

PERTURBATION THEORY MACHINE LEARNING MODELS FOR BIOLOGICAL
ACTIVITY OF VITAMINS DERIVATIVES AND NANOPARTICLES SYSTEMS

RICARDO SANTANA CABELLO

UNIVERSIDAD PONTIFICIA BOLIVARIANA
ESCUELA DE INGENIERÍA ESCUELA DE ECONOMÍA, ADMINISTRACIÓN Y NEGOCIOS
DOCTORADO EN GESTIÓN DE LA TECNOLOGÍA Y LA INNOVACIÓN
MEDELLÍN 2021

UNIVERSIDAD DE DEUSTO
DOCTORADO EN INGENIERÍA PARA LA SOCIEDAD DE LA INFORMACIÓN
Y DESARROLLO SOSTENIBLE
BILBAO 2021

PERTURBATION THEORY MACHINE LEARNING MODELS FOR BIOLOGICAL
ACTIVITY OF VITAMINS DERIVATIVES AND NANOPARTICLES SYSTEMS

RICARDO SANTANA CABELLO

Trabajo de grado para optar al título de
Doctor en Gestión de la Tecnología y la Innovación

Asesores

Dr. Humberto González Díaz,
Dr. Enrique Onieva Caracuel and Dra.
Piedad Gañán Rojo

UNIVERSIDAD PONTIFICIA BOLIVARIANA
ESCUELA DE INGENIERÍA ESCUELA DE ECONOMÍA, ADMINISTRACIÓN Y NEGOCIOS
DOCTORADO EN GESTIÓN DE LA TECNOLOGÍA Y LA INNOVACIÓN
MEDELLÍN 2021

UNIVERSIDAD DE DEUSTO
DOCTORADO EN INGENIERÍA PARA LA SOCIEDAD DE LA INFORMACIÓN
Y DESARROLLO SOSTENIBLE
BILBAO 2021

10 de junio de 2021

Ricardo Santana Cabello

“Declaro que este trabajo de grado no ha sido presentado con anterioridad para optar a un título, ya sea en igual forma o con variaciones, en ésta o en cualquiera otra universidad”. Art. 92, parágrafo, Régimen Estudiantil de Formación Avanzada.

Firma



PERTURBATION THEORY MACHINE LEARNING
MODELS FOR BIOLOGICAL
ACTIVITY OF VITAMINS DERIVATIVES
AND NANOPARTICLES SYSTEMS

by

Ricardo Santana Cabello

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy, within the PhD “Doctorado
en Gestión de la Tecnología y la Innovación”

Supervised by Dr. Humberto González-Díaz,
Dr. Enrique Onieva Caracuel
and Dra. Piedad Gañán Rojo



Universidad de Deusto



Universidad Pontificia Bolivariana

PTML MODELS FOR BIOLOGICAL ACTIVITY OF VITAMINS DERIVATIVES AND NANOPARTICLES SYSTEMS

by
Ricardo Santana Cabello

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy, within the PhD “Doctorado
en Gestión de la Tecnología y la Innovación”

Supervised by Dr. Humberto González Díaz,
Dr. Enrique Onieva Caracuel
and Dra. Piedad Gañán Rojo

The candidate

The supervisor

**Humberto
Gonzalez
Diaz**

Digitally signed by
Humberto Gonzalez Diaz
DN: cn=Humberto
Gonzalez Diaz, c=ES,
o=UPVEHU,
ou=UPVEHU,
email=humberto@upvehu.es
Reason: Thesis
Dissertation
Date: 2021.02.24
09:15:31 +01'00'

The supervisor

The supervisor

Bilbao, January 2021

Ptml Models for Biological Activity of Vitamins Derivatives and Nanoparticles Systems

Author: Ricardo Santana Cabello
Advisor: Humberto González Díaz
Advisor: Enrique Onieva Caracuel
Advisor: Piedad Gañán Rojo

Text printed in Bilbao

First edition, January 2021

As he faced the sun,
he cast no shadow.

Abstract

The development of nanotechnology for the industry has been noticeable in recent years given the possibility of improving the functions of materials. In sectors of great importance, such as pharmacological, food and cosmetic sectors, research projects have been implemented in order to improve the characteristics of these consumer products.

Although there are currently published researches on the application of nanotechnology, a vast field remains to be explored to understand how the matter behaves in the nanoscale in an integral way. It exists a gap of considerable knowledge about the biological behavior of nanosystems and therefore their efficient development and regulation to safeguard public safety.

This thesis is focused on the design of nanosystems. It is intended to know the components that can best integrate a nanosystem through the development of predictive models. The use of *in silico* methods gives information of the design of the nanosystems and is aligned to the green chemistry principles. These predictions can be used to complement the information provided and analyzes the agencies that ensure public safety, both in the European Union and abroad. The development of models for different types of nanosystems as well as regulatory exploration has resulted in six research articles collected in this research.

First, an exploration was carried out on the regulation of nanomaterials in the context of the European Union. It started with the reference regulation in the field of nanotechnology, such as European Regulation 1223/2009 on cosmetics. Subsequently, the study of food regulations and the analysis of how nanotechnology is regulated continued. To conclude the first part, a study was also carried out on the regulation of nanotechnology in the field of pharmacology and the role of Machine

Learning.

Once the regulatory field has been explored and the difficulty in drafting and enforcing a nanotechnology regulation, predictive models are proposed to help design nanosystems and to complement information to make decisions for risk management. The chosen nanosystems incorporate vitamin derivatives given their importance in the state of the art and the challenge posed by the creation of more complex systems but with great potential for improvement from the biological point of view. Initially a model for vitamin derivatives is developed. After, model of nanosystems consisting of metal oxide nanoparticles with surface agent and the vitamin derivatives. Third, a model for the prediction of biological activities of anticancer nanosystems is developed. Fourth, the previous model is improved using Machine Learning techniques

Resumen

El desarrollo de nanotecnología para la industria ha sido notorio en los últimos años dada la posibilidad de mejorar las funciones de los materiales. En sectores de gran importancia, como el farmacológico, alimenticio y cosmético se ha diseñado y ejecutado proyectos de investigación que mejoren las características de estos productos de consumo.

Actualmente existen investigaciones sobre la aplicación de la nanotecnología, sin embargo, un vasto campo queda por explorar para entender cómo se comporta la materia en la nanoescala de manera integral. Una laguna de conocimiento considerable sobre el comportamiento biológico de los nanosistemas y por tanto su eficiente desarrollo y regulación para salvaguardar la seguridad pública.

Esta tesis está enfocada al diseño de nanosistemas. De esta forma se pretende conocer los componentes que mejor pueden integrar un nanosistema mediante el desarrollo de modelos predictivos. Dichas predicciones pueden ser utilizadas para complementar la información entregada y analiza a las agencias que velan por la seguridad pública, tanto de la Unión Europea como fuera. El desarrollo de los modelos para diferentes tipos de nanosistemas así como la exploración regulatoria ha resultado en seis artículos de investigación recogidos en esta investigación.

En primer lugar, se realizó una exploración sobre la regulación de los nanomateriales en el contexto de la Unión Europea. Se comenzó con la regulación de referencia en el ámbito de la nanotecnología, como es el Reglamento Europeo 1223/ 2009 sobre cosméticos. Posteriormente, se continuó con el estudio de las regulaciones en materia de alimentos y el análisis de cómo se regula la nanotecnología en la Unión Europea. Para concluir la primera parte, se adelantó igualmente un estudio sobre la regulación de la nanotecnología en el campo de la farmacología y el

papel del Machine Learning en el mismo contexto de referencia.

Una vez explorado el campo regulatorio e identificada la dificultad para redactar y hacer cumplir una regulación dada la falta de información, se proponen modelos predictivos que ayuden a diseñar nanosistemas y para complementar información con la que tomar decisiones para la gestión del riesgo. Los nanosistemas elegidos incorporan derivados de vitaminas dado su importancia en el estado del arte y por el reto que supone la creación de sistemas más complejos, pero con gran potencial de mejora desde el punto de vista biológico. Inicialmente se desarrolla un modelo para derivados de vitaminas. A continuación, se presenta un modelo de nanosistemas conformados por nanopartículas de óxido metálicos con agente superficial y el derivado de la vitamina. En tercer lugar, se desarrolla un modelo para la predicción de actividades biológicas de nanosistemas anticancerígenos. En cuarto lugar, se mejora el modelo anterior utilizando diferentes técnicas de Aprendizaje Automático.

Acknowledgements

There are many people without whom this work would not have been possible. First, I would like to thank my co-advisors, Humberto González Díaz, Enrique Onieva Robin Zuluaga and Piedad Gañán. This work could have not been conceived without their guidance, support and dedication. With the same gratitude, I would like to mention Sonia Arrasate and Robin Zuluaga, given their great support, professionalism and human qualities.

I also dedicate this work to the NANOCELIA network and COLCIENCIAS, for their support and confidence in my person and research.

I take the opportunity to remember the research group of Tulane University, and the special time I spent. I want to make special mention to Professor Matthew Montemore, who served as an inspiration, for his kindness and wisdom, at different times in my life.

I do not want to forget to express my gratitude for the support and welcome I received at the National Nanotechnology Laboratory (LANOTEC), especially José Vega and his research team.

I want to express my gratitude to Zulamita Zapata for her inspiration, intelligence and solidarity.

Finally, I also like to dedicate this work to my family. It would have been impossible without their energy, solidarity, lucidity and support in every step I have taken. This work is especially for all of them, for trusting me all this time, for their unconditional and genuine love and their contribution to my happiness every day.

Ricardo Santana

January, 2021.

