 doi:10.20944/preprints202202.0305.v1

Data Descriptor

Theoretical calculations and analysis method of the physico-chemical properties of phytochemicals to predict gastrointestinal absorption

Luis A. Vélez[1], Yamixa Delgado[1,*], Yancy Ferrer-Acosta[2], Ivette J. Suárez-Arroyo[2], Priscilla Rodríguez[1], Daraishka Pérez[1]

[1] San Juan Bautista School of Medicine, Urb. Turabo Gardens PR-172 Caguas, Puerto Rico, USA 00726; luisvl@sanjuanbautista.edu (L.A.V.); ydelgado@sanjuanbautista.edu (Y.D.); priscillamrt@sanjuanbautista.edu (P.R.); daraishkapc@sanjuanbautista.edu (D.P.)

[2] Universidad Central del Caribe, School of Medicine, Ave. Laurel 100, Sta. Juanita, Bayamón, Puerto Rico, USA 00960; yancy.ferrer@uccaribe.edu (Y.F.A); (ivette.suarez@uccaribe.edu (I.J.S.A.)

* Correspondence: ydelgado@sanjuanbautista.edu; Tel.: (1+787-743-3038 X. 3220)

Abstract: The discovery of more bioactive compounds for non-invasive administration has been the goal of research groups focused on pharmacotherapy. Phytonutrients have always been attractive for researchers because they are a significant source of bioactive phytochemicals, but it is challenging to determine which components show high biomedical activity. However, based on the chemical structure of these active compounds, their physicochemical properties can be calculated to predict the probability of gastrointestinal (GI) absorption after oral administration. Indeed, different researchers have proposed several rules (e.g., Lipinski's, Veber's, Ghose's, and Muegge's rules) to attain these predictions for synthetic compounds. Most phytochemicals do not fully comply with these rules even though they show high bioactivity and high GI absorption experimentally. Here, we provide a detailed methodology using web-based platforms to determine the physicochemical properties of five phytochemicals in the phytonutrients ginger, echinacea, and tobacco. Furthermore, we analyzed the calculated data and established a protocol based on the combination and integration of these rules, plus other extended parameter ranges, to reliably predict the GI absorption of natural compounds. The presented data and methodology can be beneficial for evaluating bioactive natural compounds as potential drug candidates and predicting their oral bioavailability in patients.

Dataset: N/A

Dataset License: license under which the dataset is made available (CC0, CC-BY, CC-BY-SA, CC-BY-NC, etc.)

Keywords: Physicochemical properties; Gastrointestinal absorption; Lipinski's rules; Veber's rules; Ghose's rules; Muegge's rules

1. Summary

The physicochemical properties of a drug candidate can be used to understand and predict its absorption and, therefore, increase the chances of having a biological effect after oral consumption on the physiology of the human body. These properties can be theoretically calculated as drug design and discovery guidelines to understand drug candidates. This way, we can identify potential toxic molecules, therapeutic and non-therapeutic drugs [1]. Several researchers have created different pharmacokinetic rules to aid in predicting whether a compound is likely to be absorbed and is

 doi:10.20944/preprints202202.0305.v1

readily permeable to the GI tract. Among these rules is Lipinski's rule of five (L-Ro5), Ghose filter (GF), Veber's rule (VR), and Muegge's rule (MR). Since their creation, up to the present, they appear to be useful in predicting absorption in the GI tract and compound drug-likeness. Companies widely use them to create novel synthetic drugs [2]. However, there is little knowledge regarding rules specific to the development of drugs from natural products. In this area, we found studies that incorporate the concept of molecular complexity (Cm) [3] and the extension of specific ranges in the previously proposed rules [4].

To perform this analysis, we provide a detailed methodology using web-based platforms to determine the physicochemical properties of five bioactive compounds in three phytonutrients. We selected phytochemicals from echinacea and ginger due to the recent increase in their consumption to strengthen the immune system as we face the COVID-19 pandemic [5-6]. Furthermore, these phytonutrients are currently being studied in a clinical trial against COVID-19 in Iran [7]. On the other hand, we selected tobacco due to the various functions of its phytochemicals. For example, nicotine is well-known for its addictive and cancer-proliferative effect on the human body [8], while cembranoids within the same leaves are known for their anticancer and neuroprotective effects [9]. In addition, we included in our analyses three well-known compounds as theoretical controls (aspirin, doxorubicin, and ascorbic acid (vitamin C)) to validate our method and GI absorption prediction for natural products.

Obtaining these calculations is free of charge, and the data is acquired from trusted scientific platforms. The different properties are calculated using PubChem, SwissADME, and ChemSpider-ACD/Labs, and then analyzed to predict the GI absorption of each phytochemical. The presented data and methodology can be instrumental in evaluating bioactive natural compounds as potential drug candidates and predicting their bioavailability in patients. As important as these data are, obtaining them is relatively fast and inexpensive and can be readily done at the initial stages of assessment.

2. Data Description

Lipinski's rule of five (L-Ro5) states that poor absorption and permeation are more likely when the molecular weight (MW) is over 500 Da, lipophilicity (LogP) and hydrogen-bond donors (HBD) are more than 5, and there are more than ten hydrogen-bond acceptors (HBA) [10]. The Ghose filter (GF) attempts to improve prediction by stating that high absorption is likely, with the following criteria: MW of 160 to 480 Da, a logP of -0.4 to 5.6, a molar refractivity (A) of 40 to 130, and a total number of atoms (TNA) of 20 to 70 [11]. Veber's rule (VR) further increases the criteria for bioavailability with less than ten rotatable bonds (RB) and a polar surface area (PSA) no greater than 140 [12]. Muegge's rule changed the properties ranges and included other parameters to differentiate between drug-like and nondrug-like compounds. These are MW (200-600), LogP (-2 – 5), PSA≤150, number of rings (NR)≤7, number of carbons (NC)>4, number of heteroatoms (NH)>1, RB≤15, HBD≤5, HBA≤10 [13]. Lipophilicity (LogP) is one of the most important properties for all these rules, defined as the partition coefficient ratio of a compound between the hydrophobic and hydrophilic phases [14]. Other researchers have proposed that the lipophilicity of the ionizable groups at pH 7.4, called LogD, is much more critical for physiological absorption [15]. However, recent literature showed that the determination of LogD is not quite easy because the calculated pKa and LogD values are, in

 doi:10.20944/preprints202202.0305.v1

some cases, very different from those found experimentally [16]. Hence, we included LogD values, but they will not be used for predictions.

