




Article

Compatibility Model between Encapsulant Compounds and Antioxidants by the Implementation of Machine Learning

Juliana Quintana-Rojas ^{1,*}, Rafael Amaya-Gómez ^{2,*} and Nicolas Ratkovich ¹

¹ Department of Chemical & Food Engineering, Universidad de los Andes, Cra. 1E No. 19a-40, Bogotá 111711, DC, Colombia; n.rios262@uniandes.edu.co

² Department of Industrial Engineering, Universidad de los Andes, Cra. 1E No. 19a-40, Bogotá 111711, DC, Colombia

* Correspondence: j.quintana@uniandes.edu.co (J.Q.-R.); r.amaya29@uniandes.edu.co (N.R.)

Abstract: The compatibility between antioxidant compounds (ACs) and wall materials (WMs) is one of the most crucial aspects of the encapsulation process, as the encapsulated compounds' stability depends on the affinity between the compounds, which is influenced by their chemical properties. A compatibility model between the encapsulant and antioxidant chemicals was built using machine learning (ML) to discover optimal matches without costly and time-consuming trial-and-error experiments. The attributes of the required antioxidant and wall material components were recollected, and two datasets were constructed. As a result, a tying process was performed to connect both datasets and identify significant relationships between parameters of ACs and WMs to define the compatibility or incompatibility of the compounds, as this was necessary to enrich the dataset by incorporating decoys. As a result, a simple statistical analysis was conducted to examine the indicated correlations between variables, and a Principal Component Analysis (PCA) was performed to reduce the dimensionality of the dataset without sacrificing essential information. The K-nearest neighbor (KNN) algorithm was used and designed to handle the classification problems of the compatibility of the combinations to integrate ML in the model. In this way, the model accuracy was 0.92, with a sensitivity of 0.84 and a specificity of 1. These results indicate that the KNN model performs well, exhibiting high accuracy and correctly classifying positive and negative combinations as evidenced by the sensitivity and specificity scores.

Keywords: antioxidant compounds; encapsulant compounds; wall materials; decoys; principal component analysis; machine learning; K-nearest neighbors; compatibility model



Citation: Quintana-Rojas, J.; Amaya-Gómez, R.; Ratkovich, N. Compatibility Model between Encapsulant Compounds and Antioxidants by the Implementation of Machine Learning. *Algorithms* **2024**, *17*, 412. <https://doi.org/10.3390/a17090412>

Academic Editors: Frank Werner, Cátia Vaz and Alexandre P. Francisco

Received: 13 June 2024

Revised: 27 August 2024

Accepted: 11 September 2024

Published: 17 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Encapsulation is widely used in numerous industries, including food, cosmetics, biology, agriculture, and pharmacy [1–3]. The nature of the encapsulation technique involves stabilizing active compounds using building systems to preserve their physical, chemical, and biological properties under various conditions. There are two main kinds of encapsulation techniques: physical and chemical. The first uses techniques that influence intensive properties, such as temperature and pressure [3], whereas the second, chemical encapsulation, leverages the active compounds' chemical properties [4]. Due to the method's complexity, significant obstacles exist [5] despite the technique's extensive range of applications. One of these obstacles is the compatibility between the active compound and the shell material (encapsulant compound) [6,7]. Among the widespread use of the technique and the varying amounts of compounds, antioxidant compounds (ACs) are significant for their properties in the abovementioned industries.

Antioxidants include polyphenols, carotenoids, anthocyanins, catechins, vitamins, and polyunsaturated fatty acids. They can be found in fruits, vegetables, cereals, and plants [8,9]. These compounds serve multiple functions. Plants, for example, offer structural support and defense against environmental factors such as biotic and abiotic stress,

ultraviolet radiation, and pathogens. On the other hand, consumers benefit from these compounds' properties, which protect against noncommunicable diseases (NCDs) by acting as regulators of cellular processes such as enzyme inhibition, gene expression, and protein phosphorylation. These compounds affect the sensory qualities of fruits and vegetables, such as bitterness, color, and flavor, resulting in a unique sensory profile for each product based on the compound present [10]. Due to their capabilities, the bioactive compounds mentioned above can be utilized in various industries. Nevertheless, these compounds' capacity is constrained by various obstacles that must be overcome. Nowadays, the bioavailability, bioaccessibility, and bioactivity of antioxidant compounds are issues due to negative extrinsic factors such as poor stability of compounds in various mediums [11], lack of transportation vehicles, and challenging absorption [12]. Subsequently, encapsulation could solve these concerns, promote stability, and pursue maintenance or increase the stated capabilities [13].

Compatibility between the active and encapsulant compounds (WM) is crucial in encapsulation. This relevance is due to the affinity between the compounds, influenced by various chemical properties. Most of the time, suitable matches are complex, requiring costly and time-consuming trial-and-error experiments [6]. To bridge this gap, the design and implementation of a computational tool could support decision-making processes regarding the time and resources required for this purpose in experimental development [14]. Various tools could be proposed in this regard; however, there have been remarkable advancements in machine learning (ML) for classification problems in recent years. Several research studies have considered machine learning (ML) to address the issues of determining compatibility between substances. Qi et al. suggested utilizing artificial intelligence (AI) and machine learning (ML) to facilitate drug development and identify protein–protein interactions and drug target interactions [15]. D'Souza et al. examined the application of Deep Learning (DL) in several cheminformatics methodologies for predicting the binding affinity between chemicals and proteins, focusing on the drug–target interaction. Liu et al. used an unambiguous multiple linear regression (MLR) algorithm to build QSAR models of the estrogen receptor binding affinity [16]. Rege et al. used a Support Vector Machine (SVM) regression with bootstrapping for creating and validating QSAR models to examine the DNA-binding characteristics of a library of aminoglycoside-polyamine compounds [17]. Rosas-Jiménez et al. employed ML algorithms (MLR, KNN, and RF) to construct quantitative structure–activity relationship (QSAR) models for cruzain inhibitors. These models use molecular descriptors to forecast the biological activity of the inhibitors accurately [18]. Krenn et al. investigated the application of generative models, specifically variational autoencoders (VAEs), in creating new molecules and predicting their properties using QSAR models [19]. Using computational methods to anticipate the binding affinity between compounds and targets significantly increases the likelihood of identifying lead compounds by minimizing the need for wet-lab studies. ML and DL methods employing ligand-based and target-based strategies have been utilized to forecast binding affinities, resulting in time and cost savings in drug discovery endeavors.

Nevertheless, ML is not limited to the pharmaceutical industry when predicting compatibility [20]. Piotrowsky discovered that scientific research on the corrosion of aerosol canisters is extensive but inadequate, and only some of the physical and chemical principles can be used to predict corrosion accurately. Additionally, other possible issues are influenced by many parameters and formula components. Piotrowsky's paper utilized a data-driven methodology to address these restrictions and limits to forecast the compatibility between a novel product's formulation and packaging. The model exemplifies a classification method, a subset of supervised machine learning; it utilizes input values provided for training to derive a conclusion and produce an output that classifies a dataset into distinct categories [21]. Periwal et al. proposed a use case for ML in predicting compound compatibility. They assessed the functional similarity between natural compounds and approved drugs by combining various chemical similarity metrics and physiochemical properties using a machine learning approach [22]. This study exemplifies the parallels be-

tween the present and previous studies since both investigations employ machine learning approaches to examine chemical data and generate predictions. By incorporating a range of chemical similarity measures and physical features, the machine learning approach enables a thorough evaluation of compound compatibility.

Therefore, considering a computational tool incorporating ML capabilities could be advantageous [23]. This work seeks a model with self-learning capabilities and compartmental adaptability of the prototype since ML is a data analysis model that automates the construction of analytic models. The K-nearest neighbors (KNN) algorithm is contemplated to construct a classification model that determines the compatibility between ACs and WMs, given its recognized capability in similar classification problems. To ascertain the compounds' eligibility for encapsulation, the model aims to assess the compatibility of the compounds by analyzing the relationships between their molecular descriptors.

The paper is structured as follows: Section 2 involves gathering and organizing data. This process includes assessing two databases to analyze the connections between the encapsulation process and various antioxidants and encapsulant compounds. Afterward, the model parameters are determined using the RDKit cheminformatics toolkit to extract the molecular descriptors of each compound. A review of the existing research considers the associations between the compatibility and incompatibility between substances. To increase the variety of encapsulant chemicals and the number of possible combinations between them, the computational tool LIDEB's Useful Decoys (LUDe) is used to build decoys. Section 3 establishes the connections between ACs and WMs by identifying comparison and difference relations, which are determined throughout the cleaning process. Section 4 involves conducting statistical analysis to examine the interrelationships between variables using tools such as histograms, scatter plots, and Pearson and Kendall correlations. In Section 5, a Principal Component Analysis (PCA) is performed to decrease the complexity of the dataset and assess the influence of individuals and variables in each principal component. Section 6 involves the implementation of K-nearest neighbors (KNN) to predict the compatibility of the combinations of ACs and WMs. Finally, Section 7 presents some concluding remarks and insights for further developments.

2. Data Collection and Curation Process

Data collection and curation involve acquiring and organizing raw data from multiple sources and eliminating discrepancies or missing data. The information is then converted into a format that is appropriate for analysis. Furthermore, the provided relationships can offer valuable insights into the data during the preprocessing step and help draw appropriate conclusions about the connections between compounds.

2.1. Data Collection and Parameters

The first database includes recompiled information on antioxidant compounds (AC-DB), while the second corresponds to wall materials (WM-DB). The compounds for the AC-DB were extracted from Phenol-Explorer: Database of Polyphenol Content in Foods [24]; their selection was narrowed based on the number of data found for each compound. Compounds were selected by assigning a value of 30 or higher to the number of discovered data. On the other hand, the wall materials were extracted from multiple books, such as Bioactive Carbohydrate Polymers [25] and Materials for Encapsulation. Chapter 7. Food Processing and Encapsulation Technologies for Active Food Ingredients and Wall Material Selection for Spray Drying [26].

For selecting the appropriate parameters, it is crucial to remember that the wall material must form a cohesive film with the core material and be chemically compatible and non-reactive with the core material [7]. In this regard, RDKit, an open-source cheminformatics toolkit, was utilized to collect the parameters that will shape the model [27]. First, the compounds' canonical smiles (CS) were extracted, and then, using the CS, molecular descriptors could be determined by implementing RDKIT in Python. These molecular descriptors represent the physical and chemical properties of a molecule mathematically. There are

various types of molecular descriptors, including unidimensional (1D), bi-dimensional (2D), and three-dimensional (3D) descriptors.

The molecular descriptor's complexity determines these; similarly, the 1D descriptors are the simplest to calculate. They represent the information calculated with the molecular formula of the compound, such as the type of atoms, functional group, and molecular weight. The 2D descriptors include information regarding the molecule's size, shape, and electronic distribution. The 3D descriptors, such as the Polar Surface Area (PSA) and intramolecular hydrogen bonding, are the most complicated and require the 3D structure of the molecule to be calculated. In this manner, the following molecular descriptors were extracted: chemical formula, molecular weight, logP, Hydrogen Bond Donor (HBD), Hydrogen Bond Acceptor (HBA), Polar Surface Area (PSA), number of heavy atoms, number of carboxyl groups, number of carbonyl groups, number of hydrogen groups, and number of acyclic groups [28,29].

Before data curation, it is necessary to analyze the recompiled information and determine the superficial relationships between the parameters of ACs and WMs. Understanding the parameters' meaning and other pertinent information is essential. For this purpose, bibliographic research was conducted for each parameter. See Appendix A.1.

