



UNIVERSIDADE D  
COIMBRA

Cláudia Filipa Soares Ferreira

**MYCOTOXINS: IDENTIFICATION AND  
CHARACTERIZATION USING MACHINE LEARNING  
AS A PLAYGROUND**

Dissertação de Mestrado na área científica de Química orientada pelo Professor Doutor Alberto Canelas Pais e pela Doutora Tânia Firmino Guerra Guerreiro da Cova apresentada ao Departamento de Química da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Setembro de 2021



Faculdade de Ciências e Tecnologia  
da Universidade de Coimbra

# Mycotoxins: identification and characterization using machine learning as a playground

Cláudia Filipa Soares Ferreira

Dissertação de Mestrado na área científica de Química orientada pelo Professor Doutor Alberto Canelas Pais e pela Doutora Tânia Firmino Guerra Guerreiro da Cova apresentada ao Departamento de Química da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.



UNIVERSIDADE D  
COIMBRA

Setembro de 2021



## Acknowledgments

---

Esta página será escrita em português apenas para que todas as pessoas a quem estou extremamente grata, possam ler e perceber.

Vou citar o senhor António Nóvoa que um dia me deu uma lição que nunca esqueci e que fala sobre o tratado da Gratidão de São Tomás de Aquino. Este tratado tem três níveis de gratidão: um superficial, um intermédio e um profundo. O primeiro trata do reconhecimento, o segundo do agradecimento e o terceiro, do vínculo. Podemos dizer obrigada de várias formas e línguas. Em inglês “thank you”, é um agradecimento do nível mais superficial em que há simplesmente o reconhecimento. Em francês “merci” ou em espanhol “gracias”, dá-se uma mercê, um nível mais intermédio, em que se dá uma graça a algo que nos fizeram ou trouxeram. E o nosso português, “obrigada”, que emite a questão do vínculo e o mais poderoso dos agradecimentos.

A todas as pessoas que irei mencionar, estou obrigada...vinculada...perante vós.

Pai e Mãe, faço, claro, isto por mim, mas não há nada que me deixe mais feliz do que ver-vos tão orgulhosos e satisfeitos com o meu percurso. Foram, são e serão a minha maior força da vida e sem dúvida que sem o vosso suporte não seria nada igual.

Tozé.... Confesso que parei a sorrir enquanto escrevia isto. Viste a alegria, o medo, a frustração, a esperança, o desespero, a satisfação e nem por um segundo hesitaste em dar-me confiança, apoio, motivação (nem mesmo quando eu não merecia). Não tenho palavras que consigam explicar o que pretendo, mas dizer-te, pelo menos, que marcaste a minha vida e o meu percurso da forma mais intensa e positiva possível.

Professor Doutor Alberto Canelas Pais e Doutora Tânia Firmino. Nada seria possível sem o vosso suporte. Tudo começa com uma oportunidade e cabe-nos a nós aproveitar ou usufruir dela. Deram-me esta e outras oportunidades, imenso conhecimento, confiança e inspiração. Tentarei aplicar todas as dicas, truques, sabedoria não só academicamente, mas no meu dia-a-dia. Toda a ajuda e paciência, não há palavras para agradecer, nem mesmo no nível mais profundo do agradecimento.

Bruna, Sofia Figueira, Kristian, Piochi, Zé Gui, Mariana Diniz, Manel, Padrinho Mágico, Rastas, Luana. Alguns poderão ficar surpreendidos por estarem aqui. A verdade é que a vida não é só a tese ou os estudos. E todos vocês de forma direta, indireta ou até sem qualquer noção do que significava para mim, me ajudaram muito. Quer nas risadas, no choro, uma simples companhia, desabafos, tudo isso não me passa despercebido e por isso, obrigada também.

Ao resto da minha família.



## Abstract

---

The design of cost-effective strategies to simultaneously identify and eliminate toxic compounds from the aquatic environment requires knowledge of relevant molecular fingerprints, interaction patterns, co-occurrence, synergistic effects, and contaminant sources, as this can be a gateway to an effective response to these societal obstacles.

This study aims to develop predictive models for molecular similarity and toxicity of mycotoxins based on molecular and physicochemical descriptors using cheminformatics tools and machine learning approaches. An efficient chemical data mining over different datasets composed by 30 and 59 selected mycotoxins described by several molecular descriptors is proposed for virtual screening of molecular similarity and toxicity prediction.

Hierarchical cluster analysis and k-means clustering revealed clusters consistent with the known mycotoxin families. PCA results show that discrimination between mycotoxins is largely determined by the selected molecular descriptors and evidence a tendency in the separation of acutely toxic mycotoxins from non-acutely toxic mycotoxins.

Supervised learning models (LDA, RF, SVM, NN) were constructed for the purpose of classification and combined with the molecular descriptors selected from PCA to improve the knowledge of the selected mycotoxins and predict their respective acute-toxicity profiles. RF proved to be the best model in the classification of mycotoxins into acutely toxic or non-acutely toxic.

This study allows the identification of relevant molecular and physicochemical descriptors for the 1) discrimination of different families of mycotoxins, 2) classification of structurally distinct mycotoxins and also those mycotoxins that are not so well described in the literature, and 3) prediction of toxicity. This creates a gateway for the subsequent classification, identification, and rapid and efficient characterization of potential new and unknown mycotoxins. Bridging the gap between multivariate physicochemical data and the ability of models to predict and address relevant mycotoxin-related phenomena, such as co-occurrence and molecular recognition, and to develop improved classification and remediation methods remains a challenge, often limited by available methodologies and experimental information.

Keywords: Mycotoxins, Machine Learning, Molecular Descriptors, Acute toxicity





## Resumo

---

O desenvolvimento de estratégias efetivas para identificar e eliminar compostos tóxicos do ambiente aquático requer um conhecimento profundo sobre os padrões moleculares e de interação, os fenômenos de coocorrência, e a origem dos contaminantes.

Este estudo consiste no desenvolvimento de modelos computacionais capazes de caracterizar a similaridade molecular e prever a toxicidade de diferentes classes de micotoxinas, baseando-se em descritores físico-químicos e moleculares e dando uso a ferramentas de aprendizagem computacional.

A análise de agrupamentos hierárquica e o método k-médias revelaram grupos consistentes com as famílias de micotoxinas já estabelecidas na literatura. A análise de componentes principais permitiu selecionar os descritores moleculares mais relevantes para discriminar diferentes famílias de micotoxinas, evidenciando tendências na classificação das micotoxinas tendo em conta a sua toxicidade aguda.

Diversos modelos de aprendizagem supervisionada (LDA, RF, SVM e NN) foram construídos sobre os descritores moleculares selecionados da análise de componentes principais, com o intuito de melhorar o conhecimento sobre as micotoxinas selecionadas e desenvolver modelos de previsão para os seus perfis de toxicidade aguda. O modelo RF provou ser o melhor modelo na classificação das micotoxinas em tóxicas de forma aguda ou não.

Este estudo permite a identificação de descritores moleculares e físico-químicos relevantes para: 1) a discriminação entre várias famílias de micotoxinas, 2) a classificação de micotoxinas estruturalmente distintas e de micotoxinas desconhecidas, e 3) previsão da respetiva toxicidade. Estabelecer a ponte entre dados físico-químicos multivariados e a capacidade dos modelos computacionais de direcionar e prever fenômenos relacionados com micotoxinas, bem como desenvolver métodos mais eficientes de classificação e remediação, são desafios atuais, cuja solução está ainda muito limitada pelas metodologias e os dados experimentais disponíveis.

Palavras-Chave: Micotoxinas, Aprendizagem Computacional, Descritores Moleculares, Toxicidade aguda.



# Table of Contents

---

List of Abbreviations .....	11
Table index .....	13
Figure Index .....	15
Chapter 1 - Introduction .....	19
Chapter 2 – Theory and Methods.....	27
2.1 Unsupervised Learning .....	28
2.1.1 Hierarchical Cluster Analysis .....	30
2.1.2 K-means Clustering .....	30
2.1.3 Principal Component Analysis .....	30
2.2 Supervised Learning .....	32
2.2.1 Linear Discriminant Analysis .....	32
2.2.2 Random Forest.....	33
2.2.3 Support Vector Machines.....	34
2.2.4 Neural Networks.....	35
2.2.4.1 Grid Search.....	36
2.3 Evaluation of the models.....	37
2.3.1 Unsupervised.....	37
2.3.2 Supervised.....	38
Chapter 3 - Database Description and Data Processing.....	41
3.1 Molecular Fingerprints .....	41
3.2 Molecular Descriptors .....	43
3.3 Class Attribution .....	47
3.4 Construction of the datasets .....	47
3.5 Feature Extraction and Selection .....	48
Chapter 4 - Results and Discussion.....	51
4.1 Molecular Similarity .....	51
4.2 Impact of Molecular Descriptors .....	58
Dataset A .....	62
Dataset B.....	66
Dataset C .....	70
Dataset D.....	73
Dataset E .....	76
4.3 Toxicity Prediction.....	79

4.3.1 Linear Discriminant Analysis .....	80
4.3.2 Random Forest.....	81
4.3.3 Support Vector Machines.....	82
4.3.4 Neural Networks.....	83
Chapter 5 – Conclusion and Future Perspectives .....	85
References.....	87
Annexes.....	94

## List of Abbreviations

---

**1D:** One-Dimensional  
**2D:** Two-Dimensional  
**ANN:** Artificial Neural Network  
**ATA:** Alimentary Toxic Aleukia  
**AUROC:** Area Under the Receiving Operating Curve  
**CDK:** Chemistry Development Kit  
**DNN:** Deep Neural Network  
**DTs:** Decision Trees  
**ECEP:** Extended Connectivity Fingerprints  
**FP:** False Positive  
**FN:** False Negative  
**GHS:** Global Harmonized System  
**GS:** Gridsearch  
**HCA:** Hierarchical Cluster Analysis  
**IDE:** Integrating Development Environments  
**JOELib:** Java based cheminformatics/computational chemistry packages  
**kNN:** k-Nearest Neighbors  
**LBFGS:** Limited-Memory Broyden-Fletcher-Goldfarb-Shanno  
**LDA:** Linear Discriminant Analysis  
**MACCS:** Molecular ACCess System  
**ML:** Machine Learning  
**NN:** Neural Networks  
**OEChem:** Open Eye for Chemical Information Processing  
**PCA:** Principal Component Analysis  
**PCs:** Principal Components  
**QSARs:** Quantitative Structure-Activity Relationships  
**RBF:** Radial Basis Function  
**ReLU:** Rectified Linear Unit  
**RF:** Random Forest  
**RFE:** Recursive Feature Elimination  
**SAR:** Structure-Activity Relationships  
**SDF:** Structure-Data File  
**SGD:** Stochastic Gradient Descent  
**SVM:** Support Vector Machines  
**TGD:** Typed Graph Distance  
**TGT:** Typed Graph Triangle  
**TP:** True Positive

**TPR:** True Positive Rate

**TN:** True Negative

**TNR:** True Negative Rate

## Table index





---

**Table 1:** Summary of the most common mycotoxin families and their respective characteristics. The structures were constructed using the online version of ChemDraw.

**Table 2:** Examples of ML applications to environmental micropollutant problems.

**Table 3:** Gridsearch combination parameters for the NN model.

**Table 4:** Example of a Confusion Matrix.

**Table 5:** Description of the molecular descriptors used in this study obtained by the CDK library. Only the resulting descriptors from the feature selection process are shown. The colors represent the category of the molecular descriptor:  hybrid;  topological;  electronic;  constitutional.

**Table 6:** Eigenvalues and evolution of the percentage of information recovery in relation to the number of principal components for the dataset composed by 30 mycotoxins and 15 molecular descriptors. The most relevant descriptors are underlined and correspond to eigenvalues greater than 1.

**Table 7:** Eigenvalues and evolution of the percentage of information recovery in relation to the number of principal components for the dataset composed by 59 mycotoxins and 24 molecular descriptors. The most relevant descriptors are underlined and correspond to eigenvalues greater than 1.

**Table 8:** Eigenvalues and evolution of the percentage of information recovery in relation to the number of principal components for the dataset composed by 59 mycotoxins and 35 molecular descriptors. The most relevant descriptors are underlined and correspond to eigenvalues greater than 1.

**Table 9:** Eigenvalues and evolution of the percentage of information recovery in relation to the number of principal components for the dataset composed by 59 mycotoxins and 28 molecular descriptors. The most relevant descriptors are underlined and correspond to eigenvalues greater than 1.

**Table 10:** Eigenvalues and evolution of the percentage of information recovery in relation to the number of principal components for the dataset composed by 59 mycotoxins and 39 molecular descriptors. The most relevant descriptors are underlined and correspond to eigenvalues greater than 1.

**Table 11:** Eigenvalues and evolution of the percentage of information recovery in relation to the number of principal components for the dataset composed by 59 mycotoxins and 12 molecular descriptors. The most relevant descriptors are underlined and correspond to eigenvalues greater than 1.

**Table 12:** Performance evaluation metrics for the LDA models. The dataset with the best performance is underlined.

**Table 13:** Performance evaluation metrics for the RF models. The dataset with the best performance is underlined.

**Table 14:** Performance evaluation metrics for the SVM models. The dataset with the best performance is underlined.

**Table 15:** Tested parameters using gridsearch that gave the best results in terms of accuracy of the NN models.

**Table 16:** Performance evaluation metrics for the NN models. The dataset with the best performance is underlined.



## Figure Index

---

**Figure 1:** Exemplification of the separation between two classes of objects by LDA. Data is projected in a new axis to maximize the separation between the 2 categories (A) by maximizing the distance between means and minimizing variation (B).

**Figure 2:** Schematic representation of the RF procedure. On the left are 3 decision trees through which an object (mycotoxin) is subjected (on the right), only progressing if given parameters are met. The instance will belong to different classes depending on the path taken.

**Figure 3:** Scheme illustrating the SVM basis. The instances are divided by a hyperplane that separates them by the greatest amount of distance feasible.

**Figure 4:** Representation of a shallow neural network (on the left) with only one hidden layer, and a deep neural network (on the right) with two hidden layers.

**Figure 5:** Representation of a hypothetical 21-bit substructure fingerprint, with 6 bits set as a result of the substructures they represent are present in the molecule.

**Figure 6:** Representation of a hypothetical 14-bit topological fingerprint. Fragments found from the starting atom (circled in red) are shown with their respective length and corresponding bit. There are two-bit collisions, which are bits that are set by more than one fragment.

**Figure 7:** Representation of a hypothetical 5-bit circular fingerprint.

**Figure 8:** Schematic representation of mycotoxins (30) present on the first dataset (on the left), the posteriorly added mycotoxins (29) (on the right) and their respective acronym and class.

**Figure 9:** A composed view of the steps for datasets construction and feature selection. Feature selection method 1 was only applied to the dataset composed by 30 mycotoxins. For the dataset composed by 59 mycotoxins, all feature selection methods were applied. nHBDon is in bold because it was one of the manually selected descriptors, but the process of feature selection also considered this descriptor.

**Figure 10:** Dendrogram representing the similarity among 30 mycotoxins considering their molecular fingerprints, constructed resorting to Ward's method over the Tanimoto distances. The different colors represent the different families: fumonisins are colored in green, dark blue represents the trichothecenes family (excepting CTN and PATL), in orange are ergot alkaloids, aflatoxins in red and in light blue are the remaining mycotoxins.

**Figure 11:** K-means cluster plot constructed over the Tanimoto distances between the 30 mycotoxin structures, represented by their molecular fingerprints.

**Figure 12:** Silhouette plot for k-means clustering constructed over the initial 30 mycotoxins' topological fingerprints.

**Figure 13:** Dendrogram representing the similarity among all 59 mycotoxins considering their molecular fingerprints, constructed resorting to Ward's method over the Tanimoto distances. In red are aflatoxins, orange are ergot alkaloids, brown are ochratoxins, fumonisins are colored in green, enniatins in pink, dark blue represents the trichothecenes family, and in light blue and yellow are the mycotoxins that are from different families or do not belong to a specific family.

**Figure 14:** K-means cluster plot used over the Tanimoto distances between the 59 mycotoxin structures, represented by their molecular fingerprints.

**Figure 15:** Silhouette plot for k-means clustering constructed over the 59 mycotoxins' topological fingerprints.

**Figure 16:** Graphical representation of the impact of molecular descriptors (*loadings*) on the first two principal components (correlation matrix) for the dataset composed by 30 mycotoxins.

**Figure 17:** Biplot representation of 30 mycotoxins described by 15 molecular descriptors on the first two principal components, recovering 46.8% of the variance. Mycotoxins are colored according to the clustering results (see Figure 10) and the molecular descriptors were selected according to feature selection method 1.

**Figure 18:** Biplot representation of 30 mycotoxins described by 15 molecular descriptors on the first two principal components, recovering 46.8% of the variance. Mycotoxins are labeled according to their acute toxicity (see Figure 8).

**Figure 19:** Graphical representation of the impact of molecular descriptors (*loadings*) on the first two principal components (correlation matrix) for the dataset composed by 59 mycotoxins and 24 molecular descriptors. Molecular descriptors resulted from feature selection method 2.

**Figure 20:** Biplot representation of 59 mycotoxins described by 24 molecular descriptors on the first two principal components, recovering 37.8% of the variance. Mycotoxins are colored according to the clustering results (see Figure 13) and the molecular descriptors resulted from feature selection method 2.

**Figure 21:** Biplot representation of 59 mycotoxins described by 24 molecular descriptors on the first two principal components, recovering 37.8% of the variance. Mycotoxins are colored according to their acute toxicity (see Figure 8).

**Figure 22:** Graphical representation of the impact of molecular descriptors (*loadings*) on the first two principal components (correlation matrix) for the dataset composed by 59 mycotoxins and 35 molecular descriptors. Molecular descriptors resulted from feature selection method 1 with the addition of biological-activity related descriptors.

**Figure 23:** Biplot representation of 59 mycotoxins described by 35 molecular descriptors on the first two principal components, recovering 48% of the variance. Mycotoxins are colored according to the clustering results (see Figure 13) and the molecular descriptors resulted from feature selection method 2 with the addition of biological-activity related descriptors.

**Figure 24:** Biplot representation of 59 mycotoxins described by 35 molecular descriptors on the first two principal components, recovering 48% of the variance. Mycotoxins are colored according to their acute toxicity (see Figure 8).

**Figure 25:** Graphical representation of the impact of molecular descriptors (*loadings*) on the first two principal components (correlation matrix) for the dataset composed by 59 mycotoxins and 28 molecular descriptors.

**Figure 26:** Biplot representation of 59 mycotoxins described by 28 molecular descriptors on the first two principal components, recovering 35.3% of the variance. Mycotoxins are colored

according to the clustering results (see Figure 13) and the molecular descriptors resulted from feature selection method 1.

**Figure 27:** Biplot representation of 59 mycotoxins described by 28 molecular descriptors on the first two principal components, recovering 35.3% of the variance. Mycotoxins are colored according to their acute toxicity (see Figure 8).

**Figure 28:** Graphical representation of the impact of molecular descriptors (*loadings*) on the first two principal components (correlation matrix) for the dataset composed by 59 mycotoxins and 39 molecular descriptors.

**Figure 29:** Biplot representation of 59 mycotoxins described by 39 molecular descriptors on the first two principal components, recovering 45.9% of the variance. Mycotoxins are colored according to the clustering results (see Figure 13) and the molecular descriptors resulted from feature selection method 1 with the addition of biological-activity related descriptors.

**Figure 30:** Graphical representation of the impact of molecular descriptors (*loadings*) on the first two principal components (correlation matrix) for the dataset composed by 59 mycotoxins and 12 biological activity descriptors. The top and the bottom plots refer to PC1 and PC2, respectively.

**Figure 31:** Biplot representation of 59 mycotoxins described by 12 molecular descriptors related to biological activity on the first two principal components, recovering 79.5% of the variance. Mycotoxins are colored according to the clustering results (see Figure 13).

**Figure 32:** Biplot representation of 59 mycotoxins described by 12 molecular descriptors related to biological activity on the first two principal components, recovering 79.5% of the variance. Mycotoxins are colored according to their acute toxicity (see Figure 8).



## Chapter 1 - Introduction

---

Micropollutants in the aquatic environment have been of great concern over the last few decades. Also called emerging contaminants, micropollutants involve a large spectrum of anthropogenic or natural substances including drugs, industrial chemicals, hormones, pesticides, mycotoxins, among others. (1,2) Although its presence in the aquatic environment is at trace concentrations (ranging from a few ng/L to several  $\mu\text{g/L}$ ), micropollutants are associated with negative effects such as endocrine disruption, short and long-term toxicity, and antibiotic resistance of microorganisms. (2) Due to their low concentrations and increasing diversity, their respective detection, analysis, and treatment becomes challenging. Currently, one class of micropollutants with great global impact are mycotoxins, toxic compounds that are naturally produced by molds during the growth or storage of plant products, and are capable of causing disease and death in both humans and animals. (3,4)

Historically, fungi are known to be responsible for some cause-effect relationships between human consumption of moldy food and illness. (3,4) Scientific interest started when the ability of fungi to carry out fermentations was discovered and there was a need to understand the toxicity of the “secondary metabolites” involved. (4) The appearance of penicillin promoted studies related to fungi, from grain storage to animal toxicity. With this, and with diseases of the time that were thought to be related to mold contamination, mycotoxins were discovered. (4) Mycotoxins are secondary metabolites produced by fungi capable of causing adverse effects in humans and animals, ranging from allergic responses to cancer and death even in low concentrations. (3–8) Some examples of the most important mycotoxin-related episodes are ergotism, one of the oldest mycotoxicosis killing hundreds of thousands of people in Europe; Alimentary Toxic Aleukia (ATA), responsible for the death of 100,000 Russian people during the Second World War, and finally aflatoxicosis, responsible of killing 100,000 young turkeys (and probably other animals and even humans) in the United Kingdom during the 60’s. (3–6,8,9) This last event opened the door to modern mycotoxicology as it was realized that the mycotoxins were not only a storage problem in grains but actually contaminated certain pre-harvest crops, sensitizing the scientific community to the possibility that other hidden mold metabolites might be deadly, and widening the problem of mycotoxins to a multidisciplinary context. (3–5,9) Hereupon, and even knowing that mycotoxins can kill, there was a more challenging problem at hand: the fact that mycotoxin effects might be evident only years after ingestion, compromising food safety that can only be determined by direct analysis of the toxic compound. (3)

Currently, climate changes, including global warming, are a trendy topic and, in fact, have accelerated the germination/growth/production of mycotoxins. (5–7,10) Regarding sociological aspects, in developing countries, poor food quality control, poor production technologies, and poor

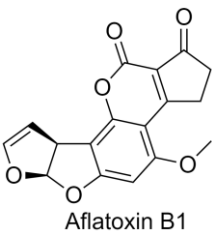
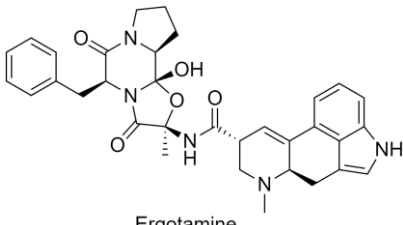
crop storage conditions, beyond the hot climate, contribute to annual agricultural and industrial losses of billions of dollars. Currently, 25% of the world's harvested crops are contaminated by mycotoxins. (6,9,10)

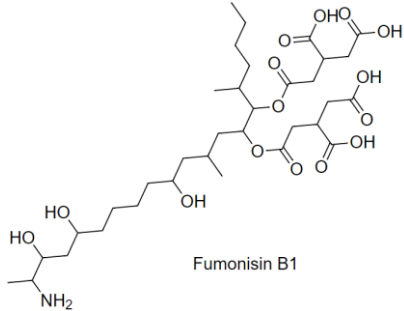
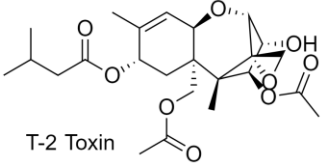
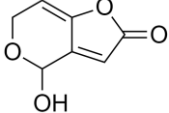
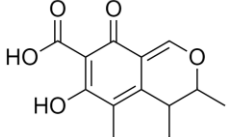
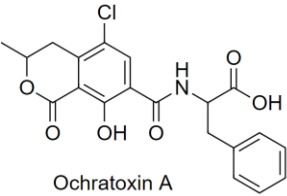
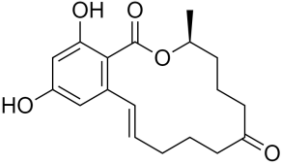
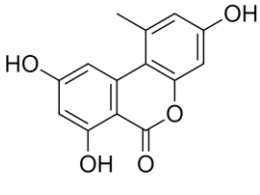
With the growing concern in the scientific community and the increase of studies in this field, the interaction between concomitantly occurring mycotoxins was discovered and until today, the consequence for the toxicity remains a challenge to face. (6) Furthermore, there are two other environmental and socioeconomic points to consider, that contribute to a higher frequency of occurrence of mycotoxin-contaminated foods/feeds. (7,9)

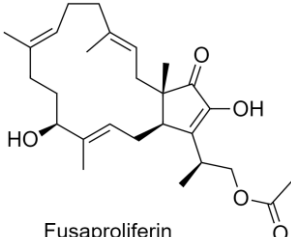
Mycotoxins are characterized by having a low molecular weight, with structures ranging from single heterocyclic rings with molecular weights of scarcely 50 Da, to groups of irregularly arranged 6 or 8 membered rings with total molecular weights greater than 500 Da, constituting a toxically and chemically heterogeneous group. (5–9) Of these biologically active metabolites, some of which have received attention in the scientific community are aflatoxins, deoxynivalenol, citrinin, trichothecenes, fumonisins, zearalenone, T-2 toxin, ochratoxins, patulin and certain ergot alkaloids, due to their sociological and agro-economic impact. (3–7,9,10) Even so, only a few are well characterized and have its effects in human or animals well established, with little information on the interaction between concomitantly occurring mycotoxins and the consequence for toxicity.

Table 1 enumerates some of the most well-known mycotoxin families and their major characteristics with one example for each family.

Table 1. Summary of the most common mycotoxin families and their respective characteristics. The structures were constructed using the online version of ChemDraw.(11)

Family	Examples	Characteristics	References
<b>Aflatoxins</b>	Aflatoxins B1, B2, G1, G2, M1  Aflatoxin B1	Aflatoxins are difuranocoumarins derivatives and consist of a coumarin nucleus to which are attached a difuran moiety in one side and either a pentene ring or a six-membered lactone ring in the other side.	(6,9,12)
<b>Ergot alkaloids</b>	Ergotamine, Ergometrine, Ergocryptine, Ergocristine, Ergocornine, Ergosine  Ergotamine	The common structural feature of ergot alkaloids is the ergoline ring, which is methylated on the N-6 nitrogen atom, substituted on C-8, and possesses a C-8, C-9 or C-10 double bond.	(9), (13)

<b>Fumonisin</b>	<p>Fumonisin A1, A2, B1, B2, B3</p>  <p>Fumonisin B1</p>	<p>Fumonisin contains 20 carbon aliphatic chain with two ester linked hydrophilic side chains. The toxic action of fumonisins is related to the competition with sphingosine in sphingolipid metabolism</p>	(5)
<b>Trichothecenes (Types A and B)</b>	<p>Type A: HT-2 Toxin, T2 Toxin Type B: Deoxynivalenol, Fus-X, Nivalenol</p>  <p>T-2 Toxin</p>	<p>Sesquiterpenoid toxins characterized by a variable number of acetoxy and hydroxyl groups, an epoxide ring at position C12-C13, and a double bond between C9 and C10.</p>	(7,9,14)
<b>Patulin</b>		<p>Heterocyclic lactone</p>	(7)
<b>Citrinin</b>		<p>Benzopyran derivative</p>	(7,15)
<b>Ochratoxins</b>	<p>Ochratoxins A, B, C, TC</p>  <p>Ochratoxin A</p>	<p>Pentaketides consisting of a dihydro- isocoumarin coupled to 8-phenylalanine</p>	(7,16)
<b>Zearalenone</b>		<p>Nonsteroidal estrogen of the resorcylic acid lactone group</p>	(6,7)
<b>Alternaria Toxins</b>	<p>Altenuene, Alternariol, Alternariol methyl ether, Alvertoxin, Tenuazonic acid</p>  <p>Alternariol</p>	<p>There are several structural types: dibenzopyrones (polyketides), perylenequinones, cyclic tetrapeptides, Anthraquinones, amine/amide metabolites and dihydroisocoumarins</p>	(17)

<p><b>Emerging <i>Fusarium</i> Mycotoxins</b></p>	<p>Fusaproliferin, Moliniformin, Beauvericin, NX-2 Toxin, Enniatins</p>  <p>Fusaproliferin</p>	<p>The structure of these emerging mycotoxins can be diverse:</p> <p>bicyclic sesterterpene (fusaproliferin); cyclic hexadepsipeptides (enniainins and beauvericin), organic acid (moniliformin)</p>	<p>(18)</p>
---	---	--	-------------

The significance of mycotoxins is based on their frequency of occurrence and on the severity of the disease (mycotoxicosis) they produce in higher vertebrates. The latter can be as diverse as the chemical structures of the compounds themselves.(4,5,7,8) Mycotoxicosis is manifested in several ways, affecting a wide range of susceptible animal species and its diagnosis can be difficult due to the similar effects produced by other agents, such as pesticides and heavy metals. (9) Also, the symptoms depend on various factors like the type of mycotoxin, age, sex, health, dietary status, and other conditions of the exposed individual. (8,9)

Mycotoxicosis can be categorized as chronic or acute. (5,7–9) Acute toxicity is characterized by a short acting time and an obvious toxic response (e.g.: deterioration of liver and kidney function) , while chronic toxicity is characterized by a low-dose exposure over a long period of time, often resulting in irreversible effects and cancers. (5,8) Furthermore, some mycotoxins can affect DNA replication, producing mutagenic or teratogenic effects. (5,9) Bennet and Klich (9) mentioned an important point: in many studies, it is not easy to interpret data on purported health effects, and regarding human and veterinarian health, mycotoxin contamination is related to chronic exposure, for example, cancer induction, kidney toxicity, immune suppression. (5,7,9) However, the best-known mycotoxin episodes are associated with acute toxicity (e.g.: aflatoxicosis, ergotism, ATA). (9) The aforementioned facts make it necessary the existence of studies demonstrating a dose-response between the mycotoxin and the disease, that are, for human purposes, epidemiological studies conducted by environmental and biological monitoring of food, air, residues, metabolites, and fluids, for example. (9)

Controlling mycotoxin contamination is a difficult task because some fungi are able to produce more than one mycotoxin and some mycotoxins are produced by more than one fungal species. (6,7) In addition, there are several factors that contribute to their growth and production (e.g.: storage, environmental and ecological conditions) and, as said before, the interaction between themselves. (6,7) Also, contamination can occur at any stage of the production process (from before harvesting to storage) which can result in a direct exposure through food consumption or indirect exposure through feed. (6,7,9) Prevention of mycotoxin production and detoxification are the two strategies to control mycotoxin contamination and include several physical, chemical, biological and food processing methods. (6,7,9)

Prevention consists in pre-harvest strategies to avoid the development of fungi and, consequently, mycotoxins. (7,9) These strategies are based on good agricultural, manufacturing,



and storage practices and environmental factors. (7,9) Decontamination/detoxification is a more complicated subject and comprises post-harvest strategies that involve chemical methods, such as oxidation, hydrolysis, absorption, etc., biological methods, with the use of biological agents, and even natural means, such as thermal insulation, radiation treatment and low temperature plasma. (7) However, some of these strategies, in particular chemical and physical methods, are not only ineffective and time-consuming, but also result in nutrient loss. (7) Beyond conventional methods, some novel and interesting approaches were performed that involve the use of nanoparticles and plant extracts. (7)

Establishing the bridge between multivariate chemical data and the ability of models to predict and deal with relevant mycotoxin-related phenomena, such as co-occurrence and molecular recognition, and also the development of improved classification and remediation procedures, is still a challenge.

With the rapid explosion of “big data” in Chemistry, Machine Learning (ML) and big-data analytics tools, as well as high-performance computing techniques, computational chemistry has significantly contributed to the discovery and characterization of new chemical entities, including drugs, reducing the cost and time required to identify lead compounds. (19–21) One of the primary uses of ML in chemical sciences is to aid researchers understanding and exploiting connections between chemical structures and activity, also known as structure-activity relationships (SAR). (19) Current search is increasingly focusing on *in silico* methods for rationalizing and predicting the occurrence and ecotoxicity of micropollutants and their degradation and transformation resulting from biological/chemical processes. (22–28)

Quantitative Structure-Activity Relationships (QSARs) models have been developed for e.g. predicting toxicity of pollutant mixtures, including pesticides, pharmaceuticals and other chemicals (22,29–31), estimating properties of persistent organic pollutants required in the evaluation of their environmental fate and risk (32,33), modeling of (i) degradation of structurally different organic pollutants, including azo dyes, heterocyclic compounds, ionic compounds, among others (27,34,35), (ii) water quality indices of alkylphenol pollutants (36), (iii) cumulative environmental endpoints for the screen, ranks and prioritization of hazardous chemicals in the environment (37), and (iv) reaction rate constants for several organic compounds in water (38), just to name a few examples.

There are two primary techniques for the application of ML: supervised and unsupervised learning. (19–21,39–41) Labels are assigned to the training data in supervised learning, and once trained, the model may predict labels for specific data inputs. (19,20,39) Regression analysis, k-nearest neighbors (kNN), Bayesian probabilistic learning, Support Vector Machines (SVM), random forests (RF), and Neural Networks (NN) are examples of supervised ML methods. (19,20,39) Machine learning approaches that fall on the category of unsupervised learning aim at identifying underlying patterns or intrinsic structures in the absence of knowledge on the outcome. Some examples of unsupervised learning techniques are Principal Component Analysis (PCA),

clustering algorithms and some supervised methods that can also support unsupervised learning. (19,20,39)

Table 2 shows several examples of ML applications in micropollutant-related problems, both supervised and unsupervised techniques.

Table 2. Examples of ML applications to environmental micropollutant problems.

