------------------------------------------------*

*ATTRIBUTES*

*self.ec\_name: the EC number & name used when writing the file*

*self.ec\_entries: the list of associated FASTA entries (FastaEcEntry objects)*

*--------------------------------------------------------------------------------------*

*FUNCTIONS*

*\_\_init\_\_: constructor for the object*

*\_\_eq\_\_: defines equality of the object*

*is\_in: determines if an identical or nearly identical object is already in the list*

*"""*

def *\_\_init\_\_*(*self*, ec\_num=None, ec\_entries=None):

*self*.ec\_name = ec\_num if ec\_num is not None else ' '

*self*.ec\_entries = ec\_entries if ec\_entries is not None else []

def *\_\_eq\_\_*(*self*, other):

return *self*.ec\_name == other.ec\_name and \

*self*.ec\_entries == other.ec\_entries

def is\_in(*self*, items):

for item in items:

if *self* == item: return True

return False

class EcCounts:

*"""*

*This object is a property of the Plant object and is used to hold each EC number and*

*the number of times it occurs in gene entries of a given plant.*

*--------------------------------------------------------------------------------------*

*ATTRIBUTES*

*self.number: the EC number*

*self.count: number of times that the EC number shows up in gene entries.*

*--------------------------------------------------------------------------------------*

*FUNCTIONS*

*\_\_init\_\_: constructor for the object*

*"""*

def *\_\_init\_\_*(*self*, number=None, count=None):

*self*.number = number if number is not None else ' '

*self*.count = count if number is not None else 0

class FastaEcEntry:

*"""*

*This object is a property of EcFastaCollection and contains the information for a*

*specific FASTA entry.*

*--------------------------------------------------------------------------------------*

*ATTRIBUTES*

*self.gene\_id: the gene ID associated with the sequence*

*self.plant: the plant that the gene is from*

*self.dna\_seq: the dna sequence/FASTA entry for the specific gene*

*--------------------------------------------------------------------------------------*

*FUNCTIONS*

*\_\_init\_\_: constructor for the object*

*\_\_eq\_\_: defines equality of the object*

*is\_in: determines if an identical or nearly identical object is already in the list*

*simple: returns a formatted string*

*"""*

def *\_\_init\_\_*(*self*, gene=None, dna=None, plant=None):

*self*.gene\_id = gene if gene is not None else ' '

*self*.dna\_seq = dna if dna is not None else ' '

*self*.plant = plant if dna is not None else ' '

def *\_\_eq\_\_*(*self*, other):

return *self*.gene\_id == other.gene\_id and \

*self*.dna\_seq == other.dna\_seq and \

*self*.plant == other.plant

def is\_in(*self*, items):

for item in items:

if *self* == item: return True

return False

def simple(*self*):

return *self*.dna\_seq

### Library file: fconstants.py

*# custom project library imports*

from flib.futil import \*

from flib.data\_types import Flav

*# other imports*

import os

import sys

SEP = os.sep *# get the right slash. / for linux & mac, \ for windows*

*# misc strings*

CSV = '.csv'

GDATA = 'Gene\_data\_'

JKEY = 'obj'

NIX = ''

NL = '\n'

SP = ' '

TF = '.txt'

URL\_DBGET = 'https://www.kegg.jp/dbget-bin/www\_bget?-f+-n+n+'

*# output directories*

DIR\_CHEM = SEP + 'Chemical\_Data'

DIR\_FASTA = SEP + 'FASTA\_Data'

DIR\_GENE = SEP + 'Gene\_Data'

DIR\_RGENE = DIR\_GENE + SEP + 'Raw\_Gene'

DIR\_RFASTA = DIR\_FASTA + SEP + 'Raw\_FASTA'

*# pathways for the json data (and the name of the json object used in all files)*

JSON\_DIR = '..' + SEP + 'json\_data' + SEP

DIR\_FJSON = 'fjson' + SEP

JSON\_FLAVS = DIR\_FJSON + 'flav\_names.json'

JSON\_FLAV\_REL = DIR\_FJSON + 'flav\_related.json'

JSON\_FLAV\_SYN = DIR\_FJSON + 'flav\_syns.json'