2.2. Data Curation Process

A literature review was conducted to determine which compounds have previously exhibited compatibility or incompatibility for encapsulation to establish the relationship between ACs and WMs. This review was conducted to establish a standard for comparing the values of the parameters for each compound.

Song et al. [30] discovered that certain peach polyphenols are stable after being encapsulated in vitro and in vivo in sodium alginate matrices using Pickering high internal phase emulsions. Thus, it was assumed that the ACs listed in Table 1 were compatible with sodium alginate as a wall material. Consequently, a comparison of the case-specific parameters led to the establishment of the following conclusions:

- In two out of three instances, the molecular weight of the WM was greater than the AC.
- The logP of the WM was always less than the logP of the AC.
- The numbers of HBAs and HBDs of the WM were consistently more significant than those of the AC.
- In two out of three instances, the PSA of the WM was greater than that of the AC.
- The HAC of the WM was consistently more significant than that of the AC.
- In two out of three instances, the CGCs and CnGCs of the AC were zero and one, respectively; both parameters were greater or equal for the WM.
- The HGCs and AGCs of the AC were always smaller than the AGCs and AGCs of the WM.
- There is a direct proportional relationship between the value of the WM and the PSA such that as the WM increases, so does the PSA.

In this model, the compatibility between AC and WM is evaluated solely based on their chemical and physical interactions. The influence of the encapsulation type, operational conditions, and other chemicals was not studied.

Other WMs did not exhibit the same response, and the active component's encapsulation was lost [30] due to temperature and other factors. Under particular conditions, encapsulating a specific AC with maltodextrin was inefficient. This result led to the conclusion that, under normal conditions, maltodextrin cannot be compatible with certain compounds [31]. Curdlan oligosaccharide, which requires a binding agent and electrostatic charges to encapsulate compounds, encountered a similar situation; however, it was determined that despite the curdlan's high encapsulation efficiency, it could not achieve the same results on its own [32].

Table 1 shows combinations of compounds that, according to the literature, were deemed incompatible for encapsulation [31,32]. Thus, the following conclusions were reached:

- In half of the instances, the AC had a greater MW than the WM.
- In every instance, the logP of the WM was less than that of the AC.
- The HBA of the WM was consistently more significant than that of the AC. The WM had greater values than the AC for the HBD four out of five times.
- In half of the instances, the PSA of the AC was higher than that of the WM.
- In all cases, the CGCs and CnGCs of the AC were more significant than or equal to the values of the WM.
- The HGCs and AGCs of the WM were always more significant than those of the AC.

Table 1. Comparison of success [30] and failure [31,32] cases and their AC and WM parameters.

Success	Compound	MW	LogP	HBA	HBD	PSA	HAC	CGC	CnGC	HGC	AGC
1	Catechin (AC)	290.27	1.46	1	5	110.38	21	0	0	7	3
	Sodium Alginate (WM)	418.23	−4.10	12	6	192.44	52	1	1	17	32
2	Estragole (AC)	148.21	3.23	0	0	9.23	11	0	0	2	1
	Sodium Alginate (WM)	418.23	−4.10	12	6	192.44	52	1	1	17	32
3	P-coumaric (AC)	164.16	1.50	2	2	57.53	12	1	1	2	0
	Sodium Alginate (WM)	418.23	−4.10	12	6	192.44	52	1	1	17	32
Failure	Compound	MW	LogP	HBA	HBD	PSA	HAC	CGC	CnGC	HGC	AGC
1	P-coumaric (AC)	164.16	1.50	2	2	57.53	12	1	1	2	0
	Maltodextrin (WM)	342.30	−4.70	11	8	190.00	23	0	0	14	8
2	Vainillin (AC)	168.15	1.17	4	2	66.76	12	1	1	4	0
	Maltodextrin (WM)	342.30	−4.70	11	8	190.00	23	0	0	14	8
3	Curcumin (AC)	368.39	2.95	2	2	93.06	27	0	4	6	3
	Curdlan oligosaccharide (WM)	180.16	−2.60	6	5	110.00	12	0	0	7	16

2.3. Decoys Implementation

Due to the need for more information regarding successful and unsuccessful cases of compound compatibility, implementing decoys was required to increase the quantity of data and strengthen it. Decoys are putatively inactive compounds generated from a set of active compounds; these molecules have not been tested against a molecular target. However, due to their structural features, they are unlikely to bind to the target and have a high affinity. They can be used to validate virtual screening tools and protocols. LUDe (LIDEB's Useful Decoys), a tool developed by the Laboratory of Investigation and Development of Bioactives (LIDEB) at the National University of La Plata, was used to generate these molecules [33].

For generating these decoys, the WMs from Table 1 were used as active compounds to generate these decoys; additionally, scleroglucan was used as the wall material before generating the decoys. This procedure was due to the compatibility and incompatibility that the compound showed with different ACs as demonstrated in Ref. [34]. When encapsulating antioxidant compounds, focusing on the wall materials rather than the antioxidants themselves is essential. The wall materials are the main protective barrier, shielding the antioxidants from oxygen, moisture, and light, thus ensuring their stability and effectiveness [35].

It is crucial to ensure that the wall materials are compatible because any lack of compatibility could jeopardize the structural strength of the enclosed product, resulting in early deterioration or leaking and ultimately making the encapsulation process inefficient [36]. The physical and chemical characteristics of wall materials substantially impact these processes. Concentrating on these materials makes it possible to find the most appropriate candidates that can be consistently employed in different settings [37].

While antioxidant chemicals are important for their functional advantages, they are usually not the primary concern in decoys. This is because their major activity is within the core of the enclosed structure, and their interaction with wall materials is typically limited. The primary focus in the encapsulation process is to prioritize the structural and barrier properties of the wall materials to achieve optimal efficacy and efficiency [38].

This method is more economical since wall materials are generally utilized in greater amounts than antioxidants, making their compatibility a crucial consideration for financial reasons [39]. After confirming the compatibility of the wall materials, attention can be directed towards enhancing the stability and functioning of the antioxidant chemicals present in the encapsulated product.

Parameters with (\pm) denotation indicate that the value could adapt to a higher or lower value depending on the magnitude selected (Table 2). For instance, a limit of (\pm) 10 was established for the molecular weight. As a result, the molecular weight of the decoys generated will increase or decrease the magnitude by ten of the MW of the active compound. It will apply to all compounds with limits denoted by the symbol (\pm). Alternatively, the fingerprint length was set to 1024; it refers to the size in bits of a molecule's chemical space [40]. The similarity metric was selected based on input from Tanimoto Similarity. This can be described as follows: A' and B' are fragments of A and B fingerprint molecule fingerprints. AB is defined as the set of fingerprints shared by both molecules. The Tanimoto coefficient ranges from 0 to 1, where 0 indicates that the fingerprints of the molecules contain no similar bits, and 1 indicates that the fingerprints are identical.

Table 2. Physicochemical features of the decoy [33].

Physicochemical Features	Limits
Molecular weight	± 10
logP	± 0.5
Rotable bonds	± 2
Number of H Acceptors	± 1
Number of H Donors	± 1
Fingerprint ratio	± 3
Fingerprint length	1024
Similarity metric	Tanimoto Similarity
Maximum similarity allowed	1
Limit of the fraction of the Maximum Common Substructure	1
Maximum similarity allowed between decoys and any of the actives	0.7

Consequently, $(A, B) = (AB)/(A + B - AB)$, the similarity finds all formulas for molecule A with a Tanimoto coefficient more significant than a predetermined threshold [41]. The threshold value determines the degree of similarity between molecules. The maximum allowed similarity and the fraction of the maximum common substructure were set to 1 bit. The maximum allowed similarity between decoys and active compounds was set to 0.7 bits.

In Table A2 (See Appendix A.2), 32 decoys are listed for the four wall materials presented. Due to the canonical smiles generated for each compound and the calculation of the MW using the LUDe program, the MW of the sodium alginate decoys does not match the limit established. However, this was not considered a problem since the molecular weight was examined and found consistent with the proposed canonical smile.

3. Relations Stated between AC and WM

Based on the conclusions drawn from the relationship between WMs and ACs in the successful and unsuccessful compatibility cases (Table 1), see Section 2.2, relations for merging the WM-DB and AC-DB were defined as shown in Table 3. Combinations were made between the 32 WMs, which included active compounds and decoys, and the 8 ACs, resulting in 288 total combinations. There are two data types in the new dataset; the first relates to a comparison between parameters and takes a binary value of one if the comparison satisfies the requirement and zero otherwise. The second one refers to a delta between the parameters, the difference between the evaluated WM and AC parameters.

Table 3. Relations stated between ACs and WMs.

Parameter	Description	Comparison	Quantification
C-MW	MW comparison	$MW_{PC} > MW_{WM}$	If comparison is True C-MW = 0 Else: C-MW = 1
Δ MW	MW difference between PC and WM	-	Δ MW = $\text{abs}(MW_{PC} - MW_{WM})$
C-logP	logP comparison	$\log P_{PC} > \log P_{WM}$	If comparison is True C-logP = 1 Else: C-logP = 0
Δ logP	logP difference between PC and WM	-	Δ logP = $\text{abs}(\log P_{PC} - \log P_{WM})$
C-HBA	HBA comparison	$HBA_{PC} > HBA_{WM}$	If comparison is True C-HBA = 0 Else: C-HBA = 1
Δ HBA	HBA difference between PC and WM	-	Δ HBA = $\text{abs}(HBA_{PC} - HBA_{WM})$
C-HBD	HBD comparison	$HBD_{PC} > HBD_{WM}$	If comparison is True C-HBD = 0 Else: C-HBD = 1
Δ HBD	HBD difference between PC and WM	-	Δ HBD = $\text{abs}(HBD_{PC} - HBD_{WM})$
C-PSA	PSA comparison	$PSA_{PC} > PSA_{WM}$	If comparison is True C-PSA = 1 Else: C-PSA = 0
Δ PSA	PSA difference between PC and WM	-	Δ PSA = $\text{abs}(PSA_{PC} - PSA_{WM})$
C-HAC	HAC comparison	$HAC_{PC} > HAC_{WM}$	If comparison is True C-HAC = 1 Else: C-HAC = 0
Δ HAC	HAC difference between PC and WM	-	$\text{abs}(HAC_{PC} - HAC_{WM})$
C-CGC	CGC comparison	$CG_{PC} > CG_{WM}$	If comparison is True C-CGC = 1 Else: C-CGC = 0
Δ CGC	CGC difference between PC and WM	-	Δ CGC = $\text{abs}(CG_{PC} - CG_{WM})$
C-CnGC	CnGC comparison	$CnG_{PC} > CnG_{WM}$	If comparison is True C-CnGC = 1 Else: C-CnGC = 0
Δ CnGC	CnGC difference between PC and WM	-	Δ CnGC = $\text{abs}(CnGC_{PC} - CnGC_{WM})$
C-HGC	HGC comparison	$HG_{PC} > HG_{WM}$	If comparison is True C-HGC = 1 Else: C-HGC = 0
Δ HGC	HGC difference between PC and WM	-	Δ HGC = $\text{abs}(HG_{PC} - HG_{WM})$
C-AGC	AGC comparison	$AG_{PC} > AG_{WM}$	If comparison is True C-AGC = 1 Else: C-AGC = 0
Δ AGC	AGC difference between PC and WM	-	Δ AGC = $\text{abs}(AG_{PC} - AG_{WM})$
-	Zero counts	-	Number of zeros of all the descriptors above
-	Classification	-	If $\text{zeros}_{counts} > 2 \rightarrow$ Case = 0 (Failure) Else: Case = 1 (Success)

4. Exploratory Analysis

Before statistical analysis, the dataset containing 288 observations and 22 variables was separated into success and failure cases. Case-sensitive variables facilitate the separation of a dataset into two distinct subsets. The first dataset describes the compatibility of the compound, while the second dataset describes their incompatibility. As a result, a pairs panel graph was created, which helped illustrate descriptive statistics (Figures 1 and 2). It is a scatter plot of matrices (SPLOM) with bivariate scatter plots below the diagonal, histograms on the diagonal, and Pearson's correlation above the diagonal [42].