References	Main Goal	Approach	Conclusion
(42)	Estimating micropollutant concentration without resorting to stable isotope labels.	Deep learning and machine learning models (RF, SVM, ANN) replacing stable isotope labels by natural organic matter information	The trained models showed accurate training and validation results for the estimation of five micropollutant concentrations. The study demonstrated potential for an alternative, rapid and economic solution to measure micropollutants.
(43)	Predict O <sub>3</sub> and ·OH exposures and consequently micropollutants abatement during ozonation.	RF algorithm to output oxidant exposures from water quality input variables.	The developed models showed useful to predict the abatement of micropollutants in drinking water and wastewater ozonation processes and to optimize the O <sub>3</sub> dose for remediation procedures. Using higher-resolution fluorescence data as input variable resulted in more accurate predictions.
(44)	Aquatic Toxicity Prediction for Polar and Nonpolar Narcotic Pollutants.	SVM algorithm for the classification of 190 narcotic pollutants into polar or nonpolar using molecular descriptors	The study demonstrated a possible application for the identification of the aquatic toxicity mechanism.
(45)	QSAR models to predict bioconcentration factors and median lethal concentrations.	Recursive Feature elimination (RFE) and SVM using 2D molecular descriptors of 450 diverse chemicals. Three ensemble models were constructed using three ML algorithms.	The ensemble-SVM model proved to be more stable with more accurate predictions when compared to the other models. This study allowed to identify important structures to aquatic toxicity with relevant information to future aquatic toxicology experiments.

Table 2 shows some ML techniques that have been applied to modelling micropollutants' behavior in different environmental matrices. However, to our knowledge, the use of ML to establish a computational framework for dealing with multidimensional data related to mycotoxins is very limited. (46)

The first study where ML was applied to mycotoxins was conducted by Torelli and his coworkers (47) in a two year study where an artificial neural network model was developed to predict fumonisins, deoxynivalenol and zearalenone contamination of maize at the harvest time, using seven cropping system variables (Food and Agriculture Organization class, sowing and harvest dates, crop duration, kernel moisture, European corn borer treatment and irrigation). These authors showed for the first time the potential of ML in the study of mycotoxins, emphasizing the importance of new approaches for rapid cataloging of grain lots. (47) More recently, Leggieri et al. (46) recorded the occurrence of aflatoxin B1 and fumonisins in maize fields and collected the corresponding cropping data over the years 2005-2018 in northern Italy. The authors built two deep neural network models to predict, at harvest, which maize fields were contaminated with those mycotoxins, obtaining two robust models and better results when compared to AFLA-maize and FER-maize. (46) Note that the aforementioned models (AFLA-maize and FER-maize) both use meteorological data as input variables to predict the risk of contamination of aflatoxin B1 and fumonisins above legal limits, but these models do not use a ML approach, being considered mechanistic models. (46,48)

The application of ML to mycotoxins-related phenomena, has been focused on field contamination, based on meteorological, environmental, agricultural data that reflects the growth and germination of fungi and therefore, the production of mycotoxins. (46,47)

In this study, supervised and unsupervised learning techniques were used to investigate the chemical structures of mycotoxins and infer possible relationships between their structure, family, and toxicity. Hierarchical Cluster Analysis (HCA) and k-means clustering were used for molecular structure analysis aiming at understanding the molecular similarity between the chosen mycotoxins and the respective families, using their molecular fingerprints so that relevant properties can be predicted through the clustering map, and relevant chemical patterns within and between mycotoxins families can be identified. PCA was used to obtain an overview of the relative positioning of the chemical structures using several molecular descriptors, by summarizing the respective variation into a reduced number of Principal Components (PCs), aiming to build a classification model to predict mycotoxins acute toxicity. Finally, supervised learning techniques including linear discriminant analysis (LDA), RFs, SVMs and NNs were applied to predict mycotoxins' toxicity into acutely toxic or non-acutely toxic.



## Chapter 2 – Theory and Methods

---

The scientific study of algorithms and statistical models that computer systems employ to execute a given task without being explicitly programmed is known as ML (19–21,39,40) This is attempted by deploying algorithms that aim at learning the rules that underlie a dataset through the assessment of a portion of data and subsequent model fitting for predictions, process often designated by model training. (39) ML is one of the most important and fast evolving topics in current science, with applications that range in context and methodology. (19,20,39,40) The major challenge of ML is the lack of interpretability and repeatability of ML-generated results, which later result in reproducibility problems. (39,49) Numerous methods exist within each technique and those methods may differ in prediction accuracy, training speed, or number of variables they can handle, which makes it necessary to ensure that the algorithms are appropriate for the task and type of data available. (20) ML approaches are usually more effective when applied to large amounts of data but is also important to verify its quality to ensure maximal effectiveness of the models. (20,21,39,49)

Two type of datasets should always be involved in a model construction: the training and test sets, usually formed by randomly splitting the original dataset. (49) The training set is used for the model to learn from for certain hyperparameters and the test set is used to assess the final performance of the model. (49) The key test for the effectiveness of a ML model is, in fact, the successful application to unseen data, in this case, the test set. (39,49) The test set is provided to the model once the training is complete and compared with the resulting predicted outputs. (39,49) A validation set can be used in cross-validation procedures that are very useful but are also only reliable when both training and test sets are representative of the whole dataset, which may be pose problem when the dataset is small. (39)

There is a large plethora of examples of chemical databases, including ZINC15 (50), PubChem (51), ChEMBL (52) and DrugBank (53), that offer scientists a wide variety of chemically and biologically relevant data which is extremely useful for ML applications. (19,39,54) In the present study, all mycotoxins 3D structures were downloaded from PubChem (51), the highest profile online database, in Structure-Data File (SDF) format. (55) SDF files are format files that handle a list of molecular structures with associated data between databases such as structure identifiers, experimental/physical properties, pricing information, among others. (55)

The software for performing all the analysis was constructed in R (version 3.6.3) and Python (version 3.8) programming languages which provide several tools for a variety of problems related to statistical analysis in a variety of fields, and their respective Integrating Development Environments (IDE), RStudio and Spider. (54,56) Regarding chemistry, R and Python provide a wide range of tools for statistical modeling of chemical information allowing to read, represent,

manipulate, and analyze chemical structures. (54,56) Some examples of cheminformatics tools are the Open Eye for Chemical Information Processing (OEChem), the Java based cheminformatics/computational chemistry packages (JOELib) and Chemistry development Kit (CDK). (54,56) The latter was used in this study and is an open-source Java framework for structural chemo and bioinformatics. (57) It possesses an interface with statistical package, *rJava* of R (58), necessary for the installation of the package *rcdk*, used in the present work. (56,57,59) The *rcdk* package (60) allowed us to analyze chemical data and structures in several formats, calculate molecular descriptors, identify and evaluate relevant molecular fingerprints. (56,57)

The *Cluster* package from R (61) was used to perform HCA and k-means, in order to analyze the chemical structures, and to understand the molecular similarity between mycotoxins using their molecular fingerprints, so that relevant properties can be predicted through the clustering map. The clustering results were visualized using the *Factoextra* package (62). PCA was used to obtain an overview of the relative positioning of the chemical structures by summarizing the respective variation into a reduced number of PCs, aiming at building a model for classifying new structures and identifying target properties, based on mycotoxins' molecular descriptors. PCA resulting biplots were constructed using the *FactomineR* package from R (63). R was also used to build LDA and SVM models. Molecular descriptors (features) were split into training set and test set using the *caret* package (64) and the models were build using *MASS* (65) and *caret* (64) packages, for LDA and SVM, respectively.

Python was used to construct RF and NN models and to perform a gridsearch (only for the NN model), so that several parameters that can affect the performance of the NN model could be tested. The RF and NN models were trained using the Python package "Scikit Learn" (version 0.23.1): the molecular descriptors were subjected to normalization using *StandardScaler* and the RF was performed using *RandomForestClassifier* package, while NN was performed using *MLPClassifier*. (66) The gridsearch for parameters optimization is described in detail in section 2.8 for the NN model.

## 2.1 Unsupervised Learning

Dimensionality reduction (PCA) and clustering algorithms (HCA and k-means) belong to unsupervised learning algorithms where algorithms discover and present the relevant patterns in the data, for example, through similarity measures. (19–21,40,67,68) In this type of algorithms, there are no corresponding labels (they can be introduced *a posteriori*) and hidden patterns or intrinsic rules are identified in the input data and clustered in meaningful ways. (19–21,67,68) For all the above-mentioned algorithms, spatial descriptions of the mycotoxins structures as points in the Euclidean space are required.

Clustering algorithms can be of several types, including partitioning, hierarchical, density-based and grid-based. (41,68) In this work, HCA and partitioning (k-means) algorithms were deemed sufficient. Hierarchical clustering algorithms divides a dataset by creating a hierarchy of

clusters, whereas partitioning clustering algorithms splits the data points into  $k$  partitions, where each partition represents a cluster. (41) The clustering process has fundamentally two steps: choosing the metric to evaluate whether two items are similar or not, and adopting a technique for the clusters to be formed. (67,68) Distance measures are used to evaluate the similarity/dissimilarity of the objects and vary according of the type of data in question, e.g., numeric, binary, nominal data. (67) For a function  $d$  to be considered a distance, some conditions need to be satisfied for the objects  $i, j$  and  $k$ :

$$1. \quad d(i, j) = d(j, i) \quad \forall x_i, x_j \in \mathcal{S} \quad (\text{Equation 1})$$

$$2. \quad d(i, j) \geq 0 \quad \forall x_i, x_j \in \mathcal{S} \quad (\text{Equation 2})$$

$$3. \quad d(i, j) = 0, \text{ if } i = j \quad \forall x_i, x_j \in \mathcal{S} \quad (\text{Equation 3})$$

$$4. \quad d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k) \quad \forall x_i, x_j, x_k \in \mathcal{S} \quad (\text{Equation 4})$$

The first condition tells that the dissimilarity matrix is symmetric in relation to the diagonal (Eq. 1), the second condition implies that all elements of the dissimilarity matrix are positive (Eq. 2), the third condition states that the diagonal of the dissimilarity matrix is composed by zeros (Eq. 3). The fourth condition is necessary for the distance measure to be considered a distance metric and refers to triangle inequality (Eq.4 ). (67)

There are several distance measures available in the *cluster* package (61) from R that can be used as dissimilarity/similarity measure, such as Euclidean, squared Euclidean, Manhattan, Cosine, Spearman, Minkowsli, Tanimoto, among others. (67) The distance measure used in this work was the Tanimoto distance (Eq.5) despite some work was done with the Euclidean Distance (data not shown) (Eq. 6).

$$Tc = \frac{c}{[a+b-c]} \quad (\text{Equation 5})$$

where  $Tc$  is the similarity,  $a$  and  $b$  are the number of “on” bits in molecules A and B and  $c$  is the number of “on” bits that both molecules have in common, resulting in a complete set of distances, in matrix form, for molecules A and B. (20,69–71)  $Tc$  values over a certain threshold (usually 0.85) (69) show that two compounds are similar, but do not provide information such as which chemical groups they share, for example. (19)

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{in} - x_{jn}|^2} \quad (\text{Equation 6})$$

### 2.1.1 Hierarchical Cluster Analysis

HCA is a clustering approach that uses agglomerative clustering, which merges smaller clusters into a larger one, opposed to divisive clustering, characterized by dividing a big cluster into smaller ones, to create a hierarchy of clusters. (19,41) This method uses the distance matrix criteria to cluster the data. (41)

Hierarchical clustering can be performed with different definitions of the distance between clusters, including single-linkage, complete-linkage, average-linkage, and Ward's linkage. (67,68,72) Single-linkage considers the shortest distance between each member of one cluster and any member of the other cluster to represent the distance between two clusters; complete-linkage is similar to single-linkage but considers the distance between two clusters as the longest distance between each member of one cluster and any member of the other cluster and average-linkage uses the average distance from any member of one cluster to any member of the other cluster as the distance between two clusters. (67) Ward's linkage identifies the two clusters,  $C_A$  and  $C_B$ , with sized  $n_A$  and  $n_B$ , that promote the minimum distance,  $d_{A,B}$ , between the merged clusters' centroids,  $I_A$  and  $I_B$ . (68)

$$d_{A,B} = \frac{n_A n_B}{n_A + n_B} (\mu_A - \mu_B)' (\mu_A - \mu_B) \quad (\text{Equation 7})$$

Ward's linkage was used over the Tanimoto distances (Eq. 5) between the mycotoxin structures and was chosen because is the only one among the agglomerative clustering methods that is based on sum-of-squares criterion, producing groups that minimize within-group dispersion. (68,72)

### 2.1.2 K-means Clustering

K-means is a simple unsupervised learning method that gathers the given input data through a certain number of clusters defined by  $k$  centers, that are the number of desired clusters. (40,41,73,74) The fundamental steps of the k-means algorithm are: (i) selecting arbitrarily a centroid; (ii) assigning each object to the cluster with the closest centroid, according to the Euclidean distance between them; (iii) calculating the centroid as the mean of the objects assigned to it and (iv) repeat steps 2 and 3 until no changes are visible. (41,67,75) In the end, each point of a dataset is grouped with the nearest center until no point is pending, minimizing within-group distances to the centroid. (19,40) This algorithm was chosen because is easy to interpret, simple to implement and converges fast. (67) The main problem of this algorithm is to depend heavily on the initial conditions. (75)

### 2.1.3 Principal Component Analysis

PCA is one of the oldest and most widely used dimensionality reduction technique that increases interpretability of datasets and at the same time minimizes information loss. (40,76) This



statistical method uses an orthogonal transformation to convert a collection of observations of possibly correlated variables into a set of linearly uncorrelated variables called PCs. (19,40,68,73,76) PCA can be performed based on the covariance matrix, or the correlation matrix, depending on how the variables have different units of measurement or vary along different orders of magnitude. (67,68,76) With variables with different units of measurement, covariance matrix may not be adequate because the PCs are sensitive to their order of magnitude. (67,68,76) To work around this problem, it is common to standardize the variables, centering and dividing for the standard deviation each data value, thus producing the correlation matrix. (68,76)

Considering a dataset with observations on  $p$  numerical variables, for  $n$  objects, these data values define a  $n \times p$  data matrix whose column is the vector  $x_j$  of observations of the  $j^{\text{th}}$  variable:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \quad (\text{Equation 8})$$

PCA seeks for linear combinations given by  $\sum_{j=1}^p a_j x_j$  where  $a$  is a vector of constants. The variance of such linear combination is given by  $\text{var}(Xa) = a' S a$  where  $S$  is the covariance matrix associated with the dataset and while ' denotes the respective transposed. (76) The result is a rotation of the orthogonal system of axes associated with the original variables, obtaining a new data system,  $Y$  matrix, where  $Y$  matrix columns are the new variables, the PCs. (76) These PCs are linear combinations of the  $p$  variables of  $X$  matrix, where  $j=1, \dots, p$  and  $a_{ij}$  ( $i=1, \dots, p; j=1, \dots, p$ ) are constants,

$$Y_i = a_{1j}X_1 + a_{2j}X_2 + \dots + a_{pj}X_p \quad (\text{Equation 9})$$

In this case,  $a_{1j} \dots a_{pj}$  are the *loadings*, i.e. weights of the original variables on the linear combination. (68) The coefficients of these linear combinations are determined such that the following conditions are satisfied:

$$1. \text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \dots \geq \text{Var}(Y_p) \quad (\text{Equation 10})$$

$$2. \text{Corr}(Y_i, Y_j) = 0, \forall_{ij} \quad (\text{Equation 11})$$

$$3. Y_i : a_{1j}^2 + a_{2j}^2 + \dots + a_{pj}^2 = 1 \quad (\text{Equation 12})$$

In this case,  $Y_1$  is the PC with highest variance,  $Y_2$  has the second highest variance, and so on. There are as many PCs as the number of original variables and these PCs are not correlated to each other (Eq. 11) and collectively explain the data entire variance (Eq. 12). (68,76)

A single value decomposition is used to obtain the transformation matrix  $W$ , whose elements are the *loadings* and the vector  $\lambda$  composed by the recovered variance  $\lambda_i$ , or eigenvalues, in each  $i^{\text{th}}$  main component:

$$C_x W = \lambda W \quad (\text{Equation 13})$$

Here,  $Var(Y_i) = \lambda_i$  and  $\sum_i^p \lambda_i$  gives the total variance of the data. When dealing with the correlation matrix, i.e., with normalized data,  $\sum_i^p \lambda_i$  is equal to the number of variables.

The next step of PCA is to evaluate how many PCs will be considered. For the covariance matrix, Pearson's criteria is the most popular approach to decide how many components should be retained and consists in selecting the PCs necessary to obtain 80% of the total variability. (68)

$$\frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^p \lambda_i} \geq 0.8 \quad (\text{Equation 14})$$

For the correlation matrix, the most common criterion is to retain  $p$  components for which  $\lambda_i \geq 1$  because if the  $\sum_i^p \lambda_i$  is equal to the number of variables, the corresponding mean would be 1. (68) For that reason,  $\lambda_i < 1$  are discarded and  $\lambda_i \geq 1$  are accepted. There is a third criterion that consists in plotting the eigenvalues,  $\lambda_i$ , against their ordinal numbers,  $i$ , and the number of PCs that should be retained is given by a break or a leveling of the slope of the plotted line. (77)

In this work, molecular descriptors were calculated, normalized and thus, the correlation matrix was used. PCA is represented by a biplot that shows both the *loadings* and the *scores* where the *loadings* are the weight of the original variables in the PCs (arrows in the biplots) and the *scores* are the objects in the new system of PCs (points in the biplots).

## 2.2 Supervised Learning

Supervised learning algorithms are characterized by having labels assigned to the training data and the models predict previously unknown values of data categories or continuous variables, for classification or regression problems, respectively. (19–21,39,40,73) The model choice is not an easy task, and its complexity does not always reflect better results. (40) Because of that, the chosen models for this work range from conventional algorithms such as LDA and SVM, to ensemble methods like RF and more complex algorithms such as DNN.

All supervised learning models were used for classification analysis with a ratio training/test of 0.7/0.3. The following sections will explain each model in detail.

### 2.2.1 Linear Discriminant Analysis

LDA technique (Figure 1) is a discriminating linear parametric approach that focuses on determining the optimal boundaries between classes by choosing the directions that provide the greatest separation between them. (68,73) Essentially, it finds the vectors in the variables space that best discriminate classes by building a linear combination from a set of independent variables that characterize the data to yield the greatest mean differences between the intended classes. (68) the between-class scatter matrix and the within-class scatter matrix are used to accomplish this. (68)

For all samples of all classes, the between-class scatter matrix,  $C_B$ , and the within-class scatter matrix,  $C_W$ , are given by:

$$C_B = \sum_{i=1}^c M_i \cdot (\mu_i - \mu) \times (\mu_i - \mu)^T \quad (\text{Equation 15})$$

$$C_W = \sum_{i=1}^c \sum_{x_k \in X_i} (x_k - \mu_i) \cdot (x_k - \mu_i)^T \quad (\text{Equation 16})$$

where  $M_i$  is the number of objects in class  $i$ ,  $c$  is the number of different classes,  $\mu_i$  is the mean vector of objects that belong to class  $i$  with  $x_k$  being the  $k$ th variable of that class.  $C_B$  represents the scatter of objects around the mean of each class and  $C_W$  represents the scatter of objects around the overall mean for all classes. (68) The goal of LDA is to maximize  $C_B$  and minimize  $C_W$  by projecting the coordinates of an object along a line, derived from the decision rule, and assign it to the class with the nearest center of mass. (68)

The major problem of LDA is to assume linearity of the border that discriminates between two classes in the space of attributes which is often hard to justify such functional form in advance. (67) LDA was built in R programming language using the *lda* function from the *MASS* package (65).

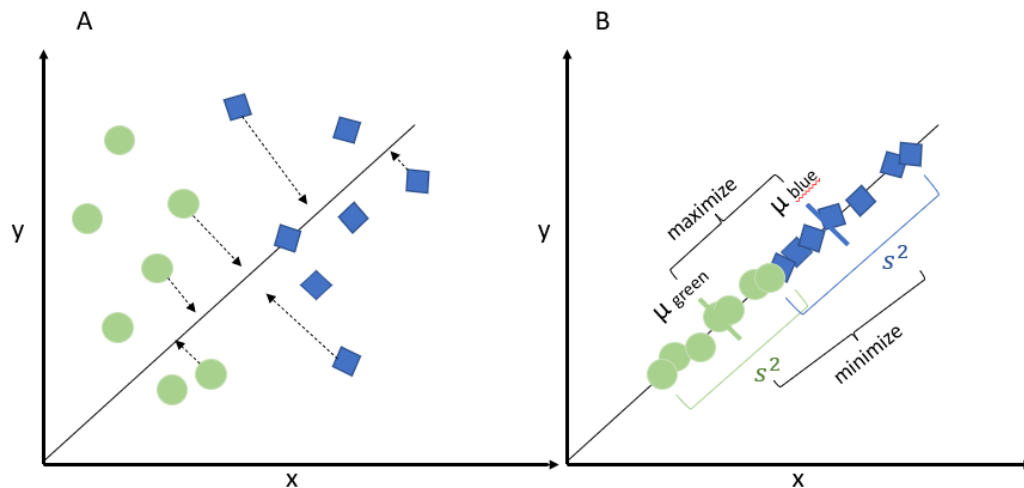


Figure 1. Exemplification of the separation between two classes of objects by LDA. Data is projected in a new axis to maximize the separation between the 2 categories (A) by maximizing the distance between means and minimizing variation (B).

### 2.2.2 Random Forest

RF was created by Breiman (78) and is an ensemble model, meaning that a prediction is generated by combining the output of several individual classifiers. (79) Essentially, multiple decision trees (DTs) are created from the training data, considering random subsets of available variables, and a majority-voting scheme is used to make classification or regression predictions for new inputs (Figure 2). (19,79) DTs are classification trees that arrange instances based on feature values in which each branch indicates a value that the node might adopt, and each node represents

a feature in an instance to be categorized. (74) Those trees are made by pulling a subset of training samples and replacing them with new ones, known as the bagging approach. (80) The model was constructed using the *RandomForestClassifier* function from “Scikit Learn” (66) and contained 100 trees.

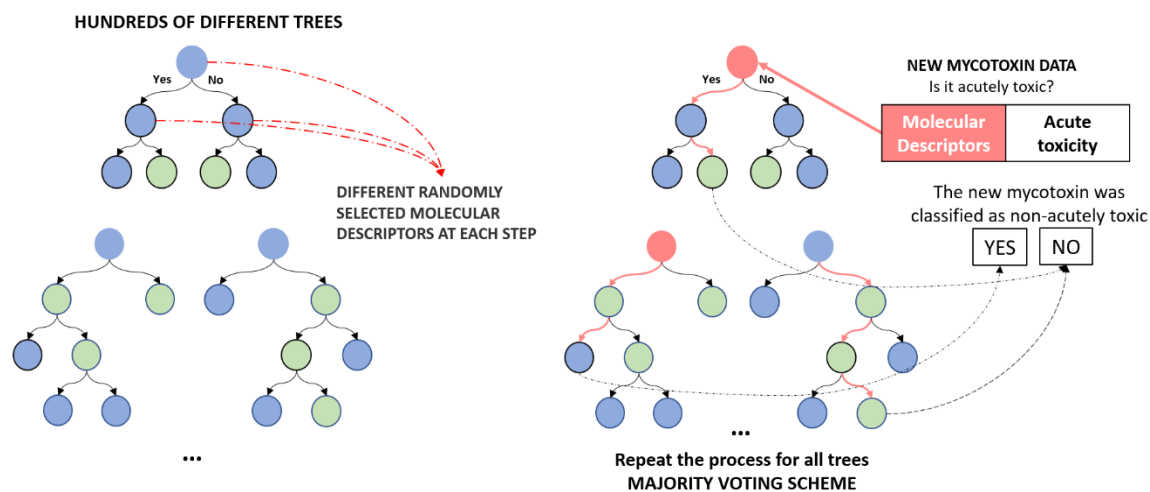


Figure 2. Schematic representation of the RF procedure. On the left are 3 DTs through which an object (mycotoxin) is subjected (on the right), only progressing if given parameters are met. The instance will belong to different classes depending on the path taken.

### 2.2.3 Support Vector Machines

This method was developed by Vapnik (81) and can be used for both regression and classification, being one of the most widely used supervised ML algorithms. (40,73,79) The fundamental idea behind support vector classification is to identify a  $n$ -dimensional hyperplane, where  $n$  is the number of features, that separates instances of two different classes of objects by the greatest amount of distance feasible. (19,40,74,79) Basically, SVM draws margins that maximize the separation between the margin and the classes, or support vectors, consequently minimizing the error (see Figure 3). (40,74)

The strength of the technique comes from the fact that during the training process, dot products are used only to refer to the input patterns which allows projecting data from a low-dimensional space to a space of higher dimension. (79,82) Kernel functions are mathematical tricks that compute the dot product of two vectors in a high-dimensional feature space without mapping the vectors to that space directly. (79) The four types of kernel functions that are typically employed in SVMs are linear, polynomial, radial basis function (RBF), and sigmoid kernels. (82,83) In this work, the *caret* package (64) was used to construct SVM models with the “svmLinear” kernel.

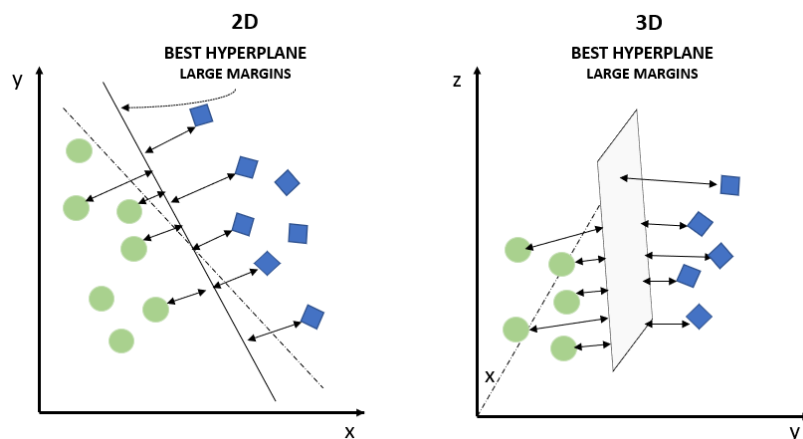


Figure 3. Scheme illustrating the SVM basis. The instances are divided by a hyperplane that separates them by the greatest amount of distance feasible.

### 2.2.4 Neural Networks

The analysis and prediction of compound properties is one of the areas that is significantly impacted by the neural network methods. (19,69) Artificial Neural Networks (ANN) and Deep Neural Networks (DNN), illustrated in Figure 4, are algorithms that endeavor to recognize intrinsic relationships in a dataset by replicating brain's activity arranging artificial neurons into input, output and hidden layers. (19,39,40,84) The hidden layers take input signals (analogous do dendrites) from other neurons, the hidden layers process and integrates and the output layer sends the resulting output. (39,40) The major difference between ANN and DNN is the depth of the network architecture. The model is considered DNN if it displays more than one hidden layer. (19,20,39,84)

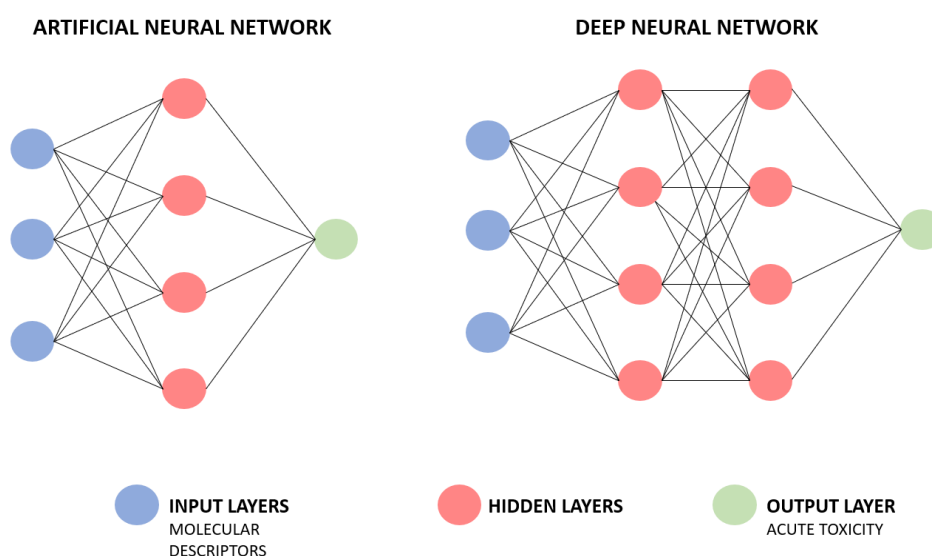


Figure 4. Representation of a shallow neural network (on the left) with only one hidden layer, and a deep neural network (on the right) with two hidden layers.

The input and activation functions of the nodes, network architecture, and the weight of each input connection are all important components of the ANN. (74) However, the activation function and the network architecture are fixed parameters, meaning that the behavior of the ANN depends fundamentally on the weights and biases. (74) The weights and biases of the ANN are usually set to random values at first, and then instances from the training set are regularly ran through the network to the net. The values for an instance's input are placed on the input units, and the net's output is compared to the instance's desired output. (74) In a NN, an activation function is what allows the network to learn complex patterns in the data and is used to get the output of a particular layer and supply it as the input for the next layer. (85) Activation functions can be linear or non-linear and some examples are sigmoid (range between 1 and 0), hyperbolic (range from -1 to 1), and the most popular Rectified Linear Unit (ReLU) activation function that range from 0 to infinity. (85) In this work, the used activation function was ReLU. In layered NN architectures, network size is determined not only by the number of layers, but also by the number of nodes in each layer and the number of connections between them. (86) There might be an unlimited number of network architectures useful to learn the properties of a dataset but is not an easy topic because the optimal architecture can only be obtained experimentally. The next topic (see Grid Search Section) shows the several network structures used, regarding the initial learning rate, maximum number of iterations, the number of layers and nodes in each layer and the optimizers; the ANNs trained were fully connected neural networks, meaning that all the nodes of on layer were connected to each node of the next layer. *MLPClassifier* function from “Scikit Learn” (66) was used to build the NN models.

#### 2.2.4.1 Grid Search

Grid Search (GS) is a technique that exhaustively scans through a series of parameters specific for an algorithm, in order to devise the best set of parameters to optimally perform a given task. ML makes use of GS to achieve the best possible model fit of a predictor on a given dataset. Commonly, the use of GS includes cross-validation upon the training dataset, in order to ensure that the parameters were adequately tuned, while still avoiding model bias. However useful, in large datasets, the deployment of GS can be problematic, as it will multiply the run time of the initial algorithm by the amount of parameter combinations inputted. For smaller datasets, on the other hand, it is still a very reliable method for optimizing the performance of single predictors. (66) This was used to correctly perform parameter optimization for the NN model, particularly to evaluate its best architecture. The tested parameters were the hidden layer sizes, the maximum number of iterations, the initial learning rate and the optimizer of the model. Table 3 shows the parameters subjected to GS with a total of 336 combinations. The optimizer is responsible for updating weights to minimize loss and the initial learning rate controls the step-size in updating those weights. (66) The loss function (difference between the network output and its expected output) is often used to evaluate the neural network performance and is frequently calculated as a gradient trough the very

popular *backpropagation algorithm*. (84) Some examples, all subjected to GS, are LBFGS, Stochastic Gradient Descent (SGD) and ADAM. The *Limited-Memory Broyden-Fletcher-Goldfarb-Shanno* (LBFGS) is derived from the BFGS algorithm that works to iteratively compute a matrix  $M_t$  that approximates the inverse of the true Hessian of the objective function. (84) The main difference between LBFGS and BFGS is that LBFGS approximates the BFGS algorithm while requiring less computer memory by storing the  $m$  last updates instead of storing the full matrix approximation. (84) Gradient descent by itself finds the minimum of a sum of functions by moving in the opposite direction of the gradient. (84) SGD differs from gradient descent in that this method would then calculate the gradient of each of the functions in the sum and evaluate all of these functions at the current point. (84) ADAM algorithm computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients. (87)

Table 3. GS combination parameters for the NN model.

Architecture	Initial Learning Rate	Maximum number of iterations	Optimizer
(3,),(5,)	0.01	250,	lbfgs
(3,3),(3,5),	0.001	300,	sgd
(5,5),	0.0005	350,	adam
(3,3,3),(5,5,5)	0.00025	400	

## 2.3 Evaluation of the models

### 2.3.1 Unsupervised

Cluster validation is used to evaluate the quality of clustering algorithm results and can be categorized in three classes: internal, external or relative. (88) Internal cluster validation evaluates the quality of a clustering procedure using internal information of the clustering without referring to external information (89), external cluster validation consists in comparing the cluster analysis results to an externally known result (88) and relative cluster validation consists in varying different parameter values for the same algorithm to evaluate the clustering procedure. (90)

In this study, to validate de k-means clustering results, the silhouette coefficient was used. The silhouette coefficient is an internal validation measure that reflects the compactness, connectedness, and separation of cluster partitions by measuring how close are the objects within the same cluster, how separated are different clusters and to what extent objects are placed in the same cluster as their nearest neighbors. (88) For an object  $i$ , the silhouette width  $S_i$  is calculated as follows:

1. For each object  $i$ , the average dissimilarity  $a_i$  between  $i$  and all other objects of the cluster in which  $i$  belongs is calculated.

2. For all other clusters  $C$  to which  $i$  does not belong, the average dissimilarity  $d(i,C)$  of  $i$  to all objects of  $C$  is calculated, where the smallest  $d(i,C)$  is defined as  $b_i$ .

3. Calculate the silhouette width defined as:

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \quad (\text{Equation 17})$$

Objects with a  $S_i$  close to 1 are very well clustered, a  $S_i$  close to 0 means that the object lies between to clusters and a negative  $S_i$  suggests that the object was placed in the wrong cluster. (88)

### 2.3.2 Supervised

A robust model must be balanced between overfitting and underfitting. (21,49) Overfitting occurs when the model becomes too complex and tends to follow small variations in the data too closely becoming susceptible to picking up random noise (21) This results in the accuracy of the training set constantly improving but in the test set reaching a plateau or declining. (20,39) In contrast, underfitting occurs when the model is highly biased and basically does not learn, for example because data are insufficiently detailed to allow the discovery of suitable rules, being unable to model the training or test data. (20,39)

In ML, performance evaluation is critical for comparing models and selecting the one that best suits the problem. Confusion matrices are often used to measure ML errors, encoding type I (false positives) and type II errors (false negatives) for each class of information to extract. (91) These matrices can be used to: (i) inspect errors for each class; (ii) detect thresholds; (iii) compare software results. (91) Table 4 exemplifies a confusion matrix where TN is True Negative, TP is True Positive, FN is False Negative, and FP is False positive.

