Brazilian Journal of
HEALTH AND PHARMACY

Natural or synthetic? Classification of common preservatives in food and drug industry by artificial intelligence

Natural ou sintético? Classificação de conservantes farmacêuticos e alimentícios por meio de inteligência artificial

Douglas Vieira Thomaz[1]; Uriel Abe Contardi[2]; Pierre Alexandre dos Santos[1]; Renê Oliveira do Couto[3]*

1. Faculdade de Farmácia (FF), Universidade Federal de Goiás (UFG), Goiânia, Goiás, Brasil.
2. Universidade Tecnológica Federal do Paraná (UTFPR), Cornélio Procópio, Paraná, Brasil.
3. Laboratório de Desenvolvimento Farmacotécnico (LADEF), Campus Centro-Oeste Dona Lindu (CCO), Universidade Federal de São João del Rei (UFSJ), Divinópolis, Minas Gerais, Brasil.

***Author to whom correspondence should be addressed:** Renê Oliveira do Couto (https://orcid.org/0000-0002-3748-3427)
Laboratório de Desenvolvimento Farmacotécnico (LADEF), Campus Centro-Oeste Dona Lindu (CCO), Universidade Federal de São João del Rei (UFSJ). Rua Sebastião Gonçalves Coelho, 400 – Bairro Chanadour. ZIP code 35501-296 – Divinópolis (MG), Brasil
Email: rocouto@ufsj.edu.br

Data de Submissão: 14/01/2022; Data do Aceite: 03/08/2022

Citar: THOMAZ et al. Natural or synthetic? Classification of common preservatives in food and drug industry by artificial intelligence. **Brazilian Journal of Health and Pharmacy**, v. 4, n. 2, p. 43-61, 2022. DOI: 10.29327/226760.4.2-4

ABSTRACT

Monoaromatic antioxidants are one of the major classes of small druggable molecules whose use is widespread as preservatives in pharmaceutical and foodstuff industry. The differentiation of these compounds according to their source is notably difficult due to their shared structural features. This work showcases how to promote the classification of natural and synthetic monoaromatic antioxidants using multivariate analysis, data mining and machine learning algorithms. Physicochemical and biopharmaceutical molecular descriptors were selected and calculated do render alignment and classification models using principal components analysis, data mining, support-vector machines (linear kernel) and multilayer perceptron. We showcased that physicochemical and biopharmaceutical molecular predictors may be suitable attributes for differentiating natural and synthetic monoaromatic antioxidants, since their outputs from multivariate analysis, data mining and machine learning algorithms generated a reliable and accurate model for prompt classification of natural and synthetic monoaromatic antioxidants. Moreover, all classification models yielded accuracies above 80%. This work therefore sheds light on the use of artificial intelligence in the development of classifiers for pharmaceutical and foodstuff applications.

Key words: Antioxidants; Cheminformatics; Multivariate Analysis; Preservatives, Pharmaceutical

RESUMO

Os antioxidantes monoaromáticos são uma das principais classes de pequenas moléculas farmacológicas cujo uso como conservantes é difundido na indústria farmacêutica e alimentícia. A diferenciação desses compostos de acordo com sua origem é notavelmente difícil devido às suas características estruturais compartilhadas. Este trabalho mostra como promover a classificação de antioxidantes monoaromáticos naturais e sintéticos usando análise multivariada, mineração de dados e algoritmos de aprendizado de máquina. Descritores moleculares físico-químicos e biofarmacêuticos foram selecionados e calculados para renderizar modelos de alinhamento e classificação usando análise de componentes principais, mineração de dados, máquinas de suporte de vetores (kernel linear) e perceptron multicamadas. Mostramos que preditores moleculares físico-químicos e biofarmacêuticos podem ser atributos adequados para diferenciar antioxidantes monoaromáticos naturais e sintéticos, uma vez que seus resultados de análise multivariada, mineração de dados e algoritmos de aprendizado de máquina geraram um modelo confiável e preciso para classificação imediata de antioxidantes monoaromáticos naturais e sintéticos. Além disso, todos os modelos de classificação apresentaram acurácias acima de 80%. Este trabalho, portanto, lança luz sobre o uso da inteligência artificial no desenvolvimento de classificadores para aplicações farmacêuticas e alimentícias.

Palavras chaves: Antioxidantes; Quimioinformática; Análise multivariada; Conservantes Farmacêuticos

1. INTRODUCTION

The secondary metabolism of plants and other organisms houses a myriad of biologically active compounds with potentially therapeutic applicability (WANG *et al.*, 2020). Although several biochemical pathways are known to foster the expression of secondary metabolites, the shikimic acid pathway is widely considered as the main contributor when the biosynthesis of antioxidants is concerned, especially upon drought stress (SOUZA *et al.*, 2021; YADAV *et al.*, 2021) mainly through phytohormones homeostasis and their signaling networks, which further initiate the biosynthesis of secondary metabolites (SMs). Among these free-radical scavenging metabolic products are monoaromatic phenolic compounds, which are acknowledged to exhibit thermodynamic feasibility to mop up reactive oxygen species, as well as interact with several bodily receptors; thereby promoting anti-inflammatory effects; gastroprotective, neuroprotective as well as anticancer activities (ALVES *et al.*, 2020; OLIVEIRA *et al.*, 2021; THOMAZ *et al.*, 2018a)

Regarding the building block of monoaromatic antioxidants, the good electron acceptor/donor behavior provided by the resonance allow these molecules to participate in charge-transfer reactions with many other chemical species (THOMAZ *et al.*, 2020). Owing to this feature, many researchers have exploited the phenolic moiety as the backbone of synthetic antioxidant compounds, thereby employing medicinal chemistry strategies such as bioisosterism and latentiation to putatively enhance the antioxidant properties of these additives in drug, foodstuff and cosmetics (ARRUDA *et al.*, 2020; MOREIRA *et al.*, 2022; NAGARAJAN *et al.*, 2020).

Both natural (MORENO *et al.*, 2019; THOMAZ *et al.*, 2018a) and synthetic (LIU, MABURY, 2020; RESENDE *et al.*, 021) monoaromatic antioxidant compounds are known to showcase fairly diverse chemical structures, what usually turns the prediction of their biological activities and physicochemical behavior a strenuous task if *in vitro* and *in vivo* tests are considered. In this sense, high-throughput computational analysis of molecular descriptors would allow preliminary evaluation of selected attributes, and better shed light on the pharmacokinetics/pharmacodynamics of these compounds; as well as their medicinal uses (THORNBURG *et al.*, 2018).

Nonetheless, *in silico* studies of small druggable molecules are becoming ever-more present in chemistry due to the leap of processing power since the transition from analog computers to digital information (AGONI *et al.*, 2020; HUANG *et al.*, 2020b). Moreover, the availability of machine learning approaches as well as online servers for small molecule-targeted cheminformatic studies such as feature alignment/ pharmacophore/ toxicophore modeling further broadens the appeal of these methods, thereby allowing readily obtainable information concerning their predicted physicochemical features, absorption/ distribution/ metabolism/ excretion/ toxicity (ADMET), as well as classification according to particular attributes (FERREIRA, ANDRICOPULO, 2019).

Considering the application of the processing power of computers into physicochemical investigations, several authors have reported the implementation of semiempirical and *ab initio* quantum chemistry calculations in order to better investigate redox features which are whether/or not involved in antioxidant capacity (ALASADY *et al.*, 2020; KUMAR *et al.*, 2021; LEUNG *et al.*, 2018).

In this sense, the extended Hückel method (EHM) has been used to determine the energies of σ and π molecular orbitals, in order to trace correlations to the thermodynamical feasibility of electron-transfer (CONTARDI *et al.*, 2020; OLIVEIRA *et al.*, 2017); while other authors made use of more computational power-demanding density-functional theory calculations (AYOUBI-CHIANEH, KASSAEE, 2020; OLIVEIRA *et al.*, 2017; RODRIGUES *et al.*, 2019).

Nevertheless, the correlation of orbital energies to general physicochemical features such as molecular weight (MW); number of hydrogen bond donors (DHb) and acceptors (AHb); as well as routable bonds (RB); topological surface area (TSA); partition coefficient (Log P) and water solubility (WS) is not as often explored (HUANG *et al.*, 2020a).

Literature describes that ADMET of small druggable molecules such as monoaromatic antioxidants is heavily determined by their chemical structure. In this sense, the gastrointestinal absorption (GI) and blood-brain barrier permeability (BBB) of natural products change according to their physicochemical features (FERREIRA, ANDRICOPULO, 2019). Nonetheless, many reports on structure-activity relationship for computer-aided drug discovery/design make use of physicochemical parameters as molecular descriptors; henceforth implementing several statistical tools to mine data and uncover patterns which describe the potential biologic activity/features of groups of compounds (LU *et al.*, 2018; USHA *et al.*, 2018).

The implementation of *in silico* classifiers is becoming more common with the advance and availability of open-sourced libraries. The accessibility of these tools allows researchers to develop data separation platforms based on statistical modeling (*e.g.*, multivariate analysis) or non-statistical approaches using artificial intelligence for several applications in science. In fact, it has been reported the development of classifiers based on distinct data processing methods, such as associating principal components (PCA) and/or hierarchical cluster analysis, as well as supervised machine learning tools such as support-vector machines (SVM) to stablish reliable cheminformatic classifiers (SOROKINA, STEINBECK, 2020). Moreover, the pharmacometrics and pharmaceutical properties of natural products such as phenolic antioxidants have been successfully explored and reliably categorized by computational approaches; what further highlights the applicability of this technology (CAPECCHI, REYMOND, 2021; LI *et al.*, 2021; NAGY *et al.*, 2022).