Note that to visualize the correlation between parameters, the variance cannot be zero ($\sigma^2 \neq 0$), which means that all observations have the same value for the evaluated parameter. The parameters above with zero variance ($\sigma^2 = 0$) were removed from their respective datasets. First, the following parameters were eliminated from the success dataset: C-MW, C-HBA, C-PSA, C-HAC, and Case. C-HBA and Case parameters were removed from the collection of failure data. The behavior of these parameters by their

respective datasets allowed for the determination of the significance of each in establishing the compounds' compatibility or incompatibility.

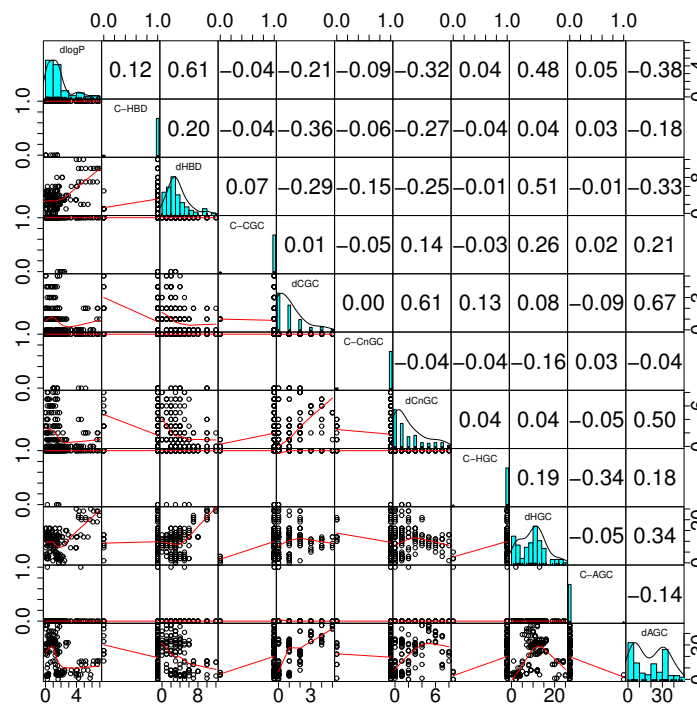


Figure 1. Pairs panel success dataset.

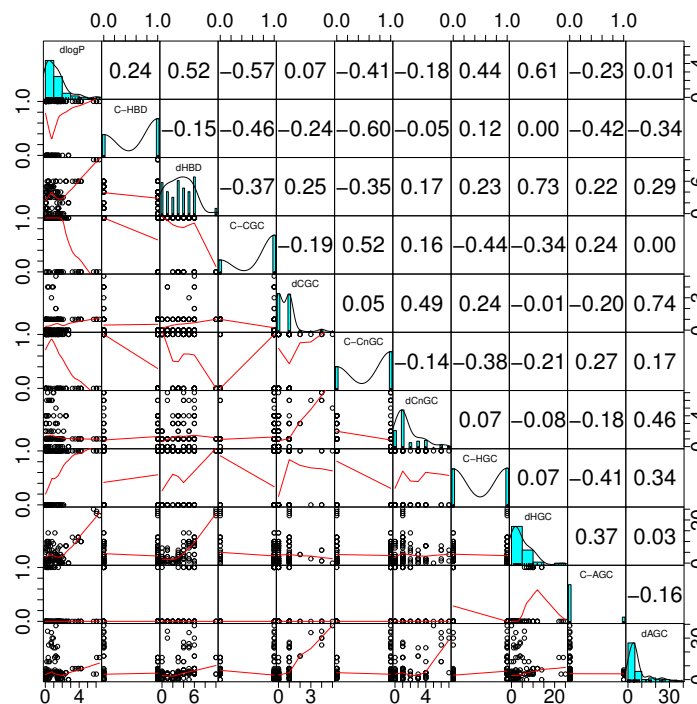


Figure 2. Pairs panel failure dataset.

4.1. Histograms

This type of graphic provides information about the representation of the data distribution, allowing the pattern, shape, and characteristics of the data to be observed. Moreover, the Gaussian distribution, represented by the curve superimposed on the histogram and

also known as the normal distribution, is a continuous probability distribution that reveals the underlying characteristics of the data.

The shape of the Gaussian curves allowed the data distribution to be determined. If the curve is bell-shaped, the data are evenly distributed around the mean, with an equal number of values on both sides as determined by the mean (μ) and standard deviation (σ) simultaneously. It may also exhibit bimodal forms or skewness in both directions. When the distribution has a longer tail on one side, the skewness indicates a greater frequency of values in the direction of the trend. Positive skewness indicates that the tail reaches higher values, whereas negative skewness indicates that the tail reaches instead lower values.

On the other hand, bimodal distributions have two distinct peaks, indicating the existence of two separate groups or phenomena. Referring to the spread of the histogram, the standard deviation, which represents the spread of the distribution, describes the range of values covered by the histogram. The greater the spread, the greater the range of values, indicating that the data tend to be more dispersed. In contrast, a narrower spread indicates a more concentrated distribution, with values clustered closely around the mean. The standard deviation quantifies the dispersion or variability of the data [43,44].

Each dataset parameter shows some similarity in their respective data behavior. The following parameters described the same data behavior for success and failure datasets, analyzed by Gaussian distribution and data spread: ΔMW shows a bimodal distribution, which refers to two different and separate clusters of data and no spread. The C-logP does not exhibit Gaussian distribution and highly spread bars, which indicates that the data are highly spread with significant variability. Referring to ΔHAC , the Gaussian distribution is bimodal and low spread, C-AGC does not exhibit a Gaussian distribution and has widely dispersed data. ΔHBA and ΔHBD , on the other hand, indicate the same Gaussian distribution for both datasets but different data spreads. ΔHBA shows a bimodal distribution and low data spread, where the information cluster is low spread. Regarding ΔHBD , the data of this parameter exhibit right negative skewness, which indicates that the data distribution is not symmetric. Most of them are located on the left side of the distribution, with fewer values to the right. $\Delta \log P$, C-HBD, ΔPSA , C-CGC, ΔCGC , C-CnGC, ΔCnG , C-HGC, and ΔAGC exhibit different Gaussian distributions expressed in Table 4, as well as low, high, and medium spread. Low means the data are not spread, medium data are regularly spread, and high means the data have a wide range of values. Lastly, ΔHGC neither share a similar Gaussian distribution nor a similar spread.

Table 4. Gaussian distribution and spread of histograms.

Parameter	Success Dataset		Failure Dataset	
	Gaussian Distribution	Spread	Gaussian Distribution	Spread
C-MW	-	-	Bimodal	High
ΔMW	Bimodal	No	Bimodal	No
C-logP	-	High	-	High
$\Delta \log P$	Bimodal	No	Right negative skewness	No
ΔHBA ¹	Bimodal	Low	Bimodal	No
C-HBD	-	High	Bimodal	High
ΔHBD	Right negative skewness	Low	Right negative skewness	Medium
C-PSA	-	-	Bimodal	High
ΔPSA	Right negative skewness	No	Bimodal	No
C-HAC	-	-	Bimodal	High
ΔHAC	Bimodal	Low	Bimodal	Low
C-CGC	-	High	Bimodal	High
ΔCGC	Right negative skewness	Medium	Bimodal	Medium
C-CnGC	-	High	Bimodal	High
ΔCnG	Right negative skewness	Medium	Bimodal	Medium
C-HGC	-	High	Bimodal	High
ΔHGC	Bimodal	Low	Right negative skewness	No
C-AGC	-	High	-	High
ΔAGC	Bimodal	No	Right negative skewness	No

¹ C-HBA is disregarded because it has identical values across all combinations in both failure and success datasets.

It is important to note that the data exhibiting a high spread pertain to the comparison parameters, which is attributable to the binomial data type of this parameter. On the other hand, the Δ parameters do not exhibit high spread in any of the cases, only no, low, or medium spread, which corresponds to the type of data, indicating that the difference between ACs and WMs is not too spread and has more consistent values.

4.2. Scatter Plots

This type of graph displays relationships between variables whose behavior can be determined by characteristics such as direction, strength, and shape. The first one determines whether there is a positive, negative, or no apparent relationship between variables; positive relationships indicate that as one variable increases, the other tends to increase as well; negative relationships indicate the opposite, where if one variable increases, the other tends to decrease; and no apparent relationships indicate that the variables are unrelated or independent. Regarding the strength, it indicates how clustered the data of both variables are; if the data exhibit a tighter cluster, it indicates a stronger relationship, whereas a decrease in this strength could indicate a moderate or weak relationship. The form that identifies whether the relationship is linear or nonlinear, shown by the trending line, is a straight trending line for a linear relationship and a curve or nonlinear trending line for a nonlinear relationship [44,45].

Thus, the scatter plots of the pair panels graph for both datasets demonstrate the same behavior for most data. The lack of apparent direction, weak strength, and linear relationships between the comparison parameters can be attributed to the nature of the data for this variable type. Furthermore, outliers are typically responsible for the poor clustering of the data. In contrast, there are usually no apparent relationships, weak strength, and nonlinear form between Δ variables. Regarding the relationship between comparison and Δ variables, there was no discernible direction, weak strength, and a nonlinear or linear shape, depending on the variables involved.

4.3. Pearson and Kendall Correlations

The Pearson correlation coefficient, frequently denoted by the symbol r , measures the linear relationship between two continuous variables. It quantifies the degree to which the variables move linearly together. The Pearson correlation coefficient ranges from -1 to $+1$, where -1 represents a perfect negative linear relationship, $+1$ represents a perfect positive linear relationship, and 0 represents the absence of a linear relationship [46,47]. In contrast, the Kendall correlation coefficient, denoted by τ (tau), is a rank-based correlation measure. It evaluates the strength and direction of the relationship between variables based on their ranks or ordinal positions. Kendall correlation applies to continuous and discrete variables, making it useful for analyzing nonlinear and nonparametric relationships [46–48].

Pearson and Kendall correlations serve the same purpose of evaluating the strength and direction of the relationship between variables, but they are not simple measures. They differ in their underlying assumptions and data processing, adding a layer of interpretability.

Figure 3 displays the strength and direction of the linear relationship between variables for Pearson correlation plots. Positive correlations are indicated by ascending values and more vibrant red hues, negative correlations by descending values and more vibrant purple hues, and no correlation by middle-ground values and white hues. For a Kendall correlation plot, Figure 4 depicts the strength and direction of the monotonic relationship between variables. Similarly, positive and negative correlations are presented, emphasizing monotonicity rather than linearity.

Figure 3a depicts a robust positive linear correlation between Δ variables. On the other hand, except for Δ HBD and Δ logP, the C-logP parameter exhibits a negative or no linear correlation with the remaining parameters. In contrast, in Figure 3b, the variables in the upper portion of the graph have a positive linear relationship, whereas the variables in the left portion from C-CGC to C-AGC and in the right lower portion have negative linear relationships. Comparing the Pearson correlation results of the success and failure

datasets reveals contradictory behaviors, indicating that the correlations referring to the comparison values vary based on the case-sensitive decision. Thus, comparison variables do not exhibit correlations for determining compatibility, whereas they exhibit a negative linear correlation for determining incompatibility.