Contents

Contents	XIII
List of Figures.....	XV
Acronyms	XVI
1) Introduction	1
1.1 Thesis Statement	4
1.2 Main research objectives	5
1.3 Research methodology	6
1.4 Publications	7
1.4.1 Paper I.....	9
1.4.2 Paper II.....	9
1.4.3 Paper III	10
1.4.4 Paper IV	11
1.4.5 Paper V	12
1.4.6 Paper VI.....	13
1.4.7 Paper VII.....	13
1.5 Outline.....	15
2) European Nanotechnology Regulation (Cosmetic sector)	17
3) European Nanotechnology Regulation (Food sector).....	53
4) European Nanotechnology Regulation (Pharmaceutic Sector).....	79
5) Modelling Vitamin Derivatives.....	107

6) Modelling systems of metal oxide nanoparticles and vitamin derivatives.....	155
7) Modelling systems DVRNs (Multiplicative operators).....	187
8) Modelling DVRNs (Metric operators and enrichment of information)	221
9) Conclusions and Future Works.....	273
9.1 Conclusions	275
9.2 Future work	278
Bibliography	281
Declaration.....	283

List of Figures

Figure 1. Relationship between the research objectives and publications.	8
Figure 2.- Nanocolorants used in European Union in the different types of cosmetic products. Information based on the catalogue reported by the EC (Version 1 (31.12.2016))	18
Figure 3.- NanoUV-filters used in European Union in the different types of cosmetic products. Information based on the catalogue reported by the EC (Version 1 (31.12.2016))	19
Figure 4.- Nanomaterials with other functions used in European Union in the different types of cosmetic products. Information based on the catalogue reported by the EC (Version 1 (31.12.2016))	20
Figure 5. Cheminformatic models workflow to predict biological activity and improve regulation application.....	54
Figure 6. General workflow to build the models for the present study	108
Figure 7. PTML data pre-processing and processing workflow proposed in this work.....	156
Figure 8. Detailed workflow to build a PTML model used in this work	188
Figure 9. Workflow to build a PTML model used in this work.....	222

Acronyms

ANN Artificial Neural Network

LDA Linear Discriminant Analysis

ML Machine Learning

PTO Perturbation Theory Operator

PTML Perturbation Theory Machine Learning

RF Random Forest

*Nothing in life is to be feared, it is only to be understood.
Now is the time to understand more, so that we may fear
less.*

Marie Curie.

CHAPTER

1

1) Introduction

Machine learning (ML) techniques have played a remarkable role in the development of nanotechnology science; Brown *et al.*¹ highlight three significant interactions between nanoscience and ML: 1) a method to infer knowledge from large nanoscience datasets, 2) application to material discovery and optimized experimental design and 3) for hardware development. For instance, Xie *et al.*² showed how ML can contribute to the search for optimal reaction parameters. Specifically, they were able to develop a model by applying XGboost with more than 90% of accuracy for crystallization propensity of metal–organic nanocapsules (MONCs). Sun *et al.*³ could infer the transfer of electrons of silver nanoparticles by using a principal component analysis and 3 layers artificial neural network with a 93% of the testing subset. These are examples, among many others, of the relevant development of ML in nanoscience field in recent years.

Nanotechnology, as transversal technology, is provoking promising studies in different fields. However, one of the most important impacts is happening in the developments on

1) INTRODUCTION

pharmaceutical discoveries. Concretely, there is a very promising type of nanosystem that has attracted the attention of research projects: cancer co-therapy drug-vitamin release nanosystems (DVRNs).⁴ These nanosystems, apart from the anticancer compound, include vitamins or vitamin derivatives. These compounds and the nanosystem helps for reduction of cancer fatigue and drug delivery.⁵ However, in terms of design, they are challenging given the possible combinations of the compounds that integrate them. In this context, screening all combinations by using *in vivo* or *in vitro* studies is expensive and time-consuming.

To develop more efficient studies, we will be developing models with Perturbation Theory Machine Learning (PTML). As a technique of ML, this technique includes the combination of ML with fundamentals of Perturbation Theory. So, when preprocessing the different datasets, we can create a reference function of the expected biological activity taking into consideration the assays with the same assays. Then we also develop Perturbation Theory Operators (PTO) with which we add the perturbation to the system. We are able to infer knowledge from datasets that have a remarkable number of labels, and solve the challenge of characteristics of big data such as high volume, velocity and variety. The PTML method has been applied in different fields. For instance, Simón-Vidal ⁶ applied PTML in order to build a model by applying General Linear Regression able to predict the yield of reactions. Da Costa *et al.* ⁷ used PTML method to predict drug-protein predictions. To do so, they run different algorithms such as ANN or RF. Bediaga *et al.* applied LDA and ANN for the discovery of desirable anticancer compounds taking as reference ChEMBL information. The resulting model has more than Sp(%) more than 90%. These are some examples of the application of PTML to develop cheminformatics studies. They are significant advances, among others.⁸⁻¹⁴

The application of ML methods gives us more information about the design of nanosystems. We can identify patterns or simply predict the most desirable compound or system. This is actually a significant advance of materials science, biology, chemics, pharmacology and food science. It implies not only improvements and saving of costs and time, but also it is aligned with the principles of green chemistry. Especially the the Rs principles: Reduce, Replace and Refine the studies that involve animal experimentations.

The fact that we can have more information about determine nanomaterials, provides of more knowledgment for risk management. The regulation of nanotechnology is different in every country, although there is harmonization by soft law or hard law, in regions such as the European Union. The processes to authorize new products including nanomaterials have been in recent years challenging given the lack of information. The possibilities in terms of design of nanomaterials are high and we still need to generate more information. In vitro and in vivo results are the most important guide we need to focus on, however predictions though machine learning permit a more efficient and green process.

This study does an initial exploration of the difficulties, especially technological, to regulate and apply the regulation in the European Union context, where possibly the development of this kind of regulations have presented an advanced state. Then, we explore the the necessity of new models to predict new vitamin derivatives and new complex nanosystems. Regarding the new nanosystems, we start with metal oxide nanosystems with vitamin derivatives. Then, we study non metal oxide nanosystems to deliver anticancer compounds. Finally, we pose an approach with non metal oxide nanosystems with anticancer compounds and vitamins information, to predict desirable nanosystems.

Hence, this study is a contribution to the development of nanosciences and materials science.

1) INTRODUCTION

1.1 Thesis Statement

This dissertation provides on one hand, an analysis of the nanotechnology regulation in the European Union context for cosmetic, food and pharmacology sectors. This analysis shows how this regulation provides specific conditions for materials on nanoscale. These provisions are challenging in cases when the current state of art is not able to determine the biological activity. In terms of risk management, agencies and European Commission must make a decision to authorized nanomaterials. The regulation does not close the door to apply Machine Learning techniques to know more about nanosystems design process and decide accordingly.

On the other hand, we provide models able to predict which compounds of nanosystems and vitamin derivatives that are desirable, taking into consideration heterogeneous data with thousands of multi-condition biological assays. These systems are used, mainly, for deliver of nutraceuticals and drugs. The positive performance of the models is shown by a high level of prediction, in terms of specificity and sensitivity, like it has not been seen in the state of art before this research.

1.2 Main research objectives

This research addresses the following main objectives to achieve the statement presented in the previous section:

1. To propose a systematic analysis of the relevance of these models for the application of European food regulations.
2. To create a model by applying PTML method to predict the biological activity of vitamin derivatives.
3. To develop a model by applying PTML method to predict the biological activity of components of nanoparticles systems.

1) INTRODUCTION

1.3 Research methodology

First Objective

The systematic research of European food regulation will be carried out through a dogmatic legislative study. To this end, the applicable European regulations will be analyzed to detect problems in application terms. Consequently, it will be assessed to apply predictive models to achieve a greater public safety.

Second Objective

Vitamin Data Pre-processing. The data points for vitamins pre-clinical assays are obtained from preclinical assays registered in ChEMBL database. Each pre-clinical assay includes a result of the value v_{ijvit} of the biological activity that the i th vitamin presents over the j th target. Specifically, v_{ijvit} varies depending on the structure of the each vitamin and the combination of the assay conditions $c_{jvit} = (c_{0vit}, c_{1vit}, c_{2vit}, \dots, c_{nvit})$. Vitamin assay conditions are c_{0vit} = the biological activity v_{ijvit} , c_{1vit} = organism of assay, c_{2vit} = target protein, etc. In order to create the PTML model, we discretize v_{ijvit} to create a binary classification taking into consideration every biological activity and its respective cutoff taken from literature.

PTML model. Classification techniques are used given the purpose of the model to predict a desirable biological effect. The model lets us predict a scoring function for the vitamin or vitamin derivative v_i combinatorial assay conditions. This PTML model takes into consideration vitamins derivatives assay conditions. We propose a linear PTML model in order to predict the biological activity and/or classify vitamin derivatives as desirable or not desirable.

Third Objective

Nanoparticle Data Pre-processing. The data for nps assays is obtained from literature. As in the case of vitamins, each preclinical assay includes a result of the value v_{ijnp} of the biological activity that the i th nps presents used over the j th target. Nps assay conditions are c_{0nm} = the biological activity v_{ijnm} , c_{1nm} = cell line, c_{2nm} = shape, etc. By analogy, we also proceed with the discretization of the v_{ijnp} to create a binary classification taking into consideration every biological activity and its respective cutoff taken from literature. Given that the database includes coated nps, there are descriptors for the core of the np D_{icore} and the coating agent D_{icoat} .

Vitamin-Nanoparticle Information Fusion. For this model, an information fusion of the results of nps tests and vitamins tests is needed, with different conditions for each set. We also apply a discretization for the pairs. The variable $f(v_{ijvit}, v_{ijnp})_{expt}$ is a function that includes the expected value of biological activity for a pair (vitamin-np), without the perturbation, with vectors of assay conditions for every combination of vitamin and nanoparticles experimental conditions and $f(v_{ijvit}, v_{ijnp})_{calc}$ takes into account the perturbation.

PTML Model. In this case, the model lets us predict a scoring function for the vitamin or vitamin derivative v_{ijvit} and the n_{pi} in the combinatorial assay conditions taking into consideration vitamins assay conditions and nps assay conditions. We propose a linear PTML model in order to predict the biological activity and/or classify pairs (vitamin-np) as desirable or not desirable.

1.4 Publications

This thesis is a collection of seven papers. The connection between these publications and research objectives are presented in **Figure 1**.