Table 4. Example of a Confusion Matrix.

<b>Predicted</b>	<b>No</b>	<b>Yes</b>
<b>Actual</b>		
<b>No</b>	TN	FP
<b>Yes</b>	FN	TP

Confusion matrices are often created by deriving advanced metrics from basic metrics, which are essentially rates of right and wrong classifications divided by the total number of objects to detect or reject. (91) Some of those rates and the most widely used are: accuracy, true positive rate or sensitivity, true negative rate or specificity, positive predicted value or precision and negative predicted value (Equations 18-22).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (\text{Equation 18})$$

$$TPR \text{ or } Sensitivity = \frac{TP}{TP+FN} \quad (\text{Equation 19})$$



$$TNR \text{ or Specificity} = \frac{TN}{FP+TN} \quad (\text{Equation 20})$$

$$Precision = \frac{TP}{TP+FP} \quad (\text{Equation 21})$$

$$Negative Predicted Value = \frac{FP}{FP+TN} \quad (\text{Equation 22})$$

Another common metric is AUROC, Area Under the Receiving Operating Curve, that is often visualized by plotting together pairs of advanced metrics, in this case, sensitivity and specificity. (91) AUROC is defined mathematically as the probability of a classifier to rank a true instance, TN or TP, higher than a randomly chosen negative one. This metric is calculated as

$$AUROC = \int_{-\infty}^{\infty} TPR(T)(-FPR'(T))dT \quad (\text{Equation 23})$$

For LDA and SVM models, constructed in R, the confusion matrix was generated using the *caret* package (64) and the AUROC curve was plotted using *pROC* package. (92) For RF and NN, performed in Python, *scikit-learn* (66) was used to calculate those metrics, inclusively AUROC. (66)



## Chapter 3 - Database Description and Data Processing

---

The format in which data is presented has significant impact on processing and learning, being featurization the process of converting raw data to something more appropriate for an algorithm. (39) During featurization, missing or spurious elements are identified and handled, which sometimes requires insights about both the scientific and learning problem and is essential to avoid ML algorithms of being misled. (39,49) This is a systematically discussed topic since there is no obvious choice of representation for the best performance. For example, in chemistry, data is not always fully assimilated or give a coherent description of structures. (39,49) The truth is that most data sources are biased (49), and this work did not escape that rule. In this context, it was not possible to use mycotoxins' experimental toxicity values reported in literature to develop the aforementioned methods not only because some studies were old, not clear, and incomplete but also because there are few mycotoxins well studied, resulting in a lot of missing points in the target variable, the acute toxicity. Detailed information about this target variable is provided on section 3.3 "Class Attribution". In this work, the dataset is composed by molecular descriptors (features, columns) and mycotoxins (objects, rows).

### 3.1 Molecular Fingerprints

Molecular fingerprints are high-dimensional vectors made up of chemical descriptor values and are widely employed in chemometric analysis and similarity-based virtual screening applications. (19,70,71) Specifically, these descriptors are presented in the form of bit string representations, consisting in numbers of bits that represent the presence or absence of a specific molecular feature and depend on the type and number of molecular descriptors and the values they capture. (69,70) Molecular ACCess System (MACCS) substructure fingerprints are 2D binary fingerprints (0 and 1) that indicate the presence or absence of certain substructure keys with each of 166 bits. (19,69) Chemical patterns up to a specified length or diameter can be extracted from a chemical graph using Daylight fingerprints and extended connectivity fingerprints (ECEP), indexing features resorting to hash functions. (19,20,69) Hash functions, in the context of molecular fingerprints, take early types of bit vector representations with associated structural features and represent that structural information in a way that promotes virtual screening by mapping data of arbitrary size to fixed-size values. (93)

Several metrics and similarity coefficients were developed to compare fingerprint representations (e.g.: Cosine similarity, Soergel distance, Manhattan distance, among others) and

the most widely used metric is the Tanimoto Coefficient (Equation 5) or Jaccard Index, used in the present work.

The technique by which the molecular representation is converted into a bit string determines the type of molecular fingerprint. (70) The majority of techniques employ a 2D molecular graph and are hence referred to as 2D fingerprints; however, certain methods, such as pharmacophore fingerprinting, may store 3D information. (19,70) From those techniques, the major ones are substructure keys-based fingerprints, topological or path-based fingerprints and circular fingerprints. (70)

Substructure keys-based fingerprints set the bits of the bit string based on the existence of specific substructures or characteristics from a list of structural keys in the compound, meaning that if a substructure is not present on that structural keys list, the features will not be represented. (70) Some examples are MAACS fingerprints, PubChem fingerprints, BCI fingerprints and Typed Graph Triangle/Typed Graph Distance (TGT/TGD) fingerprints. (94–96) Figure 5 illustrates one example of substructure keys-based fingerprints.

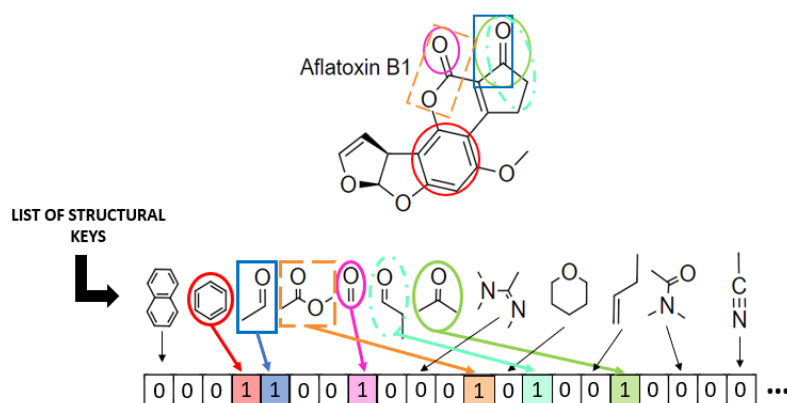


Figure 5. Representation of a hypothetical 21-bit substructure fingerprint, with 6 bits set as a result of the substructures they represent are present in the molecule.

Topological or path-based fingerprints are created by examining all of the molecule fragments that follow a (typically linear) path up to a specific number of bonds, and then hashing each of these paths to generate the fingerprint, meaning that any molecule can produce a meaningful fingerprint and its length can be adjusted. (70) These fingerprints can also be called hashed fingerprints, in which a given bit may be set by more than one feature, the so called “bit collision”. (70) Topological fingerprints were used in this study and are exemplified in Figure 6.

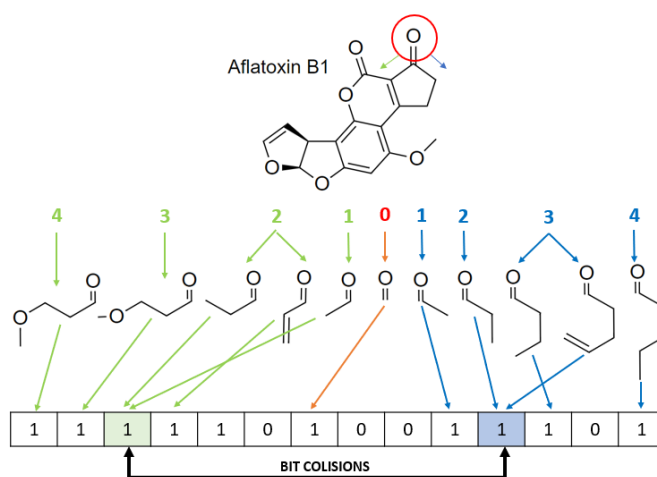


Figure 6. Representation of a hypothetical 14-bit topological fingerprint. Fragments found from the starting atom (circled in red) are shown with their respective length and corresponding bit. There are two-bit collisions, which are bits that are set by more than one fragment.

Finally, circular fingerprints, exemplified in Figure 7, are also hashed fingerprints but instead of looking for pathways in a molecule, a radius is set and the environment of each atom is recorded. (70)

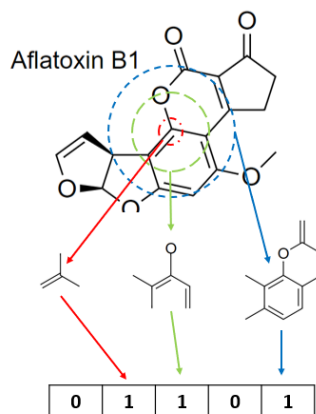


Figure 7. Representation of a hypothetical 5-bit circular fingerprint.

### 3.2 Molecular Descriptors

Molecular descriptors are used for molecular data mining and consist of numerical features extracted from molecular structures. (19,69) These descriptors may have different complexities but can be classified according to their dimensionality, depending on the molecular representations from which they are calculated. (19,69,97) One-dimensional (1D) descriptors include bulk properties and physiochemical parameters such as atom, bond or fragment counts, molecular weight and sum of atomic properties. (19,69) The most common descriptor type described in the literature are two dimensional (2D) molecular descriptors and comprise molecular profiles, topological indices and autocorrelation descriptors. (19,69) Selecting an appropriate set of molecular

descriptors is nontrivial since hundreds of molecular descriptors are available in the literature. (19,79) Also, there are several commercial software packages allowing their calculation for large compound datasets (e.g.: DRAGON, CDK, Mol<sup>2</sup>) but the major problem is how to best select descriptors beyond chemical intuition. (19,69,79). For many applications, PCA is used to reduce the dimensionality in order to choose the most important descriptors and their contributions, avoiding redundancies and correlations between descriptors. (69)

The *rcdk* package (60) groups molecular descriptors into five categories: “constitutional”, “hybrid”, “topological”, “electronic” and “geometrical”. Table 5 shows the molecular descriptors used in this study after the feature selection process, described in detail in section 3.5.

Table 5. Description of the molecular descriptors used in this study obtained by the CDK library. Only the resulting descriptors from the feature selection process are shown. The colors represent the category of the molecular descriptor:  hybrid;  topological;  electronic;  constitutional.

Molecular Descriptor		Description	References
<b>Burden Chemical Abstract Service University of Texas</b>	BCUTw,11	Eigenvalue based descriptor noted for its utility in chemical diversity. The descriptor is based on a weighted version of the Burden matrix which considers both the connectivity as well as atomic properties of a molecule. BCUTw,11 → nhigh lowest atom weighted BCUTS. BCUTc,1h → nlow highest partial charge weighted BCUTS. BCUTp,11 → nhigh lowest polarizability weighted BCUTS. BCUTc,11 → nhigh lowest partial charge weighted BCUTS.	(98–100)
	BCUTc,1h		
	BCUTp,11		
	BCUTc,11		
<b>PetitjeanNumber</b>		According to the Petitjean definition, the eccentricity of a vertex corresponds to the distance from that vertex to the most remote vertex in the graph. The distance is obtained from the distance matrix as the count of edges between the two vertices. If $r_i$ is the largest matrix entry in row $i$ of the distance matrix $D$ , then the radius is defined as the smallest of the $r_i$ . The graph diameter $D$ is defined as the largest vertex eccentricity in the graph. Petitjean Number is the value of diameter - radius. The radius-diameter diagram allows classification of the shapes of compounds and has remarkable	(101)

		properties for both graph-theoretical and geometrical shapes.	
<b>Molecular distance-edge</b>	MDEC,12	MDE descriptors are based on two fundamental structural variables, one for distance between atoms in the molecular graph and another for edges of the adjacency in the graph.	(100,102,103)
	MDEC,23	Molecular distance edge: MDEC,12 → between all primary and secondary carbons.	
	MDEC,33	MDEC,23 → between all secondary and tertiary carbons. MDEC,33 → between all tertiary carbons.	
<b>Kier Hall Smarts or Eccentric Connectivity Index</b>	khs,dsCH	A fragment count descriptor that uses electrotopological-state fragments. khs,dsCH → [CD2H](=*)-* or =CH- khs,dCH2 → [CD1H2]=* or =CH2 khs,ssO → [OD2Ho](-*)-* or -O- khs,aaO → [O,oD2Ho](:*):* or :O: khs,sCl → [CID1]-* or -Cl khs,aaaC → [C,c;D3H0](:*)(:*):* or ::C: khs,sNH2 → [ND1H2]-* or -NH2 khs,ssNH → [ND2H](-*)-* or -NH2-+ khs,ssNH → [N,nD2H](:*):* or :NH: khs,ssN → [ND3H0](-*)(-*)-* or >NH-+	(97,100,103,104)
	khs,dCH2		
	khs,ssO		
	khs,aaO		
	khs,sCl		
	khs,aaaC		
	khs,sNH2		
	khs,ssNH		
	khs,ssN		
<b>Chi-chain</b>	VCH,5	Evaluates the simple and valence chi chain descriptors of orders 3, 4, 5, 6 and 7 by finding fragments matching SMILES strings representing the fragments corresponding to each type of chain. VCH,5 → Valence chain, order 5. SCH,5 → Simple chain, order 5.	(57,100)
	SCH,5		
<b>Carbon Types</b>	C1SP3	Characterizes the carbon connectivity in terms of hybridization. C1SP3 → Singly bound carbon bound to one other carbon. C2SP2 → Doubly bound carbon to two other carbons. C3SP2 → Doubly bound carbon bound to three other carbons. C4SP3 → Singly bound carbon to four other carbons.	(57,100,103)
	C2SP2		
	C3SP2		
	C4SP3		

<b>Autocorrelation Descriptor Charge</b>	ATSc2	Autocorrelation descriptor, weighted by charges. The values are calculated considering weight equal to charges. Explain how the values of certain functions, at intervals equal to the lag d, are correlated. In this case, lag is the topological distance, and the atomic properties (weight or charge) are the functions correlated. ATSc2 → autocorrelation of a topological structure of lag 2 ATSc3 → autocorrelation of a topological structure of lag 3	(103,105)
	ATSc3		
<b>topoShape</b>		A measure of the anisotropy in a molecule.	(101)
<b>TopoPSA</b>		Calculation of topological polar surface area based on fragment contributions.	(100,103)
<b>tpsaEfficiency</b>		Polar surface area expressed as a ratio to molecular size.	(100,103)
<b>nSmallRings</b>		Total number of small rings of size 3 through 9.	(100,103)
<b>nAromRings</b>		Total number of small aromatic rings.	(100,103)
<b>nRings4</b>		Individual breakdown of 4 membered rings.	(100,103)
<b>apol</b>		Calculates the sum of the atomic polarizabilities, including implicit hydrogens.	(100)
<b>nHBDon</b>		Number of hydrogen bond donors.	(100)
<b>nHBAcc</b>		Number of hydrogen bond acceptors.	(100)
<b>nBase</b>		Basic group count descriptor.	(100,103)
<b>MW</b>		Weight of atoms of a certain element type. If no element is specified, the returned value is the molecular weight.	(100,103)
<b>XlogP</b>		Prediction of logP based on the atom type method called XlogP.	(103)
<b>AlogP</b>		Ghose-Crippen LogK <sub>ow</sub>	(100,103)
<b>Alogp2</b>		Square of AlogP.	(100,103)
<b>MLogP</b>		LogP based on the Mannhold equation using the number of carbons and hetero atoms.	(103)
<b>LipinskiFailures</b>		Number of failures of the Lipinski's Rule of 5.	(100,103)
<b>nRotB</b>		Number of non-rotatable bonds on a molecule.	(103)



### 3.3 Class Attribution

As mentioned previously, in supervised learning models classes are known. The main goal of this study is to better discriminate mycotoxins using their structures and molecular descriptors, and to predict whether they can be acutely toxic or not. The ideal would be to have experimental values of these mycotoxins in animal models. There are some experimental studies reporting toxicity values for rats, pigs, chickens, and others. (3,18,114,115,106–113) However, in order to construct a ML model, the data would have to be rigorous and there would have to exist a significant number of studies with the same animal model, route of administration, same conditions and time, which is not the case. Alternatively, mycotoxins were classified into acutely toxic (denoted as “1”) and non-acutely toxic (denoted as “0”) according to Global Harmonized System (GHS) classification. If the respective mycotoxin was classified by GHS as acutely toxic (pictogram with the skull), then the mycotoxin would be classified as 1. If not, the mycotoxins would be classified with 0, or non-acutely toxic. GHS classifies substances into five categories defined by specific cut-off values for oral, dermal, gases, vapors, dusts, and mists, based on experimental data. When no experimental data is available, GHS follows a series of steps that comprise intensive research on symptoms, similar molecules, types of animals and also several expert judgements. The GHS scheme was therefore employed in order to proceed with the analysis and bring more information about mycotoxins to the table, considering a system that accommodates the needs of the other systems, from labelling to transport.

All this information can be found on the eighth revised edition of the GHS, available at <https://unece.org/ghs-rev8-2019>.

### 3.4 Construction of the datasets

A first dataset was constructed composed of 30 mycotoxins from the most well-known families. (3–7,9,10) However, ML models work better for large datasets, and it would be interesting to evaluate not so well-known mycotoxins. With this, to the first dataset composed of 30 mycotoxins, 29 more mycotoxins were added. The complete list of mycotoxins is presented in Figure 8 with their respective assigned code and class. The classes (0 for non-acutely toxic and 1 for acutely toxic) were only added to the dataset for labelling in the unsupervised analysis and to further construct the supervised learning models.

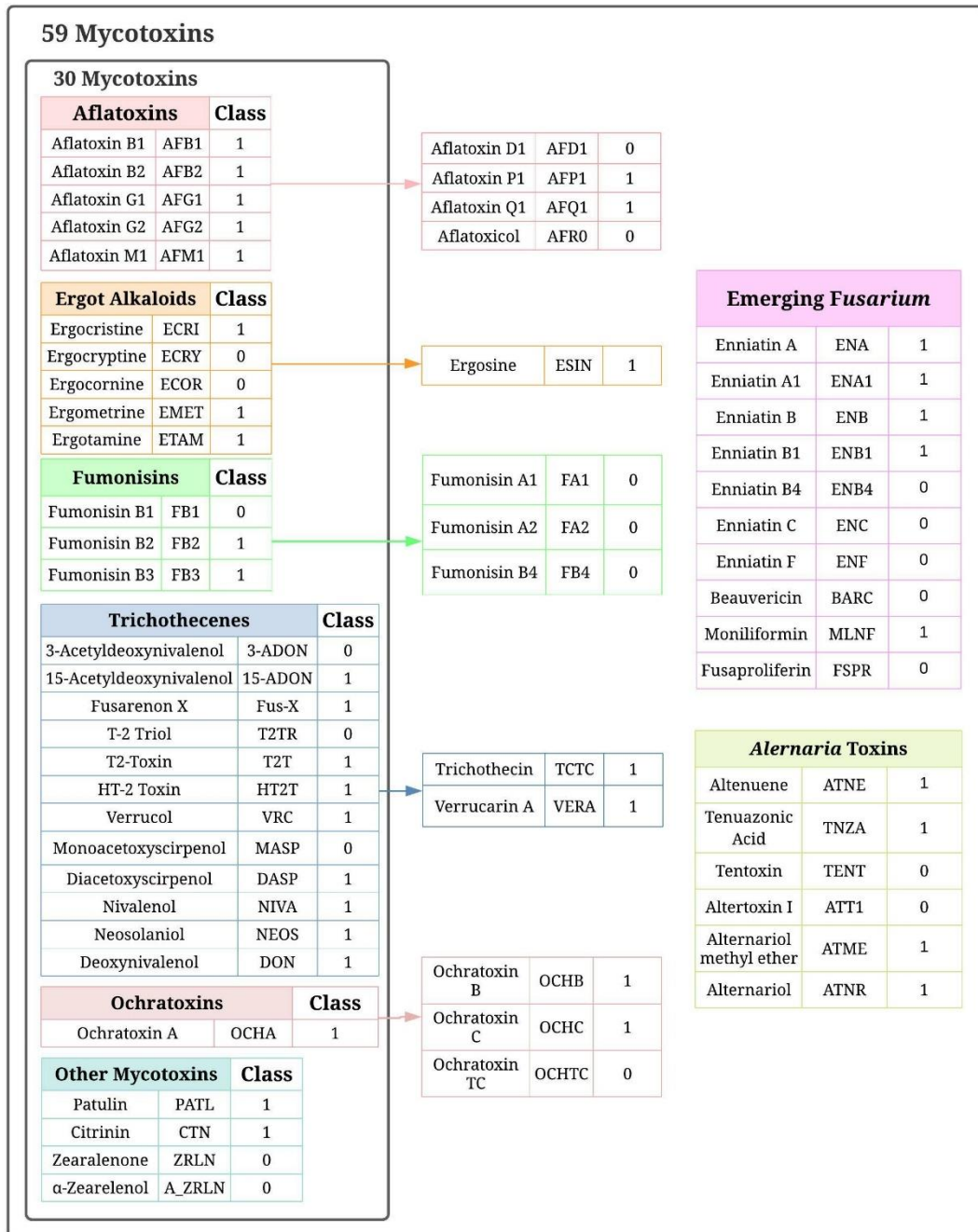


Figure 8. Schematic representation of mycotoxins (30) present on the first dataset (on the left), the posteriorly added mycotoxins (29) (on the right) and their respective acronym and class.

### 3.5 Feature Extraction and Selection

Fingerprints were computed using the *fingerprint* package (116) that handles binary fingerprint data and provides a function to construct and evaluate the distance matrix using the Tanimoto coefficient (Eq. 5).

The *rcdk* package (60) was used to calculate the molecular descriptors, resulting in a total of 287 descriptors. After the features calculation, a feature selection procedure from the *rcdk* package was performed. (60) That process consisted in: (i) removing missing values; (ii) removing correlated columns and (iii) removing constant columns. For the dataset composed with 30 mycotoxins this feature selected procedure resulted in the selection of 15 molecular descriptors and for the dataset composed by 59 mycotoxins, this procedure resulted in a final dataset of 28 molecular descriptors. However, when building the supervised learning models, some warnings arose about the existence of collinear or constant variables. For that reason, a fourth step has been performed that consisted in eliminating variables (molecular descriptors) with variance inferior to 0.005, which resulted in a different dataset with 24 molecular descriptors. This procedure was only applied to the dataset composed by 59 mycotoxins. Also, because the main goal is to relate mycotoxins' structures with their acute toxicity, another dataset was manually created with molecular descriptors related to biological activity. That selection was made according to some work in the literature (117–119) and the molecular descriptors were *HBDon*, *HBAcc*, *nRot*, *MW*, *LipinskiFailures*, *nAromRings*, *XLogP*, *MlogP*, *ALogP*, *ALogp2*, *nSmallRings* and *TopoPSA* (see Table 5). Lipinski Rule of five related descriptors were selected because they take into account the likelihood of absorption and permeation, important parameters that can directly influence the activity of every molecule, in this case mycotoxins. (120) For the sake of simplicity, Figure 9 shows the process of constructing the datasets, including the number of molecular descriptors, that resulted from feature selection.

Finally, and before constructing the models, data was normalized using the *caret* package from R (64) for LDA and SVM and using *scikit-learn* package from Python in the case of RF and NN.

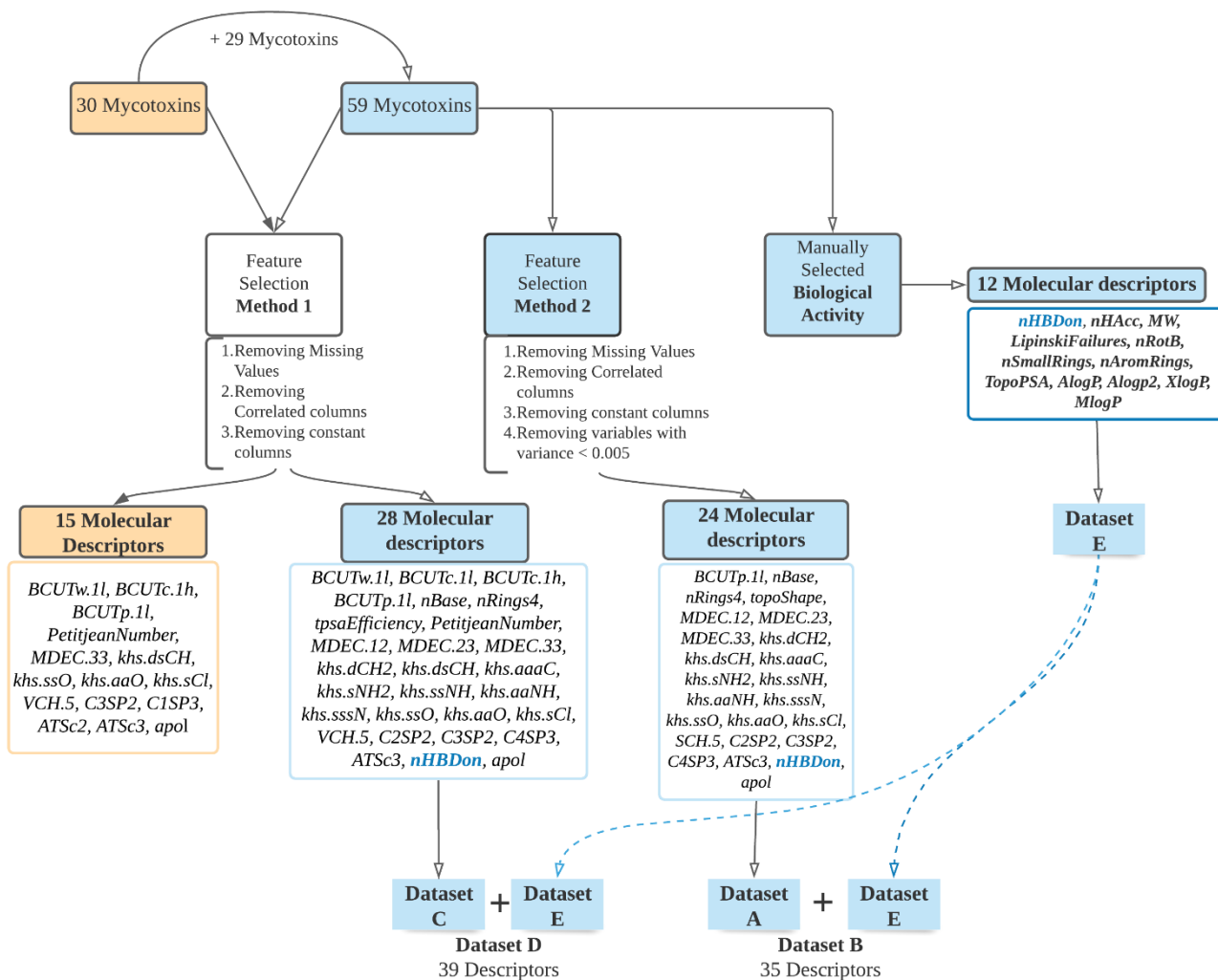


Figure 9. A composed view of the steps for datasets construction and feature selection. Feature selection method 1 was only applied to the dataset composed by 30 mycotoxins. For the dataset composed by 59 mycotoxins, all feature selection methods were applied. nHBDOn is in bold because it was one of the manually selected descriptors, but the process of feature selection also considered this descriptor.

## Chapter 4 - Results and Discussion

So far, the necessary tools to proceed with the analysis and the construction of the ML models and how those models work were presented. In short, unsupervised learning techniques, HCA and k-means, were used to evaluate the molecular similarity between the mycotoxins, based on molecular fingerprints. PCA was implemented to identify relevant molecular descriptors to discriminate the mycotoxins and their families. The relationship between the selected molecular descriptors and the acute toxicity of the mycotoxins were also assessed based on the relative positioning of mycotoxins labelled according to their acute toxicity. Finally, supervised learning models were constructed using those labels as target response to predict mycotoxins acute toxicity.

All results are presented in this section.

### 4.1 Molecular Similarity

HCA was firstly performed over 30 mycotoxins belonging to the most commonly known families. The results are presented in Figure 10.

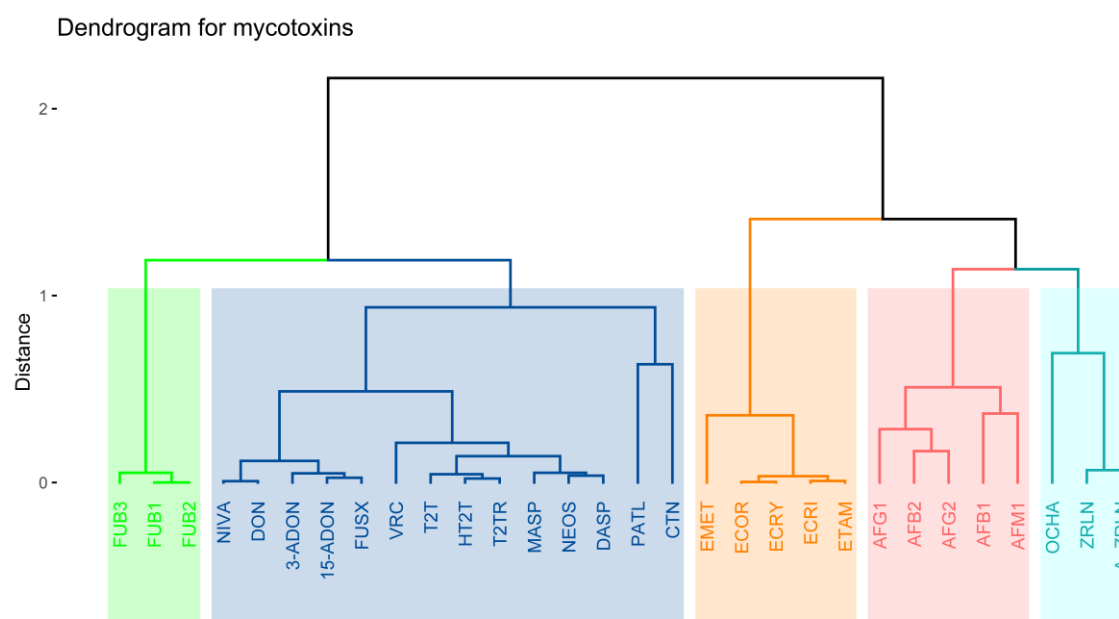


Figure 10. Dendrogram representing the similarity among 30 mycotoxins considering their molecular fingerprints, constructed resorting to Ward's method over the Tanimoto distances. The different colors represent the different families: fumonisins are colored in green, dark blue represents the trichothecenes family (excepting CTN and PATL), in orange are ergot alkaloids, aflatoxins in red and in light blue are the remaining mycotoxins.

From left to the right, we have fumonisins, trichothecenes, ergot alkaloids, aflatoxins and other mycotoxins that do not belong to a specific family.

Considering the dendrogram distances, it is visible that all families are distant from each other, suggesting that they display different characteristics. Fumonisin group (green) is the most cohesive group in this dendrogram which makes sense because fumonisins B1, B2 and B3 are closely related metabolites that only differ in small substitutions from -OH to -H or vice versa. (121) The trichothecenes family (dark blue group) is a big family of mycotoxins and can be divided in four groups (A-D). In this dataset of 30 mycotoxins, only type A and B trichothecenes were selected and, in fact, they are present of the 2 first subgroups of this family: on the left side are type B trichothecenes and on the right side are type A trichothecenes. Considering that HCA was performed using hashed fingerprints- sequences of 0 and 1 digits containing structural information- the algorithm performed well distinguishing these trichothecenes types that only differ in the presence or absence of a carbonyl group at the C-8 position. (6,14) Furthermore, PATL and CTN are not trichothecenes but were clustered together with them. Even so, it is visible their distance from the remaining trichothecenes. Although patulin (PATL) and citrinin (CTN) are far from the trichothecenes, they are very close to each other in the cluster which is in accordance with the existence of bi-toxicogenic fungal strains with the ability to produce both PATL and CTN. (122) Regarding the ergot alkaloids group (orange), there is a slight separation between ergometrine (EMET) and the other ergot alkaloids, which is explained by the fact that EMET is a lysergic acid derivative, and the remaining mycotoxins of this cluster are peptide alkaloids (also called ergopeptines) that only differ in a peptide moiety linked to the basic tetracyclic ergoline. (13) Aflatoxins are represented in the pink cluster and contain the most toxic known mycotoxin, aflatoxin B1 (AFB1). The term “B” and “G” refers to the blue and green, fluorescent colors produced by these mycotoxins under UV light and, with the exception of AFM1, all aflatoxins are the four major mycotoxins produced by mold metabolism. Aflatoxin M1 is the hydroxylated metabolite of AFB1 (5,6,123), corroborating their proximity in the cluster. The last group is composed by individual mycotoxins, ochratoxin A, zearalenone and  $\alpha$ -zearalenol. Starting with zearalenone (ZRLN) and  $\alpha$ -zearalenol (A\_ZRLN), and as the name implies, A\_ZRLN is a result of biotransformation of ZRLN carried out by animals, justifying their proximity in this cluster. OCHA was clustered together with ZRLN and A\_ZRLN and there is evidence of their co-occurrence (124) but they are still too far away on the dendrogram.

Further analysis based on K-means clustering, Figure 11, confirms the HCA results shown in the dendrogram (Figure 10). Each cluster has a centroid, and the algorithm combines distances between molecules and the group centroids.