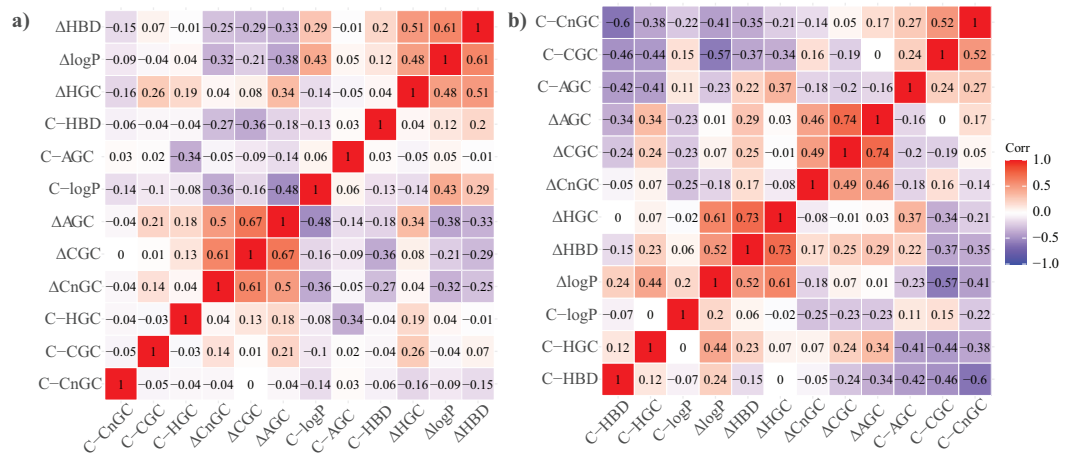


Figure 3. Pearson correlation matrices. (a) Success dataset. (b) Failure dataset.

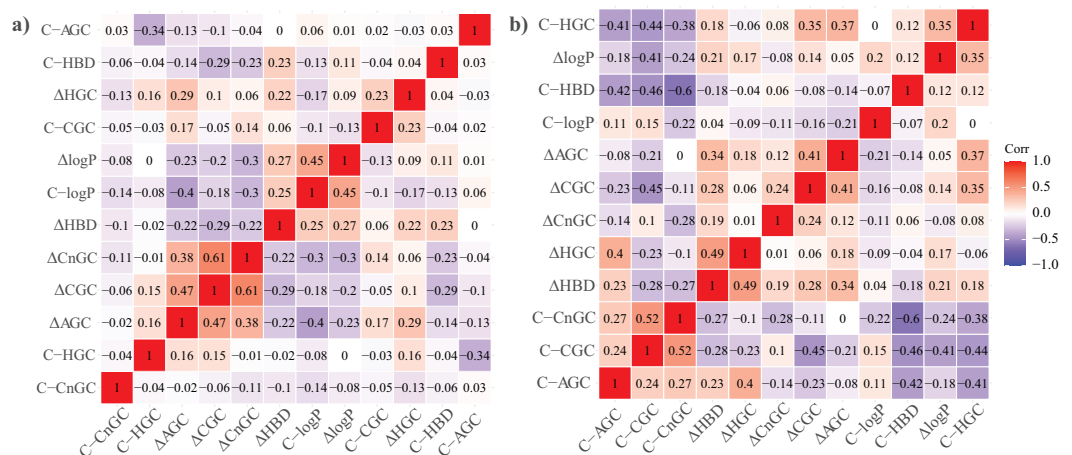


Figure 4. Kendall correlation matrices. (a) Success dataset. (b) Failure dataset.

In contrast, Kendall correlations reveal positive monotonic correlations between most delta variables and C-CGC and C-HGC. In contrast, the remaining comparison variables and $\Delta\log P$ exhibit negative monotonic or nonmonotonic correlations for Figure 4a. Regarding Figure 4b, it is determined that the correlations in the upper right portion of the graph display positive monotonic correlations between the variables and a small proportion of negative monotonic correlations dispersed throughout the entire graph. This graph displays more positive monotonic correlations than Figure 4b, which displays fewer variables with no monotonic correlations.

In conclusion, the Pearson and Kendall correlations for the analysis of the success dataset (Figures 3a and 4a) exhibit a similar distribution of correlations around the variables. In contrast, the correlations between the variables in the graphs for the failure dataset (Figures 3b and 4b) differ significantly, with Figure 3b exhibiting stronger positive correlations than Figure 4b.

5. Principal Component Analysis (PCA)

PCA is one of a series of approaches for representing high-dimensional data in a lower-dimensional, more manageable format without sacrificing too much information. PCA is one of the most straightforward and reliable dimension-reduction techniques. It is

also one of the oldest and has been repeatedly rediscovered in other domains; therefore, it is also known as the Karhunen–Loève transformation, the Hotelling transformation, the method of empirical orthogonal functions, and singular value decomposition [49]. In this manner, a PCA was performed on each dataset (success and failure) by a singular value decomposition of the centered and scaled data matrix (mean 0, variance 1) and by not using eigen on the covariance matrix [50]; the obtained results are presented in Table 5.

Table 5. PCA results: Components distribution.

	Success Dataset			Failure Dataset		
	PC1	PC2	PC3	PC1	PC2	PC3
Standard deviation	2.26	1.83	1.18	2.41	1.97	1.72
Proportion of variation	0.32	0.21	0.09	0.31	0.21	0.16
Cumulative proportion	0.32	0.53	0.61	0.31	0.51	0.67

PCA transformed the original variables into a set of principal components (PCs) by reducing the dimensionality of the variables and identifying patterns and structure in multivariate datasets [42]. The standard deviation quantifies the dispersion or spread of the data along each PC; more significant standard deviation values indicate a greater spread along the respective PC. In contrast, the variation proportion represents the total variability in the data explained by each principal component, with larger values indicating that the respective PC captures a more significant amount of variability in the data. Lastly, the cumulative proportion represents the cumulative amount of total variation explained by a set of PCs; this value provides insight into the total amount of variation captured by a combination of PCs, where higher cumulative proportion values indicate that these principal components explain a more significant proportion of the total variation in the data [42,50,51].

PC1 has the most significant standard deviation across both datasets, indicating that it captures the most data variability. PC1 explains 32% total variation in the success data and 31% in the failure data, according to the relatively high proportion of variation values for PC1. The cumulative proportion for PC1 demonstrates that it alone accounts for a significant portion of the total variation, indicating its importance in describing the overall patterns and structure of the data. PC2 has a lower standard deviation than PC1 in both datasets, indicating that it captures less variability but is still statistically significant. Despite being smaller than PC1, the PC2 proportion of variation values is still significant, indicating that PC2 explains the additional 21% of the variation in both datasets. The PC2 cumulative proportion demonstrates its contribution to capturing additional variation, enhancing the PC1 representation of the data. PC1 and PC2 represent 32% and 21% of the values, respectively, for a total coverage of 53% and 51% of the data in the success and failure datasets. PC3 has the lowest standard deviation of all principal components in both datasets, indicating that it captures the slightest variance. PC3 explains less variation than PC1 and PC2 because its proportion of variation values is the lowest. PC3 contributes to capturing additional variation, albeit to a lesser extent than PC1 and PC2.

Table 6 displays the contribution of each variable to PC1 and PC2 for the success and failure datasets. With respect to the PC1 of the success dataset, MW is the variable that contributes the most to PC1, followed by Δ HAC and Δ HBA. C-HBD, C-CnCG, and Δ logP are the variables that contribute the least to PC1. Referring to PC2, HBD is the variable that contributes the most, followed by Δ logP and Δ PSA, while C-AGC, C-CGC, and C-HGC contribute the least. Regarding PC1 of the failure dataset, the variables that contribute the most are C-MW, C-HAC, and Δ PSA, while the variables that contribute the least are CnGC, C-AGC, and C-HBD. Regarding PC2, the variables that contribute the most are Δ HAC, Δ MW, and C-AGC, while the variables that contribute the least are Δ AGC, Δ CGC, and Δ CnGC.

The data presented in Table 6 can also be represented as bar graphs indicating the percentage contribution of each variable to the PC. Figures 5 and 6 depict the contribution of the variables as bar graphs.

Table 6. PCA results: Variables contribution.

Success Dataset				Failure Dataset			
PC1		PC2		PC1		PC2	
Variable	Contribution	Variable	Contribution	Variable	Contribution	Variable	Contribution
ΔMW	17.67	ΔHBD	22.65	C-MW	11.71	ΔHAC	18.58
ΔHAC	17.37	ΔlogP	21.00	C-HAC	11.66	ΔMW	17.44
ΔHBA	12.93	ΔPSA	10.86	ΔPSA	9.60	C-AGC	10.48
ΔAGC	10.92	ΔHGC	9.14	ΔHBD	9.24	C-CnGC	9.57
ΔHGC	10.86	ΔAGC	7.22	ΔHBA	8.85	C-HBD	9.45
ΔPSA	9.69	ΔCGC	6.47	C-CGC	8.14	C-PSA	5.68
ΔCnGC	6.59	ΔCnGC	6.46	ΔlogP	6.80	C-CGC	4.83
ΔCGC	5.54	ΔHBA	5.47	ΔHGC	6.65	C-HGC	4.78
C-logP	4.50	C-logP	5.47	C-PSA	4.53	ΔPSA	4.00
C-CGC	2.54	C-HBD	2.88	C-HGC	3.86	ΔHGC	3.49
ΔHBD	0.42	C-CnGC	1.01	ΔCGC	3.74	C-logP	3.21
C-AGC	0.32	ΔHAC	0.59	ΔAGC	3.48	ΔHBA	2.37
C-HGC	0.32	ΔMW	0.27	ΔMW	2.93	ΔlogP	1.64
C-HBD	0.24	C-AGC	0.26	C-CnGC	2.53	ΔHBD	1.53
C-CnGC	0.10	C-CGC	0.18	C-logP	2.47	C-MW	1.29
ΔlogP	0.00	C-HGC	0.07	ΔHAC	1.72	C-HAC	1.24
				ΔCnGC	1.06	ΔAGC	0.42
				C-AGC	0.62	ΔCGC	1.0×10^{-3}
				C-HBD	0.41	ΔCnGC	1.2×10^{-5}

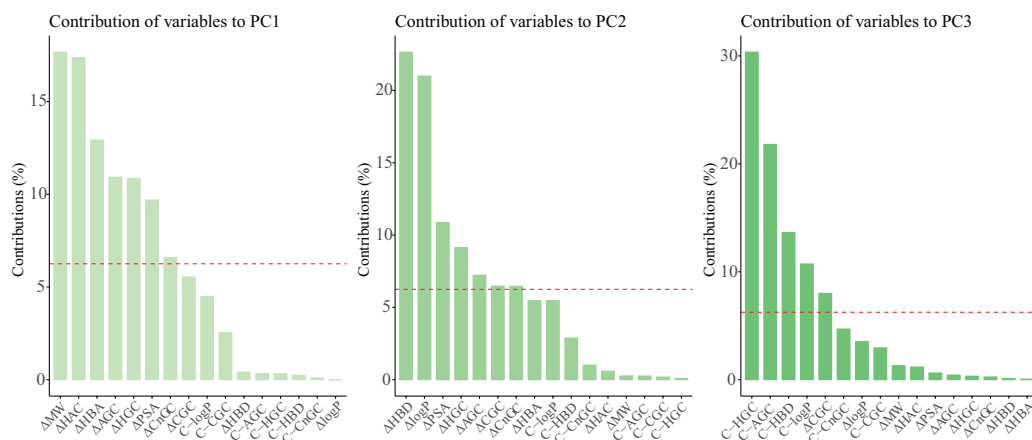


Figure 5. Success dataset contribution of variables for PC. (---) Dashed line indicates that the variables whose values are above the line account for 30% of the total contribution of the PC.