1) INTRODUCTION

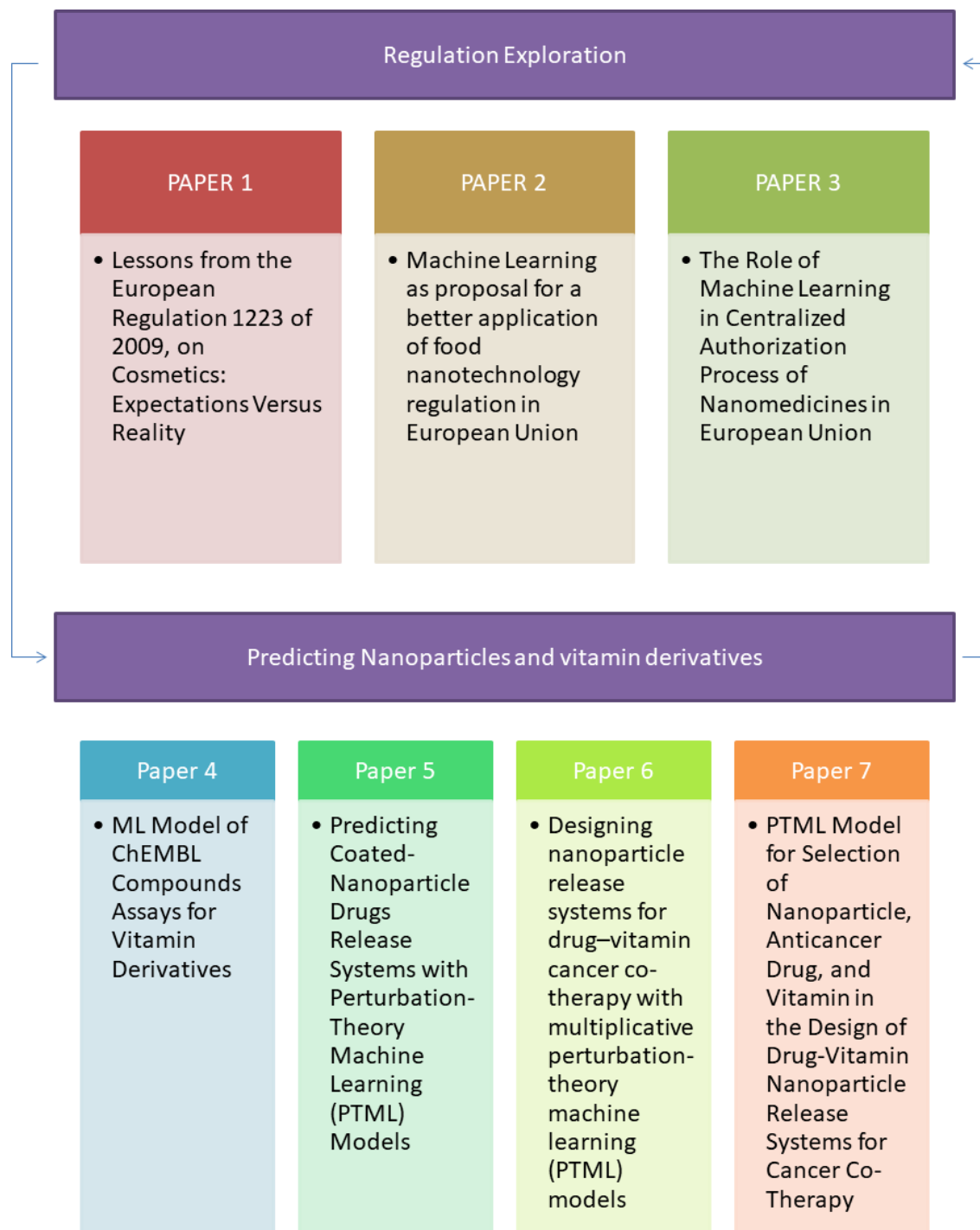


Figure 1. Relationship between the research objectives and publications.

1.4.1 Paper I

Paper I is titled “Lessons from the European Regulation 1223 of 2009, on Cosmetics: Expectations Versus Reality” and it has been published in a JCR-Q1 journal (Nanoethics). The aim of this paper is to conduct an analysis of the application of the specific rules of nanotechnology incorporated in Regulation No. 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. It has been ten years since the European Commission had issued its proposal to start the co-decision procedure to create Regulation 1223 of 2009. Although it has been praised for noting the regulatory difference of nanomaterials over the rest of the chemicals, what has been the efficacy of the standard? It is concluded that despite what it meant, the regulation has encountered technical obstacles, thus rendering the objectives relating to nanotechnology that were proposed from the European Commission unfulfilled.

Cabello, R. S., Rojo, P. G., & Zuluaga, R. (2019). Lessons from the European Regulation 1223 of 2009, on Cosmetics: Expectations Versus Reality. *NanoEthics*, 13(1), 21-35. DOI: <https://doi.org/10.1007/s11569-019-00335-6>

Nanoethics JCR IF (2018): 1.359

Q1 in History and Philosophy of Science

Q2 in InManagement of Technology and Innovation

1.4.2 Paper II

Paper II is titled “Machine Learning as proposal for a better application of food nanotechnology regulation in European Union” and it has been published in a JCR-Q1 journal (Current Topics in Medicinal Chemistry). It presents an exploration of European Union Regulation for food incorporating nanotechnology. Given the current gaps of scientific knowledge and the need of

1) INTRODUCTION

efficient application of food law, this paper makes an analysis of principles of European food law for the appropriateness of applying biological activity Machine Learning prediction models to guarantee public safety.

Santana, R., Onieva, E., Zuluaga, R., Duardo-Sánchez, A., & Gañán, P. (2020). Machine Learning as a Proposal for a Better Application of Food Nanotechnology Regulation in the European Union. *Current Topics in Medicinal Chemistry*, 20(4), 324-332.

DOI: 10.2174/1568026619666191205152538

Current Topics in Medicinal Chemistry JCR IF (2018): 3.442

Q1 in Drug Discovery

1.4.3 Paper III

Paper III is titled “The role of Machine Learning in centralized authorization process of Nanomedicines in European Union”, and it has been admitted in a JCR-Q1 journal (*Current Topics in Medicinal Chemistry*). It presents an exploration of European Drug Regulation for drug incorporating nanotechnology and how guidances can help apply centralized authorization process in European Union, by incorporating Machine Learning methods such as PTML.

Santana, R. et al., The role of Machine Learning in centralized authorization process of Nanomedicines in European Union.

Admitted in Current Topics in Medicinal Chemistry.

Current Topics in Medicinal Chemistry JCR IF (2018): 3.442

Q1 in Drug Discovery

1.4.4 Paper IV

Paper IV is titled “PTML Model of ChEMBL Compounds Assays for Vitamin Derivatives” and it has been published in a JCR-Q1 journal (Combinatorial Science). Through this study, we propose a PTML combinatorial model for ChEMBL results on biological activity of vitamins and vitamins derivatives. The linear discriminant analysis (LDA) model presented the following results for training subset a: Specificity (%) = 90.38, sensitivity (%) = 87.51, and accuracy (%) = 89.89. The model showed the following results for the external validation subset: specificity (%) = 90.58, sensitivity (%) = 87.72, and accuracy (%) = 90.09. Different types of linear and nonlinear PTML models, such as logistic regression (LR), classification tree (CT), naïve Bayes (NB), and random Forest (RF), were applied to contrast the capacity of prediction.

1) INTRODUCTION

Santana R. et al., PTML Model of ChEMBL Compounds Assays for Vitamin Derivatives, *Combinatorial Science*, vol. 22, no. 3, pp.129-141.

DOI: 10.1021/acscombsci.9b00166

Combinatorial Science (2018): 3.2

Q1 in Chemistry & Medicine

1.4.5 Paper V

Paper V is titled “Predicting Coated-Nanoparticle Drugs Release Systems with Perturbation-Theory Machine Learning (PTML) Models” and published in a JCR-Q1 journal (*Nanoscale*). we combine Perturbation Theory and Machine Learning (PTML algorithm) to train a model that is able to predict the best components (NP, coating agent, and drug) for Nanoparticle Drug Delivery Systems (DDNS) design. In so doing, we downloaded a dataset of >30 000 preclinical assays of drugs from ChEMBL. We also downloaded a nanoparticle dataset formed by preclinical assays of coated Metal Oxide Nanoparticles (MONPs) from public sources.

Santana, Ricardo, et al., Predicting Coated-Nanoparticle Drugs Release Systems with Perturbation-Theory Machine Learning (PTML) Models, *Nanoscale*, vol. 12, pp. 13471-13483.

DOI: <https://doi.org/10.1039/D0NR01849J>

Nanoscale JCR IF (2018): 6.970

Q1 in Materials Science, Nanoscience and Nanotechnology

1.4.6 Paper VI

Paper VI is titled “Designing nanoparticle release systems for drug–vitamin cancer co-therapy with multiplicative perturbation-theory machine learning (PTML) models” and it has been published in a JCR-Q1 journal (Nanoscale). It presents a PTML model able to predict biological activity of drug–vitamin release nano-systems (DVRNs). The best PTML model found showed values of specificity, sensitivity, and accuracy in the range of 83–88% in training and external validation series for >130 000 cases (DVRNs vs. ChEMBL data pairs) formed after data fusion. To the best of our knowledge, this is the first general purpose model for the rational design of DVRNs for cancer co-therapy.

Santana, R. et al., Designing nanoparticle release systems for drug–vitamin cancer co-therapy with multiplicative perturbation-theory machine learning (PTML) models, *Nanoscale*, vol. 11, no. 45, pp. 21811-21823.

DOI: <https://doi.org/10.1039/C9NR05070A>

Nanoscale JCR IF (2018): 6.970

1.4.7 Paper VII

Paper VII is titled “PTML Model for Selection of Nanoparticle, Anticancer Drug, and Vitamin in the PTML Model for Selection of Nanoparticle, Anticancer Drug, and Vitamin in the Design of Drug-Vitamin Nanoparticle Release Systems for Cancer Co-Therapy”. It has been sent a JCR-Q1 journal

1) INTRODUCTION

(Molecular Pharmaceutics). It presents a PTML model able to predict biological activity of drug–vitamin release nano-systems (DVRNs) using metric-based PTOs and then another model that incorporates information of the anticancer drug and the vitamins inside the DVRNs. We expressed all this information with perturbation theory operators and developed a qualitatively new PTML model that incorporates information of the anticancer drugs. This new model presents 96–97% of accuracy for training and external validation subsets. Furthermore, we carried out a comparative study of ML and/or PTML models published and described how the models we are presenting cover the gap of knowledge in terms of drug delivery.