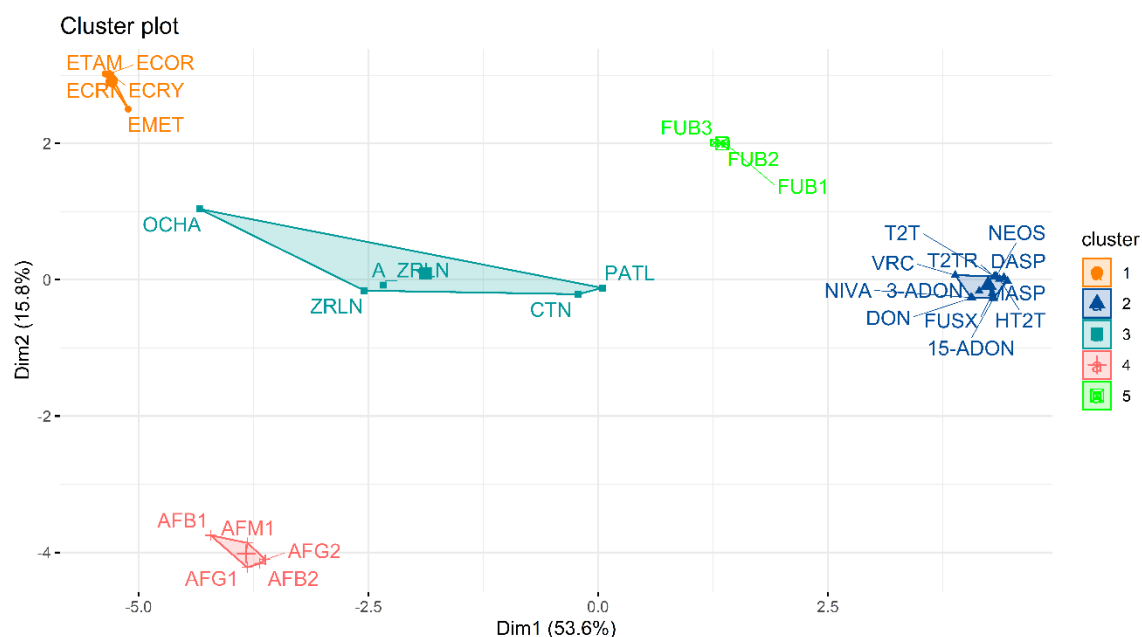


Figure 11. K-means cluster plot constructed over the Tanimoto distances between the 30 mycotoxin structures, represented by their molecular fingerprints.

Although a different algorithm was applied, the results obtained are generally in accordance with the results of the HCA.

With the exception of the light blue cluster, all other clusters show both cohesion and separation. Interestingly, PATL and CTN were now clustered together with OCHA, ZRLN and A\_ZRLN. The proximity between CTN and PATL is still visible, as well as ZRLN and A\_ZRLN. OCHA, PATL and CTN are on opposite sides of the light blue cluster and OCHA is near ergot alkaloids (orange) and CTN and PATL near fumonisins. The fact that these mycotoxins are on opposite sides of their own cluster can be explained by their different molecular weights (OCHA has a MW of 403.8, PATL of 154.1 and CTN of 250.3) and the MW of ergot alkaloids, namely ergometrine (EMET), are closer to the MW of OCHA (EMET has a MW of 325.4). Furthermore, OCHA and EMET have the same number of hydrogen bond donors and a comparable number of heavy atoms (51,94). It is curious that PATL and CTN are now on this cluster (and not in the trichothecenes as happened in HCA) since the co-occurrence of OCHA and PATL (125) and OCHA and CTN (126) has been documented. Nevertheless, they are on the opposite sides of their own cluster and can be in the decision boundary between their cluster and ergots cluster (in the case of OCHA) and fumonisins cluster (in the case of PATL and CTN).

It is also visible that ergot alkaloids (orange) are mainly distant from aflatoxins (pink) in the Y axis while the remaining groups are mostly separated in the X axis. This observation can be related to the fact that ergot alkaloids are, in general, more complex molecules, with higher molecular weights and partition coefficients as well as with more stereocenters and heavy atoms.(51,94) Relatively to the X axis, these groups are probably separated due to their significantly different structures. Fumonisins have long-chain hydrocarbon units while trichothecenes are characterized

by a variable number of acetoxy and hydroxyl groups, an epoxide at C<sub>12,13</sub> positions and a double bond between C<sub>9</sub> and C<sub>10</sub>.

To validate the clustering results, it was considered the silhouette coefficient, presented in Figure 12. Silhouette values range from -1, indicating that the mycotoxins are not in the correct cluster, to +1, indicating that the mycotoxin is far from the neighboring cluster and very close to the cluster to which is assigned.

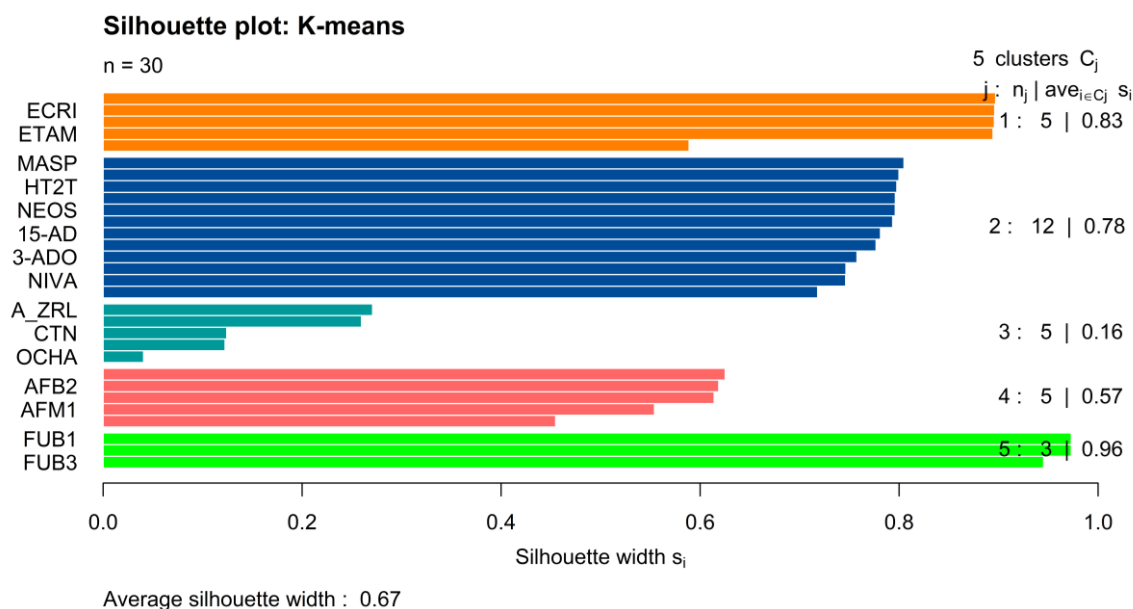


Figure 12. Silhouette plot for k-means clustering constructed over the initial 30 mycotoxins' topological fingerprints.

Excepting the light blue cluster, all other clusters show a silhouette coefficient near 1, suggesting that the similarity evaluation and group identification were performed correctly. The lower coefficient of 0.16 for the light blue cluster, is explained by the lack of cohesion within the group, visible in Figure 11, which suggests that mycotoxins in the cluster display distinct characteristics even within the group. The silhouette plot for the k-means clustering reveals a mean value of 0.67 which is considered acceptable, meaning an overall good clustering procedure.

The previous data showed that the application of clustering techniques in the most well-known mycotoxins (30 mycotoxins) resulted in their grouping according to the families defined in the literature. (5–7) That said, some not so well-known mycotoxins were added to the dataset, including mycotoxins already belonging to the families shown previously, but also mycotoxins belonging to new families. (See Figure 8) The same clustering techniques were applied, and HCA results are shown in Figure 13.



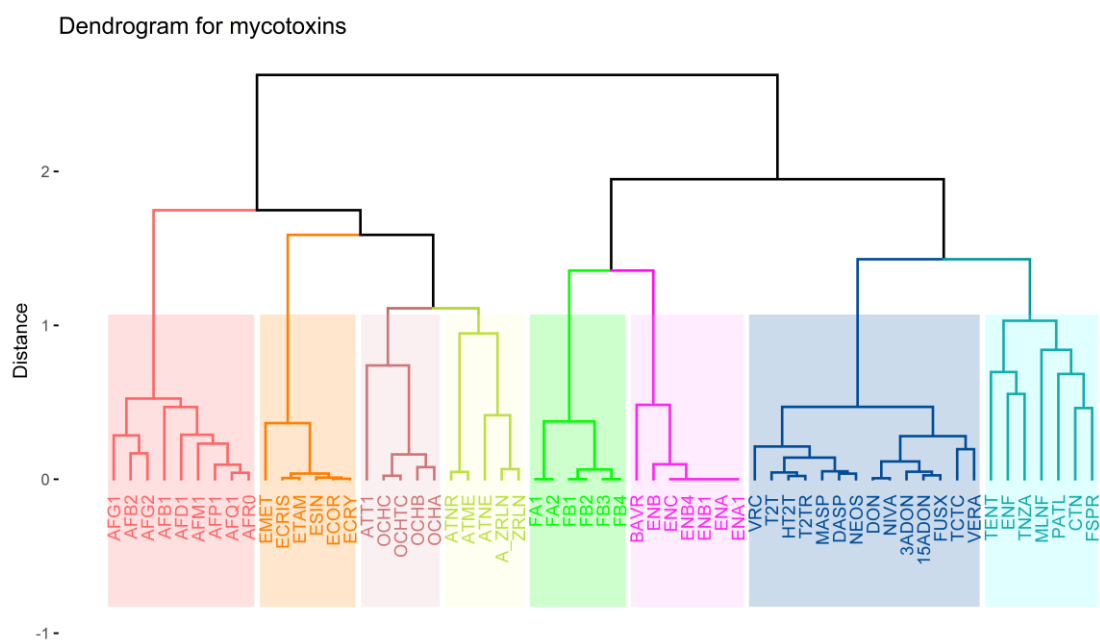


Figure 13. Dendrogram representing the similarity among all 59 mycotoxins considering their molecular fingerprints, constructed resorting to Ward's method over the Tanimoto distances. In red are aflatoxins, orange are ergot alkaloids, brown are ochratoxins, fumonisins are colored in green, enniatins in pink, dark blue represents the trichothecenes family, and in light blue and yellow are the mycotoxins that are from different families or do not belong to a specific family.

Once again, all families are distant from each other. Regarding aflatoxins group (red), and excepting the four major aflatoxins (AFB1, AFB2, AFG1 and AFG2) the remaining aflatoxins were described as mammalian biotransformation products of the major metabolites (127), such as oxidative hydroxylation (AFM1 and AFQ1) (128), O-demethylation (AFP1) (129), justifying their closeness in the cluster. ESIN, also an ergot alkaloid, ended up in the right cluster (orange) and is notoriously close to the remaining ergots. A new group was formed, in brown, composed by ochratoxins and ATT1. All ochratoxins are very close to each other, which is in agreement with the fact that OCHB and OCHC are the dechlorinated and ethyl ester derivatives of OCHA, respectively. (130) ATT1 is distant from the ochratoxins even within the group, although it is interesting that ended up clustered together. Curiously, the yellow group is only composed by 3 *Alternaria* mycotoxins, ATNR, ATME and ATNE and 2 other mycotoxins ZRLN and A\_ZRLN. ATNR and ATME are structurally very similar, reason why they are so close in the cluster. (131) However, ATNE, an *Alternaria* mycotoxin, is closer to ZRLN and A\_ZRLN, suggesting some structural similarities. The new additions in the fumonisins group (FA1 and FA2) ended up in the respective family cluster and these new fumonisins, FAs, are separated from FBs as they are N-acetates of FB1 and FB2. (132) Colored in pink is a new group composed by enniatins (ENs) and BAVR, despite one of the enniatins (ENF) is on the light blue cluster. BAVR was clustered with enniatins and indeed these mycotoxins are structurally related since several *Fusarium* species can produce ENNs, BEA or both, differing in their amino acid residues. (133) The new added trichothecenes (VERA and TCTC) were correctly clustered with the remaining trichothecenes. The light blue

group is particularly interesting because in addition to the mycotoxins already present in this group on the previous dataset (A\_ZRLN, ZRLN, PATL and CTN), MNLF, FSPR and ENF are emerging *Fusarium* mycotoxins and TENT and TNZA are *Alternaria* mycotoxins. Looking at the dendrogram in a general perspective, this cluster is composed by mycotoxins more distant from each other (within the group) than the remaining clusters. For example, the trichothecenes group (dark blue) has mycotoxins closer to each other (lower distance) than the light blue group. This is explained by the cutoff value used to plot the dendrogram. In this case, the cutoff value used was at a distance of 2.5 but if this value was a bit lower, this light blue cluster would split into two other clusters and the main result would be a dendrogram with 9 clusters formed. In contrast, if the cutoff value was a bit higher, it would result in the grouping of the yellow and brown clusters into one, and a final dendrogram with 7 clusters. Usually, this issue is assessed by determining the optimal number of clusters but, in this case, all tests gave a different optimal number of clusters and the chosen number corresponded to the number of different families present in the dataset. (See Annex I) However, this could be useful to ascertain the presence of any mycotoxin wrongly classified in the literature.

Once more, k-means clustering was performed, and the results are shown in Figure 14.

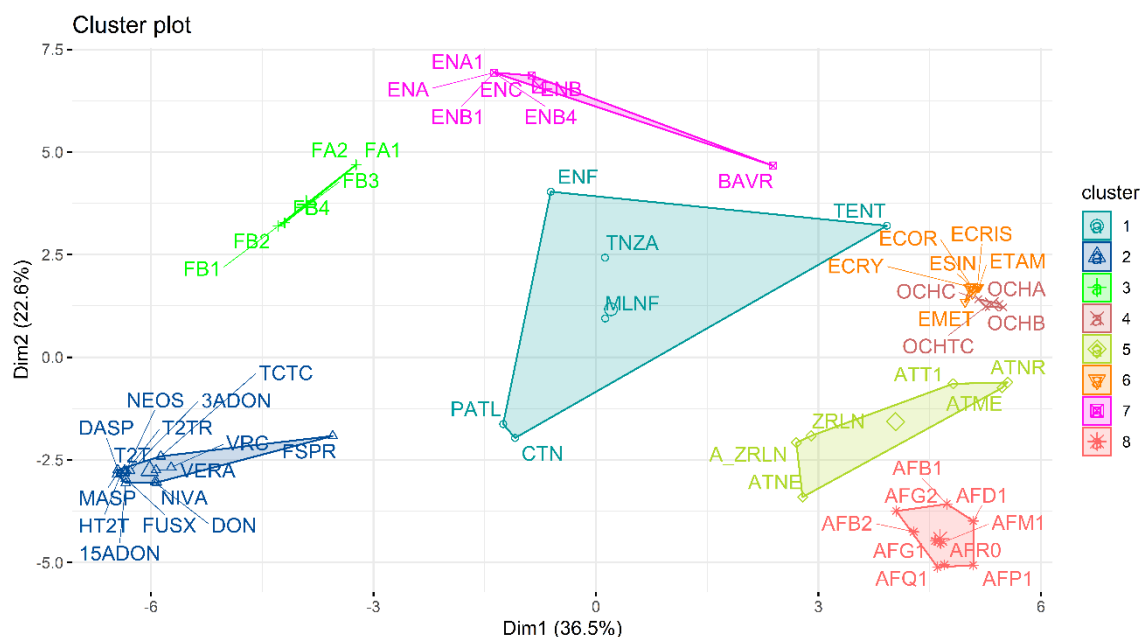


Figure 14. K-means cluster plot used over the Tanimoto distances between the 59 mycotoxin structures, represented by their molecular fingerprints.

As expected, and in agreement with the HCA results, all groups are compact, with the exception of the light blue and, in this case, the yellow cluster that show no cohesion at all. This reveals that the majority of the mycotoxins belonging to these two clusters are not structurally similar. Yet, it is visible that even though ENF is not in the same cluster of enniatins (as happened in HCA), ENF is closer to their cluster centroid than to the centroid of its own cluster. In addition, FSPR, an emerging *Fusarium* mycotoxin, was now clustered with trichothecenes, suggesting some

structural similarity between FSPR and trichothecenes. This makes sense because mycotoxins produced by *Fusarium* species include, among others, trichothecenes and zearalenone (ZRLN). (18) The light blue cluster is the less cohesive cluster which is understandable because contains mycotoxins from different families: ENF and MLNF are emerging *Fusarium* mycotoxins; TNZA and TENT are *Alternaria* mycotoxins and CTN and PATL do not belong to any specific family.

The yellow cluster, mostly composed by *Alternaria* toxins, also presents some lack of cohesion probably due to the diversity of *Alternaria* mycotoxins, that can be divided into 5 classes. ATNE, ATME and ATNR all belong to pyranones/ benzopyrones corroborating their proximity in the yellow cluster; TNZA belongs to amine/amide metabolites which is probably the reason why was clustered in the light blue group; TENT belongs to cyclic tetrapeptides being was also clustered in the light blue group and ATT1 belongs to perylenequinones (17) and was now clustered in the yellow group, in contrast with HCA where ATT1 was clustered with ochratoxins. This diversity may justify *Alternaria* mycotoxins positioning in the k-means cluster plot between yellow and light blue clusters. Even so, and similarly to ENF, it is visible that TENT is closer to the ergots cluster (orange) than to its own cluster, and ATT1 is closer to ochratoxins (brown) suggesting that these samples are on or very close to the decision boundary between these two neighboring clusters. Furthermore, and similarly to what happens in HCA, ZRLN and A\_ZRLN were clustered together with some *Alternaria* toxins in the yellow cluster and their co-occurrence has also been reported. (134) K-means also confirmed the molecular similarity between ATNR and ATME.

These clustering results were also validated using the silhouette coefficient and the resulting plot is presented in Figure 15.

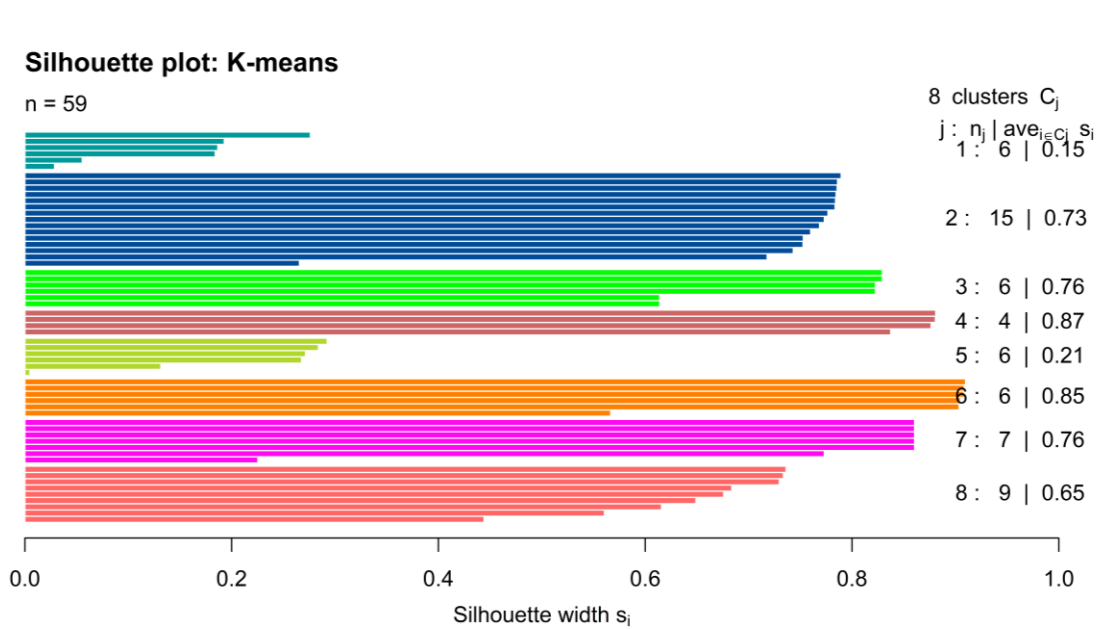


Figure 15. Silhouette plot for k-means clustering constructed over the 59 mycotoxins topological fingerprints.

As expected, the less cohesive clusters in the k-means cluster plot, light blue and yellow, contain mycotoxins with a lower silhouette coefficient (0.15 and 0.21, respectively) which means that, in fact, these mycotoxins were very close to the decision boundary between their neighboring clusters. These mycotoxins are TENT and ENF in the light blue cluster and ATT1 in the yellow cluster. As said before, ENF should have been clustered together with the other enniatins (pink), TENT with ergot alkaloids and curiously, ATT1 with ochratoxins. All the other groups show a good silhouette coefficient, meaning that they were correctly clustered.

## 4.2 Impact of Molecular Descriptors

PCA was performed in order to summarize and visualize the information in a dataset composed of 30 mycotoxins described by several molecular descriptors. The main goal is to identify the principal components along which the variation is maximal, through a dimensionality reduction with minimal loss of information, in order to understand how these mycotoxins can be described and if there is any connection with the molecular similarity results, obtained using molecular fingerprints and simple distance measures. A feature selection procedure from CDK allowed selecting 15 molecular descriptors, presented in Table 5.

Table 6 shows the eigenvalues that resulted from PCA, allowing to evaluate the principal components to be considered. PCA with the first two components (PC1 and PC2) resulted in a variability recovery of 46.8% and with a third component, ca. 65% of information variability recovered.

Table 6. Eigenvalues and evolution of the percentage of information recovery in relation to the number of principal components for the dataset composed by 30 mycotoxins and 15 molecular descriptors. The most relevant descriptors are underlined and correspond to eigenvalues greater than 1.

	<b>Eigenvalue</b>	<b>Variance %</b>	<b>Cumulative Variance %</b>
<b>Dim. 1</b>	<u>3.88</u>	<u>25.8</u>	<u>25.8</u>
<b>Dim. 2</b>	<u>3.14</u>	<u>20.9</u>	<u>46.8</u>
<b>Dim. 3</b>	<u>2.69</u>	<u>17.9</u>	<u>64.7</u>
<b>Dim. 4</b>	<u>1.86</u>	<u>12.4</u>	<u>77.1</u>
<b>Dim. 5</b>	<u>1.24</u>	<u>8.25</u>	<u>85.4</u>
<b>Dim. 6</b>	0.818	5.46	90.8
<b>Dim. 7</b>	0.540	3.60	94.4
<b>Dim. 8</b>	0.299	1.99	96.4
<b>Dim. 9</b>	0.188	1.26	97.7
<b>Dim. 10</b>	0.139	0.926	98.6
<b>Dim. 11</b>	0.111	0.737	99.4
<b>Dim. 12</b>	0.049	0.325	99.7

<b>Dim. 13</b>	0.0278	0.185	99.9
<b>Dim. 14</b>	0.0104	0.0697	99.9
<b>Dim. 15</b>	0.00877	0.0585	100

According to Table 6, 5 components would have to be retained, corresponding to components with eigenvalues superior to 1. These 5 components recover ca. 85.4% of cumulative variance. Both clustering techniques and PCA are unsupervised learning methods, but the difference now is that it is possible to understand not how similar these molecules are but what describes them and what features- molecular descriptors- are responsible for their discrimination profile. For that, the contributions of each molecular descriptor were assessed and are presented in Figure 16.

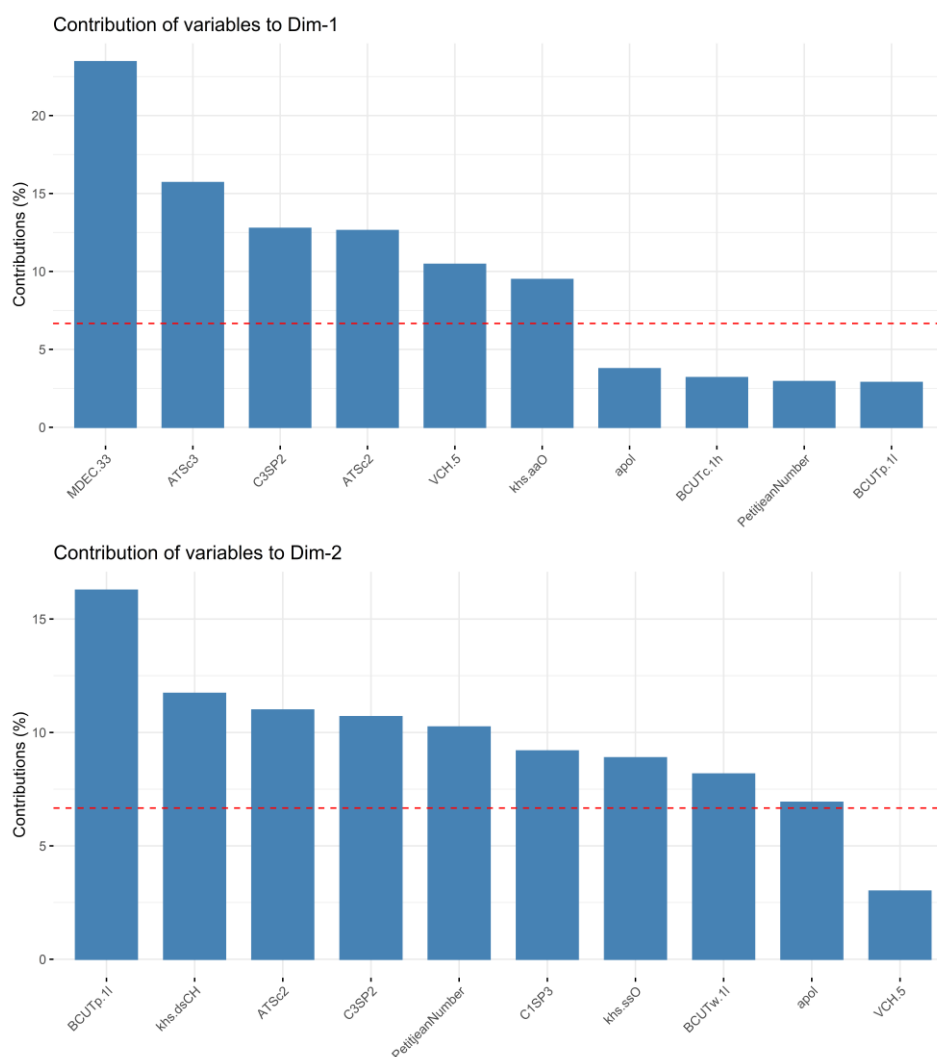


Figure 16. Graphical representation of the impact of molecular descriptors (*loadings*) on the first two principal components (correlation matrix) for the dataset composed by 30 mycotoxins. The top and the bottom plots refer to PC1 and PC2, respectively.

The molecular descriptors with the highest impact on PC1 are MDEC.33, ATSc3, C3SP2 and ATSc2. On the second component (PC2), BCUTp.11, khs.dsCH, ATSc2 and C3SP2 have the

highest impact. Surprisingly, all the molecular descriptors with highest impact on PC1 are topological but on the second component BCUTp.11, BCUTw.11 and apol are hybrid and electronic molecular descriptors, respectively. This reinforces the evolution of the topological descriptors towards higher discriminating power and better correlating ability, as described by Balaban (1998). (135)

PCA is two-dimensionally represented by a biplot composed with scores and loadings that are, respectively, the coordinates of the mycotoxins on the principal components and the weight of the original variables (molecular descriptors) on the new variables (PCs). The PCA biplot is presented in Figure 17 for the first two components, PC1 and PC2.

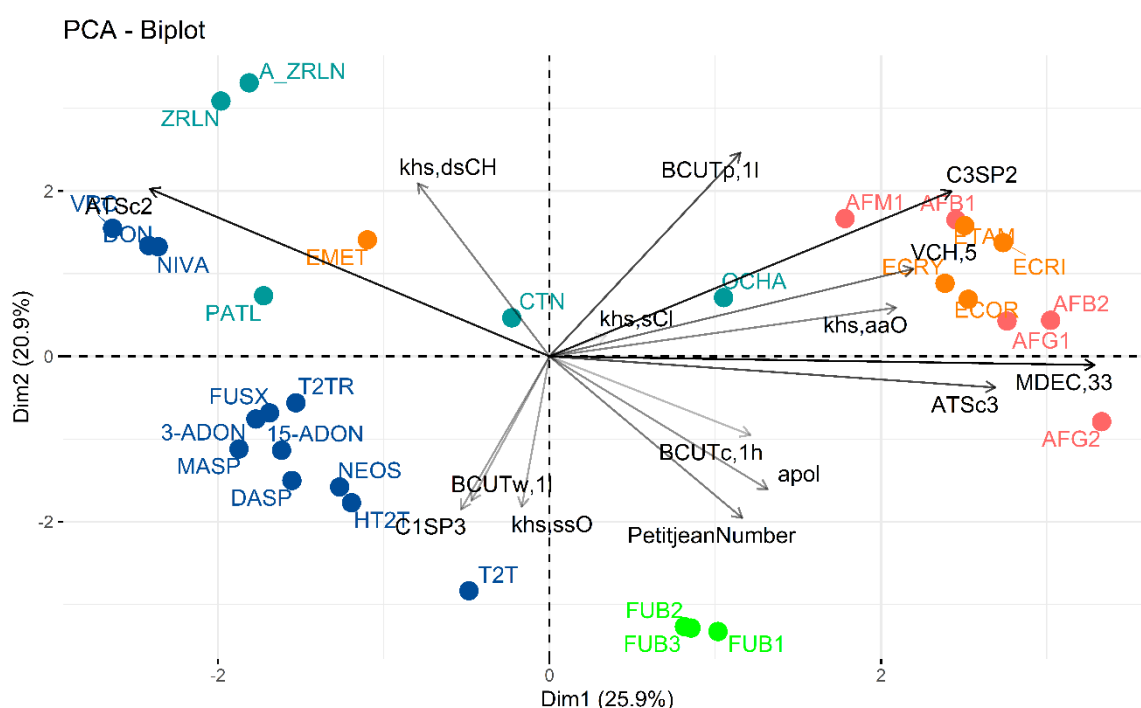


Figure 17. Biplot representation of 30 mycotoxins described by 15 molecular descriptors on the first two principal components, recovering 46.8% of the variance. Mycotoxins are colored according to the clustering results (see Figure 10) and the molecular descriptors were selected according to feature selection method 1.

The information variability recovered is ~47% but it is still visible some mycotoxin discrimination according to the already defined families in the literature, as it happened with the molecular similarity results. MDEC.33, ATSc3 and C3SP2 are mainly responsible for the positioning of aflatoxins and ergot alkaloids (pink and orange groups, respectively). This means that these two families may have in common their molecular distance edge between all tertiary carbons (MDEC.33) and also similar (i) autocorrelation of a topological structure of lag 3 (ATSc3) and (ii) numbers of doubly bound carbons bound to three other carbons (C3SP2). ATSc2 also has a high impact on PC1 but is responsible for the discrimination of some Type B trichothecenes (NIVA, DON), ZRLN and A\_ZRLN.

Most trichothecenes are influenced by BCUTw.11 and C1SP3 descriptors and fumonisins are discriminated by PetitjeanNumber and apol descriptors. Interestingly, the OCHA position is also consistent with the cluster results, considering that in both k-means cluster plot and PCA biplot representation this mycotoxin is close to ergot alkaloids. Now, it is plausible to infer that the reason for the proximity of OCHA to ergots may be due to one of the descriptors with positive impact on the principal component (MDEC.33, khs.aaO, VCH.5, C3SP2 and ATSc3).

Since the main goal of this work is to predict mycotoxins toxicity, mycotoxins were classified into acutely toxic and non-acutely toxic. The following biplot- Figure 18- is still an unsupervised learning result, only with labels to help understand if there is any relationship between molecular descriptors and the acute toxicity of mycotoxins. For detailed information about the classification procedure of mycotoxins into acutely toxic or not, please refer to the Database Description and Data Processing section.

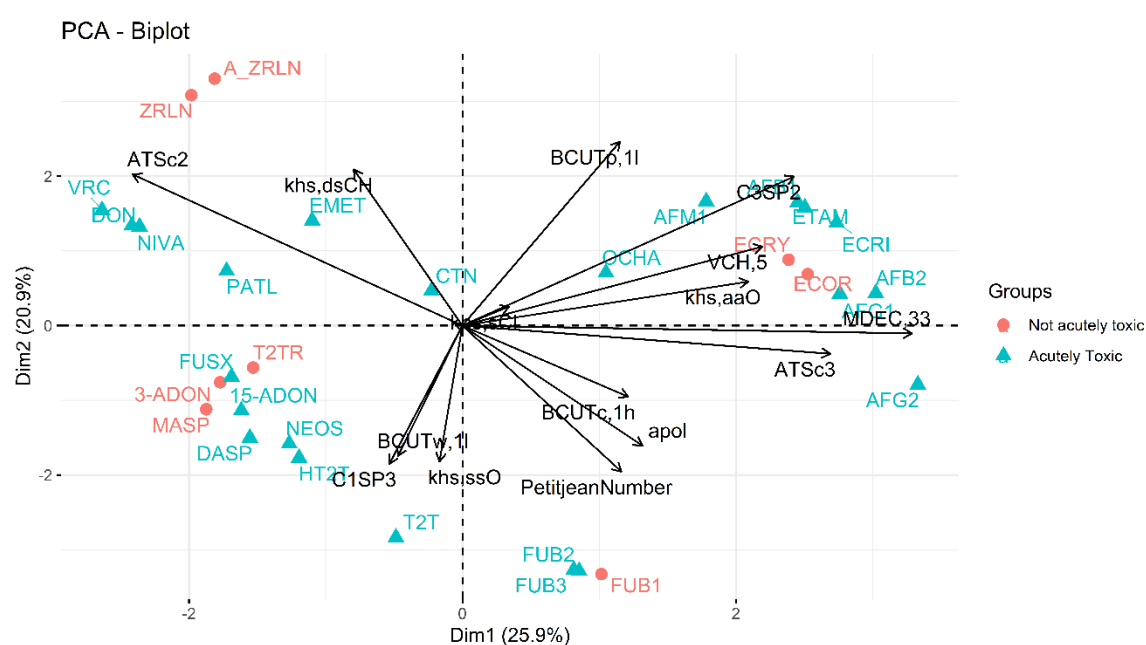


Figure 18. Biplot representation of 30 mycotoxins described by 15 molecular descriptors on the first two principal components, recovering 46.8% of the variance. Mycotoxins are labeled according to their acute toxicity (see Figure 8).

There is no visible relationship between molecular descriptors and mycotoxins' toxicity. However, considering the dataset dimensions (30 mycotoxins and 15 molecular descriptors) and the variability recovered (47%), very few information is represented in this system.

To extend the system description, and exactly as done previously to evaluate mycotoxins molecular similarity, PCA was applied to a larger dataset, composed by 59 mycotoxins. For these 59 mycotoxins, two different feature selection methods were applied and are described in the Dataset Description and Data Processing section. After that, 5 different datasets were constructed,

all of them with the same mycotoxins. Two of the five datasets correspond to one method of feature selection (dataset A and B); other two to the other feature selection method (dataset C and D) and the last one (dataset E) was created using manually added molecular descriptors related to biological activity, please consult Figure 9. Again, more detailed information about the used molecular descriptors is presented in Table 5.