The other properties in these rules mainly focus on the molecules' interactions with themselves, the solvent, and additional molecules around. On the other hand, the molecular complexity (Cm) is another property considered important to calculate GI absorption that is a rough estimate of how complicated the structure is, seen from the point of view of both the elements contained and the displayed structural features, including symmetry [17]. In general, larger compounds display greater complexity than smaller ones, but large symmetrical compounds and large compounds with low diversity of atoms are considered less complex. A recent study verified if Cm can be a useful property in medicinal chemistry by calculating Cm values for approved drugs of different major classes of synthetic and natural antibiotics. Results demonstrated that a Cm of 100-800 had favorable outcomes for absorption and permeation [18].

Interestingly, other researchers that studied hundreds of clinical orally administered drugs concluded that a larger LogP, MW, PSA, and HBA could be allowed, especially in natural products. We define these as Extended Rules (ER) for natural products. The ranges for these rules are MW$\leqslant$ 1000 Da, $-2\leqslant$LogP$\leqslant$10, HB$\leqslant$6, HBA$\leqslant$15, PSA$\leqslant$250 Å^2, and RB$\leqslant$20 [4]. It is also worth mentioning that the predictions of all these rules are established on molecules passively transported into the cells. Thus, L-Ro5, GF, VR, MR, Cm, and ER do not consider actively transported substrates by biological transporters (e.g. endocytosis) [19].

For this work, we obtained the values for molecular formula, structure figure, MW (Da), TNA, HBA, HBD, RB, Log P, Cm, PSA, and A for our theoretical predictions using the web-based platforms PubChem and SwissADME. In addition, ChemSpider/ACD Labs was used just to determine the Log D at physiological pH.

3. Methods

3.1 Data source: PubChem

The primary data source was obtained from PubChem [17]. First, the name of each phytochemical was typed into the database's search engine. Then, the program calculates and provides the results for the values of different physicochemical properties of the searched compound.

i Search for the common compound name on the PubChem engine.

ii This engine will provide the structure, molecular formula, molecular weight, LogP, HBD, HBA, RB, PSA, and Cm of the chosen compound. We included these parameters in **Table 1** for each phytochemical.

iii To determine the total number of atoms (TNA) for each compound, we manually added the number of atoms in the molecular formula.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 February 2022 doi:10.20944/preprints202202.0305.v1

PubChem Aspirin (Compound)

3 Chemical and Physical Properties

3.1 Computed Properties

Property Name	Property Value	Reference
Molecular Weight	180.16	Computed by PubChem 2.1 (PubChem release 2021.05.07)
XLogP3	1.2	Computed by XLogP3 3.0 (PubChem release 2021.05.07)
Hydrogen Bond Donor Count	1	Computed by Cactvs 3.4.8.18 (PubChem release 2021.05.07)
Hydrogen Bond Acceptor Count	4	Computed by Cactvs 3.4.8.18 (PubChem release 2021.05.07)
Rotatable Bond Count	3	Computed by Cactvs 3.4.8.18 (PubChem release 2021.05.07)
Exact Mass	180.04225873	Computed by PubChem 2.1 (PubChem release 2021.05.07)
Monoisotopic Mass	180.04225873	Computed by PubChem 2.1 (PubChem release 2021.05.07)
Topological Polar Surface Area	63.6 Å²	Computed by Cactvs 3.4.8.18 (PubChem release 2021.05.07)
Heavy Atom Count	13	Computed by PubChem
Formal Charge	0	Computed by PubChem
Complexity	212	Computed by Cactvs 3.4.8.18 (PubChem release 2021.05.07)
Isotope Atom Count	0	Computed by PubChem
Defined Atom Stereocenter Count	0	Computed by PubChem
Undefined Atom Stereocenter Count	0	Computed by PubChem
Defined Bond Stereocenter Count	0	Computed by PubChem
Undefined Bond Stereocenter Count	0	Computed by PubChem
Covalently-Bonded Unit Count	1	Computed by PubChem
Compound Is Canonicalized	Yes	Computed by PubChem (release 2021.05.07)

Figure 1. PubChem results on Aspirin properties. We selected MW, HBA, HBD, RB, Log P, PSA, and Cm from these values. They can be found following the instructions mentioned beforehand in the data source PubChem and summarized in **Table 2**.

3.2 Data source: SwissADME

Our second data source was SwissADME [20]. It uses the Simplified Molecular Input Line System (SMILES), which is a chemical notation that allows a user to represent a chemical structure in a computational way. This notation provides computing physicochemical descriptors and predicts small molecule pharmacokinetics and drug likeliness to support drug discovery [21].

i. Search for the common compound name on the PubChem engine.

ii. Identify the Canonical SMILES in the category of Computed Descriptors.

iii. Go to the SwissADME program and write the SMILES (from PubChem) in the space provided where it says "Enter a list of SMILES here" and click "Run."

iv. This program will provide the user with the compound's RB, HBD, HBA, A, PSA, Consensus Log P, GI absorption, blood-brain barrier (BBB) permeability, and P-glycoprotein (P-gp) substrate. The values of these parameters are shown in **Table 3.**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 February 2022 doi:10.20944/preprints202202.0305.v1