The contributions of each variable can also be represented as vectors, whose magnitude, direction, and angle provide insight into the behavior of the variable’s contribution to the PC. Figure 7 depicts the contribution of the variables in vectors.

In terms of the contribution and relationship between variables and their contribution to the PCA, variables with the most extended vectors contribute the most to the PC, whereas the magnitude of the angle between variables represents the correlation between variables; if the angle is slight, it indicates a positive, strong correlation between variables, indicating that they tend to vary together, whereas variables with a large angle between them have a low or negative correlation. Another factor is the direction of the respective vector, which indicates that the variable has an inverse relationship to the other variables and may have opposing effects on the PC [50,51].

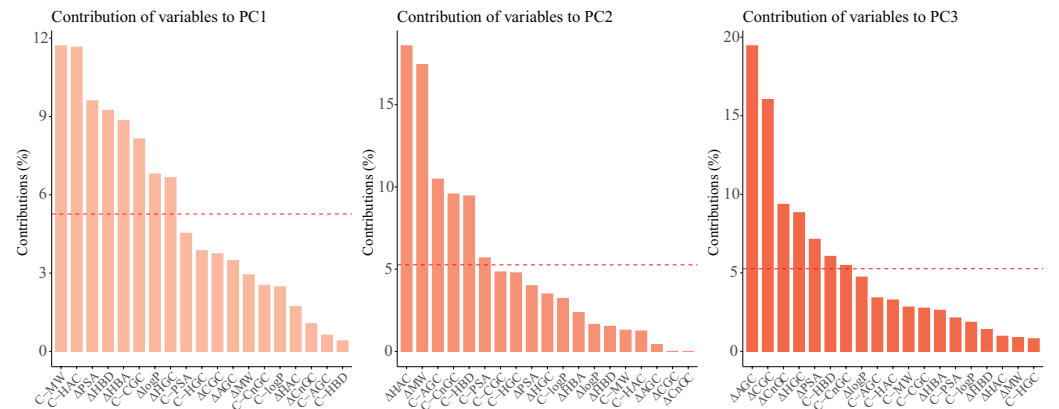


Figure 6. Failure dataset contribution of variables for PC. (---) Dashed line indicates that the variables whose values are above the line account for 30% of the total contribution of the PC.

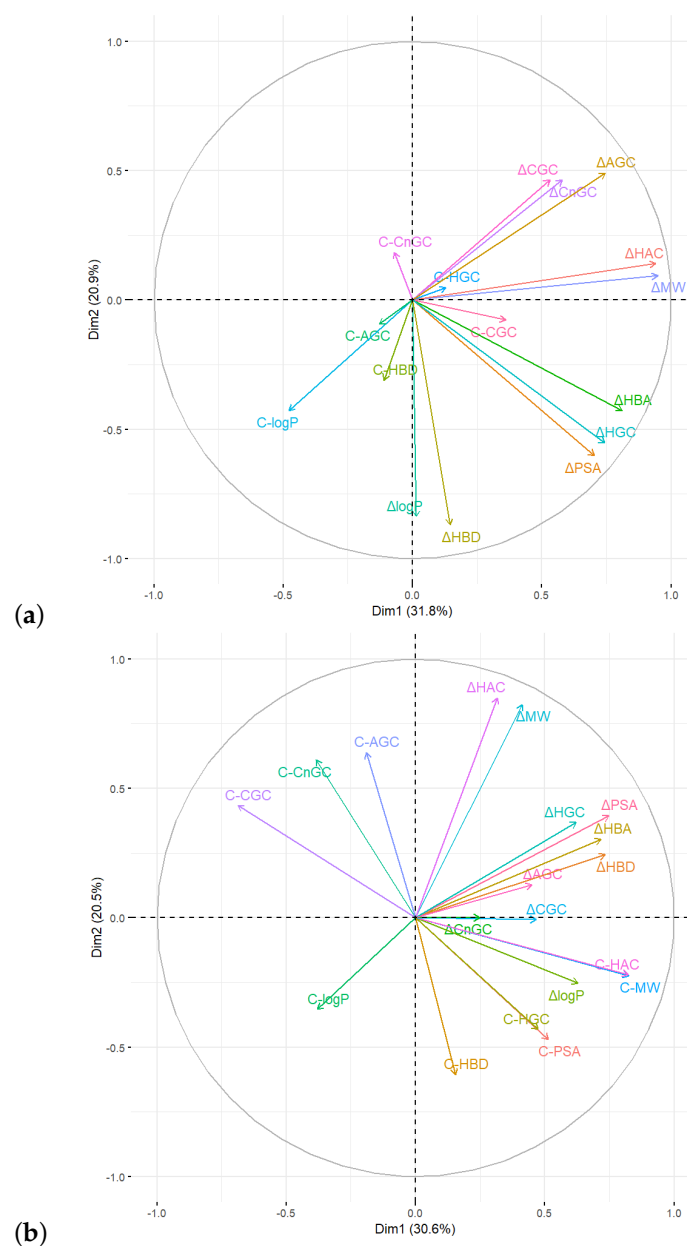


Figure 7. Vectors for contribution variables for PC. (a) Success dataset. (b) Failure dataset.

Individual contribution is another way of visualizing the contribution of the PC, representing how much each variable contributes to the PC. Nonetheless, this representation is not optimal if a variable's values are highly diverse. In this way, the individual contribution to PC was adequately accounted for in the comparison variables with only two binomial values.

Upon comparing the contribution of individuals in the success and failure datasets for the same variables, it was discovered that the individuals in the success dataset exhibit more clustering than those in the failure dataset. In contrast, the success dataset exhibits more outliers. Thus, the clustering between individuals indicates the existence of relationships between the observations [52]. The combination of ACs and WMs determines these relationships among individuals. Due to the evaluation of the compatibility or incompatibility of the combinations, an additional significant conclusion is that individuals with a zero value are less prevalent in the success dataset than in the failure dataset. It was also observed that the individuals in the success dataset have a more consistent form along the graph in all variables than those in the failure dataset. Similarly, the success dataset contains more individuals per unit area than the failure dataset.

6. Automatic Learning Approach

ML is a subfield of artificial intelligence (AI) concerned with implementing computational algorithms that improve performance based on historical data [53]. An ML strategy consists of three essential components: data, representation, and model. The data refer to previously curated dataset information that has been recompiled, followed by the representation, which is the numerical translation of the input information for use in the model, where the selection of variables or features that will comprise the model input can have a significant impact on the model's performance, and finally, the model, which is the mathematical representation of the process. The model, which can be categorized in different ways, is the mathematical representation of the process (unsupervised, supervised, active, or transfer learning). In general, the term ML can be applied to any method that implicitly models correlations within datasets [54,55].

Based on the recompiled data, the supervised learning strategy was chosen. This strategy, which involves the ML task of learning a function that translates an input to an output based on example input–output pairs, is supervised learning. It infers a function from a set of training instances labeled with training data. Supervised ML algorithms require external supervision. Train and test datasets are created from the input dataset. The training dataset contains output variables that require prediction or classification. All algorithms extract patterns from the training dataset and apply them to the test dataset for prediction or classification.

There are numerous techniques within the supervised learning methodology, including linear regression, logistic regression, K-nearest neighbors (KNN), Naïve Bayes, decision trees, and more. However, selecting a technique depends on the desired features and the data type. In this approach, the best-suited technique for the model purpose was determined by recompiling the crucial information of each technique [56]. The KNN technique was chosen based on the recompiled information about the technique that must correspond to the purpose of the model, which is determining the compatibility or incompatibility of AC and WM compounds, and the available information, which is labeled, noncontinuous, and noise-free. In this manner, the KNN technique was implemented in R, considering a data split of 70% training, 10% validation, and 20% test sets (Figure 8). The model was trained using k-fold cross-validation with hyperparameter tuning to choose the best-performing model. The validation and training sets were used in a data combination to retrain (refit) the model, and the test set was used for the classification model for the compatibility prediction.

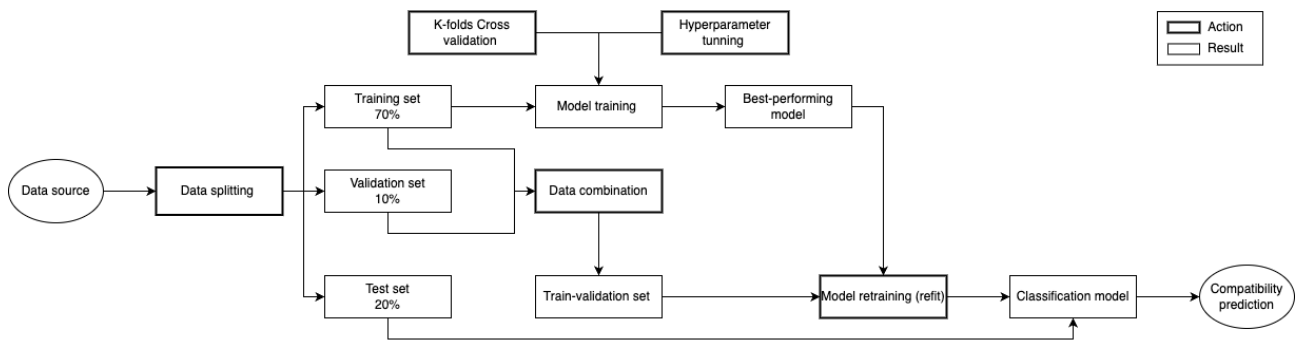


Figure 8. Classification model approach.

Initially, the binary data of the variable “case” were substituted with the strings “yes” and “no” to signify whether there was compatibility or incompatibility. The variable was subsequently transformed from a vector into a factor or category. The raw data, consisting of 288 observations and 21 variables, were divided into three datasets: training, validation, and test. The division was performed in a ratio of 70:10:20. To enhance the strength and reliability of the model, k-fold cross-validation and hyperparameter optimization were conducted simultaneously on the training dataset.

Figure 9 illustrates the implementation of the k-fold cross-validation, specifically using repeated cross-validation with 10 sets of folds. The data were divided into 10 subsets (folds) for cross-validation, and repeated cross-validation was performed three times. The model underwent training using k-1 folds and was assessed on the remaining fold [57]. The process was iterated three times, with each fold used as the validation set once. A single estimation was obtained by averaging the results from each iteration. This method reduces the variability and offers a more comprehensive performance measurement for the model [58]. The Receiver Operating Characteristic (ROC) curve was employed [59] to assess the effectiveness of various hyperparameter configurations, specifically, the number of neighbors used. The ROC curve is a visual depiction that demonstrates the diagnostic capability of a binary classifier system as the discrimination threshold changes [60]. By graphing the sensitivity (true positive rate) against the specificity (false positive rate), it was able to determine the ideal number of neighbors that maximized the area under the curve (AUC), achieving a balance between sensitivity and specificity. Figure 10 displays the ROC curve with repeated cross-validation, indicating the optimal number of neighbors as seven. The AUC value for this configuration is 0.993.