### Dataset A

The eigenvalues were extracted to understand how many principal components should be used and are shown in Table 7. PCA with the first two components resulted in a variability recovery of ~38% and with a third component, ca. 52% of information variability recovered.

Table 7. Eigenvalues and evolution of the percentage of information recovery in relation to the number of principal components for the dataset composed by 59 mycotoxins and 24 molecular descriptors. The most relevant descriptors are underlined and correspond to eigenvalues greater than 1.

	<b>Eigenvalue</b>	<b>Variance %</b>	<b>Cumulative variance %</b>
<b>Dim. 1</b>	<u>4.62</u>	<u>19.3</u>	<u>19.3</u>
<b>Dim. 2</b>	<u>4.43</u>	<u>18.5</u>	<u>37.7</u>
<b>Dim. 3</b>	<u>3.49</u>	<u>14.5</u>	<u>52.3</u>
<b>Dim. 4</b>	<u>2.42</u>	<u>10.1</u>	<u>62.3</u>
<b>Dim. 5</b>	<u>1.86</u>	<u>7.77</u>	<u>70.1</u>
<b>Dim. 6</b>	<u>1.56</u>	<u>6.49</u>	<u>76.6</u>
<b>Dim. 7</b>	<u>1.33</u>	<u>5.56</u>	<u>82.2</u>
<b>Dim. 8</b>	0.925	3.85	86.0
<b>Dim. 9</b>	0.846	3.52	89.5
<b>Dim. 10</b>	0.684	2.85	92.4
<b>Dim. 11</b>	0.525	2.19	94.6
<b>Dim. 12</b>	0.311	1.29	95.9
<b>Dim. 13</b>	0.271	1.13	97.0
<b>Dim. 14</b>	0.205	0.856	97.9
<b>Dim. 15</b>	0.185	0.771	98.6
<b>Dim. 16</b>	0.114	0.475	99.1
<b>Dim. 17</b>	0.0873	0.364	99.5
<b>Dim. 18</b>	0.0509	0.212	99.7
<b>Dim. 19</b>	0.0317	0.132	99.8
<b>Dim. 20</b>	0.0154	0.0641	99.9
<b>Dim. 21</b>	0.0137	0.0570	99.9
<b>Dim. 22</b>	0.00939	0.0391	99.9
<b>Dim. 23</b>	0.00565	0.0235	100
<b>Dim. 24</b>	7.61e-32	3.17e-31	100



In accordance with Table 7, seven components would have to be retained, with eigenvalues superior to 1. Again, the smaller dataset (15 molecular descriptors) recovered 85% of variance with five components. This dataset needs seven components to recover ~82% of variance. Molecular descriptors with highest contributions on the first two components are shown in Figure 19.

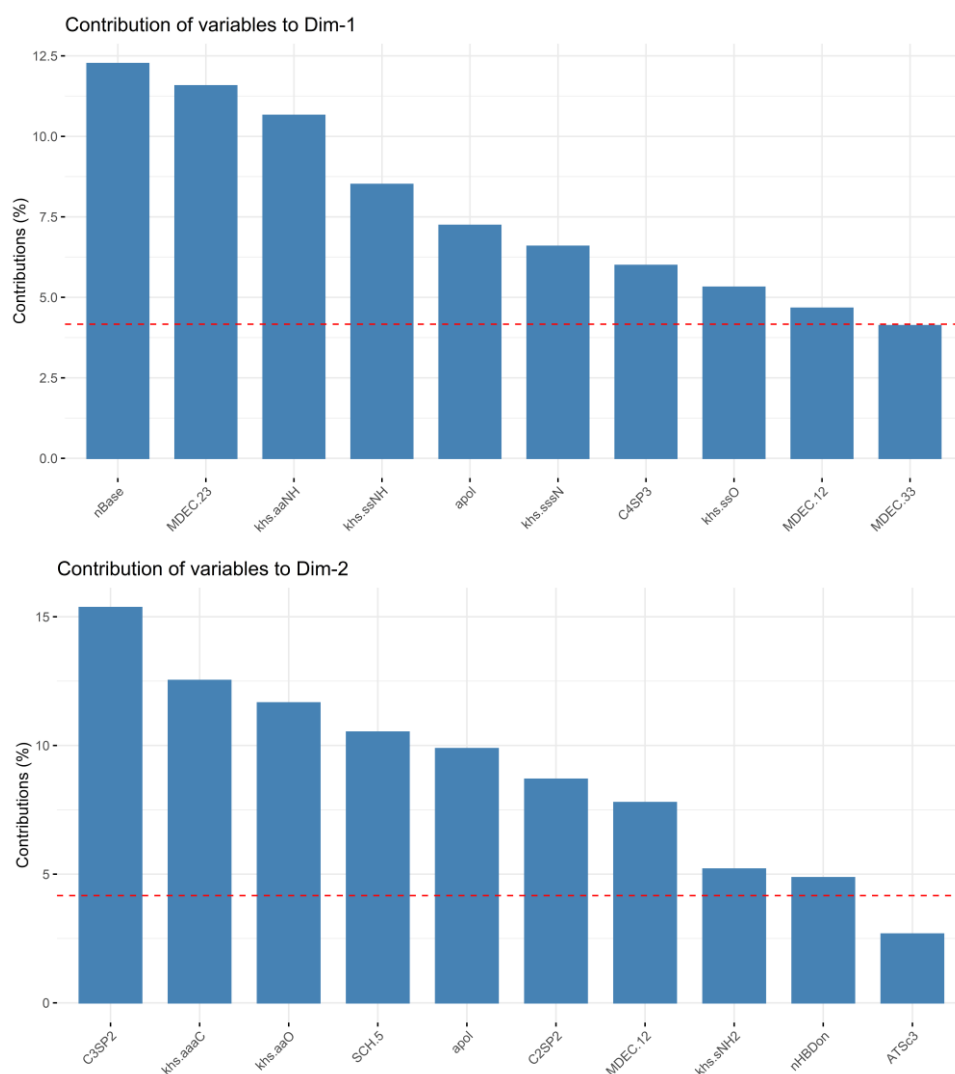


Figure 19. Graphical representation of the impact of molecular descriptors (*loadings*) on the first two principal components (correlation matrix) for the dataset composed by 59 mycotoxins and 24 molecular descriptors. The top and the bottom plots refer to PC1 and PC2, respectively.

Once again, topological descriptors dominate the higher contributions on the first two components, with the exception of *nBase*, *apol* and *HBDon* that are constitutional and electronic, respectively.

The PCA biplot for dataset A is presented in Figure 20.

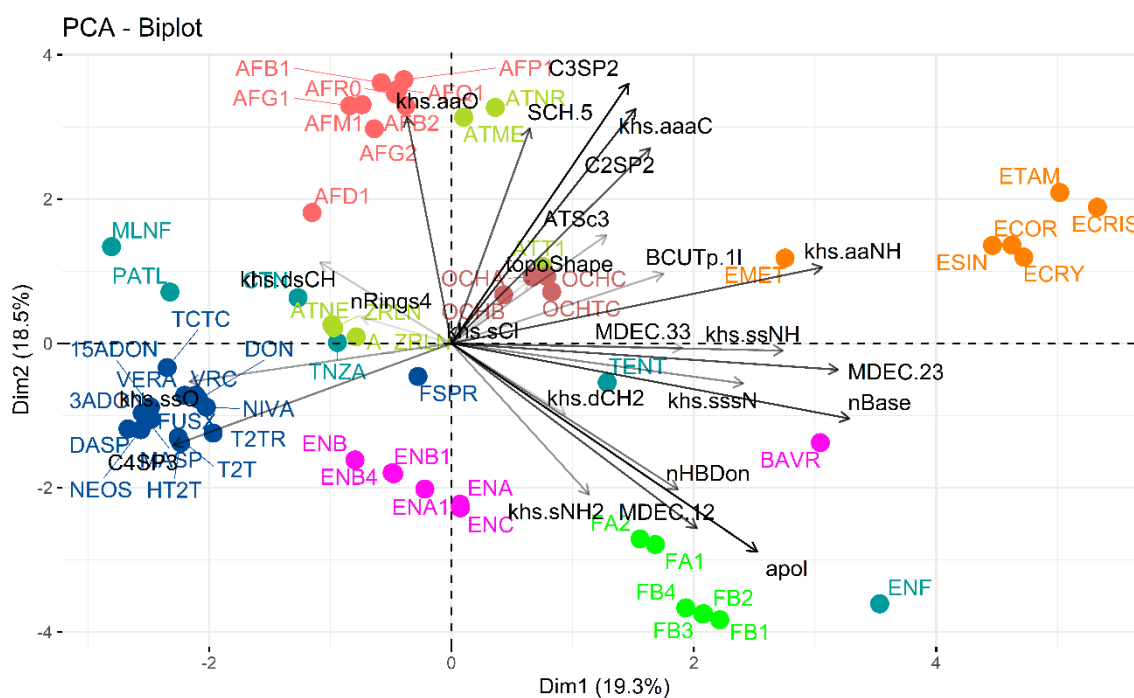


Figure 20. Biplot representation of 59 mycotoxins described by 24 molecular descriptors on the first two principal components, recovering 37.8% of the variance. Mycotoxins are colored according to the clustering results (see Figure 13) and the molecular descriptors resulted from feature selection method 2.

Mycotoxins families for this dataset are also well discriminated. In this biplot it is visible that aflatoxins (red) are mainly discriminated by *khs,aaO*, meaning that they frequently have [O,oD<sub>2</sub>H<sub>0</sub>](:\*):\* e-state fragment. The same is visible with ergot alkaloids (orange) with the *khs.aaNH* descriptor, being rich in [N,nD<sub>2</sub>H](:\*):\* fragment.

Fumonisin (green) are discriminated by *apol*, *MDEC.12*, *khs,dCh2* and *nHBDOn* and interestingly, FBs and FAs may have a slightly different *apol* and molecular distance-edge between all primary and secondary carbons (*MDEC.12* descriptor) since it is visible in the biplot a small separation between them.

The yellow group, mainly composed by *Alternaria* toxins, is widely distributed in the biplot. This is in accordance with the fact that this family presents an enormous diversity and that is reflected by the molecular descriptors. All the other families are well separated, with only a few mycotoxins in one or another family that are more distant, for example, BAVR, ENF, MLNF and EMET, which also happened in the clustering results. ENF is away from its family probably due to a different sum of atomic polarizabilities, reflected by *apol* descriptor; and it is possible that BAVR has more basic groups (*nBase* descriptor) or different MDE between secondary and tertiary carbons (*MDEC,23*) than Enniatins. EMET, a ergot alkaloid, may be influenced by *khs.ssO* or *C4SP3* descriptors that distance this mycotoxin from the other ergot alkaloids and particularly discriminate the trichothecenes' family.

However, these descriptors do not seem to have significant impact on the discrimination of ochratoxins, present in the origin of the biplot, and enniatins (pink), being one possible reason for that the low variance recovery (38%) of this PCA with the two principal components.

Figure 21 shows mycotoxins labelled into acutely toxic and non-acutely toxic to understand if these descriptors may lead to any toxicity information about these molecules.

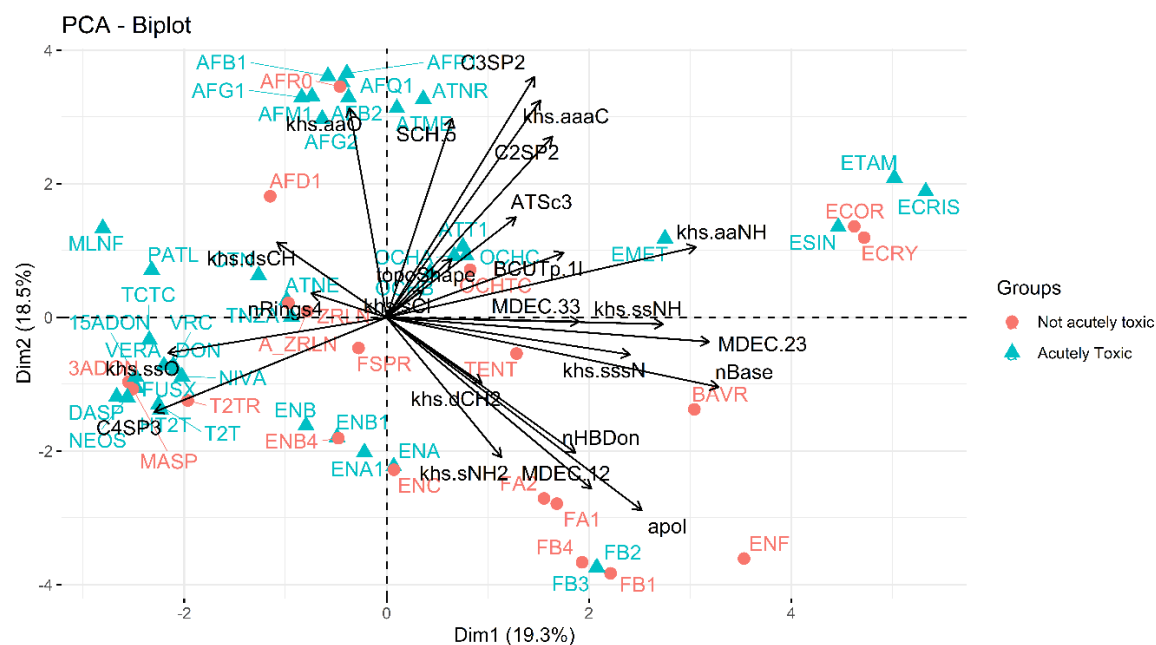


Figure 21. Biplot representation of 59 mycotoxins described by 24 molecular descriptors on the first two principal components, recovering 37.8% of the variance. Mycotoxins are colored according to their acute toxicity (see Figure 8).

On the negative side of the principal component (PC1) are present 35 of the 59 mycotoxins and on the positive side 24. Of the 35 on the left side of the biplot, 9 are non-acutely toxic meaning that only 26 % of the molecules discriminated by a negative impact on PC1 are non-acutely toxic. In contrast, the positive side of PC1 contains 46% (11 out of 24) of non-acutely toxic mycotoxins.

To evaluate if it is possible that discrimination through PC1 may contain information on toxicity profiles, to this dataset were added 11 biological activity related molecular descriptors, thus forming dataset B. One of them, *nHBDon*, was already present from the feature selection process. Exactly the same analysis was performed.

## Dataset B

Table 8. Eigenvalues and evolution of the percentage of information recovery in relation to the number of principal components for the dataset composed by 59 mycotoxins and 35 molecular descriptors. The most relevant descriptors are underlined and correspond to eigenvalues greater than 1.

	<b>Eigenvalue</b>	<b>Variance %</b>	<b>Cumulative variance %</b>
<b>Dim. 1</b>	<u>10.9</u>	<u>31.1</u>	<u>31.1</u>
<b>Dim. 2</b>	<u>5.92</u>	<u>16.9</u>	<u>48.0</u>
<b>Dim. 3</b>	<u>4.54</u>	<u>12.9</u>	<u>60.9</u>
<b>Dim. 4</b>	<u>2.80</u>	<u>8.01</u>	<u>68.9</u>
<b>Dim. 5</b>	<u>2.33</u>	<u>6.65</u>	<u>75.6</u>
<b>Dim. 6</b>	<u>1.76</u>	<u>5.02</u>	<u>80.6</u>
<b>Dim. 7</b>	<u>1.44</u>	<u>4.12</u>	<u>84.8</u>
<b>Dim. 8</b>	<u>1.13</u>	<u>3.24</u>	<u>87.9</u>
<b>Dim. 9</b>	0.953	2.72	90.7
<b>Dim.10</b>	0.778	2.22	92.9
<b>Dim.11</b>	0.590	1.69	94.6
<b>Dim.12</b>	0.505	1.44	96.1
<b>Dim.13</b>	0.328	0.938	97.0
<b>Dim.14</b>	0.241	0.689	97.7
<b>Dim.15</b>	0.212	0.606	98.3
<b>Dim.16</b>	0.153	0.436	98.7
<b>Dim.17</b>	0.121	0.346	99.1
<b>Dim.18</b>	0.0799	0.228	99.3
<b>Dim.19</b>	0.0684	0.195	99.5
<b>Dim.20</b>	0.0404	0.115	99.6
<b>Dim.21</b>	0.0362	0.103	99.7
<b>Dim.22</b>	0.0281	0.0802	99.8
<b>Dim.23</b>	0.0164	0.0470	99.8
<b>Dim.24</b>	0.0153	0.0437	99.9
<b>Dim.25</b>	0.0113	0.0323	99.9
<b>Dim.26</b>	0.00888	0.0254	99.9
<b>Dim.27</b>	0.00566	0.0162	99.9
<b>Dim.28</b>	0.00362	0.0103	99.9
<b>Dim.29</b>	0.00323	0.00924	99.9
<b>Dim.30</b>	0.00144	0.00411	100
<b>Dim.31</b>	0.000909	0.00259	100
<b>Dim.32</b>	0.000394	0.00112	100
<b>Dim.33</b>	5.06e-06	1.44e-05	100
<b>Dim.34</b>	9.32e-08	2.66e-07	100

<b>Dim.35</b>	8.58e-32	2.45e-31	100
---------------	----------	----------	-----

In fact, PCA on dataset B recovered more information variability, with 48% on the first two components and 61 % with a third component. The increase over the previous dataset was 10% in the main components. It is plausible to believe that biological activity related descriptors are a good contribution to the system. Molecular descriptors' contributions for dataset B are shown in Figure 22.

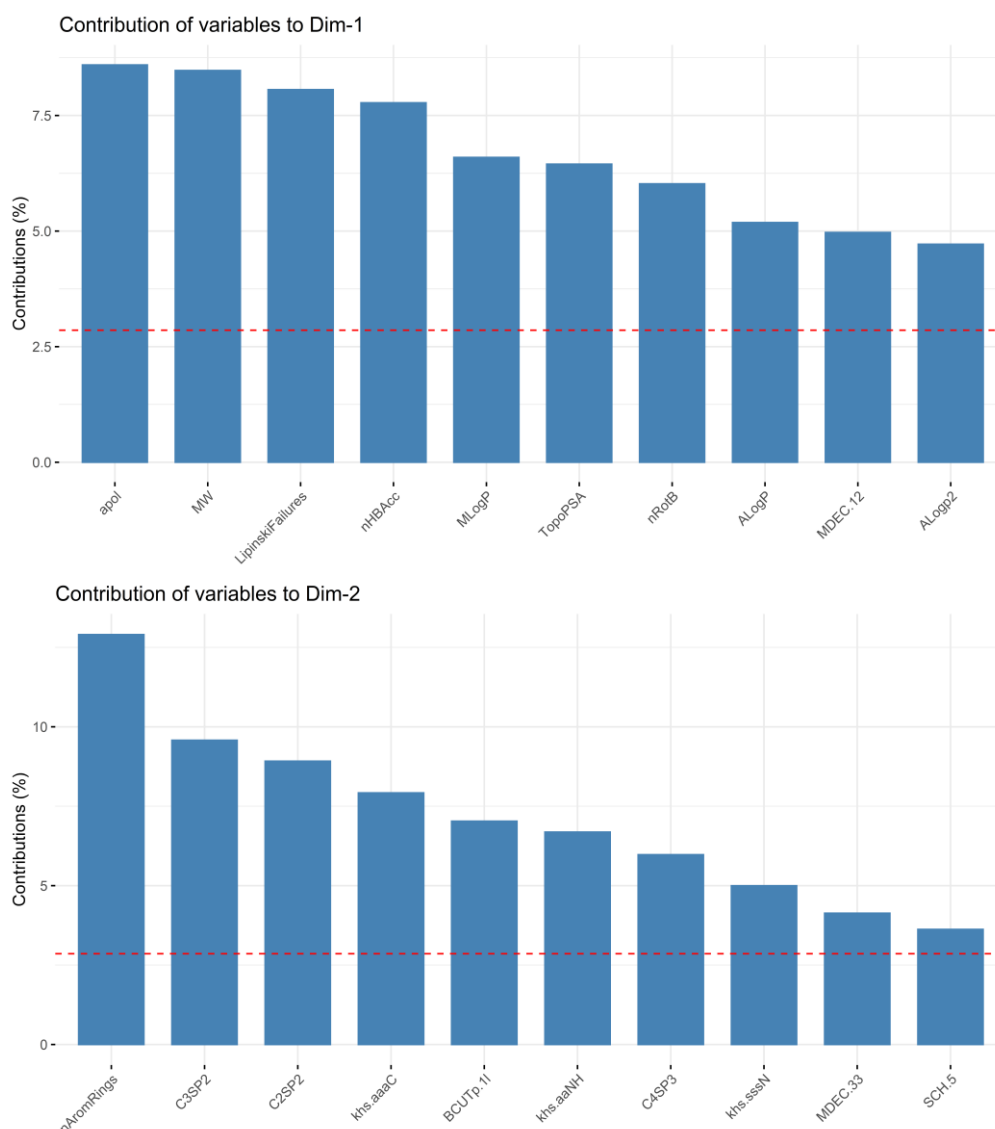


Figure 22. Graphical representation of the impact of molecular descriptors (*loadings*) on the first two principal components (correlation matrix) for the dataset composed by 59 mycotoxins and 35 molecular descriptors. The top and the bottom plots refer to PC1 and PC2, respectively.

From the descriptors with highest contributions on PC1, only *apol* and *MDEC.12* do not belong to the “biological activity group of descriptors”. Although *apol* is the descriptor that contributes the most, it is very close to MW, Lipinski Failures and *nHBAcc*. On the second





## Dataset C

Table 9 shows the resulting eigenvalues and variance for dataset C.

Table 9. Eigenvalues and evolution of the percentage of information recovery in relation to the number of principal components for the dataset composed by 59 mycotoxins and 28 molecular descriptors. The most relevant descriptors are underlined and correspond to eigenvalues greater than 1.

	<b>Eigenvalue</b>	<b>Variance %</b>	<b>Cumulative variance %</b>
<b>Dim. 1</b>	<u>5.34</u>	<u>19.08</u>	<u>19.08</u>
<b>Dim. 2</b>	<u>4.53</u>	<u>16.2</u>	<u>35.3</u>
<b>Dim. 3</b>	<u>4.28</u>	<u>15.3</u>	<u>50.5</u>
<b>Dim. 4</b>	<u>2.89</u>	<u>10.3</u>	<u>60.9</u>
<b>Dim. 5</b>	<u>2.24</u>	<u>8.01</u>	<u>68.9</u>
<b>Dim. 6</b>	<u>1.65</u>	<u>5.88</u>	<u>74.6</u>
<b>Dim. 7</b>	<u>1.52</u>	<u>5.42</u>	<u>80.2</u>
<b>Dim. 8</b>	<u>1.03</u>	<u>3.67</u>	<u>83.8</u>
<b>Dim. 9</b>	0.959	3.42	87.3
<b>Dim. 10</b>	0.698	2.49	89.8
<b>Dim. 11</b>	0.605	2.16	91.9
<b>Dim. 12</b>	0.523	1.87	93.8
<b>Dim. 13</b>	0.400	1.43	95.2
<b>Dim. 14</b>	0.316	1.13	96.3
<b>Dim. 15</b>	0.260	0.931	97.3
<b>Dim. 16</b>	0.197	0.702	97.9
<b>Dim. 17</b>	0.170	0.608	98.6
<b>Dim. 18</b>	0.126	0.449	99.0
<b>Dim. 19</b>	0.0921	0.329	99.4
<b>Dim. 20</b>	0.0599	0.214	99.6
<b>Dim. 21</b>	0.0476	0.170	99.7
<b>Dim. 22</b>	0.0270	0.0965	99.8
<b>Dim. 23</b>	0.0168	0.0599	99.9
<b>Dim. 24</b>	0.0102	0.0363	99.9
<b>Dim. 25</b>	0.00804	0.0287	99.9
<b>Dim. 26</b>	0.00560	0.0200	99.9
<b>Dim. 27</b>	0.00304	0.0108	100
<b>Dim. 28</b>	2.66e-31	9.49e-31	100



PC1 and PC2 attain ~35% of variance and a third component 50%. With this dataset, eight principal components are needed to reach 80% of the variance where both dataset A and B needed seven or six, respectively. The only differences between datasets A and C are that dataset C contains more three *BCUT* descriptors and considers *PetitJeanNumber* descriptor. Molecular descriptors contributions are shown in Figure 25.

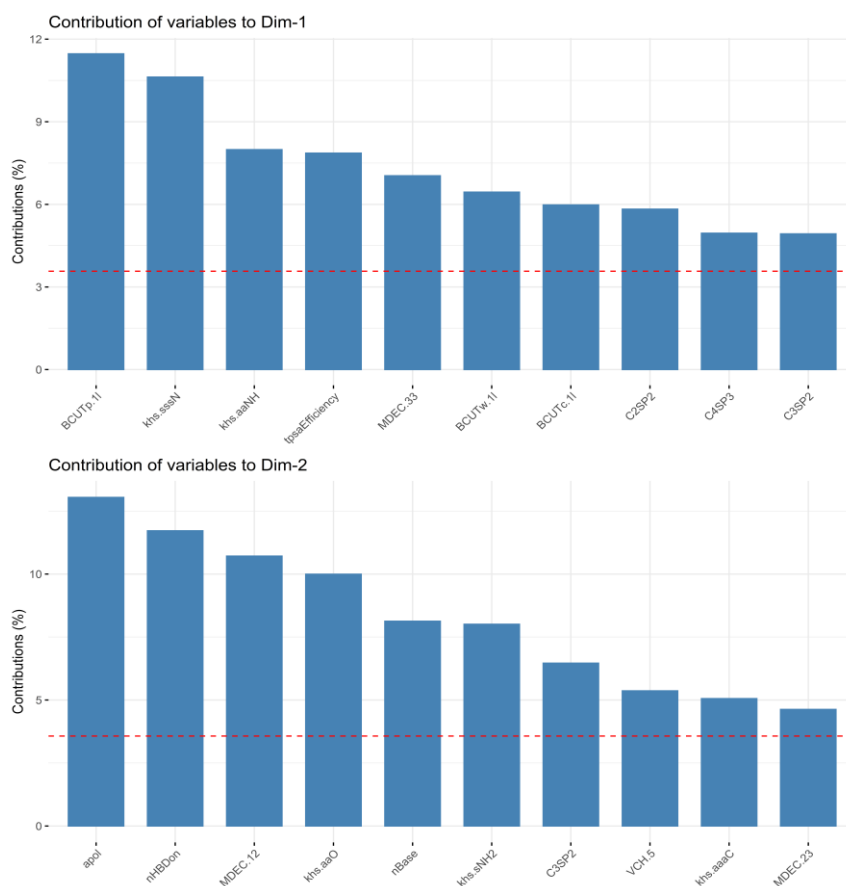


Figure 25. Graphical representation of the impact of molecular descriptors (*loadings*) on the first two principal components (correlation matrix) for the dataset composed by 59 mycotoxins and 28 molecular descriptors. The top and the bottom plots refer to PC1 and PC2, respectively.

Topological descriptors continue to dominate PC1 with *apol*, *nHBDon* and *MDEC\_12*, (electronic, constitutional, and topological) dominating the second component.

The PCA biplot for dataset C is displayed in Figure 26.

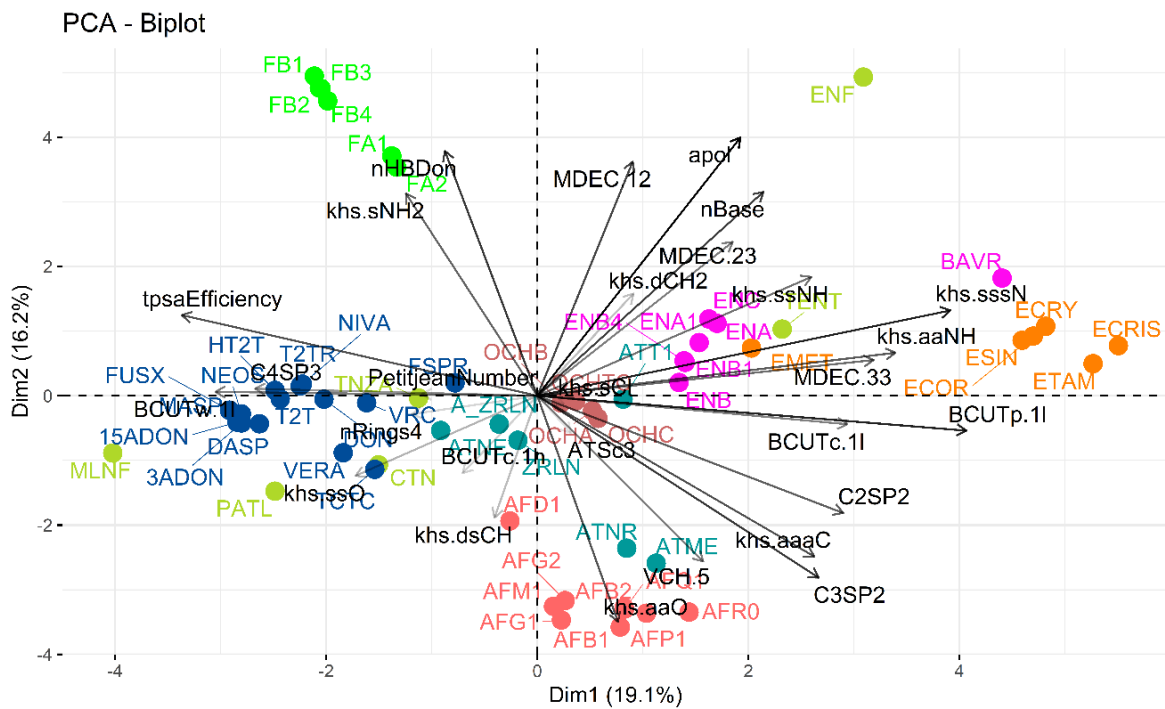


Figure 26. Biplot representation of 59 mycotoxins described by 28 molecular descriptors on the first two principal components, recovering 35.3% of the variance. Mycotoxins are colored according to the clustering results (see Figure 13) and the molecular descriptors resulted from feature selection method 1.

This biplot confirms what was seen in the previous datasets. To add to *C4SP3* descriptor, *PetitJeanNumber* and *tps*a descriptors appear to have some influence on the trichothecenes family. Ochratoxins persist in the origin of the biplot without any remarkable influence, despite the *C2SP2* descriptor has shown a very slight impact in this family, also visible in both datasets A and B. Biplot with acute toxicity labels is presented in Figure 27.

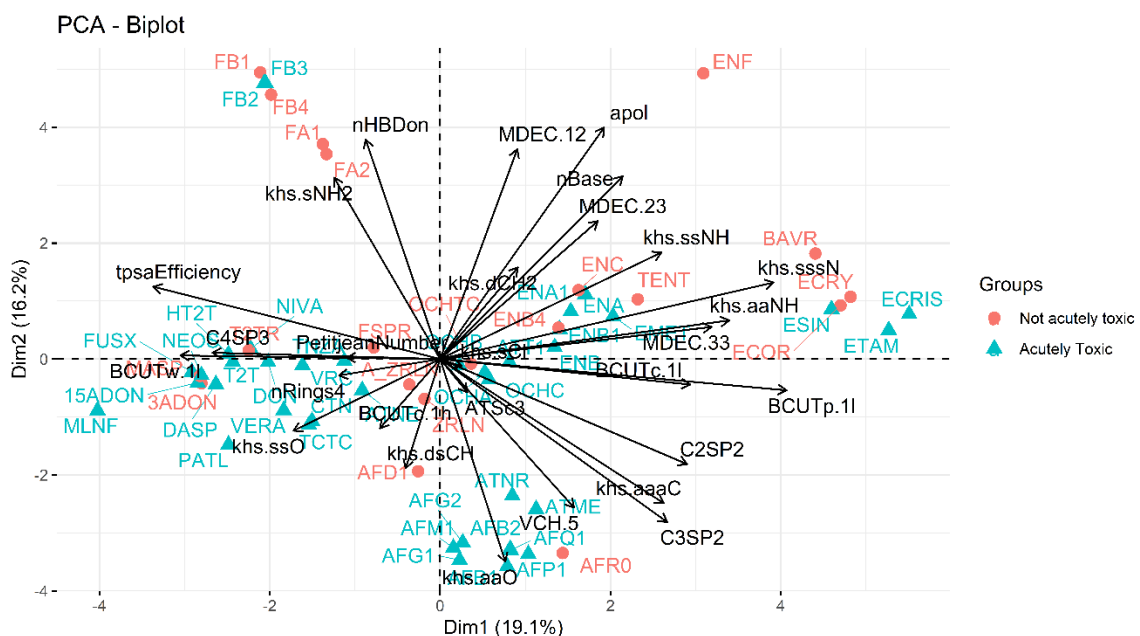


Figure 27. Biplot representation of 59 mycotoxins described by 28 molecular descriptors on the first two principal components, recovering 35.3% of the variance. Mycotoxins are colored according to their acute toxicity (see Figure 8).

In this biplot it seems that the toxicity profile is now over the second component (PC2). The positive side of PC2 contains 25 mycotoxins, 13 of which are non-acutely toxic, in other words 52%. On the negative side, only 20% of the mycotoxins are non-acutely toxic (7 in 24).

To this dataset were added biological activity descriptors, constituting dataset D.

### Dataset D

The obtained eigenvalues and variances are presented in Table 10.

Table 10. Eigenvalues and evolution of the percentage of information recovery in relation to the number of principal components for the dataset composed by 59 mycotoxins and 39 molecular descriptors. The most relevant descriptors are underlined and correspond to eigenvalues greater than 1.