JSON\_PATH\_DICT = DIR\_FJSON + 'path\_codes\_names.json'

JSON\_PATH\_LIST = DIR\_FJSON + 'path\_codes.json'

JSON\_PLANT\_DICT = DIR\_FJSON + 'plant\_names\_codes.json'

JSON\_PLANT\_DICT\_COMMON = DIR\_FJSON + 'plant\_names\_codes\_plus.json'

JSON\_PLANT\_LIST = DIR\_FJSON + 'plant\_codes.json'

JSON\_TST\_MED = DIR\_FJSON + 'test\_med.json'

JSON\_TST\_SHORT = DIR\_FJSON + 'test\_short.json'

JSON\_TST\_SINGLE = DIR\_FJSON + 'test\_single.json'

JSON\_FLAV\_NAMES = DIR\_FJSON + 'flav\_info.json'

*# lists and dictionaries made from JSON files*

flav\_list = get\_json\_data(JSON\_FLAVS, JKEY) *# list of flavonoids of interest (FOI)*

flav\_names\_info = get\_json\_data(JSON\_FLAV\_NAMES)

flav\_relatives = get\_json\_data(JSON\_FLAV\_REL) *# relatives of FOI*

flav\_synonyms = get\_json\_data(JSON\_FLAV\_SYN) *# synonyms for FOI*

path\_map\_dict = get\_json\_data(JSON\_PATH\_DICT, JKEY) *# pathway names and codes*

path\_map\_list = get\_json\_data(JSON\_PATH\_LIST, JKEY) *# pathway codes*

plant\_dict\_common = get\_json\_data(JSON\_PLANT\_DICT\_COMMON, JKEY) *# common & sci. names*

plant\_dict\_reg = get\_json\_data(JSON\_PLANT\_DICT, JKEY) *# scientific names and plant codes*

plant\_full\_list = get\_json\_data(JSON\_PLANT\_LIST, JKEY) *# full list of plant codes*

test\_med = get\_json\_data(JSON\_TST\_MED, JKEY) *# testing list of plant codes*

test\_short = get\_json\_data(JSON\_TST\_SHORT, JKEY) *# testing list of plant codes*

test\_single = get\_json\_data(JSON\_TST\_SINGLE, JKEY) *# testing list of plant codes*

plant\_dict = plant\_dict\_reg *# variable exists for ease of value change when testing*

plant\_list = plant\_full\_list *# variable exists for ease of value change when testing*

flav\_data\_lists = []

for k in flav\_names\_info:

info = flav\_names\_info.get(k)

tmp\_flav = Flav(k, [], info.get('file'), info.get('code'))

flav\_data\_lists.append(tmp\_flav)

*# keys for accessing dictionaries*

EKY = 'EC'

GKY = 'GENE'

NKY = 'NTSEQ'

OKY = 'ORTHOLOGY'

PKY = 'PLANT'

README = SEP + 'ReadMe.txt'

*# EC numbers*

E01 = 'EC:4.3.1.24'

E02 = 'EC:4.3.1.25'

E03 = 'EC:1.14.14.91'

E04 = 'EC:6.2.1.12'

E05 = 'EC:2.3.1.170'

E06 = 'EC:2.3.1.133'

E07 = 'EC:1.14.14.96'

E08 = 'EC:1.14.13.-'

E09 = 'EC:2.3.1.74'

E10 = 'EC:5.5.1.6'

E11 = 'EC:1.14.20.5'

E12 = 'EC:1.14.19.76'

E13 = 'EC:1.14.14.81'

E14 = 'EC:1.14.14.82'

E15 = 'EC:1.14.11.9'

E16 = 'EC:1.14.20.6'

E17\_1 = 'EC:1.1.1.219'

E17\_2 = 'EC:1.1.1.234'

E17\_FULL = 'EC:1.1.1.219 1.1.1.234'

E18 = 'EC:1.14.20.4'

E19 = 'EC:1.3.1.77'

E20 = 'EC:1.17.1.3'

E21 = 'EC:1.14.14.87'

E22 = 'EC:4.2.1.105'

E23 = 'EC:2.4.1.357'