Santana, Ricardo, et al., PTML Model for Selection of Nanoparticle, Anticancer Drug, and Vitamin in the Design of Drug-Vitamin Nanoparticle Release Systems for Cancer Co-Therapy, *Molecular Pharmaceutics*, vol. 17, no. 7, pp. 2612–2627

DOI: <https://doi.org/10.1021/acs.molpharmaceut.0c00308>

Molecular Pharmaceutics JCR IF (2018): 4.396

Q1 in Drug Discovery

1.5 Outline

This thesis is composed by 8 chapters. Apart from introduction and general conclusions, each of those chapters summarizes the publication that is dedicated to develop the issue in the context of the research. Each publication is self-contained. Therefore, each of them includes state-of-the-art, proposes its research questions, and describes the methodology that it applies, to solve gap of knowledge related to the biological behavior of the components of nanosystems.

After this introduction (Chapter 1), chapter 2, 3 and 4 present an exploration of the European regulation for cosmetic, food and drug sectors, respectively. Chapter 5 proposes a PTML model for vitamin derivatives. Chapter 6, presents a PTML model for nanosystems composed by metal oxide nanoparticles with or without coating agent and derivative vitamin. Chapter 7 shows a PTML model for nanosystems that include a non metal oxide nanoparticle and vitamin derivatives. Chapter 8 presents a PTML model for non metal oxide nanoparticles, with the information of the anticancer drug, and vitamin derivative.

The papers presented in this dissertation are appended at the end of each chapter.

*Equipped with his five senses, man
explores the universe around him and
calls the adventure Science*

Edwin Hubble

CHAPTER

2

2) European Nanotechnology Regulation (Cosmetic sector)

Over the last year, the growth of the use of nanomaterials in products has been notable. The sector of cosmetics, food and drugs has shown how different nanomaterials can be incorporated in order to improve their different functions. European Regulation 1223/2009, regulates cosmetics that use nanotechnology to improve characteristics. It does special focus on the safety of these products. In this chapter, we present an exploration of cosmetic regulation, to show the efficacy of this regulation in the European context.

The methodology applied is a systematic study of the Regulation 1223/2009 to

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

infer the principles in which it is based to analyze the application of the control on the cosmetics market in the European Union.

As result, we identified nanomaterials that have been applied in nanotechnology industry that have not been approved as nanomaterials according to Regulation 1223/2009. In **Figure 2**, we observe the number of mentions, which are categories of cosmetics in which the nanomaterials have been used for colorant purposes. We can observe the greater use of Silver, Titanium Dioxide and Carbon Black.

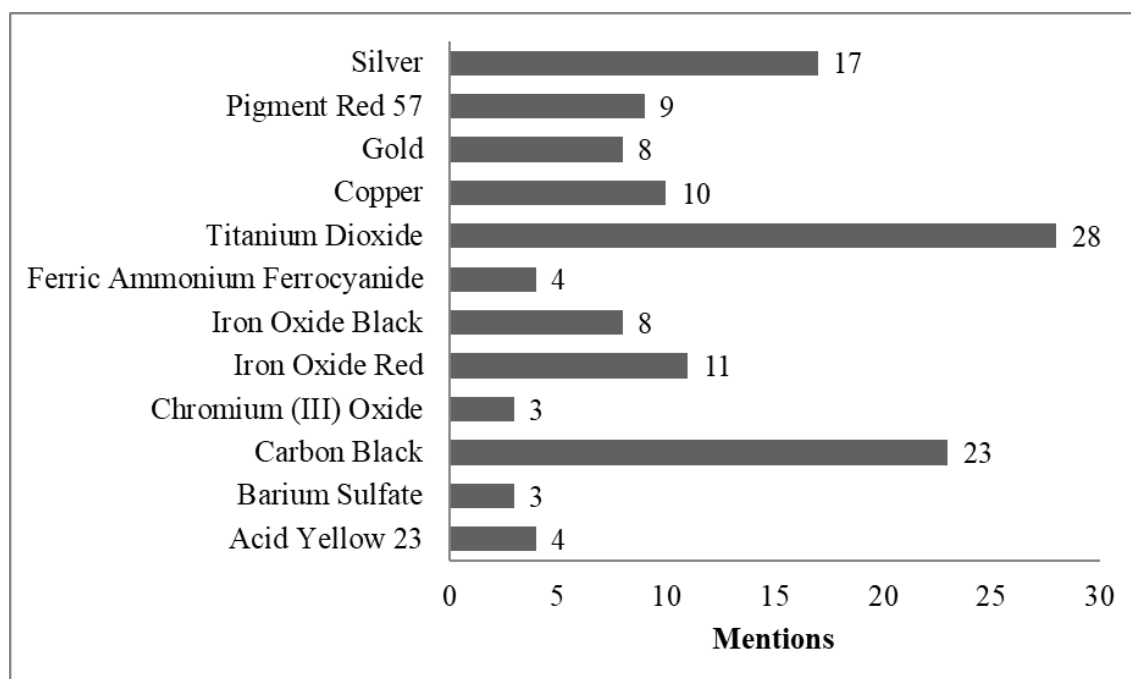


Figure 2.- Nanocolorants used in European Union in the different types of cosmetic products. Information based on the catalogue reported by the EC (Version 1 (31.12.2016))

In **Figure 3** and **Figure 4**, we find the nanomaterials that have been used UV Filters and other purposes, respectively. For better UV filter, the use of Titanium Dioxide and Zinc Oxide.

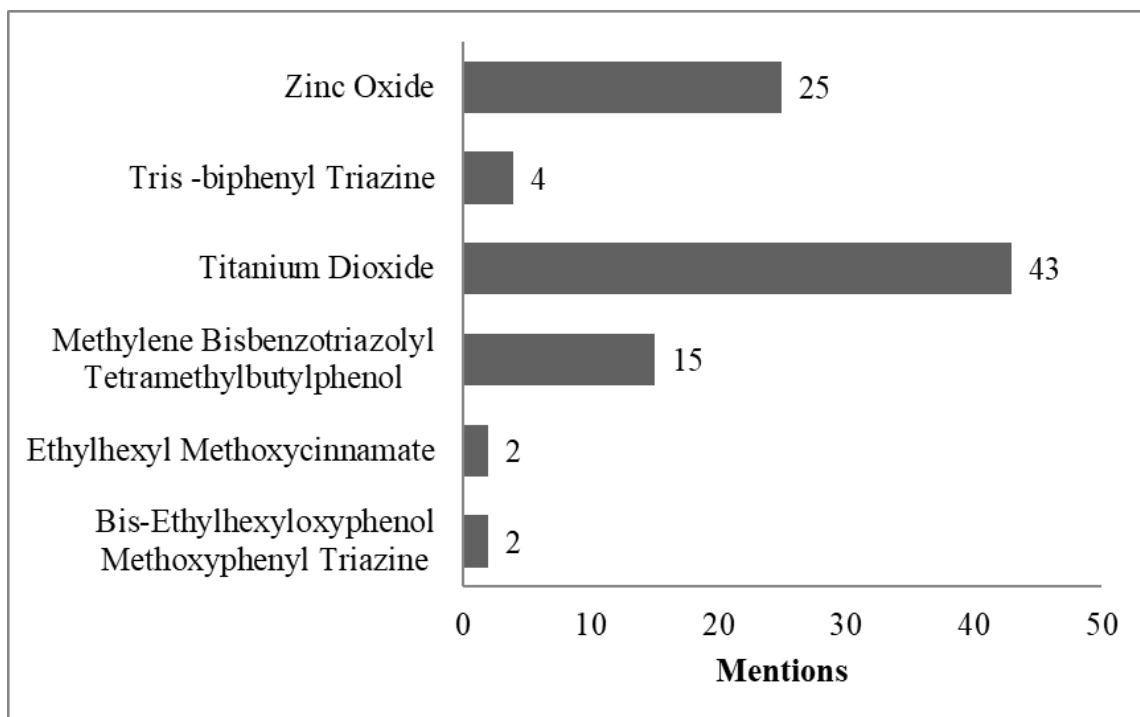


Figure 3.- NanoUV-filters used in European Union in the different types of cosmetic products. Information based on the catalogue reported by the EC (Version 1 (31.12.2016))