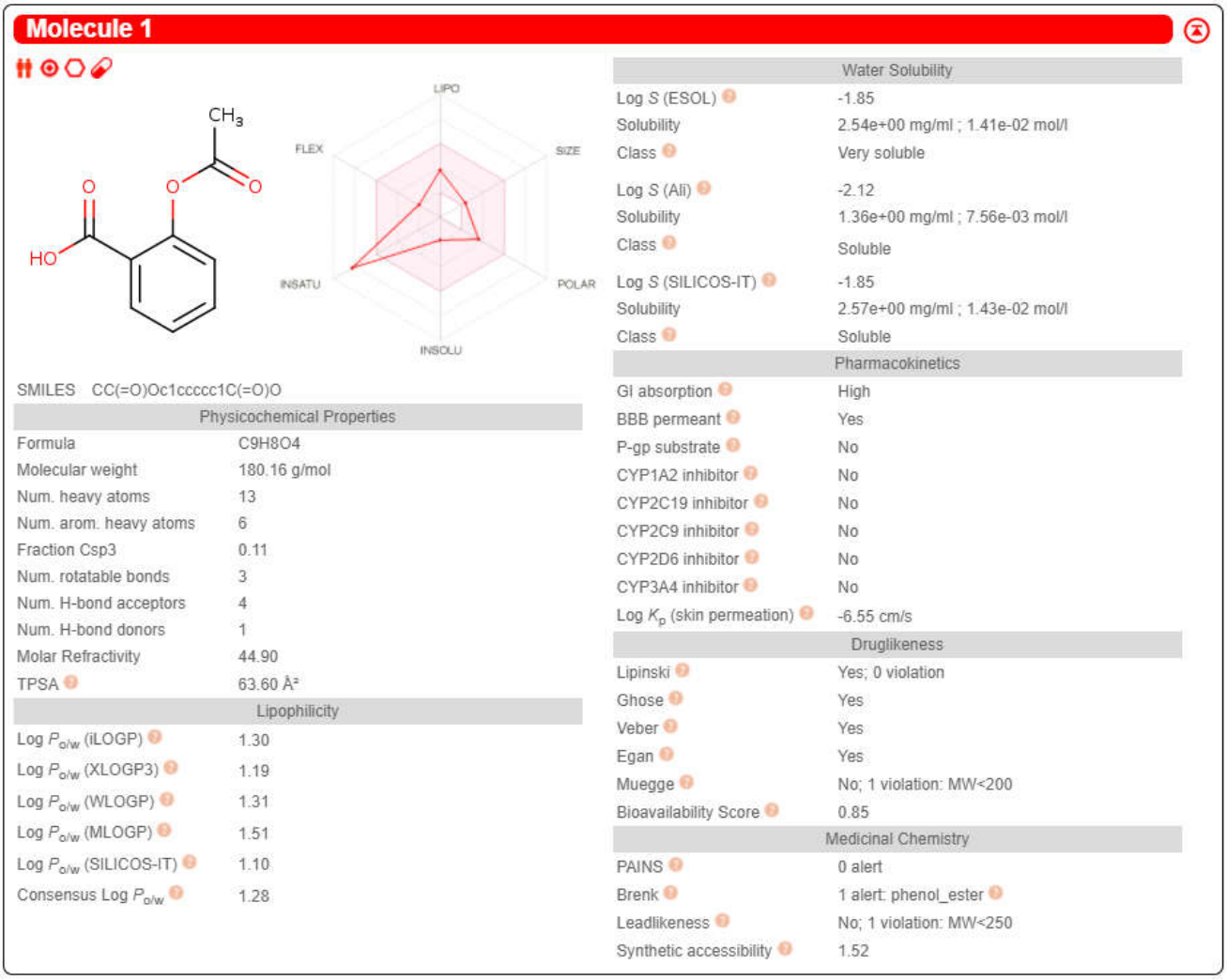


Figure 2. SwissADME results on Aspirin properties. We selected RB, HBD, HBA, A, PSA, Consensus Log P, GI absorption, BBB permeant, and P-gp substrate from these values included in **Table 3**. They can be calculated following the instructions above using the data source SwissADME.

3.3 Data source: ChemSpider

Another data source used to obtain phytochemical's parameters was ChemSpider [22]. First, the name of each phytonutrient was typed into the search engine of the database. Then, the program calculates and provides results for the values of different physicochemical properties.

i. Search for the common compound name on the ChemSpider engine.

ii. Click on the "Properties" tab.

iii. Click on the "Predicted - ACD/Labs" sub-tab.

iv. Look for the parameters ACD/LogD(pH 7.4) and A. We included these in **Table 2** for each phytochemical.

Even when this web-based program calculates the same properties as other programs, we will use the numbers provided by PubChem and SwissADME since these websites are more reliable because they are constantly monitored for updates to ensure the most accurate calculations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 February 2022 doi:10.20944/preprints202202.0305.v1

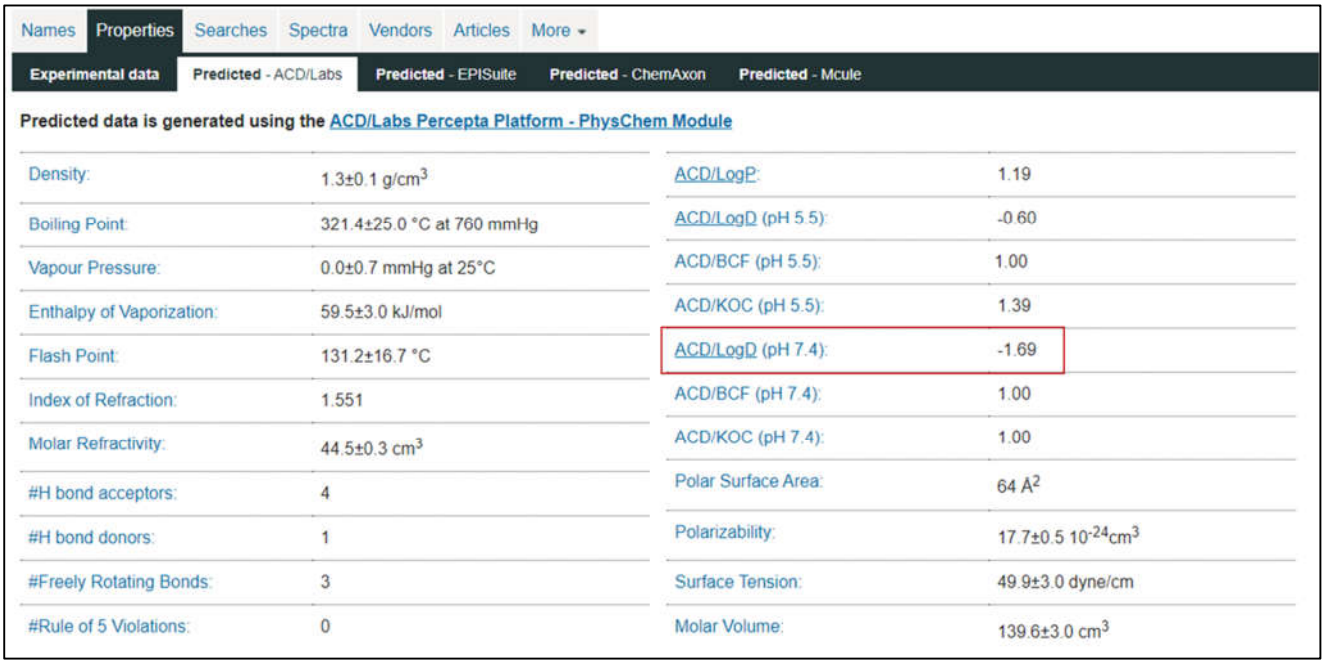


Figure 3. ChemSpider results on Aspirin properties. From these values, we selected the $LogD_{7.4}$ (red box) included in **Table 2**. These values are found following the instructions mentioned above in ChemSpider's data source.

3.4 Data Collection

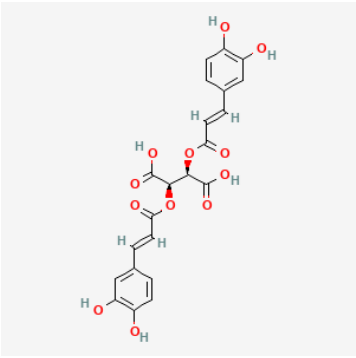
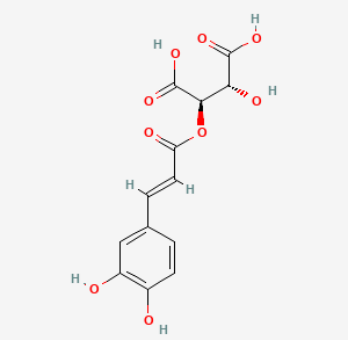
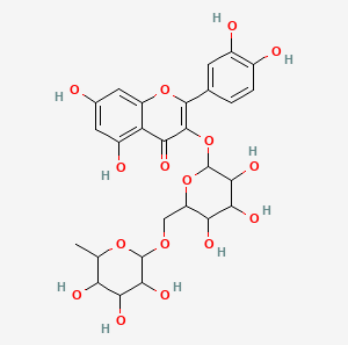
In our previous review article [23], we developed a methodology using several web-based platforms to calculate and predict the probability of GI absorption of different phytochemicals using PubChem and ChemSpider-ACD/Labs. In this article, we clearly explain our methodology using those web-based platforms and explore another platform, SwissADME, to develop the best approximation for GI absorption prediction for natural products.