Therefore, owing to the relevance of combining different techniques into the study of the physicochemical behavior and classification of natural and synthetic products of medicinal and/or industrial significance, we hereby report how to perform the cheminformatic classification of natural and synthetic monoaromatic antioxidants by multivariate analysis, data mining and machine learning algorithms. The classification tool developed in this work was based on self-fit of constrained data and is intended to show how natural products can be differentiated by their intrinsic physicochemical and biopharmaceutical features, even though their chemical structures share many similarities.

2. MATERIALS AND METHODS

2.1. Molecule selection

Herein, 20 natural monoaromatic antioxidants were selected following products of secondary plant metabolism from the shikimic acid biosynthetic pathway. The selected natural products were: gallic acid (GAc); ellagic acid (EAc); phloroglucinol (PG); benzoic acid (BAc); *p*-hydroxy benzoic acid (*p*-HBAc); gentisic acid (GenAc); 3,4 dihydroxy benzoic acid (3,4 di-HBAc); *p*-amino benzoic acid (p-ABAc); salicylic acid (SalAc); vanillic acid (VanAc); syringic acid (SyrAc); phenyl pyruvic acid (PhPAc); *p*-hydroxy phenyl pyruvic acid (p-HPhPAc); phenylalanine (PhAl); tyrosine (Tyr); cinnamic acid (CinAc); *p*-hydroxy cinnamic acid (p-HCinAc); 3,4 dihydroxy cinnamic acid (3,4 di-HCinAc); ferulic acid (FerAc); and synapinic acid (SynAc).

Moreover, 20 synthetic monoaromatic antioxidants commonly used in foodstuff and cosmetics were selected, namely: methyl paraben (MetPar); ethyl paraben (EtPar); propyl paraben (ProPar); isopropyl paraben (IsoProPar); butyl paraben (ButPar); isobutyl paraben (IsoButPar); heptyl paraben (HepPar); butyl hydroxy toluene (BHT); butylated hydroxy anisole (BHA); *t*-butyl hydroquinone (TBHQ); 2,4,5 tri-hydroxy butyrophenone (THBP); *o*-cresol (*o*-Cre); *m*-cresol (*m*-Cre); *p*-cresol (*p*-Cre); chlorocresol (ClCre); 2 phenoxy ethanol (PhEt); pyrogallol (PyrGa); propyl gallate (PropGa); ethoxyquin (EtQuin); and 2,4 dichloro phenoxy acetamide (2,4-DA).

2.2. Data gathering and pretreatment

The information regarding the physicochemical and pharmacokinetic properties of the selected compounds was gathered from PubChem database (KIM *et al.*, 2021), which was used to retrieve the isomeric (when available) or canonical simplified molecular-input line-entry system (SMILES) of each compound. Moreover, PubChem built-in features were also used to compute MW with PubChem 2.1; Log P with XLogP3 3.0; as well as DHb, AHb, RB and TSA with Cactvs 3.4.6.11. Thereafter, each SMILES was individually inputted into pkCSM database, wherein GI and BBB were also computed (PIRES *et al.*, 2015).

The SMILES-string of each compound was converted to a three-dimensional rendering of their chemical structures and submitted to steric energy minimization procedures, which involved force field approaches from classic molecular mechanics (MM2) and assisted model building and energy refinement (AMBER) toolsets (NETO *et al.*, 2019), which are detailed further in the methods section. All calculations were performed on Chem3D® Software and UCSF Chimera software (version 1.13.1) (PETTERSEN *et al.*, 2004). The resulting model (.mol2 extension) underwent editing whereupon charges were assigned using Biovia Discovery Studio® software. Manual corrections regarding aromatic bonds were also conducted and all structures were thoroughly reviewed before further experiments.

2.3. Alignment determination and rendering

To calculate the alignment of the compounds to investigate their shared structural features, a pharmacophore-modeling algorithm was used. Therefore, all treated structures were added to a single .mol2 extension file using UCSF Chimera software, and submitted to PharmaGist Webserver (SCHNEIDMAN-DUHOVNY *et al.*, 2008). The work conditions were: 5 output pharmacophores; minimum of 3 features in the predicted model; and the following weightings for contributor modelling: 3.0 for aromatic rings; 1.0 for charge (anion/cation); 1.5 for hydrogen bond (donor/acceptor); and 0.3 for hydrophobic contributors. Thereafter, the calculated models which presented the highest amount of hits (alignments) and highest score were rendered in 3D using Biovia Discovery Studio® software.

2.4. EHM and determination of molecular orbital energies

The semiempirical EHM was applied to all molecules aiming to evaluate the energy-gap between orbitals, which can be associated to the thermodynamic feasibility of redox reactions (RODRIGUES *et al.*, 2019). Following a previously described protocol (THOMAZ *et al.*, 2022), the EHM calculations were performed in order to compute the energies of the highest occupied molecular orbitals (HOMO) and lowest unoccupied molecular orbitals (LUMO).

The energy gap between HOMO and LUMO (ΔE) is inversely proportional to the redox reactivity of molecules, what therefore provides a quantitative insight on the required energy to promote charge transfer in the frontier orbitals. All calculations were conducted *in silico* post steric energy minimization by force field and classic molecular mechanics-based approaches (*i.e.,* MM2 and AMBER). Furthermore, the ΔE ($n_{Orbital} = 0$) was expressed in eV.

2.5. Energy refinement by MM2 and AMBER

To standardize the handling of cheminformatic data in this work, all molecules underwent energy refinement by MM2 and AMBER. As previously described (THOMAZ *et al.*, 2022), MM2 is a force field-based method which reliably reproduces the geometry of molecules at equilibrium by implementing a large set of continuously refined parameters, which are updated according to data regarding individual atoms and classes of organic compounds (HALGREN, 1992; PONDER, RICHARDS, 1987) a subject to which little attention has been given to date. We first show that the commonly used Lennard-Jones and Exp-6 potentials fail to account for the high quality rare-gas data but that a relatively simple distance-buffered potential (Buf-14-7, eq 10. This approach was selected to preliminarily minimize the steric energy of the compounds herein investigated and was accompanied by the application of AMBER. The minimum root-mean square gradient herein used to optimize the structures was of 0.010.

2.6. *Statistical modeling*

In this work, multivariate statistics in the form of PCA was performed (JOLLIFFE, CADIMA, 2016). This approach was selected to minimize dimensions basing on variance/correlation matrix, and to segregate observational variables based on their shared features. The two first components of the PCA were extracted and graphically represented as a biplot of the loadings and scores, with eigenvectors also represented therewithin. Moreover, a data mining approach was also used in the form of a feature selection algorithm employing MW; Log P; DHb; AHb; RB; TSA; WS; GI and BBB as continuous predictive variables, while ΔE was selected as a continuous dependent variable.

The results were expressed as importance table containing the f and p-values of the mined data. The f-value was herein used to indicate which variable impacted the most on the model and was calculated by the ratio of the variation between the dataset means and within the datasets. Furthermore, statistically significant difference was attributed to $p < 0.05$, and the calculations were carried out in Statistica® 12 software (StatSoft, Oklahoma, USA).

2.7. Machine learning algorithm

2.7.1. Support-Vector Machine (SVM)

The supervised machine learning model SVM was herein used to provide classification of the full dataset (CERVANTES *et al.*, 2020). This method consists of a non-probabilistic binary linear classifier which uses hyperplanes for multidimensional analysis. In this sense, the datasets associated to each attribute are scattered into the hyperplane and selected kernel-vectors are used to replace every dot product to better segregate/classify data. The kernel function herein selected was linear, being the code written in python language. The training set was randomly selected from the train dataset which encompassed 70% of all data, and the validation was carried out in the remaining 30% data.

Regarding SVM hyperparameters, the random state for the kernel implementation was of 0, and the random state of the selected test group was screened between 0 to 3. The final selected random state was the one which presented better response in the confusion matrix. This was performed to reduce the intrinsic bias associated to random selection in the small dataset herein used. The accuracy of the model, standard deviation and the confusion matrix were calculated and rendered in python. The following libraries were used in this work: sklearn; matplotlib and numpy, while plotly was used in graphics rendering.

Two SVM works were carried out in this work. The first considered as inputs all attributes, namely: MW; Log P; DHb; AHb; RB; TSA; WS; GI; BBB, and ΔE; being the outputs the main classes: natural and synthetic monoaromatic antioxidants. The second SVM work considered the dimension-reduced first two PC from the PCA model encompassing all data. This was performed to evaluate if the dimension-reduction promoted by the PCA would putatively improve the accuracy of the model. Results were presented as accuracy and standard deviation values, as well as confusion matrixes.

2.7.2. Multilayer Perceptron (MLP)

The MLP approach was herein used to allow data classification from inputted attributes (TANG *et al.*, 2016). MLP is a feedforward artificial neural network which uses backpropagation-based supervised machine learning, thereby allowing the use of layers of nodes (*i.e.*, input, hidden and output layers) whereupon non-linear activation functions are assigned (NAGY *et al.*, 2022). The work conditions for MLP implementation were alpha of 1e-05; automatic batch size; beta1 of 0.9; beta2 of 0.999; epsilon of 1e-08, sizes of hidden layers of (5,2); constant learning rate and beginning at 0.001. Moreover, the total number of iterations was of 500; the momentum was set to 0.9; the power was set to t = 0.5; random state set to 1 and validation fraction to 0.1. Furthermore, the training set encompassed 70% of all data, and the validation was carried out in the remaining 30% data. The following libraries were used in this work: sklearn; matplotlib and numpy, while plotly was used in graphics rendering.