E24 = 'EC:1.3.1.117'

E\_DEC = 'EC:2.3.1.70'

E\_GGT = 'EC:2.4.1.74'

E\_SOA = 'EC:2.3.1.30'

E\_V1G = 'EC:2.4.1.136'

### Library file: futil.py

import json

import os

import re

SEP = os.sep *# get the right slash. / for linux, \ for windows*

JKEY = 'obj'

def get\_json\_data(file\_name, key=None):

*"""*

*This function uses the python JSON library in order to parse JSON files into usable*

*python objects. Can return lists or dictionaries, depending on the JSON file's*

*structure.*

*"""*

if key is None: key = JKEY

data = ''

try:

with *open*(file\_name) as jsonFile:

data = json.load(jsonFile)

except *FileNotFoundError*:

file\_name = '..' + SEP + file\_name

with *open*(file\_name) as jsonFile:

data = json.load(jsonFile)

return data[key]

def remove\_dupes(dupe\_list):

*"""*

*removes duplicate elements*

*"""*

unique\_list = [] *# creates an empty list*

for item in dupe\_list:

*# adds item to empty list if it's not already in the list*

if item not in unique\_list: unique\_list.append(item)

return unique\_list

def unique\_element\_list(list\_name, index):

*"""*

*find unique EC numbers so have a generic function and run it*

*be careful as there are cases of one less item - use "last" to fix that problem here*

*"""*

original\_index = index

element\_list = []

*# assigns the string "last" to the very last list in the list of lists*

for i in list\_name:

*# finds unique EC num not in the list & adds it to the list*

if original\_index == 'last': index = *len*(i) - 1

if i[*int*(index)] not in element\_list: element\_list.append(i[*int*(index)])

return element\_list

def list\_partition(seq, num):

*"""*

*This function takes in a list and then splits it into as many parts as specified*

*by parameter num*

*"""*

avg = *len*(seq) / *float*(num)

out = []

last = 0.0

while last < *len*(seq):

out.append(seq[*int*(last):*int*(last + avg)])

last += avg

return out

def write\_readme(main\_dir, readme, init\_time, fasta\_path, gene\_path):

*"""*

*Creates the ReadMe file*

*"""*

*print*('- creating README...')

with *open*(main\_dir + readme, 'w') as readme\_doc:

readme\_doc.write("KEGG\_v1p1.py\n")

readme\_doc.write(init\_time.strftime("%m-%d-%Y") + "\n")

readme\_doc.write(main\_dir + "\n")

readme\_doc.write('This script creates a series of files related to the genes ' +

'associated with plant flavonoids from various species of ' +

'plants. This script first creates the MasterCount and ' +

'MasterList files; the MasterCount counts the number genes ' +

'each plant species have that correspond with each EC number; ' +

'while the MasterList lists every gene with number for each ' +

'plant specie. These are located in ' + os.getcwd() + '. The ' +

'script also creates files that only contains the genes of a ' +

'single plant species biochemical pathway which are located in ' +

gene\_path + '. The script also creates a Master FASTA files ' +

'which contains the DNA sequence of each gene and FASTA files ' +

'organized by EC number, these are located in ' + fasta\_path)

readme\_doc.close()

def save\_file(lists\_to\_write, output\_dir, current, sep=', '):

*"""*

*This function takes a list or list of lists and then writes its contents to a file.*

*"""*

os.chdir(current)

writedoc = *open*(output\_dir, 'w') *# open file to be written*

for line in lists\_to\_write:

for item in line:

item = *str*(item).replace('\n', '') *# removes the new lines in each list of list*

if item == '':

writedoc.write('-') *# if the list in the list of list is empty writes a dash*

else:

writedoc.write(item) *# write the entry in the list of lists to the file*

writedoc.write(sep) *# tab delimited; use "," for csv files*

writedoc.write('\n')

writedoc.close()

def write\_append(path, content, write\_over=None, skip=None):

*"""*

*This function takes a filename and the contents to be written to a file. If the*

*file doesn't exist, it is created then written to. If it does exist, then it is*

*appended to. The optional arg write\_over is used for when the file exists and the*