	<b>Eigenvalue</b>	<b>Variance %</b>	<b>Cumulative variance %</b>
<b>Dim. 1</b>	<u>11.0</u>	<u>28.2</u>	<u>28.2</u>
<b>Dim. 2</b>	<u>6.91</u>	<u>17.7</u>	<u>45.9</u>
<b>Dim. 3</b>	<u>5.07</u>	<u>13.0</u>	<u>58.9</u>
<b>Dim. 4</b>	<u>3.32</u>	<u>8.51</u>	<u>67.5</u>
<b>Dim. 5</b>	<u>2.72</u>	<u>6.98</u>	<u>74.5</u>
<b>Dim. 6</b>	<u>1.79</u>	<u>4.60</u>	<u>79.1</u>
<b>Dim. 7</b>	<u>1.64</u>	<u>4.19</u>	<u>83.3</u>
<b>Dim. 8</b>	<u>1.16</u>	<u>2.99</u>	<u>86.3</u>
<b>Dim. 9</b>	<u>1.00</u>	<u>2.57</u>	<u>88.8</u>
<b>Dim. 10</b>	0.86	2.21	91.0
<b>Dim. 11</b>	0.63	1.61	92.7
<b>Dim. 12</b>	0.62	1.58	94.2
<b>Dim. 13</b>	0.49	1.27	95.5
<b>Dim. 14</b>	0.348	0.892	96.4
<b>Dim. 15</b>	0.325	0.833	97.2
<b>Dim. 16</b>	0.267	0.686	97.9
<b>Dim. 17</b>	0.195	0.499	98.4
<b>Dim. 18</b>	0.156	0.394	98.8
<b>Dim. 19</b>	0.114	0.292	99.1

<b>Dim. 20</b>	0.0829	0.213	99.3
<b>Dim. 21</b>	0.0749	0.192	99.5
<b>Dim. 22</b>	0.0490	0.126	99.6
<b>Dim. 23</b>	0.0349	0.0894	99.7
<b>Dim. 24</b>	0.0310	0.0796	99.8
<b>Dim. 25</b>	0.0208	0.0534	99.8
<b>Dim. 26</b>	0.0160	0.0411	99.9
<b>Dim. 27</b>	0.0109	0.0279	99.9
<b>Dim. 28</b>	0.00878	0.0225	99.9
<b>Dim. 29</b>	0.00668	0.0171	99.9
<b>Dim. 30</b>	0.00509	0.0131	99.9
<b>Dim. 31</b>	0.00383	0.00982	99.9
<b>Dim. 32</b>	0.00199	0.00511	99.9
<b>Dim. 33</b>	0.00134	0.00344	99.9
<b>Dim. 34</b>	0.00105	0.00270	99.9
<b>Dim. 35</b>	0.000457	0.00117	99.9
<b>Dim. 36</b>	0.000248	0.000637	99.9
<b>Dim. 37</b>	4.55e-06	1.17e-05	100
<b>Dim. 38</b>	7.69e-08	1.97e-07	100
<b>Dim. 39</b>	1.56e-31	4.01e-31	100

These results are very similar to those on dataset B and as it happened with dataset A and B, when biological activity descriptors were added, and compared to dataset C, the explained variance increased by 10% in the first two components, reaching ca. 46%. The third component increased information variability to 59%.

The contributions of the variables are displayed in the next Figure.

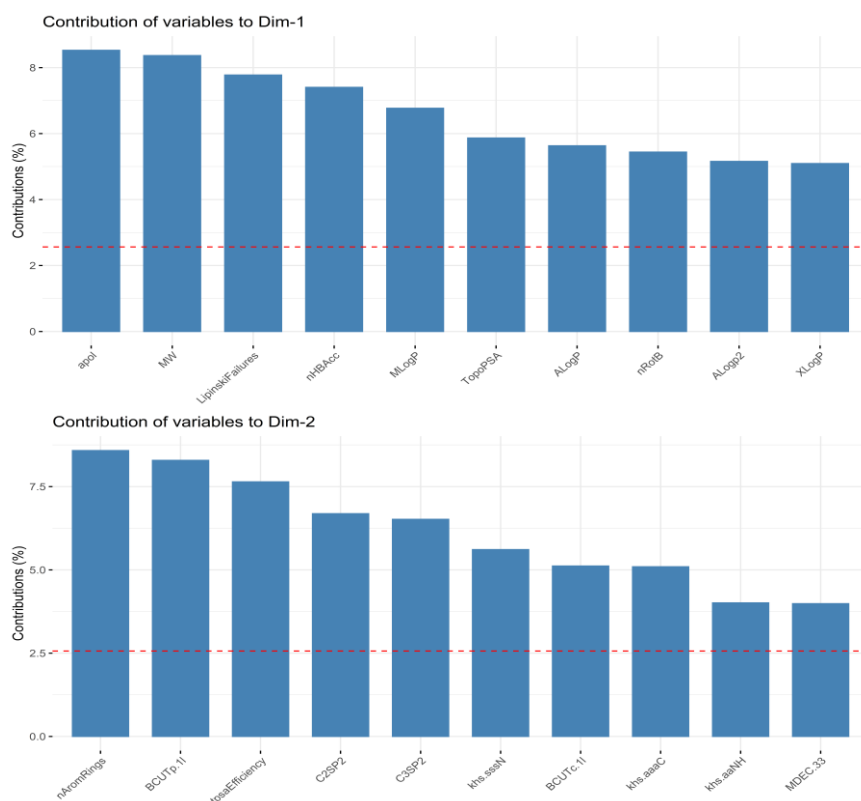


Figure 28. Graphical representation of the impact of molecular descriptors (*loadings*) on the first two principal components (correlation matrix) for the dataset composed by 59 mycotoxins and 39 molecular descriptors. The top and the bottom plots refer to PC1 and PC2, respectively.

Similarly to what happened from dataset A to dataset B, on the first component are only biological activity descriptors, with the exception of *apol*. *nAromRings* is again the descriptor with the highest contribution on PC2, followed by BCUTp.11, *tpsaEfficiency* and carbon types descriptors (*C2SP2* and *C3SP2*). Considering the explained variance and descriptors contributions in both datasets B and D, it is expectable that the PCA biplot of dataset D may be identical to dataset B.

PCA biplot for dataset D is shown below, Figure 29.

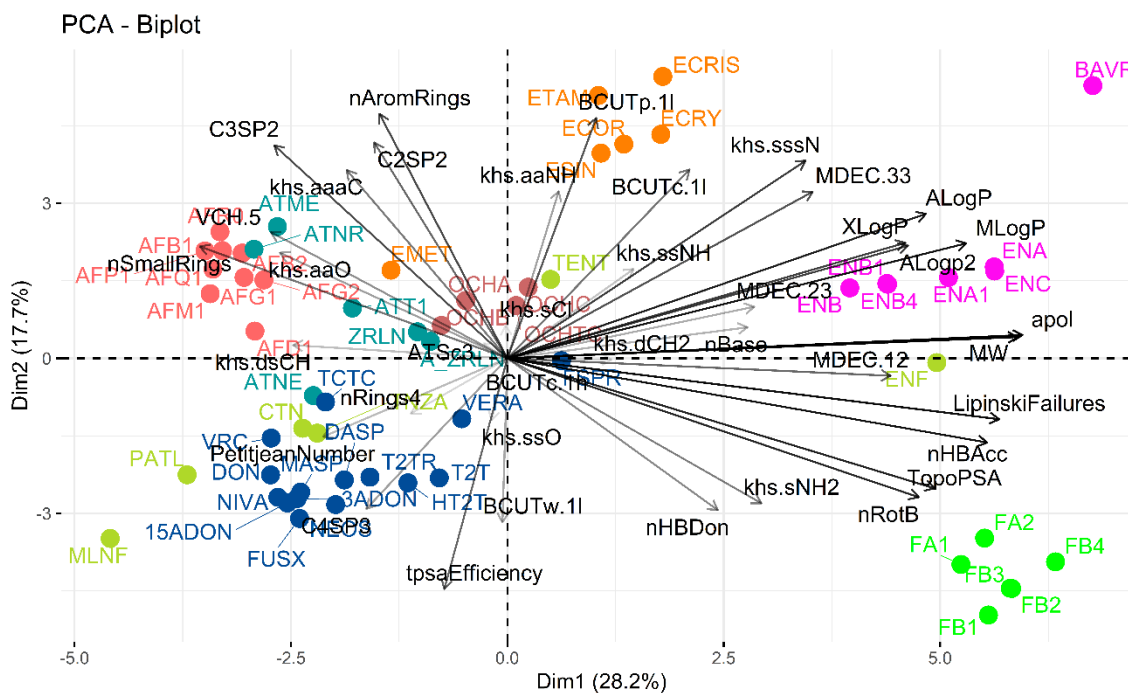


Figure 29. Biplot representation of 59 mycotoxins described by 39 molecular descriptors on the first two principal components, recovering 45.9% of the variance. Mycotoxins are colored according to the clustering results (see Figure 13) and the molecular descriptors resulted from feature selection method 1 with the addition of biological-activity related descriptors.

All that was said to dataset B in the PCA biplot is applied here. Biplots are very similar, which means that it is not worthy to use feature selection method 1 (see Figure 9) because it yields lower information recovery and requires more descriptors without conveying additional information. There is no necessity to show the same biplot only with different labels because the result would be the same.

Finally, to confirm if the biological activity descriptors really have the capacity to discriminate mycotoxins according to their toxicity profile, dataset E was constructed, only with those descriptors.

### Dataset E

As said before, this dataset is composed by a few descriptors that are: *nHBDon*, *nHBAcc*, *MW*, *XLogP*, *LipinskiFailures*, *nRotB*, *ALogP*, *MLogP*, *Alogp2*, *nSmallRings*, *nAromRings* and *TopoPSA*. The eigenvalues and explained variance are shown in Table 11.

Table 11. Eigenvalues and evolution of the percentage of information recovery in relation to the number of principal components for the dataset composed by 59 mycotoxins and 12 molecular descriptors. The most relevant descriptors are underlined and correspond to eigenvalues greater than 1.

	<b>Eigenvalue</b>	<b>Variance %</b>	<b>Cumulative variance %</b>
<b>Dim. 1</b>	<u>7.01</u>	<u>58.4</u>	<u>58.4</u>
<b>Dim. 2</b>	<u>2.54</u>	<u>21.2</u>	<u>79.6</u>
<b>Dim. 3</b>	<u>1.18</u>	<u>9.84</u>	<u>89.4</u>
<b>Dim. 4</b>	0.729	6.08	95.5
<b>Dim. 5</b>	0.192	1.59	97.1
<b>Dim. 6</b>	0.125	1.04	98.2
<b>Dim. 7</b>	0.0741	0.618	98.8
<b>Dim. 8</b>	0.0717	0.597	99.4
<b>Dim. 9</b>	0.0319	0.267	99.6
<b>Dim. 10</b>	0.0250	0.208	99.8
<b>Dim. 11</b>	0.0154	0.128	99.9
<b>Dim. 12</b>	0.00299	0.0249	100

With less principal components, the information recovered is, naturally, much higher, since the dimension of the system is lower. The first two components attain ~80% of the variability and a third component reaches 89%. The contributions are shown in Figure 30.

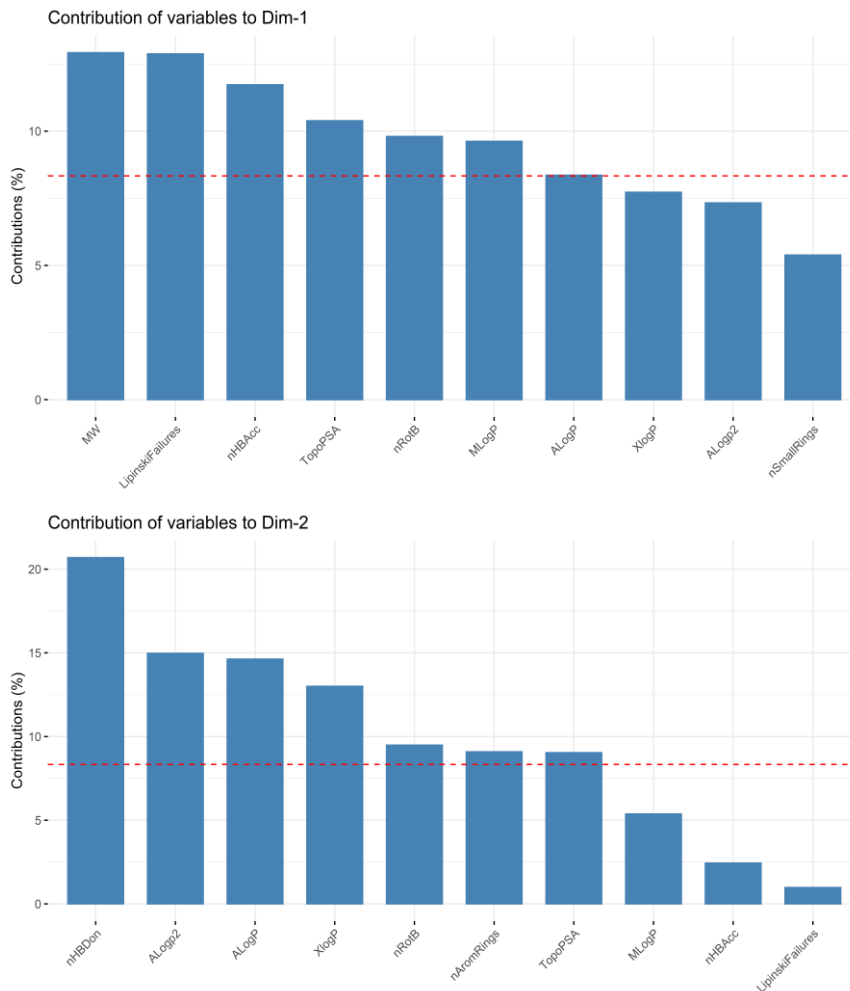


Figure 30. Graphical representation of the impact of molecular descriptors (*loadings*) on the first two principal components (correlation matrix) for the dataset composed by 59 mycotoxins and 12 biological activity descriptors. The top and the bottom plots refer to PC1 and PC2, respectively.

On the first component, *MW*, *LipinskiFailures*, *nHBAcc* and *TopoPSA* give a higher contribution to the system. On the second component, *NHBDon*, *LogP*-related descriptors contribute the most. The impact of these descriptors on the discrimination profile of the 59 mycotoxins was assessed and is present in Figure 31.

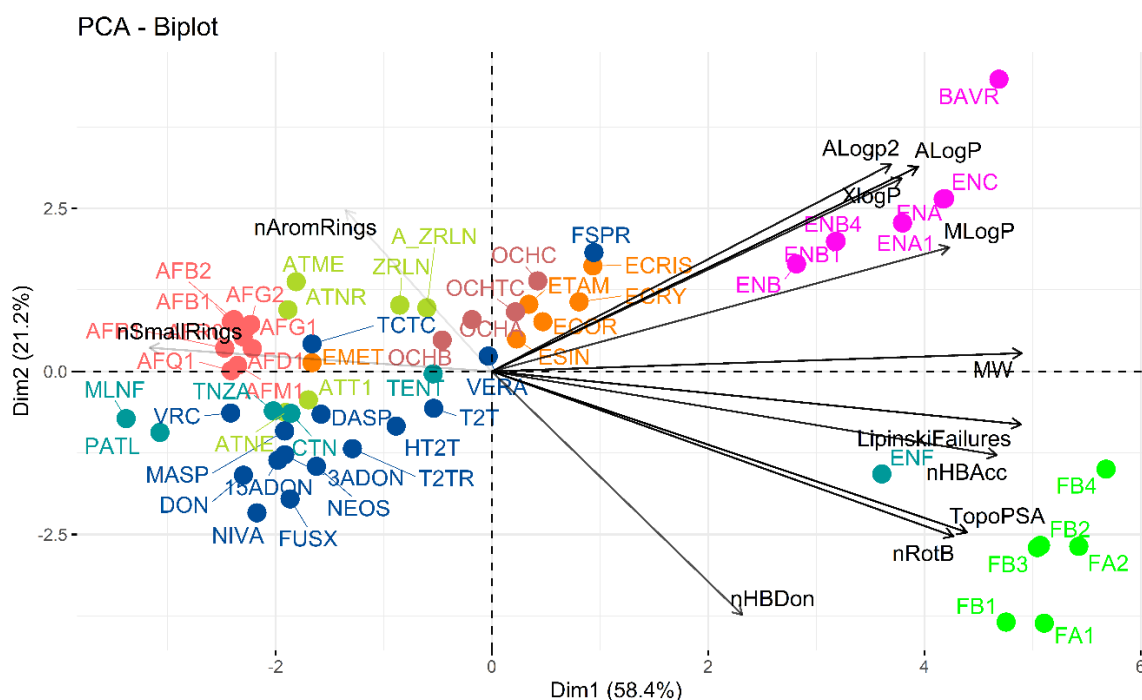


Figure 31. Biplot representation of 59 mycotoxins described by 12 molecular descriptors related to biological activity on the first two principal components, recovering 79.5% of the variance. Mycotoxins are colored according to the clustering results (see Figure 13).

The first noticeable aspect is that the families are no longer discriminated as seen in the previous datasets, with the exception of enniatins (pink) and fumonisins (green) that are separated along the second component. This suggests that constitutional descriptors, with the exception on *apol*, *HAcc* and *HBDOn* that are electronic, are not sufficient to discriminate families and that topological descriptors are necessary for that discrimination to happen. It seems that mycotoxins belonging to fumonisins and enniatins are on the same side of the biplot mainly because of their molecular weight, which is in accordance with the molecular similarity results. *AlogP* is the molecular descriptor responsible for the enniatins position on the positive side of PC2 and *TopoPSA* for the position of fumonisins on the negative side of PC2.

However, the main goal of this work is to find out relationships between molecular descriptors and acute toxicity, so that those descriptors can be use in a classification model.



The biplot with the mycotoxins labelled according to their acute toxicity is presented in Figure 32.



Figure 32. Biplot representation of 59 mycotoxins described by 12 molecular descriptors related to biological activity on the first two principal components, recovering 79.5% of the variance. Mycotoxins are colored according to their acute toxicity (see Figure 8).

There are 22 mycotoxins on the positive side of PC1 and 37 on the negative side. From the 37 mycotoxins on the negative side, only 8 are non-acutely toxic, meaning that 21,6% of those mycotoxins are non-acutely toxic. On the other side, 12 of 22, ~54% of the mycotoxins are non-acutely toxic. The positive side of the PC1 has less mycotoxins than the negative side and has more than double of non-acutely toxic mycotoxins.

However, this information was already obtained with dataset B. This means that to construct a model, it is better to use dataset B because not only those descriptors are able to give some information about the toxicity profile of the mycotoxins, but also because they can discriminate the families, ending up representing the system better, even though the variability recovery is lower.

### 4.3 Toxicity Prediction

So far it has been seen that the biological activity reflected by some of the molecular descriptors do not contribute significantly to the discrimination of the mycotoxins families. In other words, all biplots, including the dataset composed by 30 mycotoxins, showed, in general, a good discrimination between mycotoxin families. Regarding mycotoxins acute toxicity, the dataset composed by 30 mycotoxins showed no relationship between mycotoxins positioning on the biplot

and their acute toxicity. However, all the other datasets showed that there is some tendency for more toxic mycotoxins to gather on one side of PC1, and less toxic on the other.

To construct the predictive models, classes 0 (non-acutely toxic) and 1 (acutely toxic) were assigned to each mycotoxin (see Figure 4). The major problem at this stage was the fact that there are significantly more acutely toxic mycotoxins than non-acutely toxic, which can result later in an imbalanced distribution of classes in the training and test set.

To remember, all datasets were scaled, and the ratio train/test was 70/30.

#### 4.3.1 Linear Discriminant Analysis

Table 12 summarizes the LDA results for each dataset.

Table 12. Performance evaluation metrics for the LDA models. The dataset with the best performance is underlined.

	<b>30 Mycotoxins</b>	<b>Dataset A</b>	<b>Dataset B</b>	<b>Dataset C</b>	<b>Dataset D</b>	<b>Dataset E</b>
<b>Train/test: 0.7/0.3</b>						
<b>Accuracy</b>						
Train	<u>0.91</u>	0.88	0.98	0.95	0.98	0.83
Test	<u>0.75</u>	0.24	0.18	0.59	0.53	0.59
<b>Recall</b>						
Train	<u>0.94</u>	0.96	0.96	0.96	0.96	0.93
Test	<u>0.83</u>	0.36	0.18	0.73	0.55	0.82
<b>Precision</b>						
Train	<u>0.94</u>	0.87	1.00	0.96	1.00	0.84
Test	<u>0.83</u>	0.40	0.29	0.67	0.67	0.64
<b>Specificity</b>						
Train	<u>0.83</u>	0.71	1.00	0.93	1.00	0.64
Test	<u>0.50</u>	0.00	0.17	0.33	0.50	0.17
<b>AUROC</b>						
Train	<u>0.89</u>	0.84	0.98	0.95	0.98	0.79
Test	<u>0.67</u>	0.18	0.17	0.53	0.52	0.49

Overall, the LDA model shows overfitting for all datasets. Overfitting is less pronounced for the dataset composed by 30 mycotoxins and more pronounced for datasets A and B, with 24 and 35 molecular descriptors, respectively. It is also visible that for dataset A the specificity is 0. The positive class when building the LDA model was 1, being the mycotoxin acutely toxic. Since the specificity is the true negative rate (see Equation 20) the test set for dataset A did not contain negative class (0, not acutely toxic mycotoxins) which resulted in a specificity of 0.

LDA showed that the 30 mycotoxins dataset allowed to successfully predict the acute toxicity, followed by datasets C and D. For now, datasets A, B and E showed not to be appropriate.

It was seen before that even though dataset E recovered higher variance in PCA (Table 11) and reflects important biological activity parameters, those molecular descriptors were not sufficient to discriminate the different mycotoxin families. This can be a reason for the poor performance of the LDA model for this dataset since it does not provide a good description of the system.

### 4.3.2 Random Forest

The next model evaluated was the RF model, and the respective results are presented in Table 13.

Table 13. Performance evaluation metrics for the RF models. The dataset with the best performance is underlined.

	30 Mycotoxins	Dataset A	Dataset B	Dataset C	Dataset D	Dataset E
<b>Train/test: 70/30</b>						
<b>Accuracy</b>						
Train	1.00	1.00	1.00	<u>1.00</u>	<u>1.00</u>	1.0
Test	0.44	0.78	0.78	<u>0.83</u>	<u>0.83</u>	0.78
<b>Recall</b>						
Train	1.00	1.00	1.00	<u>1.00</u>	<u>1.00</u>	1.00
Test	0.80	1.00	0.92	<u>1.00</u>	<u>1.00</u>	0.92
<b>Precision</b>						
Train	1.00	1.00	1.00	<u>1.00</u>	<u>1.00</u>	1.00
Test	0.50	0.76	0.80	<u>0.81</u>	<u>0.81</u>	0.80
<b>Specificity</b>						
Train	1.00	1.00	1.00	<u>1.00</u>	<u>1.00</u>	1.00
Test	0.00	0.20	0.40	<u>0.40</u>	<u>0.40</u>	0.40
<b>AUROC</b>						
Train	1.00	1.00	1.00	<u>1.00</u>	<u>1.00</u>	1.0
Test	0.40	0.60	0.66	<u>0.70</u>	<u>0.70</u>	0.66

The first noticeable aspect is that RF results are much better than LDA. In this model, datasets C and D displayed the best performance, followed by datasets E and B. Some overfitting is also visible but is not so pronounced as in LDA. The specificity is low for all datasets, happening in the dataset composed by 30 mycotoxins what happened with dataset A in the LDA. Considering the dimension of all datasets, datasets C and D provided good results. Also, RF model demonstrated that the 28 molecular descriptors identified in dataset C are sufficient to achieve these performance values, since dataset D compiles those 28 and the biological activity results and the results were similar.

It is important to emphasize that since we are dealing with imbalanced datasets, looking at one metric can be misleading, and a close look should be given to all metrics. In general, few false negatives existed as it can be seen through recall metric with 0 false negatives in datasets A, C and

D. Some false positives existed in all datasets, which can be seen through the precision and specificity. The AUROC tells how good the model can distinguish correctly acutely toxic mycotoxins from non-acutely toxic mycotoxins. In this case, there is a 70% chance (for datasets C and D) for the model do correctly distinguish the classes.

### 4.3.3 Support Vector Machines

The results of SMV model are present in Table 14.

Table 14. Performance evaluation metrics for the SVM models. The dataset with the best performance is underlined.

	30 Mycotoxins	Dataset A	Dataset B	Dataset C	Dataset D	Dataset E
<b>Train/test: 70/30</b>						
<b>Accuracy</b>						
Train	<u>0.86</u>	0.88	0.90	0.90	0.90	0.74
Test	<u>0.88</u>	0.47	0.47	0.47	0.47	0.64
<b>Recall</b>						
Train	<u>1.00</u>	1.00	1.00	1.00	1.00	0.89
Test	<u>1.00</u>	0.73	0.64	0.64	0.64	0.91
<b>Precision</b>						
Train	<u>0.84</u>	0.85	0.88	0.88	0.88	0.76
Test	<u>0.86</u>	0.57	0.58	0.58	0.58	0.67
<b>Specificity</b>						
Train	<u>0.50</u>	0.64	0.71	0.71	0.71	0.43
Test	<u>0.50</u>	0.00	0.17	0.17	0.17	0.17
<b>AUROC</b>						
Train	<u>0.75</u>	0.82	0.86	0.86	0.86	0.66
Test	<u>0.75</u>	0.36	0.40	0.40	0.40	0.54

SVM results were very similar to those obtained with LDA with the dataset composed by 30 mycotoxins giving the best results. So far, all models showed that dataset A (24 molecular descriptors) is not adequate to predict the acute toxicity of mycotoxins. The 30 mycotoxins dataset does not show overfitting unlike the remaining datasets. Datasets B, C and D all produced the same results and all show overfitting. Dataset E, as happened previously, presents intermediate results, but with SVM gave better results than all the other datasets with 59 mycotoxins. Dataset E yielded better results than datasets B and D (35 and 39 molecular descriptors) since both contain those descriptors. SVM did not presented itself as a good approach to predict mycotoxins acute toxicity.

### 4.3.4 Neural Networks

Finally, a neural networks based model was constructed following a gridsearch procedure that tested various parameters and gave the combination with best accuracy for each dataset. The resulting gridsearch parameters combinations are shown in Table 15.

Table 15. Tested parameters using gridsearch that gave the best results in terms of accuracy of the NN models.

Dataset	Best Parameters
<b>30 Mycotoxins</b>	Hidden layer sizes: (5,5); initial learning rate: 0.0005; maximum number of iterations: 300; solver: “adam”
<b>Dataset A</b>	Hidden layer sizes: (5,5,5); initial learning rate: 0.01; maximum number of iterations: 300; solver: “adam”
<b>Dataset B</b>	Hidden layer sizes: (5,5,5); initial learning rate: 0.01; maximum number of iterations: 250; solver: “lbfgs”
<b>Dataset C</b>	Hidden layer sizes: (3,3,3); initial learning rate: 0.01; maximum number of iterations: 250; solver: “adam”
<b>Dataset D</b>	Hidden layer sizes: (3,3,3); initial learning rate: 0.001; maximum number of iterations: 250; solver: “sgd”
<b>Dataset E</b>	Hidden layer sizes: (3,3); initial learning rate: 0.01; maximum number of iterations: 300; solver: “sgd”

All datasets gave best results recurring to deep neural networks with the majority consisting in 3 hidden layers with 5 or 3 nodes in each one (datasets A, B, C and D, respectively, Table 15). For the smaller datasets, simpler architectures with 2 hidden layers with 5 or 3 nodes in each gave the best results (dataset composed by 30 mycotoxins and E, respectively).

Table 16. Performance evaluation metrics for the NN models. That dataset with the best performance is underlined.

	30 Mycotoxins	Dataset A	Dataset B	Dataset C	Dataset D	Dataset E
<b>Train/test: 70/30</b>						
<b>Accuracy</b>						
Train	0.90	1.00	0.88	0.95	0.63	<u>0.73</u>
Test	0.55	0.61	0.78	0.72	0.72	<u>0.78</u>
<b>Recall</b>						
Train	1.00	1.00	1.00	0.92	1.00	<u>0.96</u>
Test	1.00	0.77	0.92	0.85	1.00	<u>0.85</u>

<b>Precision</b>						
Train	0.89	1.00	0.84	1.00	0.63	<u>0.71</u>
Test	0.56	0.71	0.80	0.79	0.72	<u>0.85</u>
<b>Specificity</b>						
Train	0.50	1.00	0.67	1.00	0.00	<u>0.33</u>
Test	0.00	0.20	0.40	0.40	0.00	<u>0.60</u>
<b>AUROC</b>						
Train	0.75	1.00	<u>0.83</u>	0.96	0.50	<u>0.65</u>
Test	0.50	0.48	<u>0.66</u>	0.62	0.50	<u>0.72</u>

In contrast to all other models, dataset composed only by biological activity descriptors had the best performance using DNN. In all metrics this dataset showed a better performance in the test set than the training set, which most likely attests to the quality of the features as it excludes overfitting. On the other hand, this might lead us to consider underfitting which might be explained by the narrow hidden layers, in this particular case with only three nodes. This is due to the small overall sampling and suggests that larger amounts of information might benefit from being trained with broader hidden layer architectures. Dataset B also had better results and is visible some overfitting but not significant.

In short, the dataset composed by 30 mycotoxins provided better results with LDA and SVM models, datasets C and D worked better with the RF model and NN resulted better for datasets B and E. RF results for datasets C and D exceeded the results of all other models and dataset A is not good to perform supervised learning. The unbalanced dataset problem remains but it is a new step for the understanding and the application of supervised learning techniques on such diverse and spectacular molecules that are mycotoxins. Dataset C, that mainly differs from dataset A in the amount of BCUT descriptors, gave, in all models, better results than dataset A suggesting that those BCUT descriptors can have some impact on the toxicity prediction. The addition of the biological activity descriptors to both datasets A and C (forming datasets B and D) showed not to be significant on the supervised learning results, although the NN model for dataset E, only composed by those descriptors, gave better results. All things considered, dataset C and D gave the best results with the RF model and since dataset D has more molecular descriptors, dataset C performed better in this task.

## Chapter 5 – Conclusion and Future Perspectives

---

Algorithms based on unsupervised ML methods were constructed for different datasets, allowing to establish the correspondence between known mycotoxins in its literature-defined groups and validate the characterization of the mycotoxins in the pool of those not so well described in the literature. Some mycotoxins were not in their respective families suggesting some structural similarities with mycotoxins from different families, one aspect that may suggest possible literature misclassifications.

PCA suggested that topological molecular descriptors are pivotal on the discrimination of mycotoxins according to their families and constitutional molecular descriptor are not sufficient to represent that information. PCA also revealed that acutely toxic mycotoxins are tendentially on the same side of the biplot, indicating that some of the molecular descriptors (*C4SP3*, *khs.ssO*, *khs.dsCH*, *khs.aaO*, *nRings4*, *nSmallRings*) that separate them, may have some influence on their acute toxicity.

Supervised learning models suggested that dataset A is not a good dataset to use for the toxicity prediction of mycotoxins, but very satisfactory results were obtained with RF model for datasets C and D. Considering the size of the datasets, it is actually plausible to say that the results were good.

It is necessary to emphasize that this is a pioneering study, being the very first one compiling structural information about mycotoxins and using that information to execute ML tasks. This work also needs to be seen as an alert for the mycotoxins problem since there is information on some of their effects in some humans and animals but there are absolutely no clues about their impact on the aquatic environment. Also, the reported health effects on humans and animals are restricted to the most known mycotoxins and the consequences of the vast majority remains unknown.

Furthermore, and although ML is very powerful and widely used to reduce the cost of experiments in several areas of expertise, in this case, having meticulous experimental studies performed would be the best solution to achieve better predictive results, including in regression tasks that immediately could say, for example, not if a mycotoxin is toxic or not but how toxic could be. With the climate changes it is very likely that mycotoxin contamination could increase in several areas of the world, making it urgent to address this problem, for the sake of lives.