Two MLP works were carried out in this work. The first considered as inputs all attributes, namely: MW; Log P; DHb; AHb; RB; TSA; WS; GI; BBB, and ΔE; being the outputs the main classes: natural and synthetic monoaromatic antioxidants. The second MLP work considered the dimension-reduced first two PC from the PCA model encompassing all data. This was performed to evaluate if the dimension-reduction promoted by the PCA would putatively improve the accuracy of the model. Results were presented as accuracy and standard deviation values, as well as confusion matrixes.

3. RESULTS AND DISCUSSION

3.1. Alignment modeling and data preparation for multivariate analysis

The first step in this study comprised the collection and treatment of data for the alignment investigation and multivariate analysis. Therefore, the collected datasets were segregated into two major groups, each containing 20 molecules; thereby resulting in 40 molecules in total.

The first dataset corresponded to monoaromatic antioxidants of natural origin, while the second dataset encompassed the synthetic antioxidants. Each dataset was composed of two files: *i.* the MM2 and AMBER-treated .mol2 extension file containing all chemical structures (which was submitted to PharmaGist webserver); and *ii.* the numeric data containing the SMILES-ID, selected molecular descriptors and corresponding ΔE of each compound.

The highest scoring alignment models for each dataset are presented in Figure 1, while the physicochemical/pharmacokinetic data is presented in Table 1.

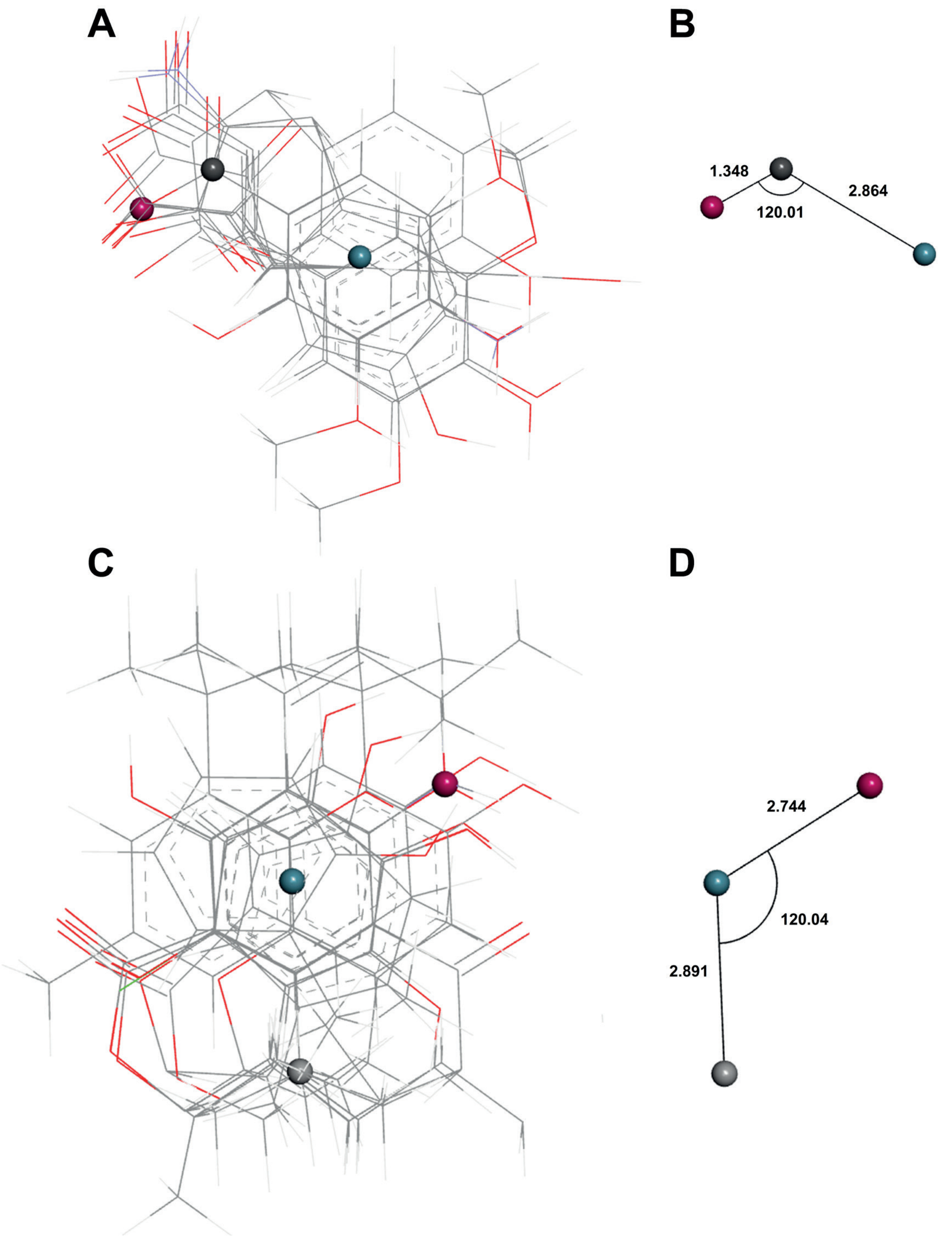

Figure 1. Frontal view of the highest scoring alignment model of natural (**A**) monoaromatic phenolic antioxidants and their respective distances in Å (**B**), 18 of the 20 inputted molecules showcased hit (score: 22.000). Frontal view of the highest scoring alignment model of synthetic monoaromatic antioxidants (**C**) and their respective distances in Å (**D**), 17 of the 20 inputted molecules showcased hit (score: 20.365). Aromatic contributors in cyan, hydrogen bond contributors in red, positive charge contributors in black and grey.

Table 1. Abbreviature, SMILES-ID, physicochemical and pharmacokinetic data of the natural and synthetic monoaromatic antioxidants herein investigated.

Name	MW*	Log P*	DHb*	AHb*	RB*	TSA*	WS$^{\#}$	GI$^{\#}$	BBB$^{\#}$	ΔE†
GAc	170.12	0.7	4	5	1	98	-2.56	43.374	-1.102	-8.016
EAc	302.19	1.1	4	8	0	134	-3.181	86.684	-1.272	-6.827
PG	126.11	0.2	3	3	0	60.7	-1.408	83.549	-0.466	-14.874
Bac	122.12	1.9	1	2	1	37.3	-1.738	100	-0.22	-8.713
***p*-HBAc**	138.12	1.6	2	3	1	57.5	-1.877	83.961	-0.334	-9.461
GenAc	154.12	1.6	3	4	1	77.8	-2.009	80.078	-0.697	-8.010
3,4 di-HBAc	154.12	1.1	3	4	1	77.8	-2.069	71.174	-0.683	-8.447
***p*-ABAc**	197.14	0.8	2	3	1	63.3	-1.907	81.966	-0.389	-8.988
SalAc	138.12	2.3	2	3	1	57.5	-1.808	83.887	-0.334	-8.733
VanAc	168.15	1.4	2	4	2	66.8	-1.838	78.152	-0.38	-8.318
SyrAc	198.17	1	2	5	3	76	-2.223	73.076	-0.191	-8.069
PhPAc	164.16	1.3	1	3	3	54.4	-2.016	80.367	-0.173	-4.348
***p*-HPhPAc**	180.16	0.9	2	4	3	74.6	-1.902	75.043	0.019	-4.358
PhAl	165.19	-1.5	2	3	3	63.3	-2.89	76.21	-0.271	-10.608
Tyr	181.19	-2.3	3	4	3	83.6	-2.89	73.014	-0.698	-10.653
CinAc	148.16	2.1	1	2	2	37.3	-2.608	94.833	0.446	-6.958
***p*-HCinAc**	164.16	1.5	2	3	2	57.5	-2.378	93.378	-0.225	-7.161
3,4 di-HCinAc	180.16	1.2	3	4	2	77.8	-2.33	69.407	-0.647	-6.170
FerAc	194.18	1.5	2	4	3	66.8	-2.817	93.685	-0.239	-6.116
SynAc	224.21	1.5	2	5	4	76	-2.869	93.064	-0.247	-5.847
MetPar	152.15	2	1	3	2	46.5	-1.881	89.457	-0.222	-9.001
EtPar	166.17	2.5	1	3	3	46.5	-2.098	93.728	0.352	-8.912
ProPar	180.2	3	1	3	4	46.5	-2.409	93.328	0.303	-8.873
IsoProPar	180.2	2.8	1	3	3	46.5	-2.501	94.581	0.273	-8.870
ButPar	194.23	3.6	1	3	5	46.5	-2.735	92.708	0.288	-8.859
IsoButPar	194.23	3.4	1	3	4	46.5	-2.785	93.564	0.267	-8.851
HepPar	236.31	4.8	2	3	8	46.5	-3.934	92.856	-0.504	-8.848
BHT	220.35	5.3	1	1	2	20.2	-4.834	91.904	0.434	-12.88
BHA	180.24	3.2	1	2	2	29.5	-2.774	92.762	0.361	-11.841
TBHQ	166.22	2.8	2	2	1	40.5	-1.682	91.427	0.388	-11.912
THBP	196.20	1.9	3	4	3	77.8	-2.096	92.812	-0.868	-7.358
***o*-Cre**	108.14	2	1	1	0	20.2	-0.961	93.067	0.348	-13.256
***m*-Cre**	108.14	2	1	1	0	20.2	-0.961	93.067	0.348	-13.435
***p*-Cre**	108.14	1.9	1	1	0	20.2	-0.961	93.067	0.348	-12.837
ClCre	142.58	3.1	1	1	0	20.2	-1.539	91.406	0.289	-11.250
PhEt	138.16	1.2	1	2	3	29.5	-0.742	85.558	-0.125	-12.729
PyrGa	126.11	0.5	3	3	0	60.7	-1.408	83.549	-0.441	-12.997
PropGa	212.20	1.8	3	5	4	87	-2.113	92.439	-1.132	-7.779