*content needs to be written over.*

*"""*

try:

file = *open*(path, 'x')

file.close()

file = *open*(path, 'w')

file.write(content)

file.close()

except *FileExistsError*:

if write\_over:

file = *open*(path, 'w')

file.write(content)

file.close()

elif skip:

pass

else:

file = *open*(path, 'a')

file.write(content)

file.close()

def is\_http\_error(msg):

*"""*

*Checks if a string is an HTTP error.*

*"""*

if *str*(msg).strip() in HTTP\_ERRS:

return False

else:

return True

def init\_dir(dir\_path):

*"""*

*Initializes directories for keggv2.py*

*"""*

*# replaced WindowsError with OSError for more general usage.*

*# try to make data directories and handle any errors*

try:

os.mkdir(dir\_path)

except *OSError*:

pass

def quick\_fetch(pattern, line):

*"""*

*Fetch one item from re.findall and return as a string.*

*"""*

out = ''

try:

out = re.findall(pattern, line)[0]

except *IndexError*:

out = ''

return out

def multi\_replace(line, pairs):

*"""*

*This function makes multiple string replacements.*

*"""*

for pair in pairs: line = line.replace(pair[0], pair[1])

return line

def skin(line):

*"""*

*Simple function that removes ALL whitespace from a string*

*"""*

return ''.join(line.split()).strip()

### Library file: prediction\_logic.py

*# custom project library imports*

from flib.fconstants import \*

*# take the label, which is the same as the function name & then call the function.*

def flav\_check(label, ec\_list):

try:

res = label(ec\_list)

except *KeyError*:

res = False

return res

*# returns true if at least 1 arg is in the list*

def or\_in(items, \*args):

for a in args:

if a in items:

return True

return False

*# returns true only if all args are in the list*

def and\_in(items, \*args):

for a in args:

if a not in items:

return False *# all values must be present*

return True

*# The prediction functions. There is one for numerous compounds not being predicted*

*# as they are precursors to the compounds of interest, which makes the predictions*

*# a little more readable. Function names are either the 3 letter code assigned*

*# to a specific compound by the Protein Data Bank (PDB) or are alternative codes*

*# for compounds that either do not have a PDB code or do not have a variable-friendly*

*# PDB code (for example, the PDB code for Epicatechin cannot be used because it is*

*# 28E and the python interpreter does not allow functions to start with a number)*

*# Function order is determined by a compounds requirements as functions cannot*

*# be called before they are written.*

def tca(e): *# cinnamic acid || precursor*

return or\_in(e, E01, E02)

def hc4(e): *# p-coumaric acid || precursor*

return tca(e) and (E03 in e)

def nca(e): *# cinnamoyl-coa || precursor*

return tca(e) and (E04 in e)

def pich(e): *# pinocembrin chalcone || new*

return nca(e) and (E09 in e)

def pino(e): *# pinocembrin || new*

return pich(e) and (E10 in e)

def chsn(e): *# chrysin || new*

return pino(e) and (E11 in e)

def pban(e): *# pinobanksin || new*

return pino(e) and (E15 in e)

def galn(e): *# galangin || new*

return pban(e) and (E16 in e)

def wca(e): *# p-coumaroyl-coa || precursor*

return (nca(e) and (E03 in e)) or (hc4(e) and (E04 in e))

def g50(e): *# phloretin || new*

return wca(e) and (and\_in(e, E24, E09))

def narc(e): *# naringenin chalcone || new*

return wca(e) and (E09 in e)

def nar(e): *# naringenin || mh thesis*

return narc(e) and (E10 in e)

def dgen(e): *# 2-Hydroxy-2,3-dihydrogenistein || new*

return nar(e) and (E21 in e)

def hesp(e): *# hesperetin || new*

return nar(e)

def apif(e): *# apiforol || new*

return nar(e) and (or\_in(e, E17\_1, E17\_FULL, E17\_2))

def agi(e): *# apigenin || mh thesis*

return nar(e) and or\_in(e, E11, E12)

def lu2(e): *# luteolin || mh thesis*

return agi(e) and (or\_in(e, E13, E14))

def fca(e): *# caffeoyl-coa || precursor*

return wca(e) and (and\_in(e, E06, E07) or (E08 in e))