Table 1 presents the common compound name, structure, molecular formula, and molecular weight for the selected plants' top five bioactive compounds obtained from PubChem. This information should not vary on any platform since it is already established in the literature. **Table 2** is a collection of physicochemical properties for those bioactive compounds from our main database, PubChem. As we mentioned before, LogD values were added as supplementary information but were not used for predictions. In **Table 3**, we determined the physicochemical properties of the bioactive compounds of the selected plants using our second main database, SwissADME. It is important to mention that these calculations from SwissADME may have minor variations for the same property compared to PubChem, mostly due to differences in the algorithms. In addition to the properties mentioned by L-Ro5, GF, VR, MR, this program provides more information, including blood-brain barrier (BBB) permeation, targeting the P-glycoprotein (P-gp), and GI absorption. The P-gp transporters, expressed in the intestinal epithelium and cancer cells, and the decrease in cellular uptake of their substrates [24]. This property is important to study in chemoresistance and cancer therapeutics because Pgp is the key efflux pump of chemotherapeutic drugs. Furthermore, the permeability of a drug to the BBB is significant for laboratories working on brain therapies because brain-targeted drugs must have the capacity to cross this barrier to target neurological disorders. Interestingly, SwissADME predicts that aspirin and ascorbic acid are highly absorbable even when both show violations to the rules. In addition to the rules, we understand that this program mainly

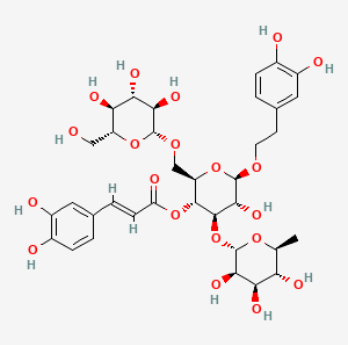
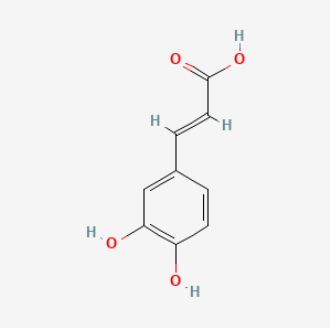
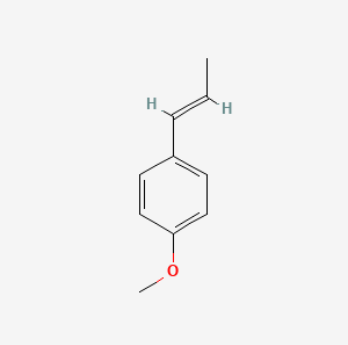
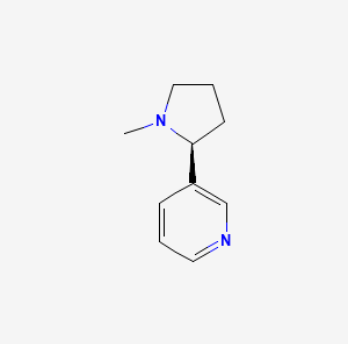
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 February 2022 doi:10.20944/preprints202202.0305.v1

adjusts the GI absorption predictions using the BOILED-Egg algorithm model [25]. This model makes its predictions by constructing two ellipses using the coordinates of the PSA (0 – 142.1 Å^2) and the LogP (-2.3 – 6.8) ranges. In this way, the properties from Tables 1, 2, and 3 facilitate the GI absorption predictions described in **Table 4,** using the ranges from L-Ro5, GF, VR, MR, ER, and Cm.

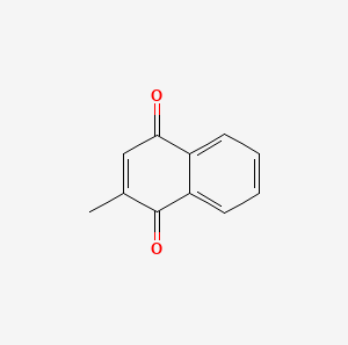
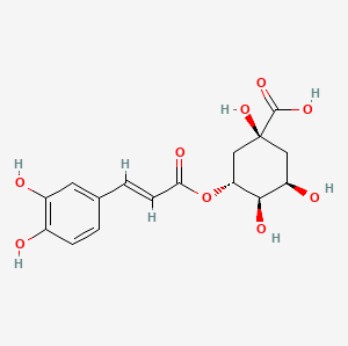
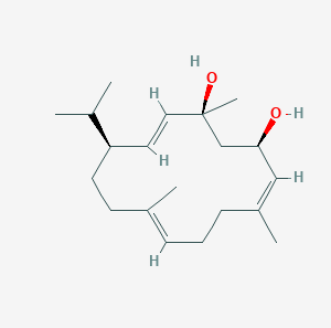
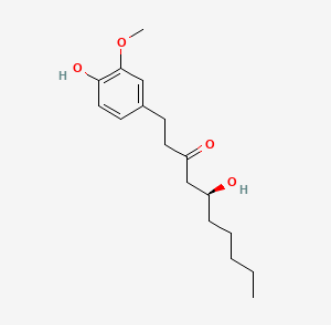
Table 1. Name, structure, molecular formula, and molecular weight of the main bioactive compounds in echinacea, tobacco, and ginger

Name	Structure / Molecular formula	MW (Da)
	Echinacea	
Cichoric acid	 $C_{22}H_{18}O_{12}$	474.4
Caftaric acid	 $C_{13}H_{12}O_9$	312.2
Quercetin -3-O-rutinoside	 $C_{27}H_{30}O_{16}$	610.5