Name	MW*	Log P*	DHb*	AHb*	RB*	TSA*	WS$^{\#}$	GI$^{\#}$	BBB$^{\#}$	ΔE†
EtQuin	217.31	3.1	1	2	2	21.3	-3.365	92.118	0.403	-9.527
2,4-DA	220.05	2.2	1	2	3	52.3	-2.131	91.855	-0.04	-11.186

* MW was computed by PubChem 2.1 and is expressed in g mol^{-1}; Log P was computed by XLogP3 3.0; DHb/AHb/ RB and TSA (in Å^2) were computed by Cactvs 3.4.6.11; # Data retrieved from PkCSM, WS is expressed in numeric Log mol l^{-1}; GI is expressed in %; BBB is expressed in numeric Log BB. † ΔE in eV = (HOMO-LUMO) energy values, which were obtained by EHM. Data from synthetic monoaromatic phenolic antioxidants.

The alignment calculation evidenced that 18 of the 20 inputted molecules of the monoaromatic antioxidants of natural origin superimposed in a high-scored model (*i.e.,* 22.000); while 17 of the 20 inputted synthetic antioxidants superimposed in a high-scored model (i.e., 20.365). Moreover, the main contributors in each model were negatively charged (hydrogen bond donors), positively charged and aromatic; being the angle between them of ≈ 120º in both models.

As can be seen in Table 1, the SMILES-ID of each compound was particular and singular to each of them, which is an expected finding since this string describes their chemical structure. Regarding the physicochemical and pharmacokinetic features, MW ranged from 108.14 g mol^{-1} to 302.19 g mol^{-1}; Log P ranged from 0.2 to 5.3; DHb ranged from a single donor to 4 donors; AHb ranged from a single acceptor to 8 acceptors; RB ranged from no routable bond to 8 routable; TSA ranged from 20.2 Å^2 to 134 Å^2; WS ranged from -4.834 Log mol l^{-1} to -0.742 Log mol l^{-1}; GI ranged from 43.374% to 100%; BBB ranged from -1.102 Log BB to 0.446 Log BB; and ΔE ranged from -14.874 eV to -4.348 eV.

Monoaromatic phenolic antioxidants are known to exhibit antioxidant behavior due to their thermodynamic feasibility to undergo oxidation (ALVES *et al.*, 2020) it was evaluated the phenolic content, redox behavior and antioxidant capacity of several selected teas and tisanes from Brazilian market. The samples were classified as simple (single herb, thereby donating one proton and one electron to stabilize free radicals/reactive oxygen species (THOMAZ *et al.*, 2018b). This process is favored by the electron donor-acceptor properties of aromatic compounds, as well as to the influence of the hydroxyl group, which acts as a ring activator by increasing the electron density of the aryl moiety (THOMAZ *et al.*, 2020). In this sense, the presence of aromatic and negatively charged contributor in the modeling is an expected finding.

Moreover, when taking into account that the highly electronegative oxygen in the hydroxyl promotes a permanent dipole on the molecule due to its inductive effect on adjacent carbon atoms; the presence of a positively-charged contributor is also an expected finding (CONTARDI *et al.*, 2020). This same effect may also be promoted by the aromatic ring on atoms directly bounded to it, what also explains the different positions of the positively charged contributors in both models.

3.2. Multivariate analysis

To correlate the information of all datasets, PCA was performed. Therefore, all physicochemical/pharmacokinetic descriptors were inputted as variables, while the names of the molecules were selected as observational labels for the scoring plot of the model. PCA was performed in three separate experiments, namely: *i.* a simultaneous calculation encompassing all datasets (*i.e.,* natural and synthetic monoaromatic antioxidants); *ii.* a calculation encompassing the dataset of the natural monoaromatic antioxidants; and *iii.* a calculation encompassing the dataset of the synthetic monoaromatic antioxidants. Results are presented as PCA biplots and correlation matrixes in Figure 2 and Table 2, respectively.

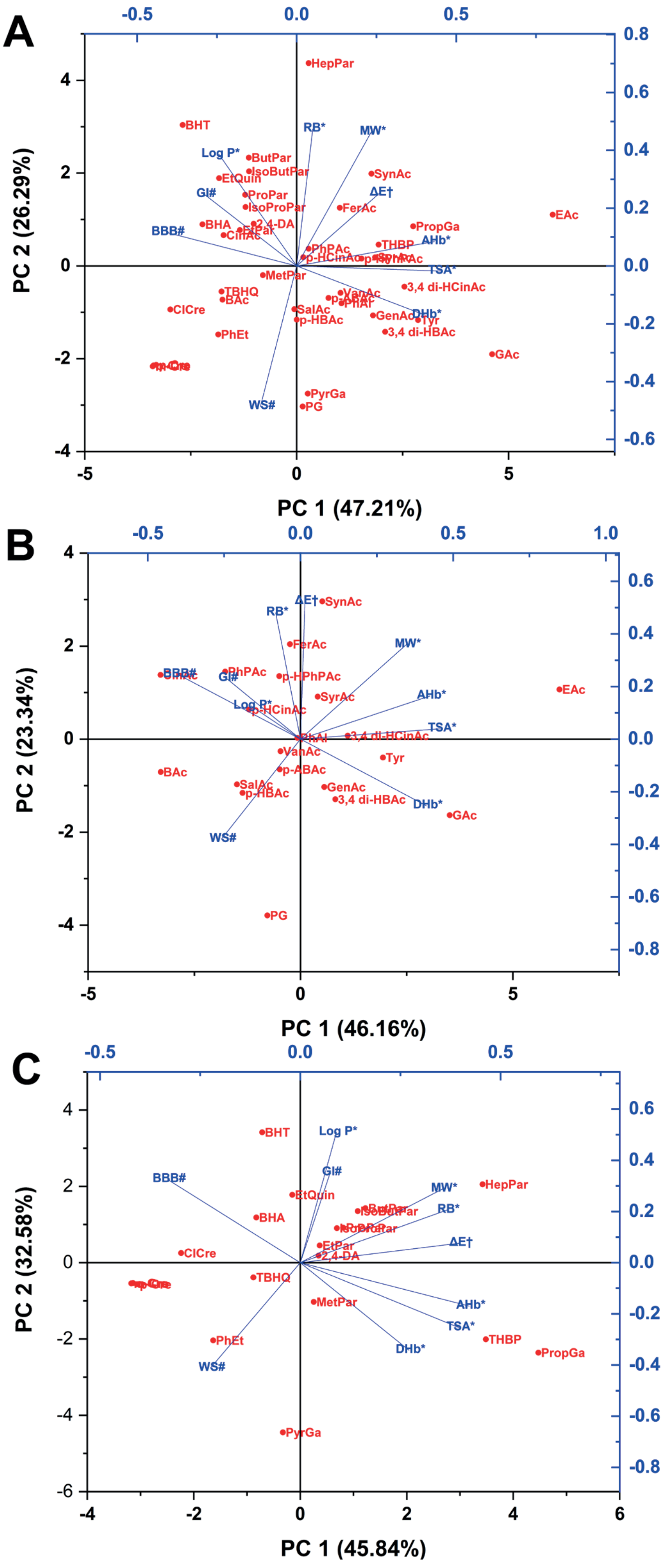

Figure 2. PCA biplot of all datasets (**A**), biplot of the data from monoaromatic antioxidants of natural origin (**B**), and the dataset from synthetic monoaromatic antioxidants (**C**).

Table 2. Correlation matrix of the physicochemical and pharmacokinetic data of all monoaromatic antioxidants herein investigated.