def erdc(e): *# eriodictyol chalcone || new*

return (nar(e) and or\_in(e, E13, E14)) or (fca(e) and (E09 in e))

def erd(e): *# eriodictyol || mh thesis || same as erch for now*

return erdc(e)

def lutf(e): *# luteoforol || new*

return erd(e) and (or\_in(e, E17\_1, E17\_FULL, E17\_2))

def dtri(e): *# dihydrotricetin || new*

return erd(e) and (E13 in e)

def myf(e): *# tricetin || new*

return lutf(e) and (E13 in e) or (dtri(e) and (E11 in e))

def dkam(e): *# dihydrokaempferol || new*

return nar(e) and (E15 in e)

def lpel(e): *# leucopelargonidin || new*

return dkam(e) and (E17\_1 in e)

def pelr(e): *# pelargonidin || new*

return lpel(e) and (E18 in e)

def ezel(e): *# epiafzelechin || new*

return pelr(e) and (E19 in e)

def azel(e): *# afzelechin || new*

return lpel(e) and (E20 in e)

def kmp(e): *# kaempferol || mh thesis*

return dkam(e) and (E16 in e)

def dque(e): *# dihydroquercetin || new*

return (dkam(e) and or\_in(e, E13, E14)) or (erd(e) and (E15 in e))

def lcyn(e): *# leucocyanidin || new*

return dque(e) and or\_in(e, E17\_1, E17\_FULL)

def que(e): *# quercetin || mh thesis*

return (kmp(e) and or\_in(e, E13, E14)) or (dque(e) and (E16 in e))

def kxn(e): *# catechin || mh thesis*

return lcyn(e) and (E20 in e)

def hwb(e): *# cyanidin || mh thesis*

return lcyn(e) and (E18 in e)

def ec(e): *# epicatechin || mh thesis*

return hwb(e) and (E19 in e)

def dmyr(e): *# dihydromyricetin || new*

return (dque(e) and (E13 in e)) or (erd(e) and and\_in(e, E13, E15))

def myc(e): *# myricetin || mh thesis*

return (dmyr(e) and (E16 in e)) or (que(e) and (E13 in e))

def ldel(e): *# leucodelphinidin || new*

return dmyr(e) and or\_in(e, E17\_1, E17\_FULL)

def dlm(e): *# delphinidin || new*

return ldel(e) and (E18 in e)

def gc(e): *# gallocatechin || mh thesis*

return ldel(e) and (E20 in e)

def egt(e): *# epigallocatechin || mh thesis*

return dlm(e) and (E19 in e)

def gen(e): *# genistein || mh thesis*

return nar(e) and and\_in(e, E21, E22)

def bun(e): *# butein || mh thesis*

return wca(e) and (or\_in(e, E05, E09))

def butn(e): *# butin || new*

return bun(e) and (E10 in e)

def hcc(e): *# isoliquiritigenin || mh thesis*

return bun(e)

def dfv(e): *# liquiritigenin || new*

return hcc(e) and (E10 in e)

def gban(e): *# garbanzol || new*

return dfv(e) and (E15 in e)

def fstn(e): *# dihydrofisetin || new*

return (gban(e) and (E14 in e)) or (butn(e) and (E15 in e))

def df74(e): *# 7,4'-dihydroxyflavone || new*

return dfv(e) and (or\_in(e, E11, E12))

def tiso(e): *# 2,7,4'-Trihydroxyisoflavanone || new*

return dfv(e) and (E21 in e)

def daid(e): *# daidzein || new*

return tiso(e) and (E22 in e)

def tnon(e): *# 6,7,4'-Trihydroxyflavanone || new*

return dfv(e) and (E08 in e)

def tet2(e): *# 2,6,7,4'-Tetrahydroxyisoflavanone || new*

return tnon(e) and (E21 in e)

def hdai(e): *# 6-Hydroxydaidzein || new*

return tet2(e)

def ggt(e): *# Glycosaminoglycan galactosyltransferase || enzyme*

return E\_GGT in e

def dec(e): *# Deleted entry || enzyme*

return E\_DEC in e

def soa(e): *# Serine O-acetyltransferase || enzyme*

return E\_SOA in e

def v1g(e): *# vanillate 1-glucosyltransferase || enzyme*

return E\_V1G in e

## Data files

### flav\_names.json

{

"obj": [

"Apigenin", "Butein", "Catechin", "Cyanidin", "Epicatechin", "Epigallocatechin",

"Eriocitrin", "Eriodictyol", "Gallocatechin", "Genistein", "Isoliquiritigenin",

"Kaempferol", "Luteolin", "Myricetin", "Naringenin", "Quercetin"