## References

---

1. Schwarzenbach, R. P., Escher, B. I., Fenner, K., Hofstetter, T. B., Johnson, C. A., Von Gunten, U., & Wehrli, B. (2006). The challenge of micropollutants in aquatic systems. *Science*, *313*(5790), 1072-1077.
2. Luo, Y., Guo, W., Ngo, H. H., Nghiem, L. D., Hai, F. I., Zhang, J., ... & Wang, X. C. (2014). A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment. *Science of the total environment*, *473*, 619-641.
3. Stoloff, L. (1976). Occurrence of mycotoxins in foods and feeds. *Mycotoxins and other fungal related food problems*, 23-50.
4. Richard, J. L. (2007). Some major mycotoxins and their mycotoxicoses—An overview. *International journal of food microbiology*, *119*(1-2), 3-10.
5. Pitt, J. I. (2000). Toxigenic fungi and mycotoxins. *British medical bulletin*, *56*(1), 184-192..
6. Zain, M. E. (2011). Impact of mycotoxins on humans and animals. *Journal of Saudi chemical society*, *15*(2), 129-144.
7. Agriopoulou, S., Stamatelopoulou, E., & Varzakas, T. (2020). Advances in occurrence, importance, and mycotoxin control strategies: Prevention and detoxification in foods. *Foods*, *9*(2), 137.
8. JW, B. (1987). Mycotoxins, mycotoxicoses, mycotoxicology and Mycopathologia. *Mycopathologia*, *100*(1), 3-5.
9. Bennett, J. W., & Klich, M. (2003). Mycotoxins. *16*(3), 497–516.
10. Medina, A., Akbar, A., Baazeem, A., Rodriguez, A., & Magan, N. (2017). Climate change, food security and mycotoxins: Do we know enough?. *Fungal biology reviews*, *31*(3), 143-154.
11. Cousins, K. R. (2011). Computer review of ChemDraw ultra 12.0.
12. Benkerroum, N. (2020). Aflatoxins: Producing-molds, structure, health issues and incidence in Southeast Asian and Sub-Saharan African countries. *International journal of environmental research and public health*, *17*(4), 1215.
13. Krska, R., & Crews, C. (2008). Significance, chemistry and determination of ergot alkaloids: A review. *Food Additives and Contaminants*, *25*(6), 722-731.
14. Rocha, O., Ansari, K., & Doohan, F. M. (2005). Effects of trichothecene mycotoxins on eukaryotic cells: a review. *Food additives and contaminants*, *22*(4), 369-378.
15. Dickman, K. G., & Grollman, A. P. (2010). Nephrotoxicity of Natural Products: Aristolochic Acid and Fungal Toxins.
16. Pohland, A. E., Nesheim, S., & Friedman, L. (1992). Ochratoxin A: a review (technical report). *Pure and Applied chemistry*, *64*(7), 1029-1046..
17. Escrivá, L., Oueslati, S., Font, G., & Manyes, L. (2017). Alternaria mycotoxins in food and feed: an overview. *Journal of Food Quality*, 2017.
18. Jestoi, M. (2008). Emerging Fusarium-mycotoxins fusaproliferin, beauvericin, enniatins, and moniliformin—A review. *Critical reviews in food science and nutrition*, *48*(1), 21-49..
19. Lo, Y. C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug discovery today*, *23*(8), 1538-1546.
20. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., ... & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, *18*(6), 463-477.
21. Keith, J. A., Vassilev-Galindo, V., Cheng, B., Chmiela, S., Gastegger, M., Müller, K. R., & Tkatchenko, A. (2021). Combining machine learning and computational chemistry for predictive insights into chemical systems. *arXiv preprint arXiv:2102.06321*.
22. Martin, O., Scholze, M., Ermler, S., McPhie, J., Bopp, S. K., Kienzler, A., ... & Kortenkamp, A. (2021). Ten years of research on synergisms and antagonisms in chemical mixtures: A systematic review and quantitative reappraisal of mixture studies. *Environment International*, *146*, 106206.
23. Tan, Y., Cheng, Z., Liu, Y., Gao, X., Liu, S., & Shen, Z. (2021). Quantum parameter analysis of the adsorption mechanism by freshly formed ferric hydroxide for synthetic dye

- and antibiotic wastewaters. *Chemosphere*, 280, 130577.
24. Gu, L., Zhu, T., & Chen, M. (2021). Modeling polyurethane foam (PUF)-air partition coefficients for persistent organic pollutants using linear and non-linear chemometric methods. *Journal of Environmental Chemical Engineering*, 105615.
  25. Gui, B., Xu, X., Zhang, S., Wang, Y., Li, C., Zhang, D., ... & Zhao, Y. (2021). Prediction of organic compounds adsorbed by polyethylene and chlorinated polyethylene microplastics in freshwater using QSAR. *Environmental Research*, 197, 111001.
  26. Singh, A. K., Bilal, M., Iqbal, H. M., & Raj, A. (2021). Trends in predictive biodegradation for sustainable mitigation of environmental pollutants: Recent progress and future outlook. *Science of The Total Environment*, 144561.
  27. Li, M., Mei, Q., Wei, B., An, Z., Sun, J., Xie, J., & He, M. (2021). Mechanism and kinetics of ClO<sup>-</sup>-mediated degradation of aromatic compounds in aqueous solution: DFT and QSAR studies. *Chemical Engineering Journal*, 412, 128728.
  28. Cordero, J. A., He, K., Janya, K., Echigo, S., & Itoh, S. (2021). Predicting formation of haloacetic acids by chlorination of organic compounds using machine-learning-assisted quantitative structure-activity relationships. *Journal of Hazardous Materials*, 408, 124466.
  29. Bureš, M. S., Ukić, Š., Cvetnić, M., Prevarić, V., Markić, M., Rogošić, M., ... & Bolanča, T. (2021). Toxicity of binary mixtures of pesticides and pharmaceuticals toward *Vibrio fischeri*: Assessment by quantitative structure-activity relationships. *Environmental Pollution*, 275, 115885.
  30. Voigt, M., & Jaeger, M. (2021). Structure and QSAR analysis of photoinduced transformation products of neonicotinoids from EU watchlist for ecotoxicological assessment. *Science of The Total Environment*, 751, 141634.
  31. Bureš, M. S., Cvetnić, M., Miloloža, M., Grgić, D. K., Markić, M., Kušić, H., ... & Ukić, Š. (2021). Modeling the toxicity of pollutants mixtures for risk assessment: a review. *Environmental Chemistry Letters*, 1-27.
  32. Sabljic, A. (2001). QSAR models for estimating properties of persistent organic pollutants required in evaluation of their environmental fate and risk. *Chemosphere*, 43(3), 363-375.
  33. Miyawaki, T., Nishino, T., Asakawa, D., Haga, Y., Hasegawa, H., & Kadokami, K. (2021). Development of a rapid and comprehensive method for identifying organic micropollutants with high ecological risk to the aquatic environment. *Chemosphere*, 263, 128258.
  34. Zhu, H., Guo, W., Shen, Z., Tang, Q., Ji, W., & Jia, L. (2015). QSAR models for degradation of organic pollutants in ozonation process under acidic condition. *Chemosphere*, 119, 65-71.
  35. Zhang, G., Xie, M., Zhao, J., Wei, S., Zheng, H., & Zhang, S. (2021). Key structural features that determine the selectivity of UV/acetylacetone for the degradation of aromatic pollutants when compared to UV/H<sub>2</sub>O<sub>2</sub>. *Water Research*, 196, 117046.
  36. Kim, J. H., Gramatica, P., Kim, M. G., Kim, D., & Tratnyek, P. G. (2007). QSAR modelling of water quality indices of alkylphenol pollutants. *SAR and QSAR in Environmental Research*, 18(7-8), 729-743.
  37. Gramatica, P., Papa, E., & Sangion, A. (2018). QSAR modeling of cumulative environmental end-points for the prioritization of hazardous chemicals. *Environmental Science: Processes & Impacts*, 20(1), 38-47.
  38. Zheng, S., Li, C., & Wei, G. (2020). QSAR modeling for reaction rate constants of e<sup>-</sup>aq<sup>-</sup> with diverse organic compounds in water. *Environmental Science: Water Research & Technology*, 6(7), 1931-1938.
  39. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, 559(7715), 547-555.
  40. Mahesh, B. (2020). Machine Learning Algorithms-A Review. *International Journal of Science and Research (IJSR)*, 9, 381-386.
  41. Mann, A. K., & Kaur, N. (2013). Review paper on clustering techniques. *Global Journal of Computer Science and Technology*.
  42. Baek, S. S., Choi, Y., Jeon, J., Pyo, J., Park, J., & Cho, K. H. (2021). Replacing the internal standard to estimate micropollutants using deep and machine learning. *Water Research*, 188, 116535.

43. Cha Cha, D., Park, S., Kim, M. S., Kim, T., Hong, S. W., Cho, K. H., & Lee, C. (2020). Prediction of Oxidant Exposures and Micropollutant Abatement during Ozonation Using a Machine Learning Method. *Environmental Science & Technology*, 55(1), 709-718.
44. Ivanciuc, O. (2003). Aquatic toxicity prediction for polar and nonpolar narcotic pollutants with support vector machines. *Internet Electron. J. Mol. Des*, 2, 195-208.
45. Ai, H., Wu, X., Zhang, L., Qi, M., Zhao, Y., Zhao, Q., ... & Liu, H. (2019). QSAR modelling study of the bioconcentration factor and toxicity of organic compounds to aquatic organisms using machine learning and ensemble methods. *Ecotoxicology and environmental safety*, 179, 71-78.
46. Camardo Leggieri, M., Mazzoni, M., & Battilani, P. (2021). Machine Learning for Predicting Mycotoxin Occurrence in Maize. *Frontiers in microbiology*, 12, 782.
47. Torelli, E., Firrao, G., Bianchi, G., Saccardo, F., & Locci, R. (2012). The influence of local factors on the prediction of fumonisin contamination in maize. *Journal of the Science of Food and Agriculture*, 92(8), 1808-1814.
48. Battilani, P. (2016). Recent advances in modeling the risk of mycotoxin contamination in crops. *Current Opinion in Food Science*, 11, 10-15.
49. Artrith, N., Butler, K. T., Coudert, F. X., Han, S., Isayev, O., Jain, A., & Walsh, A. (2021). Best practices in machine learning for chemistry. *Nature Chemistry*, 13(6), 505-508.
50. Sterling, T., & Irwin, J. J. (2015). ZINC 15—ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11), 2324-2337.
51. Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., ... & Bryant, S. H. (2016). PubChem substance and compound databases. *Nucleic acids research*, 44(D1), D1202-D1213.
52. Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., ... & Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic acids research*, 45(D1), D945-D954.
53. Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... & Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46(D1), D1074-D1082.
54. Cao, D. S., Xu, Q. S., Hu, Q. N., & Liang, Y. Z. (2013). ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics*, 29(8), 1092-1094.
55. Williams, A. J. (2008). A perspective of publicly accessible/open-access chemistry databases. *Drug discovery today*, 13(11-12), 495-501.
56. Guha, R. (2007). Chemical informatics functionality in R. *Journal of Statistical Software*, 18(1), 1-16.
57. Willighagen, E. L., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliaskova, N., ... & Steinbeck, C. (2017). The Chemistry Development Kit (CDK) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of cheminformatics*, 9(1), 1-19.
58. Urbanek, S., Urbanek, M. S., & JDK, S. J. (2021). Package ‘rJava’.
59. Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R., & Willighagen, E. L. (2006). Recent developments of the chemistry development kit (CDK)-an open-source java library for chemo-and bioinformatics. *Current pharmaceutical design*, 12(17), 2111-2120.
60. Guha, R., & Cherto, M. R. (2017). rcdk: Integrating the CDK with R.
61. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2014). cluster: Cluster Analysis Basics and Extensions. R package version 2.1. 2.
62. Kassambara, A., & Mundt, F. (2017). Package ‘factoextra’. *Extract and visualize the results of multivariate data analyses*, 76.
63. Husson, F., Josse, J., Le, S., Mazet, J., & Husson, M. F. (2016). Package ‘FactoMineR’. *An R package*, 96, 698.
64. Kuhn, M. (2009). The caret package. *Journal of Statistical Software*, 28(5).
65. Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). Package ‘mass’. *Cran r*, 538, 113-120.
66. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine*

- Learning research*, 12, 2825-2830.
67. Boulicaut, J. F., & Jeudy, B. (2005). The Data Mining and Knowledge Discovery Handbook, chapter Constraint-based data mining.
  68. Cova, T. F., Pereira, J. L., & Pais, A. A. (2013). Is standard multivariate analysis sufficient in clinical and epidemiological studies?. *Journal of biomedical informatics*, 46(1), 75-86.
  69. Bajorath, J. (2001). Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *Journal of chemical information and computer sciences*, 41(2), 233-245.
  70. Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. *Methods*, 71, 58-63.
  71. Raymond, J. W., & Willett, P. (2002). Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *Journal of computer-aided molecular design*, 16(1), 59-71.
  72. Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?. *Journal of classification*, 31(3), 274-295.
  73. Saranya, T., Sridevi, S., Deisy, C., Chung, T. D., & Khan, M. A. (2020). Performance analysis of machine learning algorithms in intrusion detection system: A review. *Procedia Computer Science*, 171, 1251-1260.
  74. Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3), 128-138.
  75. Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2), 451-461.
  76. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.
  77. Kanyongo, G. Y. (2005). Determining the correct number of components to extract from a principal components analysis: A Monte Carlo study of the accuracy of the scree plot. *Journal of Modern Applied Statistical Methods*, 4(1), 13.
  78. Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
  79. Goldman, B. B., & Walters, W. P. (2006). Machine learning in computational chemistry. *Annual Reports in Computational Chemistry*, 2, 127-140.
  80. Belgiu, M., & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, 114, 24-31.
  81. Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5), 988-999.
  82. Patle, A., & Chouhan, D. S. (2013, January). SVM kernel functions for classification. In *2013 International Conference on Advances in Technology and Engineering (ICATE)* (pp. 1-9). IEEE.
  83. Kavzoglu, T., & Colkesen, I. (2009). A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, 11(5), 352-359.
  84. Karlsson, L., & Bonde, O. (2020). A Comparison of Selected Optimization Methods for Neural Networks.
  85. Sharma, S., & Sharma, S. (2017). Activation functions in neural networks. *Towards Data Science*, 6(12), 310-316.
  86. Kavzoglu, T. (1999, September). Determining optimum structure for artificial neural networks. In *Proceedings of the 25th Annual Technical Conference and Exhibition of the Remote Sensing Society* (pp. 675-682). Remote Sensing Society Nottingham UK Cardiff, UK.
  87. Kingma, D. P., & Ba, J. L. (2015, May). Adam: A method for stochastic gradient descent. In *ICLR: International Conference on Learning Representations* (pp. 1-15).
  88. Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning* (Vol. 1). Sthda.

89. Halkidi, M., & Vazirgiannis, M. (2008). A density-based cluster validity approach using multi-representatives. *Pattern Recognition Letters*, 29(6), 773-786.
90. Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data set. *Journal of statistical software*, 61(1), 1-36.
91. Beauxis-Aussalet, E., & Hardman, L. (2014). Visualization of confusion matrix for non-expert users. In *IEEE Conference on Visual Analytics Science and Technology (VAST)-Poster Proceedings*.
92. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., ... & Robin, M. X. (2021). Package 'pROC' (Vol. 56, pp. 1-71). 2012-09-10 09: 34.
93. Arlitt, R., Manion, C., Stone, R., Campbell, M., & Tumer, I. (2015, August). Using Molecular Fingerprinting to Infer Functional Similarity in Engineered Systems. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (Vol. 57175, p. V007T06A022). American Society of Mechanical Engineers.
94. Bolton, E. E., Wang, Y., Thiessen, P. A., & Bryant, S. H. (2008). PubChem: integrated platform of small molecules and biological activities. In *Annual reports in computational chemistry* (Vol. 4, pp. 217-241). Elsevier.
95. Barnard, J. M., & Downs, G. M. (1997). Chemical fragment generation and clustering software. *Journal of chemical information and computer sciences*, 37(1), 141-142.
96. Sheridan, R. P., Miller, M. D., Underwood, D. J., & Kearsley, S. K. (1996). Chemical similarity using geometric atom pair descriptors. *Journal of chemical information and computer sciences*, 36(1), 128-136.
97. Roy, K., & Mitra, I. (2012). Electrotopological state atom (E-state) index in drug design, QSAR, property prediction and toxicity assessment. *Current computer-aided drug design*, 8(2), 135-158.
98. Stanton, D. T. (1999). Evaluation and use of BCUT descriptors in QSAR and QSPR studies. *Journal of chemical information and computer sciences*, 39(1), 11-20.
99. Pearlman, R. S., & Smith, K. M. (1999). Metric validation and the receptor-relevant subspace concept. *Journal of Chemical Information and Computer Sciences*, 39(1), 28-35.
100. Serino, T., Takigawa, Y., Nakamura, S., Huang, M., Ono, N., & Kanaya, S. (2019). [Special Issue for Honor Award dedicating to Prof Kimito Funatsu] Chemoinformatics Approach for Estimating Recovery Rates of Pesticides in Fruits and Vegetables. *Journal of Computer Aided Chemistry*, 20, 92-103.
101. Petitjean, M. (1992). Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *Journal of chemical information and computer sciences*, 32(4), 331-337.
102. Liu, S., Cao, C., & Li, Z. (1998). Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance-edge (MDE) vector,  $\lambda$ . *Journal of chemical information and computer sciences*, 38(3), 387-394.
103. Bruder, M., Polo, G., & Trivella, D. B. (2020). Natural allosteric modulators and their biological targets: Molecular signatures and mechanisms. *Natural product reports*, 37(4), 488-514.
104. Butina, D. (2004). Performance of Kier-hall E-state descriptors in quantitative structure activity relationship (QSAR) studies of multifunctional molecules. *Molecules*, 9(12), 1004-1009.
105. Ray, S., Mondal, S., Ray, S. D., & Roy, P. P. (2014). Role of antioxidants on docetaxel-induced in vitro lipid peroxidation using malondialdehyde as model marker: an experimental and in silico approach. *Medicinal Chemistry Research*, 23(10), 4436-4446.
106. Flajs, D., & Peraica, M. (2009). Toxicological properties of citrinin. *Arhiv za higijenu rada i toksikologiju*, 60(4), 457.
107. Mutoh, A., Ishii, K., & Ueno, Y. (1988). Effects of radioprotective compounds and anti-inflammatory agents on the acute toxicity of trichothecenes. *Toxicology letters*, 40(2), 165-174.
108. Cope, R. B. (2018). Trichothecenes. In *Veterinary Toxicology* (pp. 1043-1053). Academic

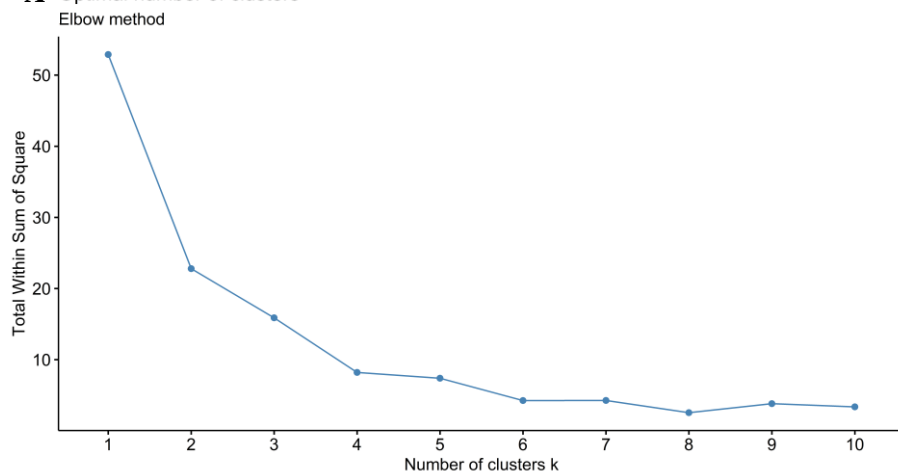
- Press.
109. Hussein, H. S., & Brasel, J. M. (2001). Toxicity, metabolism, and impact of mycotoxins on humans and animals. *Toxicology*, *167*(2), 101-134.
  110. Ueno, Y. (1980). Trichothecene mycotoxins mycology, chemistry, and toxicology. In *Advances in nutritional research* (pp. 301-353). Springer, Boston, MA.
  111. Jordan, W. H., Carlton, W. W., & Sansing, G. A. (1978). Citrinin mycotoxicosis in the rat. I. Toxicology and pathology. *Food and cosmetics toxicology*, *16*(5), 431-438.
  112. Zinedine, A., Soriano, J. M., Molto, J. C., & Manes, J. (2007). Review on the toxicity, occurrence, metabolism, detoxification, regulations and intake of zearalenone: an oestrogenic mycotoxin. *Food and chemical toxicology*, *45*(1), 1-18.
  113. Schuda, P. F., Potlock, S. J., & Wannemacher Jr, R. W. (1984). Trichothecenes, 1: the synthesis of 4-deoxyverrucarol from verrucarol and diacetoxyscirpenol. *Journal of natural products*, *47*(3), 514-519.
  114. Thompson, W. L., & Wannemacher Jr, R. W. (1986). Structure-function relationships of 12, 13-epoxytrichothecene mycotoxins in cell culture: comparison to whole animal lethality. *Toxicon*, *24*(10), 985-994.
  115. Wogan, G. N., Edwards, G. S., & Newberne, P. M. (1971). Structure-activity relationships in toxicity and carcinogenicity of aflatoxins and analogs. *Cancer Research*, *31*(12), 1936-1942.
  116. Guha, R., & Guha, M. R. (2018). Package ‘fingerprint’.
  117. Ericson, E., Gebbia, M., Heisler, L. E., Wildenhain, J., Tyers, M., Giaever, G., & Nislow, C. (2008). Off-target effects of psychoactive drugs revealed by genome-wide assays in yeast. *PLoS genetics*, *4*(8), e1000151.
  118. Xue, L., & Bajorath, J. (2000). Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *Journal of Chemical Information and Computer Sciences*, *40*(3), 801-809.
  119. Xue, Y., Li, H., Ung, C. Y., Yap, C. W., & Chen, Y. Z. (2006). Classification of a diverse set of Tetrahymena pyriformis toxicity chemical compounds from molecular descriptors by statistical learning methods. *Chemical research in toxicology*, *19*(8), 1030-1039.
  120. Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, *23*(1-3), 3-25.
  121. Scott, P. M. (2012). Recent research on fumonisins: a review. *Food additives & contaminants: part A*, *29*(2), 242-248.
  122. Ostry, V., Malir, F., Cumova, M., Kyrova, V., Toman, J., Grosse, Y., ... & Ruprich, J. (2018). Investigation of patulin and citrinin in grape must and wine from grapes naturally contaminated by strains of *Penicillium expansum*. *Food and Chemical Toxicology*, *118*, 805-811.
  123. Prandini, A., Tansini, G. I. N. O., Sigolo, S., Filippi, L. A. U. R. A., Laporta, M., & Piva, G. (2009). On the occurrence of aflatoxin M1 in milk and dairy products. *Food and chemical toxicology*, *47*(5), 984-991.
  124. Ibáñez-Vea, M., González-Peñas, E., Lizarraga, E., & De Cerain, A. L. (2012). Co-occurrence of aflatoxins, ochratoxin A and zearalenone in barley from a northern region of Spain. *Food chemistry*, *132*(1), 35-42.
  125. Assunção, R., Martins, C., Dupont, D., & Alvito, P. (2016). Patulin and ochratoxin A co-occurrence and their bioaccessibility in processed cereal-based foods: A contribution for Portuguese children risk assessment. *Food and Chemical toxicology*, *96*, 205-214.
  126. Vrabcheva, T., Usleber, E., Dietrich, R., & Märtilbauer, E. (2000). Co-occurrence of ochratoxin A and citrinin in cereals from Bulgarian villages with a history of Balkan endemic nephropathy. *Journal of agricultural and food chemistry*, *48*(6), 2483-2488.
  127. Lee, S. E., Campbell, B. C., Molyneux, R. J., Hasegawa, S., & Lee, H. S. (2001). Inhibitory effects of naturally occurring compounds on aflatoxin B1 biotransformation. *Journal of agricultural and food chemistry*, *49*(11), 5171-5177.
  128. Guo, Y., Qin, X., Tang, Y., Ma, Q., Zhang, J., & Zhao, L. (2020). CotA laccase, a novel

- aflatoxin oxidase from *Bacillus licheniformis*, transforms aflatoxin B1 to aflatoxin Q1 and epi-aflatoxin Q1. *Food chemistry*, 325, 126877.
129. Stoloff, L., Verrett, M. J., Dantzman, J., & Reynaldo, E. F. (1972). Toxicological study of aflatoxin P1 using the fertile chicken egg. *Toxicology and applied pharmacology*, 23(3), 528-531.
  130. Chu, F. S., & Wilson, B. J. (1973). Studies on ochratoxins. *CRC critical reviews in toxicology*, 2(4), 499-524.
  131. Pfeiffer, E., Schebb, N. H., Podlech, J., & Metzler, M. (2007). Novel oxidative in vitro metabolites of the mycotoxins alternariol and alternariol methyl ether. *Molecular nutrition & food research*, 51(3), 307-316.
  132. Scott, P. M. (1993). Fumonisin. *International Journal of Food Microbiology*, 18(4), 257-270.
  133. Liuzzi, V. C., Mirabelli, V., Cimmarusti, M. T., Haidukowski, M., Leslie, J. F., Logrieco, A. F., ... & Mulè, G. (2017). Enniatin and beauvericin biosynthesis in *Fusarium* species: Production profiles and structural determinant prediction. *Toxins*, 9(2), 45.
  134. Gambacorta, L., Magistà, D., Perrone, G., Murgolo, S., Logrieco, A. F., & Solfrizzo, M. (2018). Co-occurrence of toxigenic moulds, aflatoxins, ochratoxin A, *Fusarium* and *Alternaria* mycotoxins in fresh sweet peppers (*Capsicum annuum*) and their processed products. *World Mycotoxin Journal*, 11(1), 159-174.
  135. Balaban, A. T. (1998). Topological and stereochemical molecular descriptors for databases useful in QSAR, similarity/dissimilarity and drug design. *SAR and QSAR in Environmental Research*, 8(1-2), 1-21.
  136. Roy, K., Das, R. N., & Popelier, P. L. (2014). Quantitative structure–activity relationship for toxicity of ionic liquids to *Daphnia magna*: Aromaticity vs. lipophilicity. *Chemosphere*, 112, 120-127.
  137. P Patel, A. B., Shaikh, S., Jain, K. R., Desai, C., & Madamwar, D. (2020). Polycyclic aromatic hydrocarbons: sources, toxicity and remediation approaches. *Frontiers in Microbiology*, 11, 2675.
  138. Pollastri, M. P. (2010). Overview on the Rule of Five. *Current protocols in pharmacology*, 49(1), 9-12.

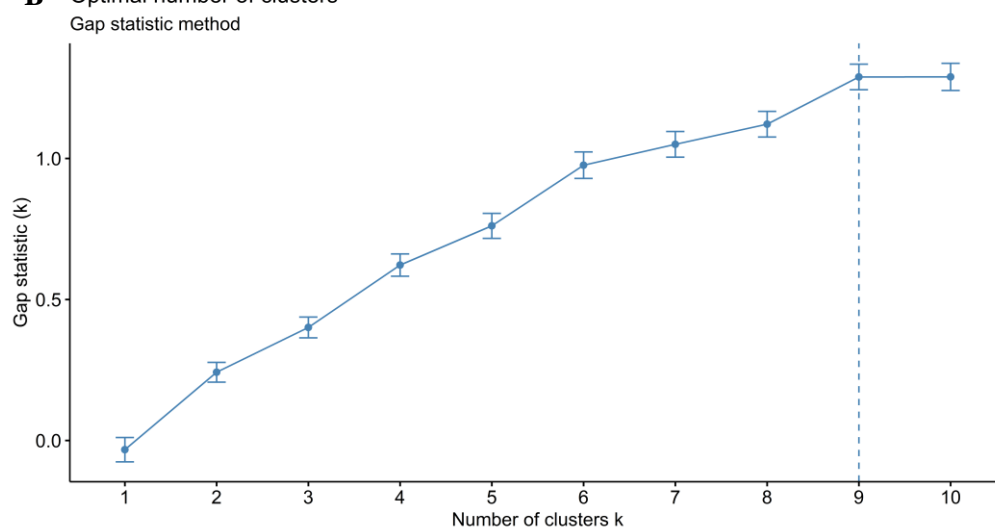
# Annexes

Annex 1. Evaluation of the optimal number of clusters using within-sum-of-squares (A), silhouette (B), and gap statistic techniques (C) using the initial 30 mycotoxins molecular fingerprints.

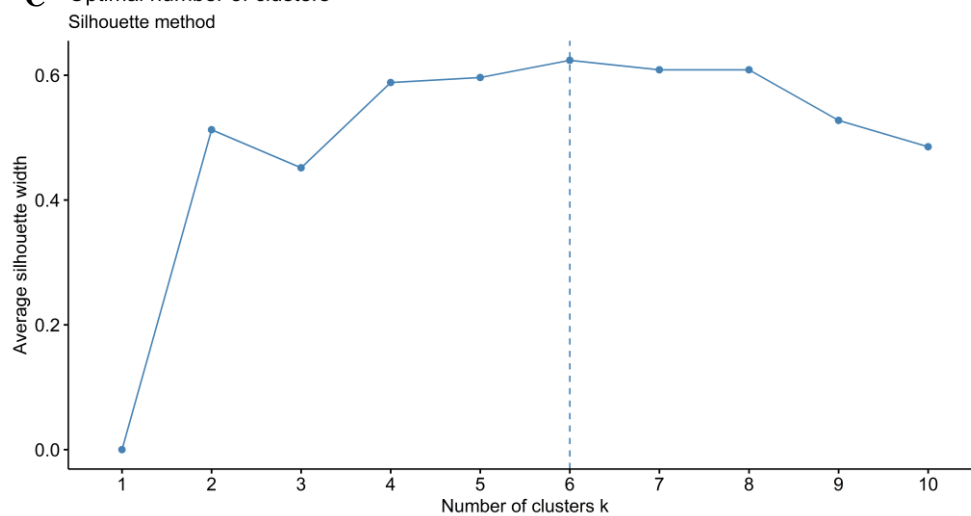
**A** Optimal number of clusters



**B** Optimal number of clusters



**C** Optimal number of clusters





Annex 2. Code implementation for all clustering results on the 30 mycotoxins molecular fingerprints.

```

#confirming cdk version
cdk.version()
#Setting my working directory
setwd("C:/Users/claude/OneDrive/Ambiente de Trabalho/LAB/mycotoxins_project/Script_mycotoxins")
getwd()

#Loading the sdf files of the mycotoxins
all= load.molecules(c("aflatoxina.sdf", "3acdon.sdf", "fumonisinab1.sdf", "fumonisinab2.sdf",
"15acdon.sdf", "fumonisinab3.sdf", "aflatoxinab2.sdf", "fusx.sdf", "aflatoxinag1.sdf",
"monoacetoscir.sdf", "aflatoxinag2.sdf", "neosol.sdf", "nivalenol.sdf", "don.sdf",
"ota.sdf", "diacetoscir.sdf", "patulin.sdf", "ergocris.sdf", "ergocor.sdf", "t2toxin.sdf",
"ht2toxin.sdf", "ergocryp.sdf", "aflatoxinam1.sdf", "t2triol.sdf", "ergomet.sdf",
"verrucarol.sdf", "ergotam.sdf", "zeanone.sdf", "zealenol.sdf", "citrinin.sdf"))

all

#calculating molecular fingerprints
fps <- lapply(all, get.fingerprint, type='standard') #hashed fingerprints
fps
#construction of the similarity matrix between molecular fingerprints
fp.sim <- fp.sim.matrix(fps, method='tanimoto')
fp.sim
#Converting the similarity matrix into a dissimilarity matrix
fp.dist <- 1 - fp.sim
fp.dist
#setting row names to the matrix
row.names(fp.dist)<- c("AFB1", "3-ADON", "FUB1", "FUB2",
"15-ADON", "FUB3", "AFB2", "FUSX", "AFG1", "MASP",
"AFG2", "NEOS", "NIVA", "DON", "OCHA",
"DASP", "PATL", "ECRI", "ECOR", "T2T", "HT2T", "ECRY", "AFM1", "T2TR", "EMET",
"VRC", "ETAM", "ZRLN", "A_ZRLN", "CTN")

#evaluating the number of optimal clusters
#within cluster sum of squares method
png(file = "onc_wss_30.png", width=9, height=5, units="in", res=1200) #creating an high resolution image
fviz_nbclust(fp.dist, kmeans, method = "wss") +
  labs(subtitle = "Elbow method")
dev.off()
#Average silhouette method
png(file = "onc_silh_30.png", width=9, height=5, units="in", res=1200)
fviz_nbclust(fp.dist, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")
dev.off()
#Gap statistic
#nboot = 50 to keep the function speedy.
png(file = "onc_gapstat_30.png", width=9, height=5, units="in", res=1200)
fviz_nbclust(fp.dist, kmeans, nstart = 2,, method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")
dev.off()

#setting the dissimilarity matrix into a distance matrix
fp.dist<- as.dist(fp.dist)
#grouping objects through the Ward's method
res.hc <- hclust(fp.dist, method = "ward.D2" )

library(factoextra)
png(file = "HCA_30.png", width=9, height=5, units="in", res=1200) #plotting HCA dendrogram
fviz_dend(res.hc, k=5, cex=0.7, color_labels_by_k = TRUE, rect = TRUE, main= "Dendrogram for mycotoxins",
k_colors = c("#00FF00", "#004c99", "#FF8000", "#FF6666", "#009999"), rect_fill= TRUE,
rect_border = c("#00FF00", "#004c99", "#FF8000", "#FF6666", "#66FFFF"), ylab = "Distance")
dev.off()

#setting the dissimilarity matrix into a distance matrix
fp.dist<- as.dist(fp.dist)
#grouping objects through the Ward's method
res.hc <- hclust(fp.dist, method = "ward.D2" )

library(factoextra)
png(file = "HCA_30.png", width=9, height=5, units="in", res=1200) #plotting HCA dendrogram
fviz_dend(res.hc, k=5, cex=0.7, color_labels_by_k = TRUE, rect = TRUE, main= "Dendrogram for mycotoxins",
k_colors = c("#00FF00", "#004c99", "#FF8000", "#FF6666", "#009999"), rect_fill= TRUE,
rect_border = c("#00FF00", "#004c99", "#FF8000", "#FF6666", "#66FFFF"), ylab = "Distance")
dev.off()

```

```

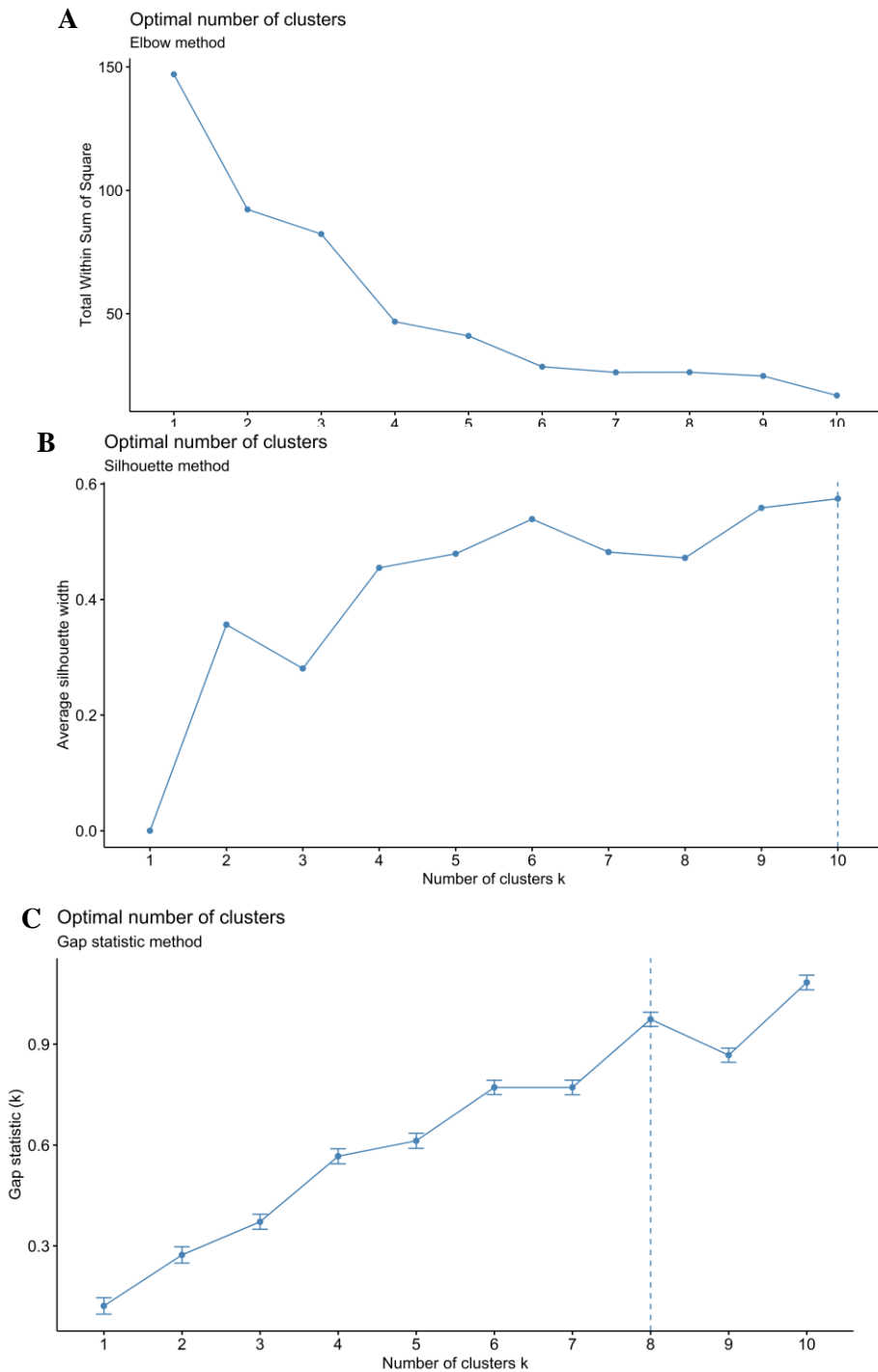
km.res <- kmeans(fp.dist, 5, nstart = 3) #performing k means
km.res
png(file = "kmeans_30.png", width=9, height=5, units="in", res=1200)
fviz_cluster(km.res, data = fp.dist, ellipse.type = "convex", repel = TRUE, #plotting k means
             palette= c( "#009999", "#FF6666", "#004c99", "#FF8000", "#00FF00" ),ggtheme = theme_minimal())
dev.off()

sil <- silhouette(km.res$cluster, dist(fp.dist)) #Evaluating kmeans with the silhouette coefficient
row.names(sil)<- c("AFB1", "3-ADON", "FUB1", "FUB2", "15-ADON", "FUB3", "AFB2", "FUSX", "AFG1", "MASP", "AFG2",
                 "NEOS", "NIVA", "DON", "OCHA", "DASP", "PATL", "ECRI", "ECOR", "T2T", "HT2T", "ECRY", "AFM1",
                 "T2TR", "EMET", "VRC", "ETAM", "ZRLN", "A_ZRLN", "CTN")

sil
png(file = "silhouette_30.png", width=9, height=5, units="in", res=1200)
plot(sil, main = "Silhouette plot: K-means", col= c( "#FF8000", "#004c99", "#009999", "#FF6666", "#00FF00" ))
dev.off()

```

Annex 3. Evaluation of the optimal number of clusters using within-sum-of-squares (A), silhouette (B), and gap statistic techniques (C) using the initial 59 mycotoxins molecular fingerprints.