All	MW*	Log P*	DHb*	AHb*	RB*	TSA*	WS#	GI#	BBB#	ΔE†
MW*	1	0.20243	0.25879	0.5914	0.50299	0.492	**-0.74114**	0.02451	-0.29748	0.43338
Log P*	0.20243	1	-0.52278	-0.39306	0.28059	-0.53005	-0.32953	0.55336	0.50845	-0.07532
DHb*	0.25879	-0.52278	1	**0.75165**	-0.16963	**0.85873**	-0.07455	-0.62228	**-0.88467**	0.22596
AHb*	0.5914	-0.39306	**0.75165**	1	0.17115	**0.95441**	-0.25031	-0.43676	**-0.77227**	0.61332
RB*	0.50299	0.28059	-0.16963	0.17115	1	0.05424	-0.51258	0.15959	0.01233	0.38741
TSA*	0.492	-0.53005	**0.85873**	**0.95441**	0.05424	1	-0.17502	-0.55081	**-0.85905**	0.54922
WS#	**-0.74114**	-0.32953	-0.07455	-0.25031	-0.51258	-0.17502	1	-0.01996	0.04517	-0.28821
GI#	0.02451	0.55336	-0.62228	-0.43676	0.15959	-0.55081	-0.01996	1	0.53522	-0.19288
BBB#	-0.29748	0.50845	**-0.88467**	**-0.77227**	0.01233	**-0.85905**	0.04517	0.53522	1	-0.3422
ΔE†	0.43338	-0.07532	0.22596	0.61332	0.38741	0.54922	-0.28821	-0.19288	-0.3422	1
Natural	MW*	Log P*	DHb*	AHb*	RB*	TSA*	WS#	GI#	BBB#	ΔE†
MW*	1	-0.10235	0.4062	**0.86428**	0.14173	**0.78392**	-0.67687	-0.01275	-0.4235	0.38805
Log P*	-0.10235	1	-0.32196	-0.09398	-0.23839	-0.25898	0.34356	0.38227	0.26103	0.46238
DHb*	0.4062	-0.32196	1	**0.70005**	-0.43383	**0.84985**	-0.27926	-0.60195	**-0.89815**	-0.24695
AHb*	**0.86428**	-0.09398	**0.70005**	1	-0.05604	**0.95166**	-0.52922	-0.3017	-0.68746	0.23743
RB*	0.14173	-0.23839	-0.43383	-0.05604	1	-0.17233	-0.39302	0.05558	0.48795	0.46517
TSA*	**0.78392**	-0.25898	**0.84985**	**0.95166**	-0.17233	1	-0.50946	-0.46501	**-0.82208**	0.1169
WS#	-0.67687	0.34356	-0.27926	-0.52922	-0.39302	-0.50946	1	0.02778	0.23792	-0.28361
GI#	-0.01275	0.38227	-0.60195	-0.3017	0.05558	-0.46501	0.02778	1	0.52094	0.06748
BBB#	-0.4235	0.26103	**-0.89815**	-0.68746	0.48795	**-0.82208**	0.23792	0.52094	1	0.19322
ΔE†	0.38805	0.46238	-0.24695	0.23743	0.46517	0.1169	-0.28361	0.06748	0.19322	1
Synthetic	MW*	Log P*	DHb*	AHb*	RB*	TSA*	WS#	GI#	BBB#	ΔE†
MW*	1	0.62218	0.15997	0.46637	**0.74797**	0.40126	**-0.83629**	0.30785	-0.26216	0.61716
Log P*	0.62218	1	-0.32514	-0.15972	0.47976	-0.2802	**-0.85383**	0.52293	0.34485	0.1939
DHb*	0.15997	-0.32514	1	0.62452	0.11644	**0.75567**	0.04788	-0.31543	**-0.82394**	0.25397
AHb*	0.46637	-0.15972	0.62452	1	0.59867	**0.92973**	-0.15977	0.00177	**-0.73272**	**0.81796**
RB*	**0.74797**	0.47976	0.11644	0.59867	1	0.45198	-0.56768	0.25641	-0.35638	0.69041
TSA*	0.40126	-0.2802	**0.75567**	**0.92973**	0.45198	1	-0.04144	-0.0725	**-0.81561**	0.6699
WS#	**-0.83629**	**-0.85383**	0.04788	-0.15977	-0.56768	-0.04144	1	-0.32652	-0.05267	-0.36704
GI#	0.30785	0.52293	-0.31543	0.00177	0.25641	-0.0725	-0.32652	1	0.29895	0.3942
BBB#	-0.26216	0.34485	**-0.82394**	**-0.73272**	-0.35638	**-0.81561**	-0.05267	0.29895	1	-0.43452
ΔE†	0.61716	0.1939	0.25397	**0.81796**	0.69041	0.6699	-0.36704	0.3942	-0.43452	1

* MW was computed by PubChem 2.1 and is expressed in g mol^{-1}; Log P was computed by XLogP3 3.0; DHb/AHb/ RB and TSA (in Å^{2}) were computed by Cactvs 3.4.6.11; # Data retrieved from PkCSM, WS is expressed in numeric Log mol l^{-1}; GI is expressed in %; BBB is expressed in numeric Log BB. † ΔE in eV = (HOMO-LUMO) energy values, which were obtained by EHM. Data from synthetic monoaromatic phenolic antioxidants. **Bold values represent the most representative correlations.**

Results showed that the first two PCs in the calculations encompassing all data accounted for 73.5% of all variances; while the first two PCs of the calculations encompassing the datasets of natural and synthetic monoaromatic antioxidants accounted for 69.5% and 78.42%, respectively.

Concerning the PCA encompassing all datasets, Figure 2.A depicted that the descriptive eigenvectors representing Log P, GI and BBB were displaced in the II quadrant of the biplot; while those of RB, MW, ΔE and AHb were displaced in the I quadrant. Moreover, the eigenvectors of TSA and DHb were in the IV quadrant of the biplot, while WS eigenvector was located in the III quadrant. The displacement and convergence of these eigenvectors could also be noted in the correlation matrix (Table 2). Furthermore, although there was no absolute clustering of the scores in the biplot; some grouping could be suggested such as the phenylpropanoids in the I quadrant, the parabens in the II quadrant, the cresols in the III quadrant, and the short-chained phenols in the IV quadrant.

Regarding the PCA encompassing the antioxidants of natural origin, Figure 2.B, presented that the descriptive eigenvectors representing Log P, GI and BBB were also displaced in the II quadrant of the biplot, though their convergence was bigger than what seen in Figure 2.A, and RB eigenvector is now in this quadrant. The eigenvectors representing MW, ΔE and AHb were again displaced in the I quadrant, however the TSA eigenvector is now in this quadrant. Moreover, the eigenvector of DHb was in the IV quadrant of the biplot; while WS eigenvector was in the III quadrant, as in Figure 2.A. The displacement and convergence of these eigenvectors was also suggested in the correlation matrix (Table 2). Furthermore, some grouping of short-chained phenols on the III and IV quadrants could be suggested.

The PCA encompassing the antioxidants of synthetic nature presented the highest amount of accounted variance by the first two PCs, what is presented in the correlation matrix (Table 2). The eigenvectors also were differently displaced form what was seen in Figures 2.A and B, being that Log P and GI eigenvectors are now in the I quadrant, while AHb eigenvector is in the IV quadrant (Figure 2.C). Moreover, some grouping could be suggested for parabens in the I quadrant and for cresols in the III quadrant.

When taking into account that a higher partition coefficient value (*i.e.,* Log P) indicates the propension of chemicals to solubilize in less-polar media, considering an interface between two immiscible solvents (polar and less-polar) at equilibrium, it can be suggested that the convergence between Log P, GI and BBB eigenvectors is in agreement with the literature (THOMPSON *et al.*, 2012). This interpretation is since the penetration of compounds through the tissues of the GI tract and endothelial epithelium is favored by less-polar compounds (*i.e.,* greater Log P). Moreover, the very lipophilic nature of the outer-phospholipidic bilayer of both enteral and endothelial cells also suggest that small lipophilic compounds are more feasible to permeate through them, what is nonetheless widely reported (CHARALABIDIS *et al.*, 2019; MATSUMURA *et al.*, 2020).

Considering that all compounds shared a monoaromatic core, the increase of their MW also suggested higher structural complexity. In this sense, the increase in RB is an expected finding, as the increase of the atom count of non-aromatic substituents would increase the amount of frontally-overlapped orbitals (σ bonds); which are known to exhibit rotativity (KRAPP *et al.*, 2006) Pauli repulsion ΔEPauli and attractive orbital interactions ΔEorb. The energy terms are compared with the orbital overlaps at different interatomic distances. The quasiclassical electrostatic interactions between two electrons occupying 1s, 2s, 2p (σ. Moreover, owing to the fact that hydroxyl moieties are good donors and acceptors of hydrogen bonds (NETO *et al.*, 2019), some degree of correlation between their respective eigenvectors is expected; what was nonetheless hinted by the PCA and the correlation matrix.

3.3. Application of the feature selection algorithm

After multivariate analysis, all datasets were submitted to a feature selection data mining algorithm. Therefore, MW; Log P; DHb; AHb; RB; TSA; WS; GI and BBB were inputted as continuous predictive variables, while ΔE was selected as a continuous dependent variable. The results are presented in Table 3.

Table 3. Feature selection of the continuous predictors for the dependent variable ΔE† according to the data mining model. The predictors were ranked according to their F and p values.

Predictor	F-value	p-value
WS#	6.377698	0.000068
AHb*	5.836442	0.000539
MW*	3.401675	0.007892
RB*	3.028577	0.022904
TSA*	2.865562	0.019340
DHb*	1.908551	0.145647
Log P*	1.862857	0.102604
GI#	1.477862	0.222744
BBB#	1.415533	0.229178

* MW was computed by PubChem 2.1 and is expressed in g mol^{-1}; Log P was computed by XLogP3 3.0; DHb/AHb/ RB and TSA (in Å^{2}) were computed by Cactvs 3.4.6.11; # Data retrieved from PkCSM, WS is expressed in numeric Log mol l^{-1}; GI is expressed in %; BBB is expressed in numeric Log BB. † ΔE in eV = (HOMO-LUMO) energy values, which were obtained by EHM. Data from synthetic monoaromatic phenolic antioxidants.

Results presented that the feature selection data mining algorithm favored WS, AHb, MW, RB and TSA as the main contributors to the model; being that all of them presented statistically significant difference and importance (Table 3). On the other hand, DHb; Log P; GI and BBB did not present statistically significant difference.