]

}

### flav\_info.json

{

"obj": {

"2,6,7,4'-Tetrahydroxyisoflavanone": {

"code": "TET2",

"file": "TET2\_2-6-7-4-tetrahydroxyisoflavanone.csv"

},

"2,7,4'-Trihydroxyisoflavanone": {

"code": "TISO",

"file": "TISO\_2-7-4-trihydroxyisoflavanone.csv"

},

"2-Hydroxy-2,3-dihydrogenistein": {

"code": "DGEN",

"file": "DGEN\_2-hydroxy-2-3-dihydrogenistein.csv"

},

"6,7,4'-Trihydroxyflavanone": {

"code": "TNON",

"file": "TNON\_6-7-4-trihydroxyflavanone.csv"

},

"6-Hydroxydaidzein": {"code": "HDAI", "file": "HDAI\_6-hydroxydaidzein.csv"},

"7,4'-Dihydroxyflavone": {"code": "DF74", "file": "DF74\_7-4-dihydroxyflavone.csv"},

"Afzelechin": {"code": "AZEL", "file": "AZEL\_afzelechin.csv"},

"Apiforol": {"code": "APIF", "file": "APIF\_apiforol.csv"},

"Apigenin": {"code": "AGI", "file": "AGI\_apigenin.csv"},

"Butein": {"code": "BUN", "file": "BUN\_butein.csv"},

"Butin": {"code": "BUTN", "file": "BUTN\_butin.csv"},

"Catechin": {"code": "KXN", "file": "KXN\_catechin.csv"},

"Chrysin": {"code": "CHSN", "file": "CHSN\_chrysin.csv"},

"Cyanidin": {"code": "HWB", "file": "HWB\_cyanidin.csv"},

"Daidzein": {"code": "DAID", "file": "DAID\_daidzein.csv"},

"Delphinidin": {"code": "DLM", "file": "DLM\_delphinidin.csv"},

"Dihydrofisetin": {"code": "FSTN", "file": "FSTN\_dihydrofisetin.csv"},

"Dihydrokaempferol": {"code": "DKAM", "file": "DKAM\_dihydrokaempferol.csv"},

"Dihydromyricetin": {"code": "DMYR", "file": "DMYR\_dihydromyricetin.csv"},

"Dihydroquercetin": {"code": "DQUE", "file": "DQUE\_dihydroquercetin.csv"},

"Dihydrotricetin": {"code": "DTRI", "file": "DTRI\_dihydrotricetin.csv"},

"Enzyme Deleted entry": {"code": "DEC", "file": "EC\_2-3-1-70.csv"},

"Enzyme Glycosaminoglycan galactosyltransferase": {

"code": "GGT",

"file": "EC\_2-4-1-74.csv"

},

"Enzyme Serine O-acetyltransferase": {"code": "SOA", "file": "EC\_2-3-1-30.csv"},

"Enzyme Vanillate 1-glucosyltransferase": {"code": "V1G", "file": "EC\_2-4-1-136.csv"},

"Epiafzelechin": {"code": "EZEL", "file": "EZEL\_epiafzelechin.csv"},

"Epicatechin": {"code": "EC", "file": "EC\_epicatechin.csv"},

"Epigallocatechin": {"code": "EGT", "file": "EGT\_epigallocatechin.csv"},

"Eriodictyol Chalcone": {"code": "ERDC", "file": "ERDC\_eriodictyol-chalcone.csv"},