Annex 4. Code implementation for all clustering results on the 59 mycotoxins molecular fingerprints.

```
rm(list=ls())
setwd("C:/Users/claud/OneDrive/Ambiente de Trabalho/LAB/mycotoxins_new/script")
#Loading molecules
all= load.molecules(c( "3adon.sdf", "15adon.sdf", "aflatoxina.sdf", "aflatoxinab2.sdf",
" aflatoxinag1.sdf", "aflatoxinag2.sdf", "aflatoxinam1.sdf",
" aflatoxind1.sdf", "aflatoxinp1.sdf", "aflatoxinq1.sdf", "aflatoxinr0.sdf",
" altenuene.sdf", "alternariol.sdf", "alternariolmethylether.sdf",
" altertoxin1.sdf", "beauvericin.sdf", "citrinin.sdf", "diacetoscir.sdf", "don.sdf",
" enniatina.sdf", "enniatina1.sdf", "enniatinb1.sdf", "enniatinb.sdf",
" enniatinb4.sdf", "enniatic.sdf", "enniatinf.sdf", "ergocor.sdf", "ergocris.sdf",
" ergocryp.sdf", "ergomet.sdf", "ergosine.sdf", "ergotam.sdf", "fumonisina1.sdf",
" fumonisina2.sdf", "fumonisinab1.sdf", "fumonisinab2.sdf", "fumonisinab3.sdf",
" fumonisinb4.sdf", "fusaproliferin.sdf", "fusx.sdf", "ht2toxin.sdf",
" moniliformin.sdf", "monoacetoscir.sdf", "neosol.sdf", "nivalenol.sdf",
" ochratoxinb.sdf", "ochratoxinc.sdf", "ochratoxintc.sdf", "ota.sdf", "patulin.sdf",
" t2toxin.sdf", "t2triol.sdf", "tentoxin.sdf", "tenuazonicacid.sdf", "trichothecin.sdf",
" verrucarol.sdf", "zealenol.sdf", "zeanone.sdf", "verrucarina.sdf"))

fps <- lapply(all, get.fingerprint, type='standard') #getting molecular fingerprints
fp.sim <- fp.sim.matrix(fps, method='tanimoto') #computing similarity matrix
fp.dist <- 1 - fp.sim #calculating the dissimilarity matrix
fp.dist
row.names(fp.dist)<- c("3ADON", "15ADON", "AFB1", "AFB2", "AFG1", "AFG2", "AFM1", "AFD1", "AFP1", "AFQ1", "AFR0",
" ATNE", "ATNR", "ATME", "ATT1", "BAVR", "CTN", "DASP", "DON", "ENA", "ENA1", "ENB1",
" ENB", "ENB4", "ENC", "ENF", "ECOR", "ECRIS", "ECRY", "EMET", "ESIN", "ETAM",
" FA1", "FA2", "FB1", "FB2", "FB3", "FB4", "FSPR", "FUSX", "HT2T", "MLNF", "MASP",
" NEOS", "NIVA", "OCHB", "OCHC", "OCHTC", "OCHA", "PATL",
" T2T", "T2TR", "TENT", "TNZA", "TCTC", "VRC", "A_ZRLN", "ZRLN", "VERA")
```

```
#evaluating the number of optimal clusters
#within cluster sum of squares method
png(file = "onc_wss_59.png", width=9, height=5, units="in", res=1200) #creating an high resolution image
fviz_nbclust(fp.dist, kmeans, method = "wss") +
  labs(subtitle = "Elbow method")
dev.off()

#Average silhouette method
png(file = "onc_silh_59.png", width=9, height=5, units="in", res=1200)
fviz_nbclust(fp.dist, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")
dev.off()

#Gap statistic
#nboot = 50 to keep the function speedy.
png(file = "onc_gapstat_59.png", width=9, height=5, units="in", res=1200)
fviz_nbclust(fp.dist, kmeans, nstart = 2,, method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")
dev.off()

distance <- as.dist(fp.dist) #setting the dissimilarity matrix into a distance matrix
distance
res.hc <- hclust(distance, method = "ward.D2") #grouping object with ward.

km.res <- kmeans(fp.dist,8, nstart = 25) #performing kmeans
km.res
library(factoextra)
png(file = "HCA_59.png", width=9, height=5, units="in", res=1200) #plotting HCA
fviz_dend(res.hc, k=8, cex=0.7, color_labels_by_k = TRUE, rect = TRUE,
  main= "Dendrogram for mycotoxins",
  k_colors = c( "#FF6666", "#FF8000", "#CC6666", "#AFD829",
" #00FF00", "#FF09F0", "#004c99", "#009999"),
  rect_fill= TRUE,
  rect_border = c("#FF6666", "#FF8000", "#E6B4BD", "#FBFFB7",
" #00FF00", "#FA1F8", "#004c99", "#66FFFF"),
  ylab = "Distance")
dev.off()

png(file = "kmeans_59.png", width=9, height=5, units="in", res=1200) #plotting kmeans
fviz_cluster(km.res, data = fp.dist, ellipse.type = "convex", repel = TRUE,
  palette= c( "#009999", "#004c99", "#00FF00", "#CC6666",
" #AFD829", "#FF8000", "#FF09F0", "#FF6666"), ggtheme = theme_minimal())
dev.off()
```

```

sil <- silhouette(km.res$cluster, dist(fp.dist)) #calculating silhouette coefficient
row.names(sil) <- c("3ADON", "15ADON", "AFB1", "AFB2", "AFG1", "AFG2", "AFM1", "AFD1", "AFP1", "AFQ1", "AFR0",
"ATNE", "ATNR", "ATME", "ATT1", "BAVR", "CTN", "DASP", "DON", "ENA", "ENA1", "ENB1",
"ENB", "ENB4", "ENC", "ENF", "ECOR", "ECRIS", "ECRY", "EMET", "ESIN", "ETAM",
"FA1", "FA2", "FB1", "FB2", "FB3", "FB4", "FSPR", "FUSX", "HT2T", "MLNF", "MASP",
"NEOS", "NIVA", "OCHB", "OCHC", "OCHTC", "OCHA", "PATL",
"T2T", "T2TR", "TENT", "TNZA", "TCTC", "VRC", "A_ZRLN", "ZRLN", "VERA")

sil
png(file = "silhouette_59.png", width=9, height=5, units="in", res=1200) #plotting silhouette results
plot(sil, main = "Silhouette plot: K-means", col= c( "#009999", "#004c99", "#00FF00", "#CC6666",
"#AFD829", "#FF8000", "#FF09F0", "#FF6666"))
dev.off()

```

Annex 5. Code implementation for all PCA results on the dataset composed by 30 mycotoxins and 15 molecular descriptors.

```

rm(list=ls())
cdk.version()
all= load.molecules(c("aflatoxina.sdf", "3acdon.sdf", "fumonisinab1.sdf", "fumonisinab2.sdf",
"15acdon.sdf", "fumonisinab3.sdf", "aflatoxinab2.sdf", "fusx.sdf",
"aflatoxinag1.sdf", "monoacetoscir.sdf", "aflatoxinag2.sdf", #loading mycotoxins
"neosol.sdf", "nivalenol.sdf", "don.sdf", "ota.sdf", "diacetoscir.sdf",
"patulin.sdf", "ergocris.sdf", "ergocor.sdf", "t2toxin.sdf", "ht2toxin.sdf",
"ergocryp.sdf", "aflatoxinam1.sdf", "t2triol.sdf", "ergomet.sdf",
"verrucarol.sdf", "ergotam.sdf", "zeanone.sdf", "zealenol.sdf", "citrinin.sdf"))

descNames <- unique(unlist(sapply(get.desc.categories(), get.desc.names)))#getting all molecular desc
dc <- get.desc.categories() #Getting descriptors categories
allDescs <- eval.desc(all, descNames, verbose = T) #calculating the molecular descriptors for all
allDescs #30 mycotoxins
dim(allDescs) #the dimension of the dataset is 30x287

allDescs <- allDescs[, !apply(allDescs, 2, function(x) any(is.na(x)))] #removing missing values
dim(allDescs) #222x30

allDescs <- allDescs[, !apply(allDescs, 2, function(x) length(unique(x)) == 1)] #removing duplicated
dim(allDescs) #153x30 columns

#perceber
r2 <- which(cor(allDescs)^2 > .6, arr.ind=TRUE) #removing correlated columns
r2 <- r2[r2[,1] > r2[,2], ]
allDescs_1 <- allDescs[, -unique(r2[,2])]
dim(allDescs_1)
View(allDescs_1)

row.names(allDescs_1) <- c("AFB1", "3-ADON", "FUB1", "FUB2", "15-ADON", "FUB3", "AFB2", "FUSX", "AFG1", "MASP",
"AFG2", "NEOS", "NIVA", "DON", "OCHA", "DASP", "PATL", "ECRI", "ECOR", "T2T", "HT2T",
"ECRY", "AFM1", "T2TR", "EMET", "VRC", "ETAM", "ZRLN", "A_ZRLN", "CTN")
class <- c(1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, #creating an ordered vector with the classes of the
1, 1, 1, 1, 1, 1, 1, 0, 1, 1, #30 mycotoxins 0 is non acutely toxic and 1 is acutely toxic
0, 1, 0, 1, 1, 1, 0, 0, 1)

class <- as.factor(class)
levels(class)[levels(class)=="0"] <- "Not acutely toxic" |
levels(class)[levels(class)=="1"] <- "Acutely Toxic"
library(FactoMineR)
res.pca <- PCA(my_descriptors_1, scale.unit = TRUE, ncp = 5, graph=FALSE)
summary(res.pca) #eigenvalues are summarized here
print(res.pca)
png(file = "PCA_30_fam.png", width=9, height=5, units="in", res=1200)
fviz_pca_biplot(res.pca, #BIPLOT- MYCOTOXINS COLORED WITH THE CLUSTERS
# Individuals
geom.ind = c("point", "text"),
fill.ind = km.res$cluster, col.ind = km.res$cluster, #CLUSTERS TO COLOS MYCOTOXINS
gradient.cols = c("#004c99", "#009999", "#00FF00", "#FF6666", "#FF8000" ),
label="all",
pointshape = 21, pointsize = 4,
addEllipses = FALSE,
repel = TRUE,
# Variables
alpha.var = "contrib",
col.var = "black",
fill.var = "black")
dev.off()

```

```

png(file = "PCA_30_tox.png", width=9, height=5, units="in", res=1200)
fviz_pca_biplot(res.pca, #BIPLOT- MYCOTOXINS COLORED WITH CLASSES
  geom.ind = c("point", "text"),
  geom.var = c("arrow", "text"),
  col.ind =class, #CLASSES TO COLOR MYCOTOXINS
  repel=TRUE,
  habillage = class,
  col.var = "black",
  pointsize=3, axes = c(1,2), mean.point=FALSE)
dev.off()

png(file = "contrib_30_PC2.png", width=9, height=5, units="in", res=1200) #GETTING THE CONTRIBUTIONS OF THE
fviz_contrib(res.pca, choice = "var", axes =2, top = 10, cex.lab=2) #MOLECULAR DESCRIPTORS, the argument
dev.off() #axes gives the component in question

```

The next code was applied to perform PCA on all datasets composed by 59 mycotoxins. The only difference is on the feature selection where a 4rth step was added (mentioned in the code) and the name of the files read. All of the names follow a pattern in which the created variable contains the number of molecular descriptors. The following code is for dataset A, composed by 24 molecular descriptors and all variables end with 24.

Annex 6. Code implementation for PCA results of the dataset composed by 59 mycotoxins and 24 molecular descriptors. The remaining datasets are not shown because the code is similar, changing the variables names in R.

```

cdk.version()
setwd("C:/Users/claud/OneDrive/Ambiente de Trabalho/LAB/mycotoxins_new/script")
getwd()

descNames <- unique(unlist(sapply(get.desc.categories(), get.desc.names)))
dc <- get.desc.categories()
get.desc.names(dc[1]) #hybrid
get.desc.names(dc[2]) #constitutional
get.desc.names(dc[3]) #topological
get.desc.names(dc[4]) #electronic
get.desc.names(dc[5]) #geometrical
all= load.molecules(c( "3adon.sdf", "15adon.sdf", "aflatoxina.sdf", "aflatoxinab2.sdf", "aflatoxinag1.sdf",
"aflatoxinag2.sdf", "aflatoxinam1.sdf", "aflatoxind1.sdf", "aflatoxinp1.sdf",
"aflatoxinq1.sdf", "aflatoxinr0.sdf", "altenuene.sdf", "citrinin.sdf", "diacetoscir.sdf",
"alternariol.sdf", "alternariolmethylether.sdf", "altertoxin1.sdf", "beauvericin.sdf",
"don.sdf", "enniatina.sdf", "enniatina1.sdf", "enniatinb1.sdf", "enniatinb.sdf",
"enniatinb4.sdf", "enniatic.sdf", "enniaticf.sdf", "ergocor.sdf", "ergocris.sdf",
"ergocryp.sdf", "ergomet.sdf", "ergosine.sdf", "ergotam.sdf", "fumonisina1.sdf",
"fumonisinab2.sdf", "fumonisina1.sdf", "fumonisina2.sdf", "fumonisina3.sdf",
"fumonisinb4.sdf", "fusaproliferin.sdf", "fusx.sdf", "ht2toxin.sdf", "moniliformin.sdf",
"monoacetoscir.sdf", "neosol.sdf", "nivalenol.sdf", "ochratoxinb.sdf", "ochratoxinc.sdf",
"ochratoxintc.sdf", "ota.sdf", "patulin.sdf", "t2toxin.sdf", "t2triol.sdf", "tentoxin.sdf",
"tenuazonicacid.sdf", "trichothecin.sdf", "verrucarol.sdf", "zealenol.sdf",
"zeanone.sdf", "verrucarina.sdf"))

drug.descs <- eval.desc(all, descNames, verbose=T)
row.names(drug.descs) <- c("3ADON", "15ADON", "AFB1", "AFB2", "AFG1", "AFG2", "AFM1", "AFD1", "AFP1", "AFQ1", "AFR0",
"ATNE", "ATNR", "ATME", "ATT1", "BAVR", "CTN", "DASP", "DON", "ENA", "ENA1", "ENB1",
"ENB", "ENB4", "ENC", "ENF", "ECOR", "ECRIS", "ECRY", "EMET", "ESIN", "ETAM",
"FA1", "FA2", "FB1", "FB2", "FB3", "FB4", "FSPR", "FUSX", "HT2T", "MLNF", "MASP",
"NEOS", "NIVA", "OCHB", "OCHC", "OCHTC", "OCHA", "PATL",

drug.descs
write.csv(drug.descs, "all_R_descs_59.csv") #writing the dataset in a csv file
dim(drug.descs) #59 objects and 287 variables
allDescs <- drug.descs[, !apply(drug.descs, 2, function(x) any(is.na(x)))] #removing missing values
dim(allDescs) #59x222
write.csv(allDescs, "R_descs_without_NAs_59.csv")
allDescs <- allDescs[, !apply( allDescs, 2, function(x) length(unique(x)) == 1 )] #removing duplicated columns
dim(allDescs) #59x155
#THE MAIN DIFFERENCE IS THE NEXT STEP ON THE DATA PROCESSING
allDescs <- as.matrix(allDescs)
colVars(allDescs)
which(colVars(allDescs) < .005, arr.ind=TRUE) #WHICH COLUMNS HAVE VARIANCE <0.005
allDescs <- allDescs[, -c(1,3,4,34,38,47,61,125,126,127)] #REMOVING THOSE COLUMNS
dim(allDescs)

r2 <- which(cor(allDescs)^2 > .6, arr.ind=TRUE) #removing columns with a correlation superior to 0.6
r2 <- r2[ r2[,1] > r2[,2] , ]
allDescs_1 <- allDescs[, -unique(r2[,2])]
dim(allDescs_1)

```

```

write.csv(allDescs_1, "Descriptors_cleaned_59_variance_removed.csv") #Saving the datasets in a csv file
class<- c(0,1,1,1,1,1,0,1,1,0,1,1,1,1,0,1,1,1,1,1,1,0,0,0,1,0,1,1,1,0,0,0,1,1,0,0,1,1,1,0,
1,1,1,0,1,1,1,0,0,1,1,1,0,0,1) #creating an ordered vector with the class of each mycotoxin
class<- as.factor(class)
levels(class)[levels(class)=="0"] <- "Not acutely toxic"
levels(class)[levels(class)=="1"] <- "Acutely Toxic"
Variables_24<- read.csv("Descriptors_cleaned_59_variance_removed.csv")
Variables_24$X<- NULL
row.names(Variables_24)<- c("3ADON", "15ADON", "AFB1", "AFB2", "AFG1", "AFG2", "AFM1", "AFD1", "AFP1", "AFQ1", "AFR0",
"ATNE", "ATNR", "ATME", "ATT1", "BAVR", "CTN", "DASP", "DON", "ENA", "ENA1", "ENB1",
"ENB", "ENB4", "ENC", "ENF", "ECOR", "ECRIS", "ECRY", "EMET", "ESIN", "ETAM",
"FA1", "FA2", "FB1", "FB2", "FB3", "FB4", "FSPR", "FUSX", "HT2T", "MLNF", "MASP",
"NEOS", "NIVA", "OCHB", "OCHC", "OCHTC", "OCHA", "PATL",
"T2T", "T2TR", "TENT", "TNZA", "TCTC", "VRC", "A_ZRLN", "ZRLN", "VERA")

library(FactoMineR)
PCA_24<- PCA(Variables_24, scale.unit = TRUE, ncp = 5, graph=FALSE)
write.csv(PCA_24$eig, "59m24d_eigenvalues.csv")
View(PCA_24$ind$coord)
summary(PCA_24)
png(file = "PCA_59_24_fam.png", width=9, height=5, units="in", res=1200)
fviz_pca_biplot(PCA_24,
# Individuals
geom.ind = c("point", "text"),
fill.ind = km.res$cluster, col.ind = km.res$cluster,
gradient.cols = c( "#00FF00", "#FF6666", "#FF09F0", "#FF8000",
"#009999", "#004c99", "#AFD829", "#CC6666"),
label="all",
pointshape = 21, pointsize = 4,
addEllipses = FALSE,
repel= TRUE,
# Variables
alpha.var ="contrib",
col.var = "black",
fill.var="black")

dev.off()

#####ADDING THE BIOLOGICAL ACTIVITY DESCRIPTORS FORMING DATASET WITH 35 DESCRIPTORS]

More_Biological_activity<- cbind(Variables_24, drug.descs$nHBAcc)
More_Biological_activity<- cbind(More_Biological_activity, drug.descs$MW)
More_Biological_activity<- cbind(More_Biological_activity, drug.descs$XLogP)
More_Biological_activity<- cbind(More_Biological_activity, drug.descs$LipinskiFailures)
More_Biological_activity<- cbind(More_Biological_activity, drug.descs$nRotB)
More_Biological_activity<- cbind(More_Biological_activity, drug.descs$MLogP)
More_Biological_activity<- cbind(More_Biological_activity, drug.descs$ALogP)
More_Biological_activity<- cbind(More_Biological_activity, drug.descs$ALogp2)
More_Biological_activity<- cbind(More_Biological_activity, drug.descs$nSmallRings)
More_Biological_activity<- cbind(More_Biological_activity, drug.descs$nAromRings)
More_Biological_activity<- cbind(More_Biological_activity, drug.descs$TopoPSA)

names(More_Biological_activity)[names(More_Biological_activity)==" drug.descs$nHBAcc"]<- "nHBAcc"
names(More_Biological_activity)[names(More_Biological_activity)==" drug.descs$MW"]<- "MW"
names(More_Biological_activity)[names(More_Biological_activity)==" drug.descs$XLogP"]<- "XLogP"
names(More_Biological_activity)[names(More_Biological_activity)==" drug.descs$LipinskiFailures"]<- "LipinskiFailures"
names(More_Biological_activity)[names(More_Biological_activity)==" drug.descs$nRotB"]<- "nRotB"
names(More_Biological_activity)[names(More_Biological_activity)==" drug.descs$MLogP"]<- "MLogP"
names(More_Biological_activity)[names(More_Biological_activity)==" drug.descs$ALogP"]<- "ALogP"
names(More_Biological_activity)[names(More_Biological_activity)==" drug.descs$ALogp2"]<- "ALogp2"
names(More_Biological_activity)[names(More_Biological_activity)==" drug.descs$nSmallRings"]<- "nSmallRings"
names(More_Biological_activity)[names(More_Biological_activity)==" drug.descs$nAromRings"]<- "nAromRings"
names(More_Biological_activity)[names(More_Biological_activity)==" drug.descs$TopoPSA"]<- "TopoPSA"

rownames(More_Biological_activity)<-c("3ADON", "15ADON", "AFB1", "AFB2", "AFG1", "AFG2", "AFM1", "AFD1", "AFP1", "AFQ1",
"AFR0", "ATNE", "ATNR", "ATME", "ATT1", "BAVR", "CTN", "DASP", "DON", "ENA",
"ENA1", "ENB1", "ENB", "ENB4", "ENC", "ENF", "ECOR", "ECRIS", "ECRY", "EMET",
"ESIN", "ETAM", "FA1", "FA2", "FB1", "FB2", "FB3", "FB4", "FSPR", "FUSX",
"HT2T", "MLNF", "MASP", "NEOS", "NIVA", "OCHB", "OCHC", "OCHTC", "OCHA", "PATL",
"T2T", "T2TR", "TENT", "TNZA", "TCTC", "VRC", "A_ZRLN", "ZRLN", "VERA")

PCA_35<- PCA(More_Biological_activity, scale.unit = TRUE, ncp = 5, graph=FALSE) #PERFORMING PCA-correlation matrix
write.csv(PCA_35$eig, "59m35d_eigenvalues.csv") #getting eigenvalues
print(PCA_35)
summary(PCA_35)

```



```

png(file = "PCA_59_35_fam.png", width=9, height=5, units="in", res=1200)
fviz_pca_biplot(PCA_35,
  # Individuals
  geom.ind = c("point", "text"),
  fill.ind = km.res$cluster, col.ind = km.res$cluster,
  gradient.cols = c( "#00FF00", "#FF6666", "#FF09F0", "#FF8000",
                    "#009999", "#004c99", "#AFD829", "#CC6666"),
  label="all",
  pointshape = 21, pointsize = 4,
  addEllipses = FALSE,
  repel= TRUE,
  # Variables
  alpha.var = "contrib",
  col.var = "black",
  fill.var="black")
dev.off()
png(file = "PCA_59_35_tox.png", width=9, height=5, units="in", res=1200)
fviz_pca_biplot(PCA_35,
  geom.ind = c("point", "text"),
  geom.var = c("arrow", "text"),
  col.ind =class,
  repel=TRUE,
  habillage = class,
  col.var = "black",
  pointsize=3, axes = c(1,2), mean.point=FALSE)
dev.off()#é so mudar as componentes em axes
png(file = "contrib_59_35_PC2.png", width=9, height=5, units="in", res=1200)
fviz_contrib(PCA_35, choice = "var", axes =2, top = 10)#CONTRIBUTIONS OF THE VARIABLES
dev.off()

```

In similar with PCA, all implemented codes of supervised learning follow the same pattern: when dealing with one dataset, all variables in R end with the respective number of molecular descriptors of that dataset.

Annex 7. Code implementation for LDA results of the dataset composed by 59 mycotoxins and 24 molecular descriptors. The remaining datasets are not shown because the code is similar, changing the variables names in R and the files read.

```

Variables_24<- read.csv("Descriptors_cleaned_59_variance_removed.csv")
Variables_24$X<- NULL
row.names(Variables_24)<- c("3ADON", "15ADON", "AFB1", "AFB2", "AFG1", "AFG2", "AFM1", "AFD1", "AFP1", "AFQ1", "AFR0",
  "ATNE", "ATNR", "ATME", "ATT1", "BAVR", "CTN", "DASP", "DON", "ENA", "ENA1", "ENB1",
  "ENB", "ENB4", "ENC", "ENF", "ECOR", "ECRIS", "ECRY", "EMET", "ESIN", "ETAM",
  "FA1", "FA2", "FB1", "FB2", "FB3", "FB4", "FSPR", "FUSX", "HT2T", "MLNF", "MASP",
  "NEOS", "NIVA", "OCHB", "OCHC", "OCHTC", "OCHA", "PATL",
  "T2T", "T2TR", "TENT", "TNZA", "TCTC", "VRC", "A_ZRLN", "ZRLN", "VERA")

var_24_class<- data.frame(Variables_24, class)
var_24_class$rhs.dCH2<- NULL
levels(class)[levels(class)=="1"] <- "Acutely Toxic"
levels(class)[levels(class)=="0"] <- "Not acutely toxic"

set.seed(123)
training.samples_24 <- var_24_class$class %>%
  createDataPartition(p = 0.7, list = FALSE) #SPLITTING DATA INTO TRAIN AND TEST
train.data_24 <- var_24_class[training.samples_24, ] #42x25
test.data_24 <- var_24_class[-training.samples_24, ] #17x25

preproc.param <- train.data_24 %>%
  preprocess(method = c("center", "scale")) #SCALING DATA
# Transform the data using the estimated parameters
train.transformed_24 <- preproc.param %>% predict(train.data_24)
test.transformed_24 <- preproc.param %>% predict(test.data_24)

model_24 <- lda(class~., data = train.transformed_24)# Fit the model
predictions_24 <- model_24 %>% predict(test.transformed_24)# Make predictions

predicted.classes_24 <- predictions_24$class
observed.classes_24 <- test.transformed_24$class
#Confusion Matrix
confusionMatrix(predicted.classes_24, as.factor(observed.classes_24), positive = "1") #COMPUTING CONFUSION MATRIX
auc(observed.classes_24, as.numeric(predicted.classes_24), levels=c(1,0), direction=">") #COMPUTING AUROC

```



Annex 8. Code implementation for SVM results of the dataset composed by 59 mycotoxins and 24 molecular descriptors. The remaining datasets are not shown because the code is similar, changing the variables names in R and the files read.

```
rm(list=ls())
cdk.version()
setwd("C:/Users/claud/OneDrive/Ambiente de Trabalho/LAB/mycotoxins_new/script")
getwd()
class<- as.factor(c(0,1,1,1,1,1,0,1,1,0,1,1,1,0,1,1,1,1,1,1,0,0,0,0,1,0,1,1,1,0,0,0,1,1,0,0,0,1,1,0,0,1,1,0,0,1,1,1,0,
1,1,1,0,1,1,1,0,0,1,1,1,0,0,1))
Variables_24<- read.csv("Descriptors_cleaned_59_variance_removed.csv")
Variables_24$X<- NULL
row.names(Variables_24)<- c("3ADON", "15ADON", "AFB1", "AFB2", "AFG1", "AFG2", "AFM1", "AFD1", "AFP1", "AFQ1", "AFR0",
"ATNE", "ATNR", "ATME", "ATT1", "BAVR", "CTN", "DASP", "DON", "ENA", "ENA1", "ENB1", "ENB", "ENB4", "ENC", "ENF", "ECOR", "ECRIS", "ECRY", "EMET", "ESIN", "ETAM",
"FA1", "FA2", "FB1", "FB2", "FB3", "FB4", "FSPR", "FUSX", "HT2T", "MLNF", "MASP",
"NEOS", "NIVA", "OCHB", "OCHC", "OCHTC", "OCHA", "PATL",
"T2T", "T2TR", "TENT", "TNZA", "TCTC", "VRC", "A_ZRLN", "ZRLN", "VERA")

data_59_24<- cbind(Variables_24, class)
levels(class)[levels(class)=="0"] <- "Not acutely toxic"
levels(class)[levels(class)=="1"] <- "Acutely Toxic"

set.seed(123)
training.samples_24<- data_59_24$class %>%
  createDataPartition(p = 0.7, list = FALSE) #SPLITTING DATA INTO TRAIN AND TEST
train.data_24 <- data_59_24[training.samples_24, ] #42x36
test.data_24<- data_59_24[-training.samples_24, ] #17x36
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)

model_svm_59_24 <- train(#FIT THE MODEL
  class ~., data = train.data_24, method = "svmLinear",
  trControl = trctrl,
  tuneLength=10,
  preProcess = c("center", "scale")#SCALING DATA
)
predicted.classes_24 <- model_svm_59_24 %>% predict(train.data_24)#MAKING PREDICTIONS
confusionMatrix(predicted.classes_24, train.data_24$class, positive="1") # CALCULATING THE CONFUSION MATRIX
auc(train.data_24$class, as.numeric(predicted.classes_24), levels=c(1,0), direction=">") #CALCULATING AUROC
```

Annex 9. Code implementation for RF results of the dataset composed by 59 mycotoxins and 24 molecular descriptors. The remaining datasets are not shown because the code is similar, changing the variables names in R and the files read.

```
import os
os.chdir("C:/Users/claud/OneDrive/Ambiente de Trabalho/LAB/mycotoxins_new/script")

import pandas as pd

df = pd.read_excel("Descriptors_cleaned_59_variance_removed.csv", sep=",") #reading file
df.head()
df.drop("Unnamed: 0", axis=1, inplace=True)
df.head()
df.insert(24, "Class", [0,1,1,1,1,1,1,0,1,1,0,1,1,1,1,0,1,1,1,1,
1,1,1,0,0,0,0,1,0,1,1,1,0,0,0,1,1,0,0,1,
1,1,0,1,1,1,1,0,1,1,1,0,0,1,1,1,0,0,1], True) #inserting class on the dataset

features = list(df.columns.values)
features.remove("Class") #features do not contain the target variable

print(features)
X = df[features]
y = df['Class']
import random
random.seed(42)
from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import train_test_split #split into train and test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

scaler = StandardScaler()
scaler.fit(X_train)
X_train = scaler.transform(X_train) #scaling data
X_test=scaler.transform(X_test)
```



```

from sklearn.model_selection import GridSearchCV
from sklearn.exceptions import ConvergenceWarning
clf = GridSearchCV(model, parameters, n_jobs=-1) #performing gridsearch
clf.fit(X_train, y_train) #fitting the model

print(clf.best_params_) #printing the best parameters of gridsearch in terms of accuracy
print(clf.cv_results_)

predicted= clf.predict(X_train) #making predictions with train data
predicted_test= clf.predict(X_test) #making predictions with test data

from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score, roc_auc_score
#computing all metrics
print("Accuracy train",accuracy_score(y_train, predicted)) #% dos que acertou
print("Accuracy test",accuracy_score(y_test, predicted_test))

print("F1 train",f1_score(y_train, predicted))
print("F1 test",f1_score(y_test, predicted_test))

print("Precision train",precision_score(y_train, predicted)) #não teve falsos positivos
print("Precision test",precision_score(y_test, predicted_test))

print("Recall train",recall_score(y_train, predicted)) #sensitivity
print("Recall test",recall_score(y_test, predicted_test)) #problemas com falsos negativos

print("AUROC train",roc_auc_score(y_train, predicted))
print("AUROC test",roc_auc_score(y_test, predicted_test))

print("Recall train",recall_score(y_train, predicted, pos_label=0)) #sensitivity
print("Recall test",recall_score(y_test, predicted_test, pos_label=0))

```