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

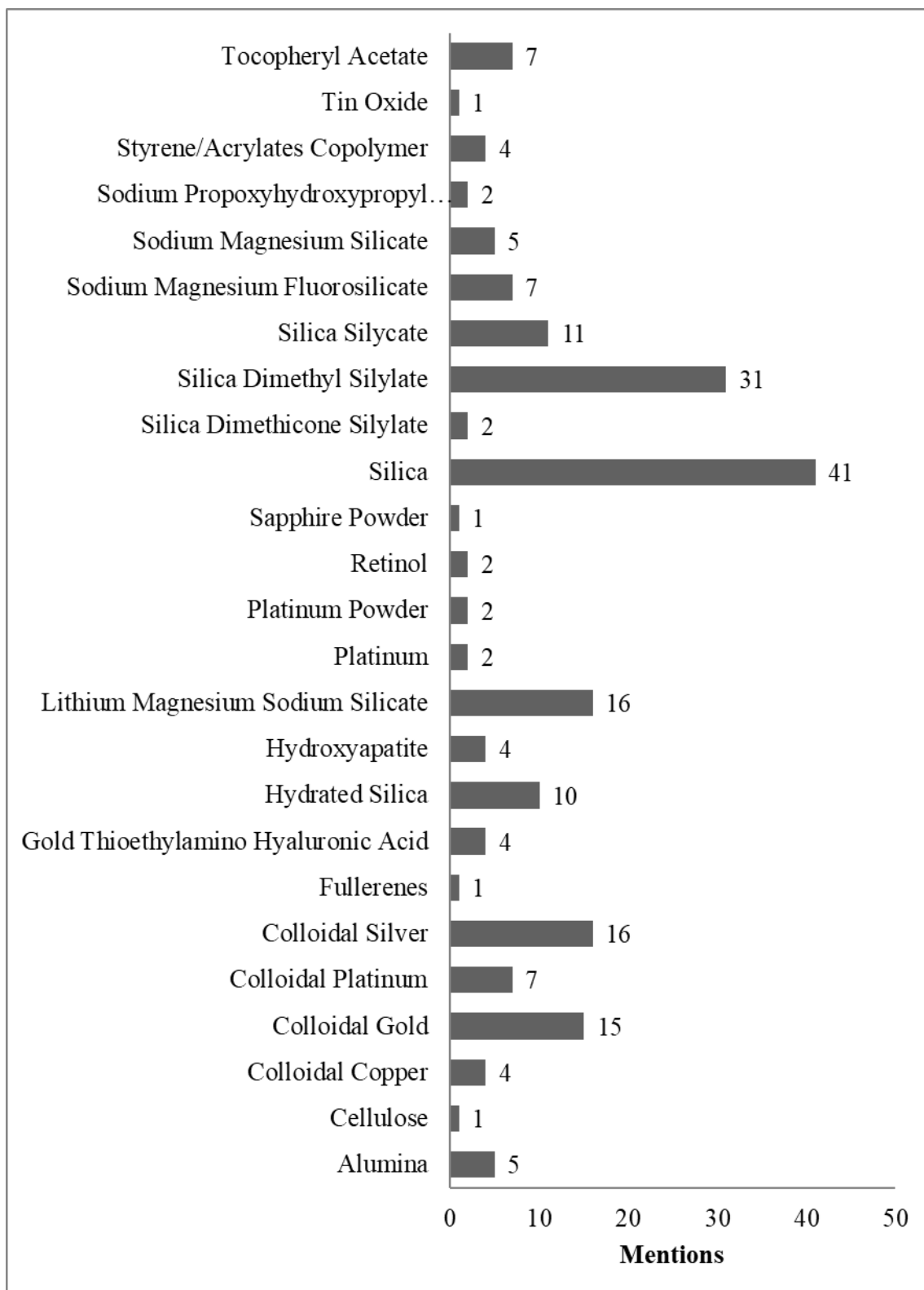


Figure 4.- Nanomaterials with other functions used in European Union in the different types of cosmetic products. Information based on the catalogue reported by the EC (Version 1 (31.12.2016))

Through this chapter we study how nanomaterials are regulated, the conditions they must fulfil to be approved. Once we explore that, we analyse if the approved nanomaterials correspond with the nanomaterials that are circulating in the market.

Lessons from the European Regulation 1223 of 2009, on Cosmetics: Expectations Versus Reality

Ricardo Santana Cabello (*) & Piedad Gañán Rojo &

Robin Zuluaga

R. Santana Cabello (*)

Grupo de Investigación sobre Nuevos Materiales, Universidad Pontificia Bolivariana, Circular 1 # 73-76,

Bloque 22B 1r piso,

Medellín, Colombia

e-mail: ricardo.santana@upb.edu.co

P. Gañán Rojo

Facultad de Ingeniería Química, Universidad Pontificia Bolivariana, Circular 1 No. 70-01, Bloque 9

Medellín, Colombia

e-mail: piedad.ganan@upb.edu.co

R. Zuluaga

Facultad de Ingeniería Agroindustrial, Universidad Pontificia Bolivariana, Circular 1 No. 70-01, Bloque

11B-411

Medellín, Colombia

e-mail: robin.zuluaga@upb.edu.co.

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

Abstract The aim of this paper is to conduct an analysis of the application of the specific rules of nanotechnology incorporated in Regulation No. 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. It has been ten years since the European Commission had issued its proposal to start the co-decision procedure to create Regulation 1223 of 2009. Although it has been praised for noting the regulatory difference of nanomaterials over the rest of the chemicals, what has been the efficacy of the standard? It is concluded that despite what it meant, the regulation has encountered technical obstacles, thus rendering the objectives relating to nanotechnology that were proposed from the European Commission unfulfilled. This finding is inferred through legal dogmatic methodologies and the identification of nanomaterials that have not been expressly approved. Nevertheless, products incorporating nanomaterials still circulate in the European market. The precepts about nanotechnology in the regulation should be reviewed because technical inconsistencies should be avoided in future regulations or applicable regulations in contexts other than Europe. Such inconsistencies exist with respect to the high level of protection of human health that should be ensured and the provisions intended to protect consumer safety. For instance, the catalog of nanomaterials in circulation does not indicate the materials that have been approved or their toxicological profiles. To date, no comparison studies have been presented between the expectations and legislative objectives set as embodied in the regulation and debated in the European Parliament involving the actual efficacy of this regulation.

Keywords Nanomaterials . Cosmetics . Regulation . Toxicity

Introduction to Nanotechnology and Regulation

Nanotechnology is a revolution in many sectors because of its intense and extensive development, especially in recent years [1–5]. As a cross-cutting technology, it is able to improve the materials involved in different processes, thereby giving rise to products that could only be created by the imagination years ago. Although the possibilities for sophistication and refinement of functionalities are high in the near future, to date, there is a significant range of nanomaterials incorporated in products that circulate in a cross-border manner in terms of jurisdiction and market [6]. These nanomaterials will be shown in the section “Commercialized Nanomaterials vs Regulated Nanomaterials”. Academia and European legislative institutions coordinate efforts to implement the most accurate

regulation [7]. Although the challenges presented by the regulation of nanotechnology have a worldwide dimension, this article focuses on the European context [8]. It is possible to adopt different models, as in the case of the USA, which is very dissimilar from the European context. The US Food and Drug Administration (FDA) issued a *Final Guidance for Industry: Safety of Nanomaterials in Cosmetic Products*, which states that “the current framework for safety assessment sufficiently robust and flexible to be appropriate for a variety of materials, including products containing nanomaterials” [9].

Although there is no doubt about the benefits of nanomaterials given their physicochemical properties, there is a significant lack of information on the adverse effects both in humans and in the environment [10], mainly given the complexity to develop toxicological profiles of the different nanomaterials in specific environments. Research into the toxicology of nanomaterials has increased profusely in recent years, and more information is available, especially in relation to toxicity testing methods and life cycle assessment [11]; however, studies to address challenges, such as universal aerosol sampler for airborne nanostructured materials, instruments to monitor waterborne engineered nanomaterials or “smart sensors” that indicate potential harm, have shown poor progress over the last ten years [12]; in addition, the tendency of nanoparticles to form agglomerates and the obstacles to isolating nanomaterials from colloidal systems among other aspects make it difficult to know the real effects of the nanomaterials in each cosmetic material [11].

In terms of regulations, there is still a long way to go, although the number of explicit regulations for nanotechnology has been increasing at the European level since the publication of the influential [13] Royal Academy of Engineering’s (RS/RAEng) report about nanoscience and nanotechnologies:

We recommend that regulatory bodies and their respective advisory committees include future applications of nanotechnologies in their horizon scanning programmes to ensure any regulatory gaps are identified at an appropriate stage [14].

Regulation No. 1223/2009 (from now on: the regulation) is the first normative instrument that includes provisions designed to regulate nanotechnology expressly at national, international, or supranational level [15] with the understanding that it must be a differentiated legal object from the rest of chemical substances [16]. The justification by

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

which its scope is limited to cosmetic products is that there were, at that time, great expectations in the involvement of this sector, and subsequent European regulations provided nanospecific frameworks for other sectors, such as Regulation No. 1169/2011, Regulation (EU) No. 10/2011, Regulation (EU) No. 10/2011, and Regulation (EU) No. 1215/2012 [13], which were created to improve the performance of its products with incorporated nanotechnology. These expectations were met. In the year 2011, the market for cosmetic and personal care products containing nanomaterials was estimated at US\$17 billion out of the US\$375 billion cosmetic and personal care industry worldwide market ([17], 36). Article 16 of the regulation states that a high level of protection of human health would be ensured with regard to cosmetic products incorporating nanomaterials.

This regulation corresponds to an update of cosmetic standards in the European Union after 35 years. The different stages of the co-decision procedure, a decision-making procedure that was replaced through the Lisbon Treaty by the current ordinary legislative procedure that gave rise to the regulation, were very significant because of the predisposition on the subject of nanomaterials of the European institutions involved, especially the European Parliament (EP).

Although it was conceived at the time as a normative success, the statements of Members of the European Parliament (MEPs) call into question whether it is currently fulfilling its objectives, as they had been defined ten years ago. In particular, the next question must be answered: Did it meet the objectives proposed, bearing in mind that the principle of a high level of protection of human health is crystallized in Article 16?

European Regulation No. 1223/2009

Gestation and Objectives of the Regulation

This section presents an analysis of the genesis and the specific objectives of the regulation concerning the standards related to nanotechnology.

The main purpose of the regulation was to harmonize and recast the Council Directive 76/768/EEC of 27 July 1976 and its subsequent 55 revisions [18]. Thus, within the framework of the “Community Lisbon Program: A strategy for the simplification of the

regulatory environment”, as well as within the strategy of the Annual Commission Policy in 2007, approximately 3500 pages of regulation in all Member States as a result of the transposition of this Directive and its subsequent amendments, would become one regulation.

The European Commission took the first step on 5 February 2008 when the proposal for a regulation (from now on: the proposal) was presented [19] to the EP and the Council of the European Union (CEU) under the leadership of Günter Verheugen who was responsible for preparing the Committee’s work on the subject. The European Commission was aware of the appropriate opportunity to incorporate substantial changes to improve safety standards and the efficiency of administrative processes. The substantive changes that were presented through the proposal were those related to the introduction of definitions, glossary of ingredients, safety assessments, reinforcement of market control, regulation of carcinogenic, mutagenic, or toxic for reproduction substances (CMR), among others.

The Commission staff working paper accompanying the proposal included the following objectives [20]:

Objective 1. To remove legal unclarities and inconsistencies. These inconsistencies can be explained by the high number of amendments (55 to date) and the complete absence of any set of definitions. This objective also includes several measures to facilitate management of the Cosmetics Directive with regard to implementing measures.

Objective 2. To remove divergences in national transposition which do not contribute to product safety but add to the regulatory burden and administrative costs.

Objective 3. To ensure that cosmetic products placed on the EU market are safe in the light of innovation in this sector.

Objective 4. To introduce a possibility in exceptional cases to regulate CMR 1, 2 substances on the basis of their actual risk.

However, the Commission staff emphasized that these objectives must not [20]:

- compromise the high level of product safety in this sector today (adverse effect 1);
- lead to changes to the arrangements for phasing out animal testing (adverse effect 2);

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

- have a negative impact on the functioning of the internal market for cosmetic products (adverse effect 3); and
- create unnecessary differences from the regulatory frameworks in non-EU states (adverse effect 4).