"Eriodictyol": {"code": "ERD", "file": "ERD\_eriodictyol.csv"},

"Galangin": {"code": "GALN", "file": "GALN\_galangin.csv"},

"Gallocatechin": {"code": "GC", "file": "GC\_gallocatechin.csv"},

"Garbanzol": {"code": "GBAN", "file": "GBAN\_garbanzol.csv"},

"Genistein": {"code": "GEN", "file": "GEN\_genistein.csv"},

"Hesperetin": {"code": "HESP", "file": "HESP\_hesperetin.csv"},

"Isoliquiritigenin": {"code": "HCC", "file": "HCC\_isoliquiritigenin.csv"},

"Kaempferol": {"code": "KMP", "file": "KMP\_kaempferol.csv"},

"Leucocyanidin": {"code": "LCYN", "file": "LCYN\_leucocyanidin.csv"},

"Leucodelphinidin": {"code": "LDEL", "file": "LDEL\_leucodelphinidin.csv"},

"Leucopelargonidin": {"code": "LPEL", "file": "LPEL\_leucopelargonidin.csv"},

"Liquiritigenin": {"code": "DFV", "file": "DFV\_liquiritigenin.csv"},

"Luteoforol": {"code": "LUTF", "file": "LUTF\_luteoforol.csv"},

"Luteolin": {"code": "LU2", "file": "LU2\_luteolin.csv"},

"Myricetin": {"code": "MYC", "file": "MYC\_myricetin.csv"},

"Naringenin Chalcone": {"code": "NARC", "file": "NARC\_naringenin-chalcone.csv"},

"Naringenin": {"code": "NAR", "file": "NAR\_naringenin.csv"},

"Pelargonidin": {"code": "PELR", "file": "PELR\_pelargonidin.csv"},

"Phloretin": {"code": "G50", "file": "G50\_phloretin.csv"},

"Pinobanksin": {"code": "PBAN", "file": "PBAN\_pinobanksin.csv"},

"Pinocembrin chalcone": {"code": "PICH", "file": "PICH\_pinocembrin\_chalcone.csv"},

"Pinocembrin": {"code": "PINO", "file": "PINO\_pinocembrin.csv"},

"Quercetin": {"code": "QUE", "file": "QUE\_quercetin.csv"},

"Tricetin": {"code": "MYF", "file": "MYF\_tricetin.csv"}

}

}

### plant\_codes.json

{

"obj": [

"adu", "aip", "aly", "aof", "apro", "ath", "atr", "ats", "bdi", "bna", "boe", "bpg",

"brp", "bvg", "cam", "cann", "ccaj", "ccav", "ccp", "cic", "cit", "cmax", "cme",

"cmo", "cmos", "cpap", "cpep", "cqi", "crb", "cre", "csat", "csl", "csv", "cvr",

"dcr", "dct", "dzi", "egr", "egu", "eus", "fve", "gab", "ghi", "gmx", "gra", "gsj",

"gsl", "han", "hbr", "ini", "jcu", "jre", "lang", "lja", "lsv", "mcha", "mdm", "mesc",

"mis", "mng", "mpp", "mtr", "mus", "nau", "nnu", "nsy", "nta", "nto", "obr", "oeu",

"olu", "dosa", "osa", "ota", "pavi", "pda", "peq", "peu", "pmum", "pop", "pper",

"ppp", "psom", "pvu", "pxb", "qsu", "rcn", "rcu", "sbi", "sind", "sita", "sly", "smo",

"soe", "sot", "spen", "tcc", "thj", "var", "vcn", "vra", "vun", "vvi", "zju", "zma",

"rsz"

]

}

### plant\_names\_codes.json

{

"obj": {

"ats": "Aegilops tauschii", "atr": "Amborella trichopoda",

"aly": "Arabidopsis lyrata", "ath": "Arabidopsis thaliana",

"adu": "Arachis duranensis", "aip": "Arachis ipaensis",

"aof": "Asparagus officinalis", "apro": "Auxenochlorella protothecoides",

"bpg": "Bathycoccus prasinos", "bvg": "Beta vulgaris",

"bdi": "Brachypodium distachyon", "bna": "Brassica napus", "boe": "Brassica oleracea",