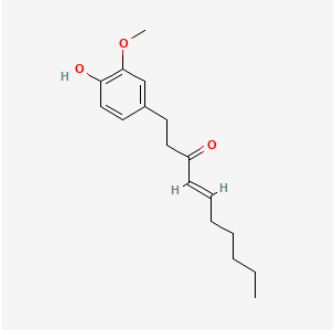
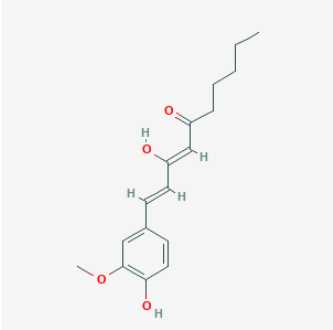
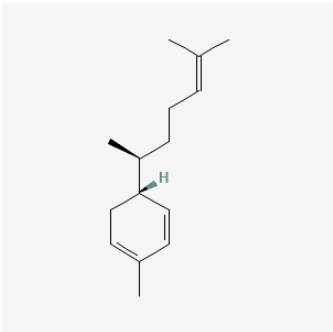
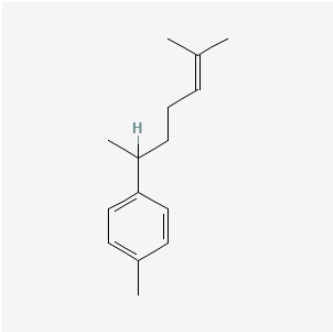
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 February 2022 doi:10.20944/preprints202202.0305.v1

Echinacoside	 $C_{35}H_{46}O_{20}$	786.7
Caffeic acid	 $C_9H_8O_4$	180.2
Tobacco		
Anethole	 $C_{10}H_{12}O$	148.2
Nicotine	 $C_{10}H_{14}N_2$	162.2

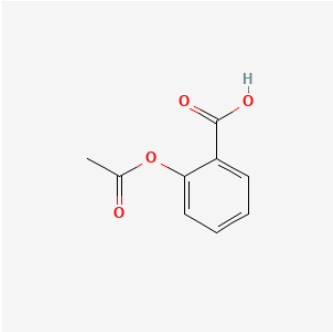
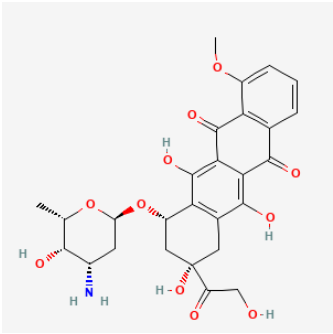
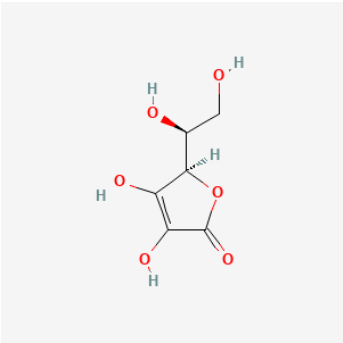
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 February 2022 doi:10.20944/preprints202202.0305.v1

Menadione	 $C_{11}H_8O_2$	172.2
Chlorogenic acid	 $C_{16}H_{18}O_9$	354.3
Cembra-2,7,11 -triene-4,6-diol	 $C_{20}H_{34}O_2$	306.5
Ginger		
6-Gingerol	 $C_{17}H_{26}O_4$	294.4

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 February 2022 doi:10.20944/preprints202202.0305.v1

6-Shogaol	 $C_{17}H_{24}O_3$	276.4
6-Dehydro gingerdione	 $C_{17}H_{22}O_4$	290.4
Zingiberene	 $C_{15}H_{24}$	204.4
α-Curcumene	 $C_{15}H_{22}$	202.3
Controls		

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 February 2022 doi:10.20944/preprints202202.0305.v1

Aspirin	 $C_9H_8O_4$	180.1
Doxorubicin	 $C_{27}H_{29}NO_{11}$	543.5
Ascorbic acid	 $C_6H_8O_6$	176.1

MW: molecular weight

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 February 2022 doi:10.20944/preprints202202.0305.v1

Table 2. Physicochemical properties for the main bioactive compounds of the selected plants from the PubChem database

Name	TNA# / HBA / HBD / RB	Log P	Log D*	A (cm^3)	PSA ($Å^2$)	Cm
			Echinacea			
Cichoric acid	52/ 12 / 6 / 11	2.00	-2.56	114.9	208	740
Caftaric acid	34 / 9 / 5 / 7	0.10	-4.39	70.8	162	458
Quercetin -3-O-rutinoside	73 / 16 / 10 / 6	-1.30	-1.75	138.2	266	1020
Echinacoside	101 / 20 / 12 / 14	-2.10	-1.06	182.1	324	1230
Caffeic acid	21 /4 /3 /2	1.20	-1.74	47.5	77.8	212
			Tobacco			
Anethole	23 / 1 / 0 / 2	3.30	3.08	48.8	9.2	121
Nicotine	26 / 2 / 0 / 1	1.20	-0.37	49.3	16.1	147
Menadione	21 / 2 / 0 / 0	2.2	2.02	47.6	34.1	289
Chlorogenic acid	43 / 9 / 6 / 5	-0.4	-3.91	82.0	165	534
Cembra-2,7,11 -triene-4,6-diol	56 / 2 / 2 / 1	4.00	5.34	94.4	40.5	431
			Ginger			
6-Gingerol	47 / 4 / 2 / 10	2.5	2.88	82.9	66.8	293
6-Shogaol	44 / 3 / 1 / 9	3.70	4.15	81.5	46.5	299
6-Dehydro gingerdione	43 / 4 / 2 / 8	4.20	3.17	84.6	66.8	373
Zingiberene	39 / 0 / 0 / 4	5.20	5.63	68.5	0	274

 doi:10.20944/preprints202202.0305.v1

α-Curcumene	37 / 0 / 0 / 4	5.40	5.20	68.3	0	190
			Controls			
Aspirin	21 / 4 / 1 / 3	1.2	-1.69	44.5	63.6	212
Doxorubicin	68 / 12 / 6 / 5	1.3	-0.79	131.5	206	977
Ascorbic acid	20 / 6 / 4 / 2	-1.6	-4.99	35.3	107	232

Log P: lipophilicity; LogD: Lipophilicity considering ionizable groups at pH 7.4; A: molar refractivity; HBD: Hydrogen bond donors, HBA: Hydrogen bond acceptors; RB: rotatable bonds; PSA: polar surface area; TNA: total number of atoms; Cm: molecular complexity.

[#] TNA was determined manually by adding the number of atoms in the molecular formula.

*LogD was determined using ChemSpider-ACD/Labs.