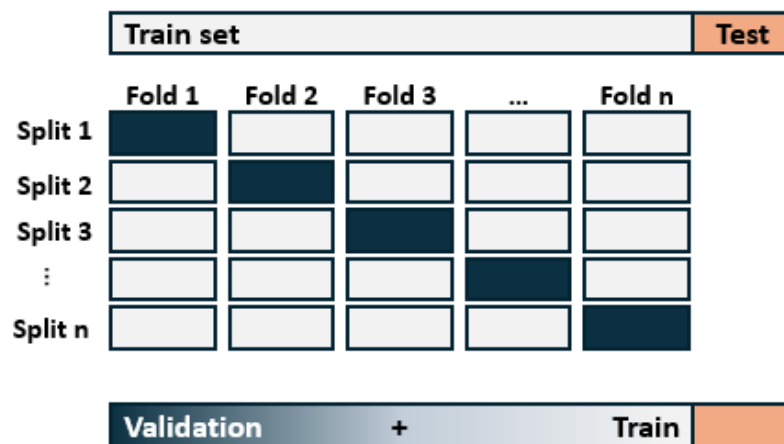


Figure 9. Implemented cross-validation approach [57].

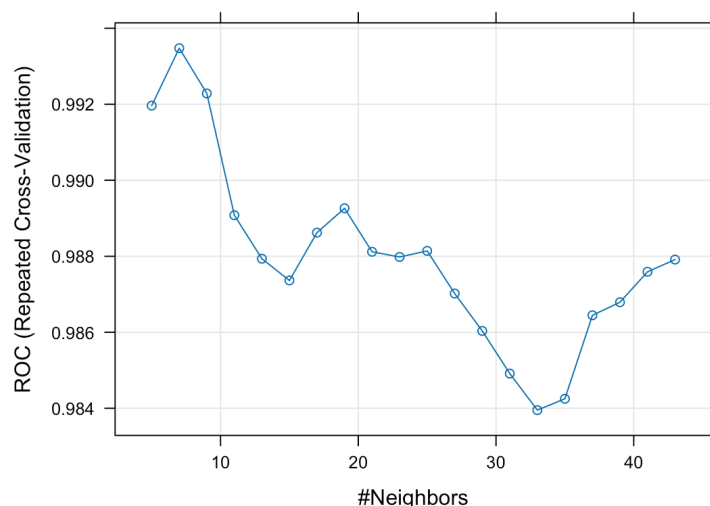


Figure 10. ROC curve. Hyperparameter optimization.

Therefore, to determine the best-performing model on the train set, the optimized model was retrained using the train–validation set, which combines the train and validation set. Ultimately, the model was applied to the unseen test set to evaluate its performance. The model’s performance was evaluated using the confusion matrix and statistical features. Table 7 presents the statistical metrics for assessing the model’s performance.

Table 7. KNN model results.

Parameter	Information
Number of K-neighbors	7
Accuracy	0.923
95% CI	(0.815, 0.979)
Kappa	0.845
McNemar’s Test <i>p</i> -Value	0.134
Sensitivity	0.84
Specificity	1
Positive Predictive Value	1
Negative Predictive Value	0.87
Prevalence	0.481
Detection Rate	0.403
Detection Prevalence	0.404
Balanced Accuracy	0.92
Positive Class	No

The confusion matrix is a tool utilized to assess the effectiveness of a classification model by comparing the predicted labels with the true labels [61]. The decomposition of the confusion matrix is as follows:

- **True Negatives (TN):** There were 21 instances in which the model accurately predicted “No”.
- **False Positives (FP):** There were 4 occurrences where the model made an incorrect prediction of “Yes” when the true label was “No”.
- **False Negatives (FN):** These refer to the instances where the model incorrectly predicted a negative outcome (“No”) when the actual label was a positive outcome (“Yes”). In this case, there were no instances of false negatives.
- **True Positives (TP):** There were 27 instances where the model accurately predicted “Yes”.

The confusion matrix assesses the classification model’s performance by comparing predicted and actual labels. The model accurately classified 21 instances as incompatible (No) and 27 instances as compatible (Yes), leading to a high overall accuracy. The model exhibited four instances of false positives, where it erroneously classified instances as “Yes” instead of “No”, and zero instances of false negatives, indicating flawless recall with

a sensitivity of 100%. The model has a specificity of 84%, accurately identifying most negative cases. Additionally, it has a precision of 87% for compatible (Yes) predictions, indicating that 27 out of 31 “Yes” predictions were correct. Overall, the model demonstrates robust performance, particularly in accurately identifying positive cases, albeit with a slight inclination to incorrectly predict compatible (Yes) cases.

The model’s accuracy, the proportion of accurate predictions relative to the total number of samples, was 0.923. The confidence interval (CI) of the accuracy shows the predicted range of values within which the accuracy will fall, in this case (0.815, 0.979), indicating a high confidence level in the estimated accuracy. Kappa denotes the Kappa statistic, which quantifies the agreement between the expected and actual class labels, considering the probability of chance agreement. A Kappa score 0.845 indicates significant agreement outside chance [62]. McNemar’s test p -value indicates the p -value from McNemar’s test, which evaluates the significance of any changes in error rates between the two models. In this instance, the p -value is 0.134, showing no statistically significant difference between the error rates of the evaluated models [63]. Sensitivity represents the actual positive rate or the proportion of positive samples accurately identified as positive [64].

In this instance, the sensitivity is 0.84, indicating that the model accurately recognizes 84% of positive samples. Specificity denotes the true negative rate or the proportion of actual negative samples accurately categorized as negative. A specificity of one suggests that the model correctly detects 100% negative samples [64]. The Positive Predictive Value is the proportion of samples anticipated to be positive that are, in fact, positive. In this instance, the positive predictive value is one, which indicates that around 100% of the anticipated positive samples are positive. A negative predictive value of 0.87 suggests that around 87.1% projected negative samples are negative [65]. Prevalence denotes the frequency or proportion of positive samples within a dataset [66]. In this instance, the prevalence is 0.481, which indicates that around 48.1% of samples are positive. The detection rate is the fraction of positive samples that the model correctly classifies as positive. A detection rate of 0.403 suggests that the model detects or identifies 40.3% of positive samples. The detection prevalence is the proportion of projected positive samples that are positive. In this instance, the detection prevalence is 0.404, which indicates that approximately 40.4% of the anticipated positive samples are positive [64]. Balanced accuracy is the average of sensitivity and specificity and measures classification performance as a whole [67]. A balanced accuracy of 0.92 means that positive and negative samples are correctly classified. The positive class column identifies the label or class regarded as positive. In this instance, “No” is the positive class. Root Mean Square Error (RMSE) is a metric utilized to assess the precision of predicted values by computing the average discrepancy between predicted probabilities and actual outcomes [61]. It is especially advantageous for assessing models that generate probabilities instead of solely class labels. The classification model obtained an RMSE of 0.228, suggesting that the model’s predicted probabilities are highly accurate and closely match the actual class labels. This demonstrates the model’s effectiveness in making dependable predictions.

Overall, the results indicate that the KNN model performs effectively, exhibiting a high and balanced accuracy. As evidenced by the sensitivity and specificity scores, it performs well in correctly categorizing positive and negative combinations. Positive and negative predictive values reflect the model’s ability to reliably predict positive and negative cases.

7. Conclusions and Recommendations

Computer modeling is the most effective method for evaluating the compatibility between active compounds (ACs) and wall materials (WMs). It eliminates the need for time-consuming trial-and-error experiments and enhances our understanding of their chemical compatibility. Utilizing decoys is beneficial for preserving the precision of parameter and variable patterns in the presence of data constraints. Data purification is an essential and critical step in implementing statistical and Principal Component Analyses. It facilitates the identification of correlations between variables and the consolidation of data while

retaining crucial information. This process is essential before model development. The developed model has proven its accuracy in compatibility assessments, highlighting the efficacy and capacity of this supervised learning approach to integrate new data. Applying model evaluation techniques like k-fold cross-validation and optimizing the number of neighbors improves the reliability and precision of the results. This approach explicitly targets concerns related to overfitting and robustness.

Future research should prioritize incorporating a broader spectrum of molecular descriptors to ensure comprehensive model validation and accurate performance evaluation. Moreover, creating a publicly accessible tool would facilitate the evaluation of compound compatibility and enhance the dataset utilized to enhance the model. By exploring alternative supervised models and optimizing the hyperparameters of each model, one can gain deeper insights into determining compatibility. By utilizing different mathematical tools in the models, a more comprehensive understanding of the data can be achieved, allowing the discovery of additional relationships that need to be considered.

Improvements in computing power and algorithm development are expected to reduce some of the current limitations, increasing the accessibility and reliability of these techniques [68]. Their predicted accuracy can be significantly improved by incorporating machine learning and artificial intelligence (AI) into molecular dynamics (MD) and docking. This is achieved by detecting patterns and correlations that may be overlooked by conventional approaches [69]. In the future, implementing molecular dynamics (MD) and docking for compatibility assessments will probably require a hybrid strategy that combines computational predictions with experimental validation. This integrated approach can exploit the advantages of both computational and empirical methodologies, providing a more thorough and effective way of evaluating compatibility. As computational techniques advance, they will become more valuable in assessing compatibility alongside traditional experimental methods and offering a greater understanding of molecular interactions.

Supplementary Materials: The databases of decoys, phenolic, and wall material registers used in this paper are available as supplementary materials in the online version <https://www.mdpi.com/article/10.3390/a17090412/s1>. These databases include the smiles obtention, the compound name (phenolic and wall material databases), and the molecular descriptors obtention, with their corresponding links for further information.

Author Contributions: Conceptualization, J.Q.-R., R.A.-G. and N.R.; methodology, J.Q.-R. and R.A.-G.; software, J.Q.-R.; validation, J.Q.-R. and R.A.-G.; formal analysis, J.Q.-R.; resources, R.A.-G. and N.R.; data curation, J.Q.-R.; writing—original draft preparation, J.Q.-R. and R.A.-G.; writing—review and editing, J.Q.-R., R.A.-G. and N.R.; visualization, J.Q.-R. and R.A.-G.; supervision, R.A.-G. and N.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Supplementary Material is available with the data implemented in this work.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Parameters

Table A1. Bibliographic review of parameters.

Parameter	Definition	Ref.
Canonical smiles	Translate the three-dimensional structure of a chemical into a string of symbols that computer software can readily interpret	[70]
Molecular weight (MW)	Mass of a single molecule of a compound in a unified atomic mass unit; it is calculated from the atomic masses of the constituents	[71]

Table A1. Cont.

Parameter	Definition	Ref.
LogP	Measured partition coefficient between two phases; it is derived from lipophilicity (lp), which is the ability of a molecule to mix with an oily substance and is measured by its interaction with a lipid. Another way to interpret logP is as the solute concentration in the organic and aqueous partitions. A compound with a negative logP value is more attracted to the aqueous phase, indicating its hydrophilic nature. If the logP value is 0, the molecule is evenly distributed across the lipid and aqueous phases. If the logP value is larger than 0, the compound is lipophilic.	[72]
Hydrogen Bond Acceptors (HBAs)	Number of hydrogen atoms that can be incorporated into a molecule due to the presence of a partial negative charge on the atom that is covalently bonded to hydrogen (mostly NOF)	[73]
Hydrogen Bond Donors (HBDs)	Number of hydrogen atoms that can be donated to a molecule when a highly polar hydrogen atom is bonded to a strongly electronegative atom, primarily, nitrogen, oxygen, or fluorine	[73]
Polar Surface Area (PSA)	Polar portion of a molecule and is proportional to the molecule's solubility	[74]
Heavy atom count (HAC)	Total number of nonhydrogen ('heavy') atoms in its chemical structure. All nonhydrogen component atoms, including carbon, nitrogen, oxygen, sulfur, and halogen, are treated equally, regardless of size.	[75]
Carboxyl groups (CGCs)	Functional groups are formed when hydroxyl (OH) and carbonyl groups are bonded to a single carbon atom (C). This forms a polar, highly electronegative, and weakly acidic group capable of hydrogen bonding through proton donation and acceptance	[76]
Carbonyl groups (CnGCs)	Consist of the bond C=O; carbon is bonded to two other atoms, in which the structure of the carbonyl bond influences the stability and reactivity of carbonyl compounds	[77]
Hydroxyl groups (HGCs)	Simple structures consisting of an oxygen atom with two lone pairs covalently bonded to a hydrogen atom. Adding a hydroxyl group to numerous organic compounds transforms them into alcohols and increases their water solubility.	[76]
Acyclic groups (AGCs)	Number of non-ring-forming linear carbon chains. These groups can reveal a molecule's size, shape, and flexibility.	[78]

Appendix A.2. Decoys

Table A2. Wall materials and respective decoys [33].