"brp": "Brassica rapa", "ccaj": "Cajanus cajan", "csat": "Camelina sativa",

"crb": "Capsella rubella", "cann": "Capsicum annuum", "cpap": "Carica papaya",

"cqi": "Chenopodium quinoa", "cre": "Chlamydomonas reinhardtii",

"cvr": "Chlorella variabilis", "ccp": "Chondrus crispus", "cam": "Cicer arietinum",

"cic": "Citrus clementina", "cit": "Citrus sinensis",

"csl": "Coccomyxa subellipsoidea", "cmo": "Cucumis melo", "csv": "Cucumis sativus",

"cmax": "Cucurbita maxima", "cmos": "Cucurbita moschata",

"cpep": "Cucurbita pepo subsp. pepo", "cme": "Cyanidioschyzon merolae",

"ccav": "Cynara cardunculus var. scolymus", "dcr": "Daucus carota",

"dct": "Dendrobium catenatum", "dzi": "Durio zibethinus", "egu": "Elaeis guineensis",

"egr": "Eucalyptus grandis", "eus": "Eutrema salsugineum", "fve": "Fragaria vesca",

"gsl": "Galdieria sulphuraria", "gmx": "Glycine max", "gab": "Gossypium arboreum",

"ghi": "Gossypium hirsutum", "gra": "Gossypium raimondii", "han": "Helianthus annuus",

"hbr": "Hevea brasiliensis", "ini": "Ipomoea nil", "jcu": "Jatropha curcas",

"jre": "Juglans regia", "lsv": "Lactuca sativa", "lja": "Lotus japonicus",

"lang": "Lupinus angustifolius", "mdm": "Malus domestica",

"mesc": "Manihot esculenta", "mis": "Micromonas commoda", "mpp": "Micromonas pusilla",

"mcha": "Momordica charantia", "mng": "Monoraphidium neglectum",

"mus": "Musa acuminata", "nnu": "Nelumbo nucifera", "nau": "Nicotiana attenuata",

"nsy": "Nicotiana sylvestris", "nta": "Nicotiana tabacum",

"nto": "Nicotiana tomentosiformis", "oeu": "Olea europaea v. sylvestris",

"obr": "Oryza brachyantha", "dosa": "Oryza sativa japonica RAPDB",

"osa": "Oryza sativa japonica RefSeq", "olu": "Ostreococcus lucimarinus",

"ota": "Ostreococcus tauri", "psom": "Papaver somniferum",

"peq": "Phalaenopsis equestris", "pvu": "Phaseolus vulgaris",

"pda": "Phoenix dactylifera", "ppp": "Physcomitrium patens subsp. Patens",

"pop": "Populus trichocarpa", "peu": "Populus euphratica", "pavi": "Prunus avium",

"pmum": "Prunus mume", "pper": "Prunus persica", "pxb": "Pyrus x bretschneideri",

"qsu": "Quercus suber", "rcu": "Ricinus communis",

"smo": "Selaginella moellendorffii", "sind": "Sesamum indicum",

"sita": "Setaria italica", "sly": "Solanum lycopersicum", "spen": "Solanum pennellii",

"sot": "Solanum tuberosum", "sbi": "Sorghum bicolor", "soe": "Spinacia oleracea",

"thj": "Tarenaya hassleriana", "tcc": "Theobroma cacao", "var": "Vigna angularis",

"vra": "Vigna radiata", "vvi": "Vitis vinifera",

"vcn": "Volvox carteri f. nagariensis", "zma": "Zea mays", "zju": "Ziziphus jujuba",

"gsj": "Glycine soja", "mtr": "Medicago truncatula", "rcn": "Rosa chinensis",

"vun": "Vigna unguiculata", "rsz": "Raphanus sativus"

}

}

### path\_codes.json

{

"obj": ["00940", "00941", "00942", "00943","00944"]

}

### path\_codes\_names.json

{

"obj": {

"00940": "phenylpropanoids",

"00941": "flavonoids",

"00942": "anthocyanins",

"00943": "isoflavonoids",

"00944": "flavones/flavonols",

"00945": "stilbenoids"

}

}

### Additional JSON files

There are some additional JSON files not listed here as they do not play a direct role in code functionality and can be found on the project GitHub.

###### 