The proposal did not mention nanomaterials concerning the adverse effect 4. This aspect is important because to ensure technological development, an adequate and predictable regulatory framework must be ensured. In the same year (2008), the European Commission issued a communication stating the following: “Current legislation covers in principle the potential health, safety and environmental risks in relation to nanomaterials” [21]. However, in the Commission staff working paper, problems that should be addressed through the new regulation were identified: (1) legal unclarity/inconsistencies and burdensome management of the Cosmetics Directive; (2) incoherent and resource-intensive transposition without adding value; (3) ensuring the safety of cosmetic products in the light of innovation; (4) addressing substances classified as CMR 1 and 2 by considering, in exceptional cases, safe exposure limits. The Commission staff working paper refers to the problem of nanomaterials in the following terms:

Use of known ingredients in nanosizes: Future innovation is likely to be based on new physical characteristics of existing substances: the most prominent example is the use of particles in micronised forms. Micronised particulars are presently in use as physical UV-filters. As such, their use in cosmetic products has to be authorised by the European Commission. However, other uses in other types of cosmetic products cannot be excluded in the future [20].

The preceding paragraph leads to the conclusion that in the European Commission, possible challenges associated with the technology based on nanomaterials were identified. On the same terms, the European Economic and Social Committee gave an opinion [22] on the relevance of the new regulation, noting the potential economic impact of developing new methods of monitoring and evaluating chemicals in cosmetics with small- and medium-sized enterprises which would generate transaction costs difficult to afford by small- and medium-sized enterprises in the European market. However, this opinion does not address the need to regulate nanomaterials expressly.

Once the proposal was presented to the EP, the parliament's committee to deal with the matter was the Committee on the Environment, Public Health and Food Safety (from now on: ENVI). On 26 February, Dagmar Roth-Behrendt was appointed as rapporteur [23].

On 8 December 2008, ENVI approved the document with 44 votes in favor, zero against, and zero abstentions with the necessary amendments to be submitted for approval by the EP [23]. Among the main amendments are those relating to nanotechnology. In the explanatory memorandum of the document presented by ENVI, the regulation of nanomaterials was justified as follows [23]

Already today nanomaterials are part of many products on the market. In 2006, the Commission estimated the amount of cosmetic products containing nanoparticles of approximately 5%

There is a wide range of definitions what is to be called a nanomaterial which mostly refer to the size of the substance. To avoid legal uncertainty, it is important to make sure what is meant by nanomaterial. Therefore, the rapporteur introduces a definition to this regulation which is based on a definition developed by the SCCP (Scientific Committee on Consumer Products) in December 2007.

Because of their small size, nanomaterials contain special and very positive characteristics but, at the same time, new risks can be created. Therefore, these products should be evaluated by the SCCP on the basis of a nanospecific safety assessment prior to their placing on the market to ensure the safety for consumers. The rapporteur suggests the introduction of a transitional period for existing products, which contain nanomaterials.

Nanomaterials that are used as colorants, preservatives, and UV-filters are already covered in Annex IV, V and VI of this regulation and already have to be positively listed by the Commission after consultation with the SCCP.

To ensure the safety of cosmetic products, an evaluation by the SCCP for all products containing nanomaterials should be required. The rapporteur therefore tables amendments which introduce a congruent procedure for all nanomaterials.

As the research on nanomaterials is still progressing, the Commission is requested to regularly review this regulation in the view of nanomaterials [23].

The next step taken during the process of creating the regulation was the debate generated in the EP. During the discussion session of 23 March 2009, one day before the vote, the MEPs added interesting aspects of nanomaterials as specific objects regulated in the proposal and the proposed objectives. These aspects are summarized in Table 1.

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

The parliamentary approval was given in the first reading on 24 March 2009 with a large majority (633 votes to 29, with 11 abstentions) [25]. The regulation would finally have the signature of the President of the EP and the CEU on 30 November 2009 [26]. It must be mentioned that no further conclusions should be drawn from the fact of the ENVI proposal being approved on the first reading because during the period from July 2009 to June 2014, 85% of the ordinary legislative procedures ended in the first reading [27].

As noted in the interventions in the EP, despite initial discrepancies, especially with regard to labeling and the reporting system of nanomaterials, the ENVI proposal was well received by the different parties, with the consideration that nanomaterials should be regulated and, as the result of such regulation, consumer safety would increase [24].

Considerations About Consumer Safety

Regarding the regulation of nanomaterials through the regulation, we can highlight the following provisions related to the increase in consumer safety:

Under the terms of the right of consumer information, the nanoscale of the incorporated material must be included on the label (article 19 of the regulation) [16]. The word “nano” must precede the name of the material. In this way, the consumer will be able to know the characteristics of the product that he is acquiring. However, additional research about the effectiveness of these labels referring to nanotechnology is needed.

The consumer’s right to information translates into an obligation assumed by the producer to provide information on the ingredients of a product. In this way, the consumer may be warned so that he can manage his consumption according to his needs. Thus, the institutions promote responsible consumption and adequate risk management. As will be seen below, the number of cosmetics not approved by the European Commission and circulating in the market is significant. If one of the few nanomaterials approved by the European Commission is included on the label of a product, the consumer has no way to check how that nanomaterial reacts with the rest of the ingredients. In short, the consumer has the information of the scale of the material that is incorporated into the product but still has little information on how that nanomaterial

behaves in that product or on the advantages and disadvantages of adding a specific nanomaterial and thus on how to manage its consumption in a responsible way.

The other main contribution was to create a system capable of storing information about the circulating nanomaterials, the cosmetics containing those nanomaterials, and the use of these cosmetics (article 16 of the regulation) [16]. After a process described below, the European Commission decided to give authorization for commercialization and described the required conditions. This case-by-case system is understandable given the particularities of each nanomaterial and the initial situation of ignorance of the toxicity of nanomaterials at that time. The task of analyzing them would be extended over the years. Furthermore, the development of standards of evaluation of toxicity by CEN/TC 352 of the European Committee for Standardization in terms of assessment of toxicity is still needed, especially about dermal exposure.

The storage and processing of these data are also done for subsequent distribution to consumers, which is the reason why the regulation establishes that a catalog of commercialized nanomaterials must be published (article 16 of the regulation) [16]. Nevertheless, the information in the catalog available to the public was not sufficient to understand the characteristics of the nanomaterials better and only referred to the types of products that include a specific nanomaterial. Therefore, this has had a limited impact on the control of commercialized nanomaterials in cosmetics as will be shown below.

Therefore, consumer protection has not been as rigorous as it was thought when the regulation was developed, and such considerations must be addressed for the internal coherence and sustainability of regulations that offer efficiency and safety.

Commercialized Nanomaterials vs Authorized Nanomaterials

The above-mentioned comments made by the MEPs refer to amendments by ENVI in the final wording of the regulation. The lack of data at that time on nanotechnology and its toxic effects was reflected in recitals 30, 65, and 31 of the regulation [16]. However, there was no doubt about the possibilities offered by the nanoscale and specifically in the cosmetics sector. For this reason, the wording tries to leave open the door for new advances and discoveries of this technology, and not to hinder a development which is especially promising to small and medium enterprises (SMEs) without large economies

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

of scale, as the European Economic and Social Committee pointed out in the abovementioned opinion.

In reaction to the discussion, a system of notifications was created to accumulate the necessary information and thus provide better monitoring of the nanomaterials used in cosmetics in the European Union.

Since the regulation was enforced, two types of notification of cosmetics are provided in case they incorporate nanomaterials. First, according to Article 13 of the regulation, a mandatory notification must be completed at the time the online registration of any cosmetics through the Cosmetic Products Notification Portal (from now on: CPNP) created on January 11, 2012 [28]. The information to be provided is (a) the category of the cosmetic product and its name or names, enabling its specific identification; (b) the name and address of the responsible person where the product information file is made readily accessible; (c) the country of origin in the case of import; (d) the Member State in which the cosmetic product is to be placed on the market; (e) the contact details of a physical person to contact if necessary; (f) the presence of substances in the form of nanomaterials and (i) their identification including the chemical name (IUPAC) and other descriptors as specified in point 2 of the Preamble to Annexes II to VI to this regulation and (ii) the reasonably foreseeable exposure conditions; (g) the name and the Chemicals Abstracts Service (CAS) or EC number of CMR substances, of category 1A or 1B, under Part 3 of Annex VI to regulation (EC) No 1272/2008; and (h) the frame formulation allowing for prompt and appropriate medical treatment in the event of difficulties.

Table 1 Interventions of MEPs during the EP session of 23 March 2009 in Strasbourg. All speeches are available on the EP website [24]

Name of the MEP	Intervention session of 23 March 2009
Dagmar Roth-Behrendt, rapporteur	<p>“The Commission also rightly recognised the fact that <i>new technologies, such as nanotechnology, need special attention, in particular when we are dealing with microscopic particles which may be able to pass through layers of skin.</i> We simply want to ensure that they present no danger. Here, too, I am satisfied that we have achieved a compromise which I can wholeheartedly support...”; “[...] <i>We did not always agree on issues such as how to deal with notification in relation to nanotechnology and what should be done in terms of labelling, but we managed to reach an excellent compromise. I am very pleased about this</i>”; “[...] <i>Labels allow consumers to make free and informed choices. Consumers have a right to be informed about nanotechnologies and to know that a specific substance contains particularly small, even microscopic particles.</i> They have the right to decide whether they want to use sun lotion and whether they want to use the sun lotion on their children. Consumers have the right to decide”.</p>
Günter Verheugen, Vice-President of the Commission	<p>“Mrs Roth-Behrendt has already spoken on the subject of nanotechnology. We have found a solution in this case, which I would like to describe as a model because this same solution will be used again later this week with regard to other important pieces of legislation. The specific provisions concerning the nanomaterials used in cosmetics introduce a mechanism for providing the necessary information before the materials are made publicly available on the market, which ensures that relevant data on safety has to be presented and the authorities have time to take any necessary safety precautions”. “In addition to ensuring products safety, the proposal improves the level of information provided to consumers. An example is the addition to the list of ingredients of information on which substances appear in nano form”.</p>
Françoise Grossetête, PPE-DE Group	<p>“We have, in fact, had much discussion about nanomaterials, which are used in cosmetics, particularly in sun protection products, and <i>which must be subject to very strict requirements in relation to safety, but without standing in the way of innovation</i>”.</p>
Daciana Octavia Sârbu, PSE Group	<p>“<i>The use of nanomaterials is a promising solution in this area, but they are to be assessed and declared safe by the Scientific Committee for products intended for consumer use, while the use of alternative methods is an initiative which must continue to be supported</i>”.</p>
Chris Davies, ALDE Group	<p>“My colleague Frédérique Ries, who cannot be with us tonight, wanted to ensure that steps were taken to try and avoid the marketing of counterfeit products, to strengthen product traceability and to tighten up restrictions on the making of false claims about the beneficial effects of these products. <i>She wanted to support clear labelling of products about the content of nanomaterials. We have made progress on all these areas</i>”.</p>
Hiltrud Breyer, Verts/ALE Group	<p>“Mr. President, protecting human health is also the primary objective when it comes to cosmetic products. <i>We are making history with this vote, which is the first time that specific regulations have been drawn up for the use of nanomaterials in cosmetic products, and we are breaking new ground.</i> I am, of course, particularly pleased to be able to say that it was an initiative of the Group of the Greens/European Free Alliance that led to this ground-breaking event. We Greens were the driving force, we placed it on the agenda, and I would like to wholeheartedly thank the rapporteur, Mrs. Roth-Behrendt, for her clear and unwavering support”. “<i>I am also pleased to be able to praise the Commission for changing its mind. Until now, it had continually stressed that the existing legislation was sufficient to guarantee the safety of nanomaterials. Now, it has clearly stated that we do indeed need specific regulations</i>”.</p>
Eva-Britt Svensson, GUE/NGL Group	<p>“The biggest stumbling block in the negotiations with the Council was precisely nanomaterials”. “<i>The agreement will entail better protection for European consumers when nanomaterials are used in hair dyes, UV filters and so on. They will undergo a safety assessment before the products are allowed onto the market and the cosmetic products industry will also need to notify the Commission of the use of nanomaterials in any of their other products</i>”.</p>