Considering that aryl-bound moieties containing highly electronegative atoms enhance the reactivity of the compound, and thereby turns it more susceptible to undergo redox reactions (CONTARDI *et al.*, 2020); the significant contribution of WS and AHb is an expected trend. This can be justified by the fact that the permanent dipole promoted by electronegative atoms such as oxygen leads to a higher feasibility of hydration (what is opposed by Lop P as well as lipophilicity-based parameters such as GI and BBB) (BANERJEE *et al.*, 1980).

3.4. Machine learning algorithms

3.4.1. SVM and MLP results

The SVM employing linear kernel function used in this work aimed to stablish a classifier model of all datasets; thereby allowing the separation of the natural and synthetic monoaromatic antioxidants according to their imputed features. Therefore, two main approaches were conducted: *i.* a linear-kernel SVM calculation was implemented on all dataset, and *ii.* a linear-kernel SVM calculation was implemented with the dataset provided by the first two PCs (*i.e.,* the ones which explained the highest amount of cumulative variance in the model).

Although the use of the first two PCs would be counter-intuitive due to the data loss of the dimension reduction provided by the PCA, it must be considered that PCA is a common tool in hierarchical clustering analysis and *in tandem* with classification approaches. Furthermore, the MLP implementation in this work was intended to evaluate if the use of an artificial neural network would provide improvement of the classification in comparison to the SVM non-probabilistic binary linear classifier. Therefore, two main MLP works were conducted akin to SVM, namely: *i.* a MLP calculation using all dataset, and *ii.* a MLP calculation using the dataset provided by the first two PCs (*i.e.,* the ones which explained the highest amount of cumulative variance in the model). Results are presented in Figure 3.

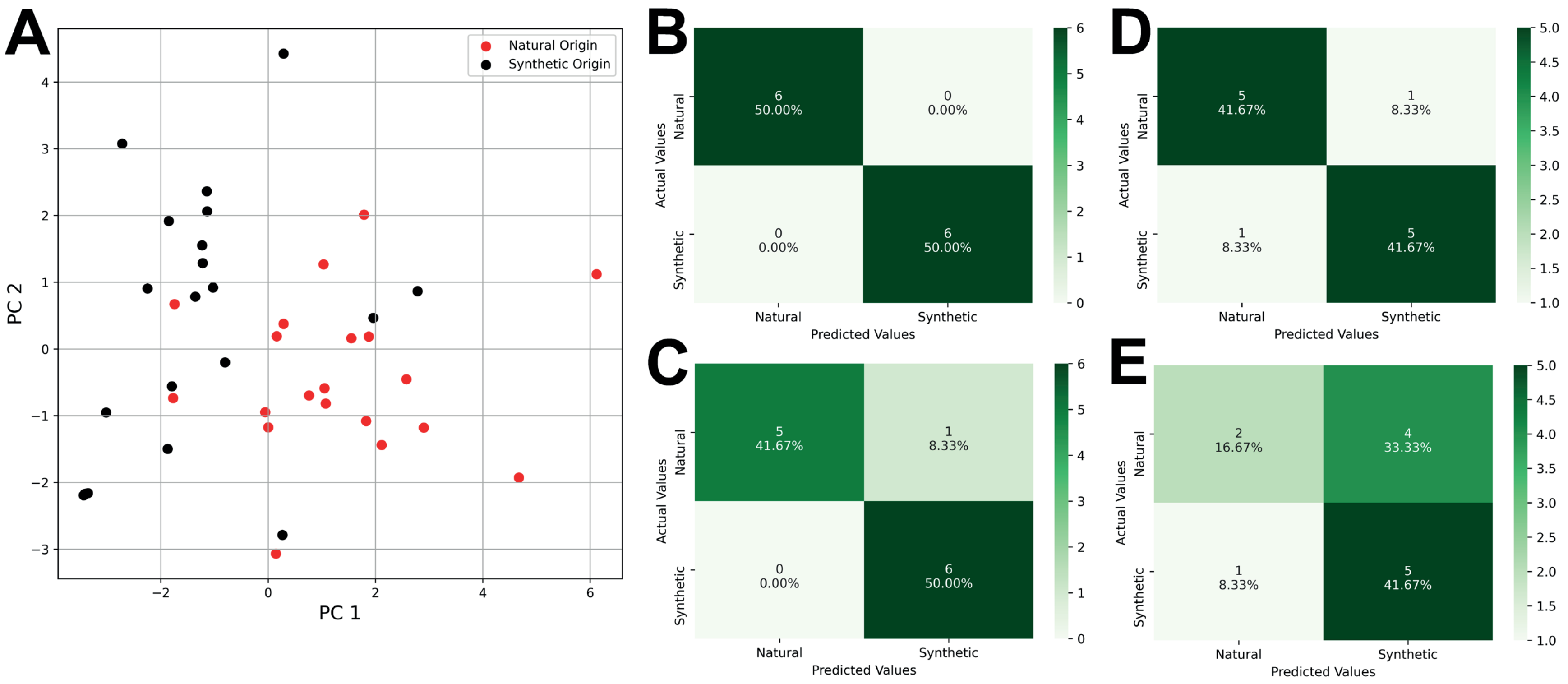

Figure 3. PCA score plot of all datasets with classification of the inputs as natural or synthetic monoaromatic antioxidants (**A**). Confusion matrix of the SVM linear kernel algorithm-results employing all attributes of the complete dataset (**B**). Confusion matrix of the SVM linear kernel algorithm-results employing the first two PCs of the PCA using all dataset as inputs (**C**). Confusion matrix of the MLP algorithm-results employing all attributes of the complete dataset (**D**). Confusion matrix of the MLP algorithm-results employing the first two PCs of the PCA using all dataset as inputs (**E**).

The PCA scoring plot allowed the partial differentiation of the scattered data, what is nonetheless a remarkable finding (Figure 3.A). The SVM accuracy and standard deviation results for both models were 85 % and 20 %; although their responses differed as presented in the confusion matrix (Figure 3.B and C). Moreover, MLP yielded 80% accuracy and 18.71% standard deviation in both works, although the confusion matrix exhibited different responses according to the imputed data (Figure 3. D and E).

The PCA scoring plot (Figure 3.A) also suggested that the information therewithin contained in the datasets could be classified, although some overlap could be noticed. In this sense, the application of SVM linear kernel-vectors and MLP would most likely allow the segregation of the datasets upon calculations using PC1 and PC2 as inputs. However, the overall results (*i.e.*, accuracy and standard deviation) were the same upon SVM and MLP using all attributes and PC1/PC2 as inputs, what is a remarkable finding. Although PCA is a statistical dimension-reduction method, this effect seemingly didn't impact the accuracy and standard deviation of both SVM and MLP models.

Regarding SVM, the confusion matrixes presented difference, since the one using all datasets exhibited 50% of hits regarding true positives; no false positive regarding natural monoaromatic antioxidants; no false positives regarding synthetic monoaromatic antioxidants, and 50% true positives regarding synthetic antioxidants. On the other hand, the confusion matrix containing the first two PCs as inputs presented 41.67 % of hits regarding true positives; 8.33% of false positive regarding natural monoaromatic antioxidants; no false positives regarding synthetic monoaromatic antioxidants, and 50% true positives regarding synthetic antioxidants.

Concerning MLP, the confusion matrixes presented that the one using all datasets exhibited 41.67% of hits regarding true positives; 8.33% of false positive regarding natural monoaromatic antioxidants; 8.33% of false positives regarding synthetic monoaromatic antioxidants, and 41.67% true positives regarding synthetic antioxidants. On the other hand, the confusion matrix containing the first two PCs as inputs presented 16.67 % of hits regarding true positives; 33.33% of false positive regarding natural monoaromatic antioxidants; 8.33% of false positives regarding synthetic monoaromatic antioxidants, and 41.67% true positives regarding synthetic antioxidants.

In this sense, it can be suggested that both SVM and MLP approaches presented similar accuracies and standard deviations; although the dimension reduction promoted by the PCA seemingly impaired the adequate classification of the compounds in natural and synthetic. Therefore, the best results were obtained by imputing all attributes in the SVM calculations, while MLP calculations with all attributes yielded the second-best results. This is an important finding to refine the application of machine learning classifiers based on physicochemical and biopharmaceutical descriptors of small molecules, as the adequate selection of the tool is critical for potential applications in discrimination analysis, such as the classification of food and drug preservatives according to their sourcing that was herein described.

Even though our results are promising, this investigation was intended as a proof-of-concept, and more studies are to be performed to further shed light on the use of mathematical and machine learning classifiers to evaluate the similarities and distinctions between datasets composed by the biopharmaceutical and physicochemical descriptors of food and drug preservatives from natural and synthetic nature.

The limitations of this work are: i) the reduced number of substances evaluated; ii) the absence of experimental data regarding their antioxidant activity; and iii) many properties that were considered are predicted values. Thus, we investigated a predictive model based on a mixture of empirical and predicted data, which increases bias.

4. CONCLUSIONS

This work showcased how to classify natural and synthetic monoaromatic antioxidants by means of multivariate analysis, data mining and machine learning algorithms. Altogether the data showcased that the alignment model suggested similarity between the major aromatic, negative and positively charged contributors. Moreover, the physicochemical and biopharmaceutical molecular descriptors remarkedly correlated with each other, thereby allowing reliable and prompt classification of natural and synthetic monoaromatic antioxidants. Furthermore, we hope that this work sheds light on the use of artificial intelligence techniques to develop classification tools for natural products in scientific and industrial applications.

5. ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

6. CONFLICTS OF INTEREST

None.

7. REFERENCES

AGONI, C.; OLOTU, F.A.; RAMHARACK, P.; SOLIMAN, M.E. Druggability and drug-likeness concepts in drug design: are biomodelling and predictive tools having their say? **Journal of Molecular Modeling**, v. 26, n. 120, p. 1-11, 2020. DOI: https://doi.org/10.1007/s00894-020-04385-6

ALASADY, S.A.; MUHAMAD, Y.H.; AHMED, R.S. Theoretical and thermodynamics studies of complexes formation between natural flavonoids and Hg (II) ion. *2393* **Systematic Reviews in Pharmacy**, v. 11, n. 12, p. 2393–2404, 2020. DOI: https://doi.org/10.31838/srp.2020.12.362

ALVES, C.B. RODRIGUES, E.S.B.; THOMAZ, D.V.; AGUIAR FILHO, A.M.; GIL, E.S.; COUTO, R.O. Correlation of polyphenol content and antioxidant capacity of selected teas and tisanes from Brazilian market. **Brazilian Journal of Food Technology**, v. 23, e2020036, p. 1-15, 2020. DOI: https://doi.org/10.1590/1981-6723.03620

ARRUDA, E.L.; JAPIASSU, K.B.; SOUZA, P.L.M.; ARAÚJO, K.C.F.; THOMAZ, D.V.; CORTEZ, A.P.; GARCIA, L.F.; VALADARES, M.C.; GIL, E.S.; OLIVEIRA, V. Zidovudine glycosylation by filamentous fungi leads to a better redox stability and improved cytotoxicity in B16F10 murine melanoma cells. **Anti-Cancer Agents in Medicinal Chemistry**, v. 20, n. 14, p. 1688-1694, 2020. DOI: https://doi.org/10.2174/1871520620666200424112504

AYOUBI-CHIANEH, M.; KASSAEE, M.Z. Novel silylphenol antioxidants by density functional theory. **Journal of the Chinese Chemical Society**, v. 67, n. 11, p. 1986–1991, 2020. DOI: https://doi.org/10.1002/jccs.202000131

BANERJEE, S.; YALKOWSKY, S.H.; VALVANI, S.C. Water solubility and octanol/water partition coefficients of organics. limitations of the solubility-partition coefficient Correlation. **Environmental Science and Technology**, v. 14, n. 10, p. 1227–1229, 1980. DOI: https://doi.org/10.1021/es60170a013

CAPECCHI, A.; REYMOND, J.L. Classifying natural products from plants, fungi or bacteria using the COCONUT database and machine learning. **Journal of Cheminformatics**, v. 13, n. 82, p. 1-11, 2021. DOI: https://doi.org/10.1186/s13321-021-00559-3

CERVANTES, J.; GARCIA-LAMONT, F.; RODRÍGUEZ-MAZAHUA, L.; LOPEZ, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. **Neurocomputing**, v. 408, p. 189-215, 2020. DOI: https://doi.org/10.1016/j.neucom.2019.10.118

CHARALABIDIS, A.; SFOUNI, M.; BERGSTRÖM, C.; MACHERAS, P. The Biopharmaceutics Classification System (BCS) and the Biopharmaceutics Drug Disposition Classification System (BDDCS): Beyond guidelines. **International Journal of Pharmaceutics**, v. 566, p. 264–281, 2019. DOI: https://doi.org/10.1016/j.ijpharm.2019.05.041

CONTARDI, U.A.; MORIKAWA, M.; THOMAZ, D.V. Redox behavior of central-acting opioid tramadol and its possible role in oxidative stress. **Medical Sciences Forum**, v. 2, n. 1, p. 1-7, 2020. DOI: https://doi.org/10.3390/CAHD2020-08557

FERREIRA, L.L.G.; ANDRICOPULO, A.D. ADMET modeling approaches in drug discovery. **Drug Discovery Today**, v. 24, n. 5, p. 1157–1165, 2019. DOI: https://doi.org/10.1016/j.drudis.2019.03.015

HALGREN, T. A. Representation of van der Waals (vdW) interactions in molecular mechanics force fields: Potential form, combination rules, and vdW parameters. **Journal of the American Chemical Society**, v. 114, n. 20, p. 7827–7843, 1992. DOI: https://doi.org/10.1021/ja00046a032

HUANG, C.; ZHOU, Y.; YANG, J.; CUI, Q.; LI, Y. A new metric quantifying chemical and biological property of small molecule metabolites and drugs. **Frontiers in Molecular Biosciences**, v. 7, n. 594800, p. 1-9, 2020a. DOI: https://doi.org/10.3389/fmolb.2020.5948

HUANG, C.; YANG, J.; CUI, Q. A simple and efficient metric quantifying druggable property of chemical small molecules. **bioRxiv**, p. 2020.07.13.199752, 2020b. DOI: https://doi.org/10.1101/2020.07.13.199752

JOLLIFFE, I.T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical Transactions of the Royal Society A - Mathematical, Physical and Engineering Sciences**, v. 374, n. 2065, p. 20150202, 2016. DOI: https://doi.org/10.1098/rsta.2015.0202.

KIM, S.; CHEN, J.; CHENG, T.; GINDULYTE, A.; HE, J.; HE, S.; LI, Q.; SHOEMAKER, B.A; THIESSEN, P.A.; YU, B.; ZASLAVSKY, L.; ZHANG, J.; BOLTON, E.E. PubChem in 2021: new data content and improved web interfaces. **Nucleic acids research**, v. 49, n. D1, p. D1388–D1395, 2021. DOI: https://doi.org/10.1093/nar/gkaa971

KRAPP, A.; BICKELHAUPT, F.M.; FRENKING, G. Orbital overlap and chemical bonding. **Chemistry - A European Journal**, v. 12, n. 36, p. 9196-9216, 2006. DOI: https://doi.org/10.1002/chem.200600564

KUMAR, N.; GUSAIN, A.; KUMAR, J.; SINGHA, R.; HOTA, P.K. Anti-oxidation properties of 2-substituted furan derivatives: A mechanistic study. **Journal of Luminescence**, v. 230, p. 117725, 2021. DOI: https://doi.org/10.1016/j.jlumin.2020.117725

LEUNG, R.; VENUS, C.; ZENG, T.; TSOPMO. A. Structure-function relationships of hydroxyl radical scavenging and chromium-VI reducing cysteine-tripeptides derived from rye secalin. **Food Chemistry**, v. 254, p. 165–169, 2018. DOI: https://doi.org/10.1016/j.foodchem.2018.01.190

LI, S.; YU, Y.; BIAN, X.; YAO, L.; LI, M.; LOU, Y.R.; YUAN, J.; LIN, H.; LIU, L.; HAN, B.; XIANG, X. Prediction of oral hepatotoxic dose of natural products derived from traditional Chinese medicines based on SVM classifier and PBPK modeling. **Archives of Toxicology**, v. 95, n. 5, p. 1683-1701, 2021. DOI: https://doi.org/10.1007/s00204-021-03023-1

LIU, R.; MABURY, S.A. Synthetic phenolic antioxidants: A review of environmental occurrence, fate, human exposure, and toxicity. **Environmental Science and Technology**, v. 54, n. 19, p. 11706–11719, 2020. DOI: https://doi.org/10.1021/acs.est.0c05077

LU, W.; ZHANG, R.; JIANG, H.; ZHANG, H.; LUO, C. Computer-aided drug design in epigenetics. **Frontiers in Chemistry**, v. 6, n. 57, p. 1–23, 2018. DOI: https://doi.org/10.3389/fchem.2018.0005

MATSUMURA, N.; HAYASHI, S.; AKIYAMA, Y.; ONO, A.; FUNAKI, S.; TAMURA, N.; KIMOTO, T.; JIKO, M.; HARUNA,Y.; SARASHINA, A.; ISHIDA, M.; NISHIYAMA, K.; FUSHIMI, M.; KOJIMA, Y.; YONEDA, K.; NAKANISHI, M.; KIM, S.; FUJITA, T.; SUGANO, K. Prediction characteristics of oral absorption simulation software evaluated using structurally diverse low-solubility drugs. **Journal of Pharmaceutical Sciences**, v. 109, n. 3, p. 1403–1416, 2020. DOI: https://doi.org/10.1016/j.xphs.2019.12.009

MOREIRA, L.K.S. SILVA, R.R.; SILVA, D.M.; MENDES, M.A.S.; BRITO, A.F.; CARVALHO, F.S.; SANZ, G.; RODRIGUES, M.F.; SILVA, A.C.G.; THOMAZ, D.V.; OLIVEIRA, V.; VAZ, B.G.; LIÃO, L.M.; VALADARES, M.C.; GIL, E.S.; COSTA, E.A.; NOËL, F.; MENEGATTI,.R. Anxiolytic- and antidepressant-like effects of new phenylpiperazine derivative LQFM005 and its hydroxylated metabolite in mice. **Behavioural Brain Research**, v. 417, p. 113582, 2022. DOI: https://doi.org/10.1016/j.bbr.2021.113582

MORENO, E.K.G.; THOMAZ, D.V.; MACHADO, F.B.; LEITE, K.C.S.; RODRIGUES, E.S.B.; FERNANDES, M.A.; CARVALHO, M.F.; OLIVEIRA, M.T.; CAETANO, M.P.; PEIXOTO, C.E.C.; ISECKE, B.G.; GIL, E.S.; ISAAC MACÊDO, Y.L. Antioxidant study and electroanalytical investigation of selected herbal samples used in folk medicine. **International Journal of Electrochemical Science**, v. 14, n. 1, p. 838–847, 2019. DOI: https://doi.org/10.20964/2019.01.82