Table 3. Physicochemical properties for the main bioactive compounds of the selected plants from the SwissADME database

Name	TNA / HBA / HBD / RB	Log P	A (cm^3)	PSA ($Å^2$)	BBB Permeant	P-gp substrate	GI absorption
			Echinacea				
Cichoric acid	52/ 12 / 6 / 11	1.01	114	208	No	Yes	Low
Caftaric acid	34 / 9 / 5 / 7	-0.23	70.6	162	No	No	Low
Quercetin -3-O-rutinoside	73 / 16 / 10 / 6	-1.12	141.4	269.43	No	Yes	Low
Echinacoside	101 / 20 / 12 / 14	-2.08	180.8	324	No	No	Low
Caffeic acid	21 /4 /3 /2	0.93	47.2	77.8	No	No	High
			Tobacco				
Anethole	23 / 1 / 0 / 2	2.79	47.8	9.2	Yes	No	High
Nicotine	26 / 2 / 0 / 1	1.50	53.1	16.1	Yes	No	High

 doi:10.20944/preprints202202.0305.v1

Menadione	21 / 2 / 0 / 0	1.98	49.1	34.1	Yes	No	High
Chlorogenic acid	43 / 9 / 6 / 5	-0.38	83.5	164.8	No	No	Low
Cembra-2,7,11 -triene-4,6-diol	56 / 2 / 2 / 1	3.93	97.1	40.5	Yes	No	High
Ginger							
6-Gingerol	47 / 4 / 2 / 10	3.13	84.6	66.8	Yes	No	High
6-Shogaol	44 / 3 / 1 / 9	3.76	82.9	46.5	Yes	No	High
6-Dehydro gingerdione	43 / 4 / 2 / 8	3.45	84.8	66.8	Yes	No	High
Zingiberene	39 / 0 / 0 / 4	4.47	70.68	0	No	No	Low
α-Curcumene	37 / 0 / 0 / 4	4.86	69.55	0	No	No	Low
Controls							
Aspirin	21 / 4 / 1 / 3	1.28	44.9	63.6	Yes	No	High[#]
Doxorubicin	68 / 12 / 6 / 5	0.44	132.7	206.1	No	Yes	Low[#]
Ascorbic acid	20 / 6 / 4 / 2	-1.28	35.1	107.2	No	No	High[#]

Log P: lipophilicity (Consensus $LogP_{o/w}$); A: molar refractivity; HBD: Hydrogen bond donors, HBA: Hydrogen bond acceptors; RB: rotatable bonds; PSA: polar surface area; TNA: total number of atoms; BBB permeant: Blood-brain barrier permeant; Pgp substrate: P-glycoprotein substrate; GI absorption: Gastrointestinal absorption predicted by the program for synthetic compounds.

[#]These theoretical GI predictions are supported by the experimental GI absorption for these synthetic and natural control compounds.

3.5 Data Analysis: Prediction of GI absorption for natural products

Considering the favorable properties based on the physicochemical parameters determined, we identified in **Table 4** any violation of the rules, L-Ro5, VR, GF, and MR. Because these rules mainly apply to study synthetic compounds, we included in **Table 4** the Cm and the ER that consider the study of natural compounds. SwissADME also predicts the probability of being absorbed through the GI, as we showed in **Table 3**. However, this program was also developed for synthetic compounds. For this reason, the probability of GI absorption for natural products shown in **Table 4** was predicted by integrating and combining the L-Ro5, GF, VR, MR, Cm, and ER. In this table, we used the following ranges delimiting these rules:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 February 2022 doi:10.20944/preprints202202.0305.v1

a. L-Ro5: HBD≤5, HBA≤10, MW≤500, logP≤5 [10];
b. GF: -0.4≤logP≤5.6, A (40 – 130), MW (160 – 480), TNA (20 – 70) [11];
c. VR: RB≤10, PSA≤140 [12];
d. Cm: 100-800 [18];
e. MR: MW (200-600), -2≤logP≤5, PSA≤150, NR≤7, NC> 4, NH > 1, RB≤15, HBD≤5, HB ≤10 [13].

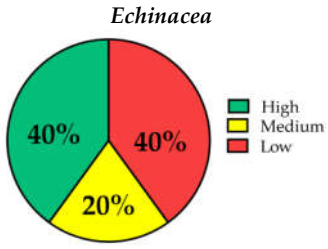
The following ranges delimited the ER:

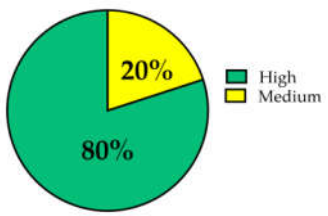
a. MW≤1000 Da, −2≤LogP≤10, HBD≤6, HBA≤15, PSA≤250 Å^2, and RB≤20 [4].

Violations to these rules affect the GI absorption. The GI predictions for natural products were manually determined by the combination of all the rules above as follows: **High:** The compound fully complies with all the rules or has a limit of 3 violations to the L-Ro5, GF, VR, or MR that are covered by the ER and Cm; **Medium:** The compound fully complies the ER and Cm but has > 3 violations to the other rules. **Low:** The compound does not fully comply with ER, Cm, and the other rules. We suggest using just the L-Ro5, GF, MR, and VR ranges for synthetic compounds.

Table 4. Combination of the rules to predict the drug-likeness and GI absorption of natural compounds

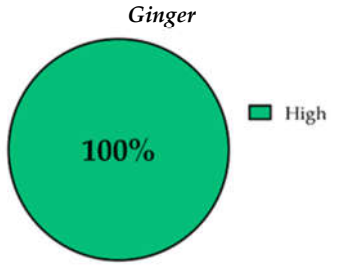
Name	L-Ro5	GF	VR	Cm	MR	ER	Predicted GI absorption	
							Natural compound	Phytonutrient (Plant)
Caffeic acid	✔	✔	✔	✔	1 / MW<200	✔	High	
Caftaric acid	✔	1 / LogP<-0.4	1 / PSA>140	✔	1 / PSA>150	✔	High	
Cichoric acid	2 / HBA>10 HBD>5	✔	2 / RB > 10 PSA>140	✔	3 / PSA>150 HBA>10 HBD>5	✔	Medium	
Quercetin-3-O-rutinoside	3 / MW>500 HBA>10 HBD>5	4 / MW>480 LogP<-0.4 A>130 TNA>70	1 / PSA>140	Cm>800	4 / MW>600 PSA>150 HBA>10 HBD>5	4 / PSA>250 TNA>70 HBA>10 HBD>5	Low	
Echinacoside	3 / MW>500 HBA>10 HBD>5	4 / MW>480 LogP<-0.4 A>130 TNA>70	2 / RB>10, PSA>140	Cm>800	5 / MW>600 LogP<-2 PSA>150 HBA>10 HBD>5	5 / PSA>250 LogP<-2 TNA>70 HBA>10 HBD>5	Low	
Nicotine	✔	✔	✔	✔	1 / MW<200	✔	High	
Menadione	✔	✔	✔	✔	1 / MW<200	✔	High	
Cembra-2,7,11-triene-4,6-diol	✔	✔	✔	✔	✔	✔	High	
Anethole	✔	1 /	✔	✔	2 /	✔	High	





Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 February 2022 doi:10.20944/preprints202202.0305.v1