Compound	MW	LogP	HBA	HBD	PSA	HAC	CGC	CnGC	HGC	AGC
Curdlan	180.16	-2.60	6	5	110	12	0	0	7	16
DC1	177.16	-3.28	6	5	102.18	12	0	0	6	4
DC2	178.14	-3.01	6	4	107.22	12	1	1	6	4
DC3	176.17	-2.99	5	6	116.8	12	0	0	4	4
DC4	177.16	-3.44	5	5	110.02	12	0	1	4	4
DC5	178.14	-3.01	6	4	107.22	12	1	1	6	4
DC6	177.16	-2.87	5	5	110.02	12	1	1	4	4
DC7	178.14	-3.63	6	5	118.22	12	0	2	5	6
DC8	182.15	-2.86	5	5	101.15	12	0	0	5	6
DC9	177.16	-3.05	5	4	113.01	12	0	1	5	4
DC10	188.18	-3.07	6	5	116.74	13	0	0	4	5
DC11	177.16	-2.87	5	5	110.02	12	1	1	4	5
DC12	185.15	-2.99	6	5	151.81	13	0	2	0	0
DC13	175.18	-2.87	5	4	84.16	12	0	0	4	5
Scleroglucan	714.50	-7.20	20	12	317	46	0	0	27	4
DS1	704.63	-8.12	21	12	362.32	48	2	2	26	16
DS2	705.69	-7.57	21	12	319.73	47	0	0	26	16
Maltodextrin	342.30	-4.70	11	8	190	23	0	0	14	8
DM1	342.30	-5.40	11	8	189.53	23	0	0	14	8
DM2	342.30	-5.40	11	8	189.53	23	0	0	14	8
DM3	342.30	-5.40	11	8	189.53	23	0	0	14	8

Table A2. Cont.

Compound	MW	LogP	HBA	HBD	PSA	HAC	CGC	CnGC	HGC	AGC
Sodium Alginate	418.23	−4.10	12	6	192.44	52	1	1	17	32
DA1	754.92	1.53	15	6	218.84	55	0	1	0	7
DA2	749.98	1.97	14	6	197.07	52	1	1	18	35
DA3	748.01	1.87	14	5	185.87	52	1	1	16	32
DA4	757.83	1.73	14	4	221.29	54	4	7	13	38
DA5	741.92	1.67	14	5	201.75	52	1	4	15	31
DA6	743.93	1.47	14	6	204.91	52	1	3	16	31
DA7	755.86	1.93	13	5	215.22	54	3	6	12	42
DA8	749.81	2.26	14	4	218.21	54	4	6	13	33
DA9	740.72	1.68	14	6	247.61	54	0	7	10	13
DA10	746.98	1.97	14	5	181.17	52	1	1	18	34
DA11	743.89	1.42	15	4	199.98	52	2	3	18	25
DA12	743.89	1.42	15	4	199.98	52	2	3	18	25
DA13	750.75	1.70	15	5	232.65	54	2	8	15	24
DA14	748.78	1.41	14	4	212.42	54	5	5	14	47

Rows shown in grey correspond with the wall material.

References

- Gürbüz, E.; Keresteci, B.; Günneç, C.; Baysal, G. Encapsulation Applications and Production Techniques in the Food Industry. *J. Nutr. Health Sci.* **2020**, *7*, 106.
- Casanova, F.; Santos, L. Encapsulation of cosmetic active ingredients for topical application—A review. *J. Microencapsul.* **2016**, *33*, 1–17. [\[CrossRef\]](#)
- Sonawane, S.; Bhanvase, B.; Sivakumar, M.; Potdar, S. Current overview of encapsulation. In *Encapsulation of Active Molecules and Their Delivery System*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 1–8. [\[CrossRef\]](#)
- Wang, B.; Akanbi, T.; Agyei, D.; Holland, B.; Barrow, C. Coacervation Technique as an Encapsulation and Delivery Tool for Hydrophobic Biofunctional Compounds. In *Role of Materials Science in Food Bioengineering*; Elsevier: Amsterdam, The Netherlands, 2018; pp. 235–261. [\[CrossRef\]](#)
- Botelho, G.; Canas, S.; Lameiras, J. Development of phenolic compounds encapsulation techniques as a major challenge for food industry and for health and nutrition fields. In *Nutrient Delivery*; Elsevier: Amsterdam, The Netherlands, 2017; pp. 535–586. [\[CrossRef\]](#)
- Muljajew, I.; Chi, M.; Vollrath, A.; Weber, C.; Beringer-Siemers, B.; Stumpf, S.; Hoepfener, S.; Sierka, M.; Schubert, U. A combined experimental and in silico approach to determine the compatibility of poly(ester amide)s and indomethacin in polymer nanoparticles. *Eur. Polym. J.* **2021**, *156*, 110606. [\[CrossRef\]](#)
- Wandrey, C.; Bartkowiak, A.; Harding, S. Materials for Encapsulation. In *Encapsulation Technologies for Active Food Ingredients and Food Processing*; Springer: New York, NY, USA, 2010; pp. 31–100. [\[CrossRef\]](#)
- Barrón-García, O.; Morales-Sánchez, E.; Ramírez Jiménez, A.; Antunes-Ricardo, M.; Luzardo-Ocampo, I.; González-Jasso, E.; Gaytán-Martínez, M. Phenolic compounds profile and antioxidant capacity of ‘Ataulfo’ mango pulp processed by ohmic heating at moderate electric field strength. *Food Res. Int.* **2022**, *154*, 111032. [\[CrossRef\]](#)
- Luana Carvalho de Queiroz, J.; Medeiros, I.; Costa Trajano, A.; Piuvezam, G.; Clara de França Nunes, A.; Souza Passos, T.; Heloneida de Araújo Morais, A. Encapsulation techniques perfect the antioxidant action of carotenoids: A systematic review of how this effect is promoted. *Food Chem.* **2022**, *385*, 132593. [\[CrossRef\]](#)
- Câmara, J.; Albuquerque, B.; Aguiar, J.; Corrêa, R.; Gonçalves, J.; Granato, D.; Pereira, J.M.; Barros, L.; Ferreira, I. Food Bioactive Compounds and Emerging Techniques for Their Extraction: Polyphenols as a Case Study. *Foods* **2020**, *10*, 37. [\[CrossRef\]](#)
- Mohsin, A.; Mat Nor, N.; Muhialdin, B.; Mohd Roby, B.; Abadl, M.; Marzlan, A.; Hussain, N.; Meor Hussin, A. The effects of encapsulation process involving arabic gum on the metabolites, antioxidant and antibacterial activity of kombucha (fermented sugared tea). *Food Hydrocoll. Health* **2022**, *2*, 100072. [\[CrossRef\]](#)
- Grgić, J.; Šelo, G.; Planinić, M.; Tišma, M.; Bucić-Kojić, A. Role of the Encapsulation in Bioavailability of Phenolic Compounds. *Antioxidants* **2020**, *9*, 923. [\[CrossRef\]](#)
- Cheng, H.; Liang, L. Characterization and Encapsulation of Natural Antioxidants: Interaction, Protection, and Delivery. *Antioxidants* **2022**, *11*, 1434. [\[CrossRef\]](#)
- Nilsson, N.J. *Introduction to Machine Learning*; Stanford University: Stanford, CA, USA, 1998; pp. 1–188.
- Qi, X.; Zhao, Y.; Qi, Z.; Hou, S.; Chen, J. Machine Learning Empowering Drug Discovery: Applications, Opportunities and Challenges. *Molecules* **2024**, *29*, 903. [\[CrossRef\]](#)
- Liu, H.; Papa, E.; Gramatica, P. QSAR Prediction of Estrogen Activity for a Large Set of Diverse Chemicals under the Guidance of OECD Principles. *Chem. Res. Toxicol.* **2006**, *19*, 1540–1548. [\[CrossRef\]](#)
- Rege, K.; Ladiwala, A.; Hu, S.; Breneman, C.M.; Dordick, J.S.; Cramer, S.M. Investigation of DNA-Binding Properties of an Aminoglycoside-Polyamine Library Using Quantitative Structure-Activity Relationship (QSAR) Models. *J. Chem. Inf. Model.* **2005**, *45*, 1854–1863. [\[CrossRef\]](#)