NAGARAJAN, S.; NAGARAJAN, R.; KUMAR, J.; SALEMME, A.; TOGNA, A.R.; SASO, L.; BRUNO, F. Antioxidant activity of synthetic polymers of phenolic compounds. **Polymers**, v. 12, n. 8, p. 1–27, 2020. DOI: https://doi.org/10.3390/polym12081646

NAGY, B.; GALATA, D.L.; FARKAS, A.; NAGY, Z.K. Application of Artificial Neural Networks in the Process Analytical Technology of Pharmaceutical Manufacturing—a Review. **The AAPS Journal**, v. 24, n. 74, 2022. DOI: https://doi.org/10.1208/s12248-022-00706-0

NETO, L.F.L. BARRUFFINI, A.C.C.; THOMAZ, D.V.; MACHADO, F.B.; MACEDO, I.Y.L. *In silico* investigation of possible caffeine interactions with common inflammation-related targets. **Journal of Applied Biology and Biotechnology**, v. 7, n. 5, p. 31-34, 2019. DOI: https://doi.org/10.7324/JABB.2019.70505

OLIVEIRA, L.A.R.; SILVA, A.C.G.; THOMAZ, D.V.; BRANDÃO, F.; CONCEIÇÃO, E.C.; VALADARES, M.C.; BARA, M.T.F.; SILVEIRA, D. The potential of vouacapanes from *Pterodon emarginatus* Vogel against COVID-19 cytokine storm. **Advanced Pharmaceutical Bulletin**, 2021. DOI: https://doi.org/10.34172/apb.2023.016

OLIVEIRA, T.L.S.; Leite, K.C.S.; Macêdo, I.Y.L.; Morais, S.R.; Costa, E.A.; Paula, J.R.; Gil, E.S. Electrochemical behavior and antioxidant activity of hibalactone. **International Journal of Electrochemical Science**, v. 12, n. 9, p. 7956–7964, 2017. DOI: https://doi.org/10.20964/2017.09.54

PETTERSEN, E.F.; GODDARD, T.D.; HUANG, C.C.; COUCH, G.S.; GREENBLATT, D.M.; MENG, E.C.; FERRIN, T.E. UCSF Chimera - A visualization system for exploratory research and analysis. **Journal of Computational Chemistry**, v. 25, n. 13, p. 1605-1612, 2004. DOI: https://doi.org/10.1002/jcc.20084

PIRES, D.E.V.; BLUNDELL, T.L.; ASCHER, D.B. pkCSM: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. **Journal of Medicinal Chemistry**, v. 58, n. 9, p. 4066–4072, 2015. DOI: https://doi.org/10.1021/acs.jmedchem.5b00104

PONDER, J.W.; RICHARDS, F.M. An efficient newton-like method for molecular mechanics energy minimization of large molecules. **Journal of Computational Chemistr**y, v. 8, n. 7, p. 1016-1024, 1987. DOI: https://doi.org/10.1002/jcc.540080710

RESENDE, D.D.F.; ALVES, G.C.S.; COUTO, R.O.; SANCHES, C.; CHEQUER, F.M.D. Can parabens be added to cosmetics without posing a risk to human health ? A systematic review of its toxic effects. **Revista de Ciências Farmacêuticas Básica e Aplicada**, v. 42, n. e706, p. 1–18, 2021. DOI: https://doi.org/10.4322/2179-443X.0706

RODRIGUES, E.S.B.; MACÊDO, I.Y.L.; LIMA, L.L.S.; THOMAZ, D.V.; CUNHA, C.E.P.; OLIVEIRA, M.T.; BALLAMINUT, N.; ALECRIM, M.F.; CARVALHO, M.F.; ISECKE, B.G.; LEITE, K.C.S.; MACHADO, F.B.; GUIMARÃES, F.F.; MENEGATTI, R; SOMERSET, V.; GIL, E.S. Electrochemical characterization of central action tricyclic drugs by voltammetric techniques and density functional theory calculations. **Pharmaceuticals**, v. 12, n. 3, p. 116, 2019. DOI: https://doi.org/10.3390/ph12030116

SCHNEIDMAN-DUHOVNY, D.; DROR, O.; INBAR, Y.; NUSSINOV, R.; WOLFSON, H.J. PharmaGist: a webserver for ligand-based pharmacophore detection. **Nucleic Acids Research**, v. 36, Web Server issue W223–W228, 2008. DOI: https://doi.org/10.1093/nar/gkn187

SOROKINA, M.; STEINBECK, C. Review on natural products databases: where to find data in 2020. **Journal of Cheminformatics**, v. 12, n. 20, p. 1- 51, 2020. DOI: https://doi.org/10.1186/s13321-020-00424-9

SOUZA, M.J.M.F.; THOMAZ, D.V.; KLOPPEL, L.L.; CARNEIRO, L.A.; ROCHA, M.C.; AGUIAR, D.V.A.; VAZ, B.G.; SOUSA, C.M.; SANTOS, P.A .Influence of organo-mineral supplementation on the production of secondary metabolites in in vitro-germinated *Bromelia balansae* Mez. **Research, Society and Development**, v. 10, n. 11, p. e411101118052, 2021. DOI: https://doi.org/10.33448/rsd-v10i11.18052

TANG, J.; DENG, C.; HUANG, G.B. Extreme learning machine for multilayer perceptron. **IEEE Transactions on Neural Networks and Learning Systems**, v. 27, p. 809-821, 2016. DOI: https://doi.org/10.1109/TNNLS.2015.2424995

THOMAZ, D.V.; PEIXOTO, L.F.; OLIVEIRA, T.S.; FAJEMIROYE, J.O.; NERI, H.F.S.; XAVIER, C.H.; COSTA, E.A.; SANTOS, F.C.AL.; GIL, E.S.; GHEDINI, P.C. Antioxidant and neuroprotective properties of *Eugenia dysenterica* leaves. **Oxidative Medicine and Cellular Longevity**, v. 2018, n. 3250908, p. 1–9, 2018a. DOI: https://doi.org/10.1155/2018/3250908

THOMAZ, D.V.; COUTO, R.O.; ROBERTH, A.O.; OLIVEIRA, L.A.R.; LEITE, K.C.S.; BARA, M.T.F.; GHEDINI, P.C.; BOZINIS, M.C.V.; LOBÓN, G.S.; GIL, E.S. Assessment of Noni (*Morinda citrifolia* L.) product authenticity by solid state voltammetry. **International Journal of Electrochemical Science**, v. 13, n. 9, p. 8983–8994, 2018b. DOI: https://doi.org/10.20964/2018.09.390

THOMAZ, D.V.; OLIVEIRA, M.G.; RODRIGUES, E.S.B.; SILVA, V.B.; SANTOS, P.A. Physicochemical investigation of psoralen binding to double stranded dna through electroanalytical and cheminformatic approaches. **Pharmaceuticals**, v. 13, n. 6, p. 108, 2020. DOI: https://doi.org/10.3390/ph13060108

THOMAZ, D.V.; COUTO, R.O.; GOLDONI, R.; MALITESTA, C.; MAZZOTTA, E.; TARTAGLIA, G.M. Redox profiling of selected apulian red wines in a single minute. **Antioxidants**, v. 11, n. 5, p. 859, 2022. DOI: https://doi.org/10.3390/antiox11050859

THOMPSON, M.D.; BEARD, D.A.; WU, F. Use of partition coefficients in flow-limited physiologically-based pharmacokinetic modeling. **Journal of Pharmacokinetics and Pharmacodynamics**, v. 39, p. 313–327, 2012. DOI: https://doi.org/10.1007/s10928-012-9252-6

THORNBURG, C.C.; BRITT, J.R.; EVANS, J.R.; AKEE, R.K.; WHITT, J.A.; TRINH, S.K.; HARRIS, M.J.; THOMPSON, J.R.; EWING, T.L.; SHIPLEY, S.M.; GROTHAUS, P.G.; NEWMAN, D.J.; SCHNEIDER, J.P.; GRKOVIC, T.; O'KEEFE, B.R. NCI program for natural product discovery: A publicly-accessible library of natural product fractions for high-throughput screening. **ACS Chemical Biology**, v. 13, n. 9, p. 2484–2497, 2018. DOI: https://doi.org/10.1021/acschembio.8b00389

USHA, T.; SHANMUGARAJAN, D.; GOYAL, A.K.; KUMAR, C.S.; MIDDHA, S.K.Recent updates on computer-aided drug discovery: Time for a paradigm shift. **Current Topics in Medicinal Chemistry**, v. 17, n. 30, p. 3296–3307, 2018. DOI: https://doi.org/10.2174/1568026618666180101163651

WANG, Z.; LI, S.; GE, S.; LIN, S. Review of distribution, extraction methods, and health benefits of bound phenolics in food plants. **Journal of Agricultural and Food Chemistry**, v. 68, n. 11, p. 3330–3343, 2020. DOI: https://doi.org/10.1021/acs.jafc.9b06574

YADAV, B.; JOGAWAT, A.; RAHMAN, M.S.; NARAYAN, O.P. Secondary metabolites in the drought stress tolerance of crop plants: A review. **Gene Reports**, v. 23, n. 101040, p. 1-14, 2021. DOI: https://doi.org/10.1016/j.genrep.2021.101040