		MW<160			MW<200 NH<2			
Chlorogenic acid	1 / HBD > 5	1 / LogP<-0.4	1 / PSA>140	✔	2 / PSA>150 HBD>5	✔	Medium	
6-Gingerol	✔	✔	✔	✔	✔	✔	High	
6-Shogaol	✔	✔	✔	✔	✔	✔	High	
6-Dehy-drogingerdione	✔	✔	✔	✔	✔	✔	High	
Zingiberene	✔	✔	✔	✔	1 / NH<2	✔	High	
α-Curcumene	✔	✔	✔	✔	1 / NH<2	✔	High	
Controls								
Aspirin	✔	✔	✔	✔	1 / MW<200	✔	N/A[#]	
Doxorubicin	3 / MW>500 HBA>10 HBD>5	2 / MW>480 A>130	1 / PSA>140	Cm>800	3 / PSA>150 HBA>10 HBD>5	✔	N/A[#]	
Ascorbic acid	✔	2 / LogP<-0.4 A<40	✔	✔	1 / MW<200	✔	High	



L-Ro5: Lipinski's rule of five; GF: Ghose filter; VR: Veber's rule; MR: Muegge's rule; ER: Extended rules; Cm: molecular complexity; Log P: Consensus LogP (lipophilicity); A: molar refractivity; HBD: Hydrogen bond donors, HBA: Hydrogen bond acceptors; RB: rotatable bonds; PSA: polar surface area; TNA: total number of atoms; NR: number of rings; NH: number of heteroatoms; NC: number of carbons; GI: Gastrointestinal.

L-Ro5: HBD≤5, HBA≤10, MW≤500, logP≤5; **GF:** logP (-0.4 – 5.6), A (40 – 130), MW (160 – 480), TNA (20 – 70); **VR:** RB≤10, PSA≤140 ; **Cm**: 100 – 800; **MR:** MW (200 – 600), logP (-2 – 5), PSA≤150, NR≤7, NC>4, NH>1, RB≤15, HBD≤5, HBA≤10; **ER:** MW≤1000 Da, LogP (-2 – 10), HBD≤6, HBA≤15, PSA≤250 Å^2, and RB≤20.

✔: Complies with all the rules

[#]Not applicable because these are not natural compounds. GI absorption is predicted for them in **Table 3**.

The GI predictions for natural products were manually determined by the combination of all the rules as follows:

High: The compound fully complies with all the rules or has 3 violations in the L-Ro5, GF, VR, or MR, covered by the ER and Cm.

Medium: The compound fully complies with the ER and Cm but has >3 violations to the other rules.

Low: The compound does not comply with the ER and Cm, and therefore neither with the other rules.

From echinacea's phytochemicals, we predict that only caffeic acid and caftaric acid will have a high GI absorption with 1 and 3 violations, respectively, and comply with Cm and ER. Cichoric acid will have a medium GI absorption because it shows 7 violations but complies with the Cm and ER. In contrast, quercetin-3-O-rutinoside and echinacoside will have a low GI absorption because they show 17 and 20 violations, respectively, including no compliance with ER and Cm. Accordingly,

 doi:10.20944/preprints202202.0305.v1

based on the five analyzed phytochemicals, the phytonutrient echinacea is predicted to show partially (high 40% / medium 40%) GI absorption.

Tobacco has four out of five phytochemicals predicted to have high GI absorption. These are the following: Nicotine and menadione, which have 1 violation; anethole which has 3 violations; and cembra-2,7,11-triene-4,6-diol with no violations. At the same time, chlorogenic acid will have a medium absorption because it shows 5 violations but complies with Cm and ER. As a result, according to these five analyzed phytochemicals, the phytonutrient tobacco is predicted to show mostly high (80%) GI absorption.

In our prediction of ginger's phytochemicals, all of them (6-gingerol, 6-shogaol, 6-dehydrogingerdione, zingiberene, and α-curcumene) will have high GI absorption. Only zingiberene, and α-curcumene showed 1 violation. Therefore, according to these five phytochemicals, the phytonutrient ginger is predicted to show a high GI absorptivity.

Moreover, three different compounds were analyzed as theoretical controls: aspirin, doxorubicin, and ascorbic acid. Aspirin, the most common over-the-counter synthetic oral analgesic used worldwide, showed good absorption in the small intestine [26]. Even though aspirin is highly absorbable, it presents one violation to the ER. Doxorubicin is a prescribed synthetic anthracycline chemotherapeutic known to exhibit poor oral bioavailability; thus, it is only administered intravenously [27]. This drug showed in our analysis low GI absorption because it violated the rules for synthetic compounds and the Cm (a total of 10 violations). Ascorbic acid is the natural well-known orally-absorbed vitamin C [28]. Although ascorbic acid is an established absorbed compound, it violates 3 rules in total, 2 from GF and 1 from MR, further supporting our analysis that natural compounds may have wider ranges than the ones proposed in the "classical rules". Based on these findings for ascorbic acid, we expanded the limit to 3 violations for high absorption. However, it is known that some clinically approved drugs for oral administration would be considered outliers because they fall out of the ranges of these rules. Furthermore, as mentioned above, these rules do not consider the active transport of molecules [19]. This is still a limitation in our GI absorptivity predictions.

In conclusion, for an initial analysis of drug-likeness and oral availability of a compound, these algorithms are an excellent method to start understanding its drug-relevant properties. Nevertheless, further in vivo studies should be conducted experimentally to confirm the predicted GI absorption.

4. User Notes (optional)

N/A

5. Patents

N/A

Supplementary Materials: N/A

Author Contributions: Coordination, Y.D.; conceptualization, Y.D.; formal analysis, L.A.V., Y.D. I.J.S.A; theoretical calculations, L.A.V., P.R., D.P.; writing—original draft preparation, Y.D., L.A.V.; writing—review and editing, Y.D., L.A.V., Y.F.A., I.J.S.A.; supervision, Y.D. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Career Development Grant (Y.D., Y.F.A.) from Sloan Scholars Mentoring Network, the Alliance Pilot Project Program from the Hispanic Alliance for Clinical

and Translational Research (Award Number U54GM133807) (Y.D., Y.F.A.), and Biomedical Fellowships from Fundación Intellectus (D.P., Y.D.).