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

Table 1 (continued)

Name of the MEP	Intervention session of 23 March 2009
Irena Belohorská (NI)	"After all, developments in chemistry and in cosmetics itself have brought enormous and <i>fundamental changes</i> . I am referring here to the use of nanomaterials, so frequently mentioned here. These can have both positive and negative effects on human health".
Horst Schnellhardt (PPE-DE)	"The use of nanomaterials has forced us to address the issue again. Within the framework of <i>preventive consumer protection, the decisions concerning labelling are welcome, while the opportunity of provisional acceptance, in view of the state of scientific discoveries is also acceptable</i> . At this point, I would also like to warn against panic-mongering, as has happened in the case of other developments, and would instead advise a scientific examination of the whole matter"
Margrete Auken (Verts/ALE)	"The most important thing is that nanomaterials have at last been included, which has been a tough fight. It is although the industry has tried to stifle the debate on the safety of nanomaterials. They would be very pleased if we would just accept these substances as unproblematic and wonderful. There has been no hint of the public concern that there has been surrounding GMOs, for example". "We are proud that nanomaterial has now been included. <i>It is to be tested, labelled and, where a number of products are concerned—UV filters, dyes and preservatives—it will now be the producer who has to guarantee safety, while the Commission is to provide detailed information and find time to monitor the rest. Finally, we have also managed to include labelling so that consumers can see what they are buying and putting on their skin</i> ".
Péter Olajos (PPE-DE)	"There are nano-applications and products intended for direct consumer use, such as clothing and food, including cosmetic products, in the case of which an inadequately circumspect approach may result in people experiencing, literally in the flesh, the potentially harmful consequences". "It is precisely for this reason that <i>it is important for people to know what kinds of preparations they are using; appropriate and detailed labelling is therefore indispensable, and the responsibility of the manufacturer is essential</i> "
Zuzana Roithová (PPE-DE)	"The extensive discussion here has focused mainly on labelling because this often misleads consumers, and I therefore warmly welcome the fact that new claims about the effects of products must be documented. There has also been a very lively discussion here—and not only here—on licensing nanomaterials and of course the elimination of carcinogenic materials from cosmetic products. <i>I do not agree that messages about the content of nanomaterials in products should take the form of warnings. It is important for us to have a list of licensed nanomaterials that are not harmful but improve the quality of a product</i> . There is, of course, no point in scaring consumers. <i>Minimum standards should ensure consumer safety. I certainly consider counterfeiting to be a serious problem and I would also like to draw attention to the limited capacity of monitoring bodies at national level to actually monitor everything</i> ". "I am delighted that the text includes a uniform definition of nanomaterials and I also welcome the fact that we will be able to amend it so that it keeps up with the latest scientific developments. I also welcome the fact that the directive will actually become a regulation and will have greater legal emphasis. I therefore welcome this piece of work and I congratulate all of the rapporteurs for managing to reach a consensus over an issue as sensitive as the introduction of cosmetic products onto the European market based on scientific developments"
Eija-Riitta Korhola (PPE-DE)	"Obsolete legislation in the cosmetics industry poses a special threat to health and the extent to which we can rely on the law. Claims about nanoparticles and cosmetic products are a good example of this. <i>Whereas the positive characteristics of nanomaterials are more or less familiar, the risks are largely uncharted. Similarly, the special characteristics of cosmetic products, which have a direct effect on the decision to purchase them, have been impossible to verify with any certainty</i> "

This is a general and therefore less detailed notification of the nature and characteristics of the incorporated nanomaterials. Carried out in practice when the product is reported online, there is a section to answer the question: Does the product contain nanomaterials? If yes, two additional questions should be answered: If the cosmetic is designed to be rinsed off or to left on as well as the route of exposure: skin, oral, or inhalation. Next, the responsible person must search on the *CosIng* database, the nanomaterial in particular that is included as an ingredient in the cosmetic. In case there is no such nanomaterial on the database, there is the option to include a new one.

Second, a notification of nanoproducts specifically applicable to products containing nanomaterials that are not regulated by Annexes IV, V, and VI is provided via the CPNP under the terms of article 16. Unlike the previous notification, this must be carried out six months before the product is marketed. Article 16 spells out the information, which the commission may recommend when it deems it appropriate, to be provided by the responsible person or the delegated person: (a) the identification of the nanomaterial including its chemical name (International Union of Pure and Applied Chemistry, IUPAC) and other descriptors as specified in point 2 of the Preamble to Annexes II to VI; (b) the specification of the nanomaterial, including the size of the particles and physical and chemical properties; (c) an estimate of the quantity of nanomaterial contained in cosmetic products intended to be placed on the market per year; (d) the toxicological profile of the nanomaterial; (e) the safety data of the nanomaterial relating to the category of cosmetic product as used in such products; and (f) the reasonably foreseeable exposure conditions. To perform these tasks, the CPNP website has a “Nanomaterials” tab on the main menu. Then, the “Notify Nanomaterial” tab is displayed, and the responsible person can include all the information required, and in most cases, the integrated system of *CosIng* facilitates the introduction of the data.

The fact that notifications are required for cosmetics and not nanomaterials is a sign of the high relevance of such materials since the toxicity of the specific nanomaterial in a solution is indicated and economic efficiency is increased because the person reporting the toxicological profile is not the manufacturer of the nanomaterial, who may not know all the nanomaterial uses, but the person who incorporates them into the product, which is more efficient since the cost of information given to the cosmetologist will be lower than that of the manufacturer.

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

Once the European Commission receives the notification that a cosmetic product includes nanomaterials, it verifies that the procedure has been carried out in an adequate way and required information is not missing. In the event that more information is needed, an electronic notification is made to the person responsible for providing such information. From the moment the information is complete, two options exist: (1) the European Commission considers that the nanomaterial does not compromise the safety of human beings and finishes the process by storing the notification in the database of the Commission or (2) the European Commission is in doubt of its safety and forwards the dossier to the Scientific Committee on Consumer Safety (from now on, SCCS) to issue an opinion under the principles of independence, transparency, and confidentiality [29].

The SCCS may request further information from the responsible person if it deems it necessary and has six months to express an opinion after the communication from the European Commission or an additional information requested is received. If the SCCS, in its opinion, considers that the nanomaterial is toxic, the European Commission *may* amend Annexes II and III of the regulation. At this point, one could ask why the regulation ultimately leaves the European Commission to decide on the modification of the Annexes, which permits a possible contradiction with the opinion of the SCCS. The answer is found in the nature of the SCCS as an advisory body which lacks the legitimacy to intervene in the supranational legal system and therefore the capacity to modify a European regulation.

Article 16 presents an unclear wording since it states that the nanomaterials regulated by Article 14 (colorants, preservatives, and ultraviolet filters) are not subject to the notification of Article 16. However, although the Annexes referred in Article 14 regulate substances used as colorants, ultraviolet filters or preservatives, they did not include nanomaterials at the time of writing. Indeed, all nanomaterials would have to be reported, regardless of the content because none had previously been authorized. The European Commission in its guide to complete the notification states:

This means that if the product contains nanomaterials included in such form in Annexes III, IV, V, or VI to Regulation (EC) No 1223/2009, it does not need to be notified under Article 16 [30].

Cosmetic products containing nanomaterials that comply with the requirements set out in Annex III shall also not be reported. As in the previous section, there were no nanomaterials included in Annex III.

It can be concluded that very few nanomaterials are regulated, and they will be analyzed below. They represent a low percentage of those already used in the market, which conflicts with the initially proposed objective of maintaining a high level of safety in relation to the consumer and the environment, because there are numerous nanomaterials in circulation of which there is no certainty about their toxicological profile.

The European Commission refers in recital 1 of Commission Regulation (EU) 2016/1143, when amending Annex VI, to incorporate titanium dioxide in its nanoform, admitting that it was not regulated until that moment:

Titanium dioxide is authorised both as a colorant under entry 143 of Annex IV to Regulation (EC) No 1223/2009 and as a UV-filter under entry 27 of Annex VI to that regulation. In accordance with point (3) of the Preamble to Annexes II to VI to Regulation (EC) No 1223/2009, the substances listed in Annexes III to VI to that regulation do not cover nanomaterials, except where specifically mentioned. Titanium dioxide (nano) is currently not regulated [26].

The European Commission also publishes a catalog of all nanomaterials reported through the CPNP to improve transparency in terms of consumption as well as to facilitate the work of the European toxicology research centers concerning nanomaterials used in the market, as indicated in the section for those incorporated as ultraviolet filter, colorants, and preservatives. As the development of test methods and knowledge on possible functionalities and risks of nanomaterials were expected to increase, the regulation instructs the commission to update this list on a regular basis:

By 11 January 2014, the Commission shall make available a catalogue of all nanomaterials used in cosmetic products placed on the market, including those used as colorants, UV-filters and preservatives in a separate section, indicating the categories of cosmetic products and the reasonably foreseeable exposure conditions. This catalogue shall be regularly updated thereafter and be made publicly available (Article 16 (10) (a) of the regulation).