18. Rosas-Jimenez, J.G.; Garcia-Revilla, M.A.; Madariaga-Mazon, A.; Martinez-Mayorga, K. Predictive Global Models of Cruzain Inhibitors with Large Chemical Coverage. *ACS Omega* **2021**, *6*, 6722–6735. [[CrossRef](#)]
19. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* **2020**, *1*, 045024. [[CrossRef](#)]
20. D'Souza, S.; Prema, K.; Balaji, S. Machine learning models for drug–target interactions: Current knowledge and future directions. *Drug Discov. Today* **2020**, *25*, 748–756. [[CrossRef](#)]
21. Piotrowski, N. Machine learning approach to packaging compatibility testing in the new product development process. *J. Intell. Manuf.* **2024**, *35*, 963–975. [[CrossRef](#)]
22. Periwal, V.; Bassler, S.; Andrejev, S.; Gabrielli, N.; Patil, K.R.; Typas, A.; Patil, K. Bioactivity assessment of natural compounds using machine learning models trained on target similarity between drugs. *PLoS Comput. Biol.* **2022**, *18*, e1010029. [[CrossRef](#)]
23. Maxwell, A.; Warner, T.; Fang, F. Implementation of machine-learning classification in remote sensing: An applied review. *Int. J. Remote. Sens.* **2018**, *39*, 2784–2817. [[CrossRef](#)]
24. Rothwell, J.A.; Perez-Jimenez, J.; Neveu, V.; Medina-Reyon, A.; M'Hiri, N.; Garcia-Lobato, P.; Manach, C.; Knox, C.; Eisner, R.; Wishart, D.S.; et al. Phenol-Explorer 3.0: A major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. *Database* **2013**, *2013*, bat070. [[CrossRef](#)]
25. Paulsen, B. *Bioactive Carbohydrate Polymers*; Springer: Dordrecht, The Netherlands, 2000; Volume 44. [[CrossRef](#)]
26. Anandharamakrishnan, C.; Ishwarya, S. Selection of wall material for encapsulation by spray drying. In *Spray Drying Techniques for Food Ingredient Encapsulation*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2015; Chapter 4; pp. 77–100. [[CrossRef](#)]
27. Landrum, G.; Tosco, P.; Kelley, B.; Ric, S.; Sriniker; Cosgrove, D.; Geddeck; Vianello, R.; NadineSchneider; Kawashima, E.; et al. rdkit/rdkit: 2023_03_1b1 (Q1 2023) Release. 2023. Available online: <https://zenodo.org/records/7828379> (accessed on 6 September 2024).
28. Winiwarter, S.; Ridderström, M.; Ungell, A.L.; Andersson, T.; Zamora, I. Use of Molecular Descriptors for Absorption, Distribution, Metabolism, and Excretion Predictions. In *Comprehensive Medicinal Chemistry II*; Elsevier: Amsterdam, The Netherlands, 2007; pp. 531–554. [[CrossRef](#)]
29. Chandrasekaran, B.; Abed, S.; Al-Attraqchi, O.; Kuche, K.; Tekade, R. Computer-Aided Prediction of Pharmacokinetic (ADMET) Properties. In *Dosage Form Design Parameters*; Elsevier: Amsterdam, The Netherlands, 2018; pp. 731–755. [[CrossRef](#)]
30. Song, Y.; Zhou, L.; Zhang, D.; Wei, Y.; Jiang, S.; Chen, Y.; Ye, J.; Shao, X. Stability and release of peach polyphenols encapsulated by Pickering high internal phase emulsions in vitro and in vivo. *Food Hydrocoll.* **2023**, *139*, 108593. [[CrossRef](#)]
31. Kak, A.; Parhi, A.; Rasco, B.; Tang, J.; Sablani, S. Improving the oxygen barrier of microcapsules using cellulose nanofibres. *Int. J. Food Sci. Technol.* **2021**, *56*, 4258–4267. [[CrossRef](#)]
32. Yan, J.K.; Wang, Z.W.; Zhu, J.; Liu, Y.; Chen, X.; Li, L. Polysaccharide-based nanoparticles fabricated from oppositely charged curdlan derivatives for curcumin encapsulation. *Int. J. Biol. Macromol.* **2022**, *213*, 923–933. [[CrossRef](#)] [[PubMed](#)]
33. Gori, D.; Alberca, L.; Rodriguez, S.; Alice, J.; Llanos, M.; Bellera, C.; Talevi, A. LIDeB Tools: A Latin American resource of freely available, open-source cheminformatics apps. *Artif. Intell. Life Sci.* **2022**, *2*, 100049. [[CrossRef](#)]
34. Balderrama, M.; Ángel, J. Phase Change Materials Encapsulation in Crosslinked Polymer-Based Monoliths: Syntheses, Characterization and Evaluation of Pullulan and Black Liquor Based-Monoliths for the Encapsulation of Phase Change Materials. Ph.D. Thesis, Université de Bordeaux, Bordeaux, France, 2018.
35. Gharsallaoui, A.; Roudaut, G.; Chambin, O.; Voilley, A.; Saurel, R. Applications of spray-drying in microencapsulation of food ingredients: An overview. *Food Res. Int.* **2007**, *40*, 1107–1121. [[CrossRef](#)]
36. Champagne, C.P.; Fustier, P. Microencapsulation for the improved delivery of bioactive compounds into foods. *Curr. Opin. Biotechnol.* **2007**, *18*, 184–190. [[CrossRef](#)]
37. Jafari, S.M.; Assadpoor, E.; He, Y.; Bhandari, B. Re-coalescence of emulsion droplets during high-energy emulsification. *Food Hydrocoll.* **2008**, *22*, 1191–1202. [[CrossRef](#)]
38. Augustin, M.A.; Sanguansri, L. *Encapsulation of Bioactives*; Springer: New York, NY, USA, 2008; pp. 577–601. [[CrossRef](#)]
39. McClements, D.J. *Future Foods: How Modern Science Is Transforming the Way We Eat*, 1st ed.; Copernicus: Cham, Switzerland, 2019.
40. Capecchi, A.; Probst, D.; Reymond, J.L. One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome. *J. Cheminform.* **2020**, *12*, 43. [[CrossRef](#)]
41. Tanimoto Similarity and Jaccard Indexes with FeatureBase. Available online: <https://www.featurebase.com/blog/tanimoto-similarity-in-featurebase> (accessed on 6 September 2024).
42. Becker, R.; Chambers, J.; Wilks, A. *The New S Language*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018. [[CrossRef](#)]
43. Rice, J.A. *Mathematical Statistics and Data Analysis*; Brooks/Cole, Cengage Learning: Boston, MA, USA, 2007.
44. Johnson, R.A.; Wichern, D.W. *Applied Multivariate Statistical Analysis*, 6th ed.; Springer: Berlin/Heidelberg, Germany, 2007. [[CrossRef](#)]
45. Lindley, D. Introduction to the practice of statistics, (3rd edition), by David S. Moore and George P. McCabe. Pp. 825 (with appendices and CD-ROM). £27.95. 1999; ISBN 0 7167 3502 4 (W. H. Freeman). *Math. Gaz.* **1999**, *83*, 374–375. [[CrossRef](#)]
46. Agresti, A.; Franklin, C.; Klingenberg, B. *Statistics: The Art and Science of Learning from Data*, 4th ed.; Pearson: London, UK, 2016.
47. Field, A. *Discovering Statistics Using IBM SPSS Statistics*, 5th ed.; SAGE Publ.: London, UK, 2017.
48. Wilcox, R. *Introduction to Robust Estimation and Hypothesis Testing*, 4th ed.; Elsevier: Amsterdam, The Netherlands, 2017.

49. Greenacre, M.; Groenen, P.; Hastie, T.; D'Enza, A.; Markos, A.; Tuzhilina, E. Principal component analysis. *Nat. Rev. Methods Prim.* **2022**, *2*, 100. [CrossRef]
50. Toutenburg, H. Mardia, K. V./Kent, J. T./Bibby, J. M., *Multivariate Analysis*. London-New York-Toronto-Sydney-San Francisco, Academic Press 1979. XV, 521 S., \$ 34.00 P/B. ISBN 0-12-471252-5. *ZAMM-J. Appl. Math. Mech./Z. Angew. Math. Mech.* **1981**, *61*, 206. [CrossRef]
51. Venables, W.; Ripley, B. *Modern Applied Statistics with S*, 4th ed.; Statistics and Computing; Springer: Berlin/Heidelberg, Germany, 2002.
52. Holland, S.M. *Principal Components Analysis (PCA)*; Technical Report; Department of Geology, University of Georgia: Athens, GA, USA, 2019.
53. Jovel, J.; Greiner, R. An Introduction to Machine Learning Approaches for Biomedical Research. *Front. Med.* **2021**, *8*, 771607. [CrossRef] [PubMed]
54. Michalski, R.; Carbonell, J.; Mitchell, T. *Machine Learning: An Artificial Intelligence Approach*; Symbolic Computation; Springer: Berlin/Heidelberg, Germany, 2013.
55. Dey, A. Machine Learning Algorithms: A Review. *Int. J. Comput. Sci. Inf. Technol.* **2016**, *7*, 1174–1179.
56. Mohamed, S.; Ashraf, R.; Ghanem, A.; Sakr, M.; Mohamed, R. *Supervised Machine Learning Techniques: A Comparison*; Technical Report; Universiti Sains Malaysia: Gelugor, Malaysia, 2022.
57. Berrar, D. Cross-Validation. In *Encyclopedia of Bioinformatics and Computational Biology*; Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C., Eds.; Academic Press: Oxford, UK, 2019; pp. 542–545. [CrossRef]
58. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995.
59. Kuhn, M. Futility Analysis in the Cross-Validation of Machine Learning Models. *arXiv* **2014**, arXiv:1405.6974.
60. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
61. Swets, J.A. Measuring the Accuracy of Diagnostic Systems. *Science* **1988**, *240*, 1285–1293. [CrossRef]
62. Xia, Y. Chapter Eleven—Correlation and association analyses in microbiome study integrating multiomics in health and disease. In *The Microbiome in Health and Disease*; Progress in Molecular Biology and Translational Science; Sun, J., Ed.; Academic Press: Cambridge, MA, USA, 2020; Volume 171, pp. 309–491. : 10.1016/bs.pmbts.2020.04.003 [CrossRef]
63. Pembury Smith, M.; Ruxton, G. Effective use of the McNemar test. *Behav. Ecol. Sociobiol.* **2020**, *74*, 133. [CrossRef]
64. Hanga, A.; Alalyani, M.; Hussain, I.; Almutheibi, M. Brief review on Sensitivity, Specificity and Predictivities. *IOSR J. Dent. Med. Sci.* **2015**, *14*, 64–68. [CrossRef]
65. Safari, S.; Baratloo, A.; Elfil, M.; Negida, A. Evidence Based Emergency Medicine Part 2: Positive and Negative Predictive Values of Diagnostic Tests. *Emergency* **2015**, *3*, 87–88.
66. Barranquero, J.; González, P.; Díez, J.; del Coz, J. On the study of nearest neighbor algorithms for prevalence estimation in binary problems. *Pattern Recognit.* **2013**, *46*, 472–482. [CrossRef]
67. García, V.; Mollineda, R.A.; Sánchez, J.S. Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions. In *Pattern Recognition and Image Analysis*; IbPRIA 2009. Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2009; pp. 441–448. _57 [CrossRef]
68. Dror, R.O.; Dirks, R.M.; Grossman, J.; Xu, H.; Shaw, D.E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41*, 429–452. [CrossRef]
69. Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2020**, *2*, 573–584. [CrossRef]
70. Usepa; Ocspp; Oppt; Rad. Sustainable Futures / P2 Framework Manual 2012 EPA-748-B12-001 Appendix F. SMILES Notation Tutorial. Technical Report. Available online: <https://www.epa.gov/sites/default/files/2015-05/documents/appendf.pdf> (accessed on 6 September 2024).
71. Speight, J. Chemical and physical properties of hydrocarbons. In *Handbook of Industrial Hydrocarbon Processes*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 387–420. [CrossRef]
72. Bhal, S.K. LogP—Making Sense of the Value. Application Note. Available online: https://www.acdlabs.com/wp-content/uploads/download/app/physchem/making_sense.pdf (accessed on 6 September 2024).
73. van Osch, D.; Dietz, C.; van Spronsen, J.; Kroon, M.; Gallucci, F.; van Sint Annaland, M.; Tuinier, R. A Search for Natural Hydrophobic Deep Eutectic Solvents Based on Natural Components. *ACS Sustain. Chem. Eng.* **2019**, *7*, 2933–2942. [CrossRef]
74. Barret, R. Importance and Evaluation of the Polar Surface Area (PSA and TPSA). In *Therapeutic Chemistry*; Elsevier: Amsterdam, The Netherlands, 2018; pp. 89–95. [CrossRef]
75. Leeson, P.; Bento, A.; Gaulton, A.; Hersey, A.; Manners, E.; Radoux, C.; Leach, A. Target-Based Evaluation of “Drug-like” Properties and Ligand Efficiencies. *J. Med. Chem.* **2021**, *64*, 7210–7230. [CrossRef] [PubMed]
76. Klecker, C.; Nair, L. Matrix Chemistry Controlling Stem Cell Behavior. In *Biology and Engineering of Stem Cell Niches*; Elsevier: Amsterdam, The Netherlands, 2017; pp. 195–213. [CrossRef]

-
77. Ouellette, R.; Rawn, J. Aldehydes and Ketones. In *Organic Chemistry Study Guide*; Elsevier: Amsterdam, The Netherlands, 2015; pp. 313–333. [[CrossRef](#)]
78. Berrick, A.J. Remarks on the Structure of Acyclic Groups. *Bull. Lond. Math. Soc.* **1990**, *22*, 227–232. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.