Data Availability Statement: N/A

Acknowledgments: This publication was made possible by the support of the San Juan Bautista School of Medicine (SJBSM) and Universidad Central del Caribe (UCC). The authors thank Dr. Estela Estapé (Director of the SJBSM Research Center and senior advisor) for her outstanding dedication and support in the writing process of this article.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the study's design, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Committee on the Design and Evaluation of Safer Chemical Substitutions: A Framework to Inform Government and Industry Decision; Board on Chemical Sciences and Technology; Board on Environmental Studies and Toxicology; Division on Earth and Life Studies; National Research Council. 5 Physicochemical properties and Environmental Fate. In *A Framework to Guide Selection of Chemical Alternatives*. National Academic Press (US): Washington (DC), 2014 Oct 29; pp. 47-68. Available from: https://www.ncbi.nlm.nih.gov/books/NBK253956/
2. Benet, L.; Hosey, C.; Ursu, O.; Oprea, TI. BDDCS, the Rule of 5 and drugability. *Adv Drug Deliv Rev*. **2016**, 101, 89-98. https://doi:10.1016/j.addr.2016.05.007
3. Kermen, F.; Chakirian, A.; Sezille, C.; et al. Molecular complexity determines the number of olfactory notes and the pleasantness of smells [published correction appears in Sci Rep. **2013**, 3, 1132]. *Sci Rep*. **2011**, 1, 206. doi:10.1038/srep00206
4. Croy, B.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral druggable space beyond the rule of 5: Insights from drugs and clinical candidates. *Cell Chem. Biol.* **2014**, 21(9), 1115-1142. https://doi.org/10.1016/j.chembiol.2014.08.013
5. Aucoin, M.; Cooley, K.; Saunders, P.; et al. The effect of *Echinacea* spp. on the prevention or treatment of COVID-19 and other respiratory tract infections in humans: A rapid review. *Adv Integr Med*. **2020**, 7(4), 203-217. https://doi:10.1016/j.aimed.2020.07.004
6. Jafarzadeh, A.; Jafarzadeh, S.; Nemati, M. Therapeutic potential of ginger against COVID-19: Is there enough evidence?. *J. Tradit. Chin. Med. Sci.* **2021**, 8(4), 267-279. https://doi.org/10.1016/j.jtcms.2021.10.001
7. Mesri, M.; (2020, April 18 -) The effect of herbal drugs (echinacea and ginger) on improvement of clinical symptoms and hospitalizations in suspected COVID-19 outpatient cases. Identifier IRCT20200415047089N1. https://en.irct.ir/trial/47234
8. Mishra, A.; Chaturvedi, P.; Datta, S.; Sinukumar, S.; Joshi, P.; Garg, A. Harmful effects of nicotine. *Indian J Med Paediatr Oncol* **2015**, 36(1), 24-31. https://doi:10.4103/0971-5851.151771
9. Rojas-Colón, L.; Dash, P.; Morales-Vías, F.; *et al.* 4R-cembranoid confers neuroprotection against LPS-induced hippocampal inflammation in mice. *J Neuroinflammation* **2021** 18, 95. https://doi.org/10.1186/s12974-021-02136-9
10. Pollastri, M.; Overview on the rule of five. *Curr Protoc Pharmacol* **2010**, 49: 9.12.1-9.12.8. https://doi.org/10.1002/0471141755.ph0912s49
11. Ghose, G.; Viswanadhan, V.; Wendoloski, John. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* **1999**, 1(1), 55-68. https://doi.org/10.1021/cc9800071
12. Veber, D.; Johnson, S.; Cheng, H.; Smith, B.; Ward, K.; Kopple, K. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, 45(12), 2615-2623. https://doi.org/10.1021/jm020017n
13. Muegge, I.; Heald, SL.; Brittelli, D. Simple selection criteria for drug-like chemical matter. *J Med Chem*. **2001**, 44(12), 1841-6. https://doi.org/10.1021/jm015507e
14. Rutkowska, E.; Pajak, K.; Jóźwiak, K. Lipophilicity--methods of determination and its role in medicinal chemistry. *Acta Pol Pharm*. 2013, 70(1), 3-18. https://pubmed.ncbi.nlm.nih.gov/23610954/
15. Tinworth, C.; Young, R. Facts, patterns, and principles in drug discovery: Appraising the rule of 5 with measured physicochemical data. *J. Med. Chem.* **2020** 63(18), 10091-10108. https://doi.org/10.1021/acs.jmedchem.9b01596
16. Winiwarter, S.; Ridderström, M.; Ungell, A.; Andersson, T.; Zamora, I. 5.22 - Use of molecular descriptors for absorption, distribution, metabolism, and excretion predictions. In *Comprehensive Medicinal Chemistry II*; Taylor J., Triggle D.; Elsevier: Netherlands, 2007; Volume 5, pp. 531-554
17. Pubchem. Available online: https://pubchem.ncbi.nlm.nih.gov/ (Accessed September 2021).
18. Böttcher, T. An Additive Definition of Molecular Complexity. *J Chem Inf Model* **2016,** 56(3), 462-470. https://doi.org/10.1021/acs.jcim.5b00723
19. Bickerton, G.; Paolini, G.; Besnard, J.; *et al.* Quantifying the chemical beauty of drugs. *Nature Chem* **2012,** 4, 90–98. https://doi.org/10.1038/nchem.1243
20. SwissADME. Available online: http://www.swissadme.ch/index.php/ (Accessed November 2021).
21. Daina, A.; Michielin, O.; Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci Rep* **2017,** 7, 42717. https://doi.org/10.1038/srep42717
22. Chemspider. Available online: http://www.chemspider.com/ (Accessed October 2021).

23. Delgado, Y.; Cassé, C.; Ferrer-Acosta, Y.; Suárez-Arroyo, I.J.; Rodríguez-Zayas, J.; Torres, A.; Torres-Martínez, Z.; Pérez, D.; González, M.J.; Velázquez-Aponte, R.A.; Andino, J.; Correa-Rodríguez, C.; Franco, J.C.; Milán, W.; Rosario, G.; Velázquez, E.; Vega, J.; Colón, J.; Batista, C. Biomedical effects of the phytonutrients turmeric, garlic, cinnamon, graviola, and oregano: A comprehensive review. *Appl. Sci.* **2021**, *11*, 8477. https://doi.org/10.3390/app11188477
24. Alqahtani, M.; Kazi, M.; Alsenaidy, M.; Ahmad, M. Advances in Oral Drug Delivery. *Front. Pharmacol* **2021**, 21. https://doi.org/10.3389/fphar.2021.618411
25. Daina, A.; Zoete, V. A BOILED-Egg to predict gastrointestinal absorption and brain penetration of small molecules. *ChemMedChem* **2016**, 11, 1117. https://doi.org/10.1002/cmdc.201600182
26. Arif, H.; Aggarwal, S. Salicylic Acid (Aspirin) [Updated **2021** Jul 15]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK519032/
27. Ahmad, N.; Ahmad, R.; Alam, M.; Ahmad, F. Enhancement of oral bioavailability of doxorubicin through surface modified biodegradable polymeric nanoparticles. *Chem Cent J*. **2018**, 12(1), 65. https://doi:10.1186/s13065-018-0434-1
28. Lykkesfeldt, J.; Tveden-Nyborg, P. The Pharmacokinetics of Vitamin C. *Nutrients*. **2019**, 11(10), 2412. https://doi:10.3390/nu11102412