Although the catalog should have been published until 11 January 2014, it was only published on 12 July 2017 [31]. After so much time has elapsed, we eventually have

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

valuable information on nanomaterials used in cosmetics in the European market. At the beginning of the catalog, it is clarified that this is not a list of authorization of nanomaterials but a merely informative catalog. The catalog classifies these nanomaterials by their main functions: colorants, ultraviolet filters, and functions other than colorants, preservatives and filters of ultraviolet rays.

The next question to be addressed is: where can we find such nanomaterials? The catalog does not identify the products one by one as described in the notification form for cosmetics containing nanomaterials but rather the types of cosmetics, such as face and body paint, shower and bath products, products for before or after sunbathing, concealers, eye contour products, eye contouring, eye shadow, skin care products, except beauty masks, beauty masks, foundation makeup, care products, lipstick products, lipsticks, make-up removers, nail care products, nail hardeners, nail sculpting products, nail polish and make-up, other facial make-up products, other make-up products, other products for the care and hardening of nails, other nail polish and nail polish remover, temporary styling products, shaving products, skin tone lightening products, soaps, sun protection products, external intimate care products, mouthwashes, shampoos, hydroalcoholic perfumes, and dentifrices, among others.

The catalog presents relevant information on the nanomaterials in circulation; however, it does not include exactly what are the approved nanomaterials and the toxicological profile of them since the regulation does not require it, although it would be advisable to give more information to the consumers and thus reduce the asymmetry of information. The following data are the nanomaterials used in the European Union and classified according to their function, and they are distributed according to the type of cosmetics and those that have been regulated.

Colorants Used in the European Market

Regarding the colorants, twelve nanomaterials have been registered as commercialized in the European market. They are classified under numerous types of products. As seen in Fig. 1, the most versatile nanomaterials are, from highest to lowest, titanium dioxide (28 types), carbon black (23 types), and nanosilver (17 types).

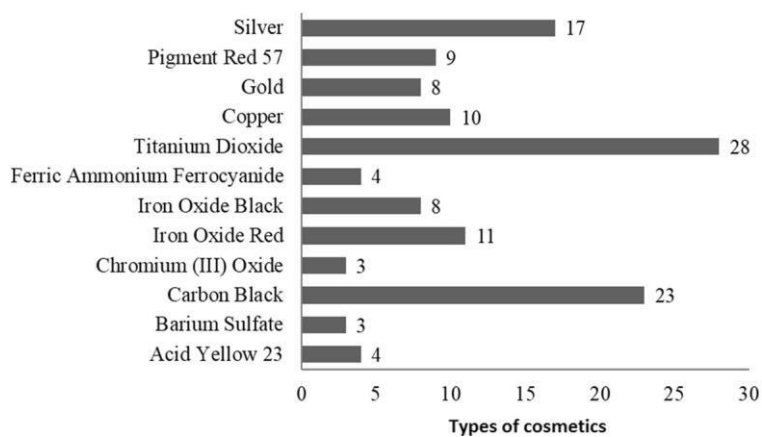
According to article 14 (1) (c), all substances used as colorant which are not included in Annex 4 and substances listed but not being applied under the conditions stated are prohibited. When a substance is classified as permissible to be used as the colorant, it does not mean that it is permitted for other uses since this will depend on how they appear in the corresponding Annexes of the Regulation.

Figure 1 shows the twelve nanocolorants that are being used in the European Union and the number of types of nanocosmetics that are being used [31]; only one of them has been expressly authorized: black coal, under reference number in Annex IV: 126a. It was integrated into Annex IV on 24 February 2017 with the intention of aligning the SCCS opinion [32], which ensured that given the data obtained from the research carried out, nanostructured black carbon with a size over 20 nm and a concentration of not more than 10% and a minimum purity of 97% does not cause adverse effects on human health when the type of exposure is cutaneous.

Due to the characteristics of the tests, the same conclusion could not be drawn in the event that such nanomaterials are inhaled.

Fig. 1 Nanocolorants used in the European Union in the different types of cosmetic products.

Information based on the catalog reported by the EC (Version 1 (31.12.2016)) [31]



2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

Regulation of UV Filters Used

Six nanomaterials have been registered as nanomaterials used as ultraviolet filters. Figure 2 shows the used distribution of the registered nanomaterials. In this case, only three out of six reported nanomaterials have been authorized and included in Annex V: Titanium dioxide, Zinc Oxide and Tris-Biphenyl Triazine.

First, titanium dioxide was covered by Annex VI through Commission Regulation 2016/1143 of 13 July 2016 [33]. In this regard, the SCCS's opinion of 22 July 2013 [34], revised on 22 April 2014 was taken into account. In this opinion, the SCCS presented the results obtained with samples of cosmetics with titanium dioxide in its nano form and ensured that it does not have adverse effects on the human being:

According to the opinion of the Scientific Committee on Consumer Safety ("SCCS") of 22 July 2013, which was revised on 22 April 2014 (2), the use of titanium dioxide (nano) as a UV-filter in sunscreens, with the characteristics as indicated in the opinion, and at a concentration up to 25% w/w, can be considered to not pose any risk of adverse effects in humans after application on healthy, intact or sunburnt skin. In addition, considering the absence of a systemic exposure, the SCCS considers that the use of titanium dioxide (nano) in dermally applied cosmetic products should not pose any significant risk to the consumer [34].

The authorized nanomaterials must have the following characteristics: purity # 99%, rutile form, or rutile with up to 5 wt% anatase, with crystalline structure and physical appearance as clusters of spherical, needle, or lanceolate shapes, median # 460 m² /cm³, coated with Silica, Hydrated Silica, Alumina, Aluminium Hydroxide, Aluminium Stearate, Stearic Acid, Trimethoxycaprylylsilane, Glycerin, Dimethicone, Hydrogen Dimethicone, Simethicone; photocatalytic activity # 10% compared to corresponding non-coated or non-doped reference, nanoparticles are photostable in the final formulation. Likewise, the Annex 6 states that in case the product is mixed with titanium dioxide in the product, the sum must not exceed the percentage of 25%. On the other hand, it established that there is no evidence to assure oral or cutaneous absorption. It also considered that it is not a safe material to use in aerosols given its lung toxicity

The second nanomaterial authorized for use in cosmetic products as a UV-filter is tris-biphenyl triazine. This modification was made through Commission Regulation 866/2014 [35] of August 8, 2014, given the opinion issued by SCCS on 20 September 2011 [36], which was due to a decree presented by SCCS on September 20:

Dermal exposure to formulations containing tris-biphenyl triazine with a mean particle size (median primary particle size) of 81 nm results in low absorption of that substance. Additionally, after oral exposure, absorption of trisbiphenyl triazine is low. No systemic effects are observed after oral or dermal exposure up to 500 mg/kg bw/day [36].

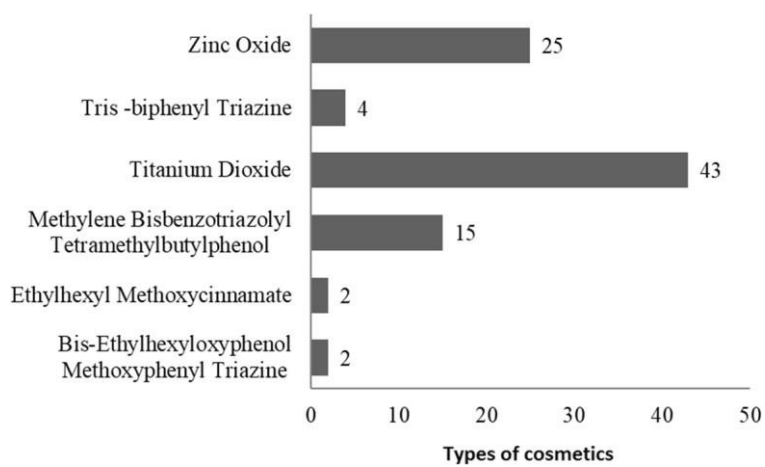
The SCCS concluded that with a concentration of less than 10 wt% as an ultraviolet filter, it is a safe substance for application to the skin, but SCCS warned of the lack of data about the possible consequences of inhalation exposure. Therefore, when it was included in Annex VI, its use as an aerosol was prohibited. In addition, Annex 6 only permitted nanomaterials with the following characteristics: median primary particle size > 80 nm, purity no. 98%, and uncoated.

The third nanomaterial authorized as a UV filter for use in cosmetic products is zinc oxide. It was amended by Commission Regulation 2016/621 [37] of 21 April 2016 on the basis of the opinion issued by the SCCS on 18 September 2012 [38] and an addendum of 23 July 2013 [39]. The same restriction indicated the required concentration of titanium dioxide, which cannot exceed 25 wt% even if mixed with non-nano-scale titanium dioxide. On the other hand, it established that there is no evidence to assure oral or cutaneous absorption. It also considered that it is not a safe material to use in aerosols given its lung toxicity.

Therefore, zinc oxide in its nanomode cannot be used in a way that exposes consumers to inhalation. The only zinc oxide nanomaterials that are allowed are those having the following characteristics: purity no. 96%, with wurtzite crystalline structure and physical appearance as clusters that are rod-like, star-like, and/or isometric shapes, with impurities (50% of the number below this diameter) > 30 nm and D1 (1% below this size > 20 nm, water solubility < 50 mg/L uncoated, or coated with triethoxycaprylyl silane, dimethicone, dimethoxydipheylsilanetriethox and caprylylsilane cross-polymer, or octyl triethoxy silane.

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

Fig. 2 Nano UV filters used in the European Union in different types of cosmetic products. Information based on the catalog reported by the EC (version 1 (31.12.2016)) [31]



Regulation for Other Commercialized Nanomaterials

Twenty-five nanomaterials used with different functions of UV filters, colorants, or preservatives have been registered. As seen in Fig. 3, the most versatile nanomaterials are (from highest to lowest) silica, silica dimethyl silylate, colloidal silver, lithium magnesium sodium silicate.

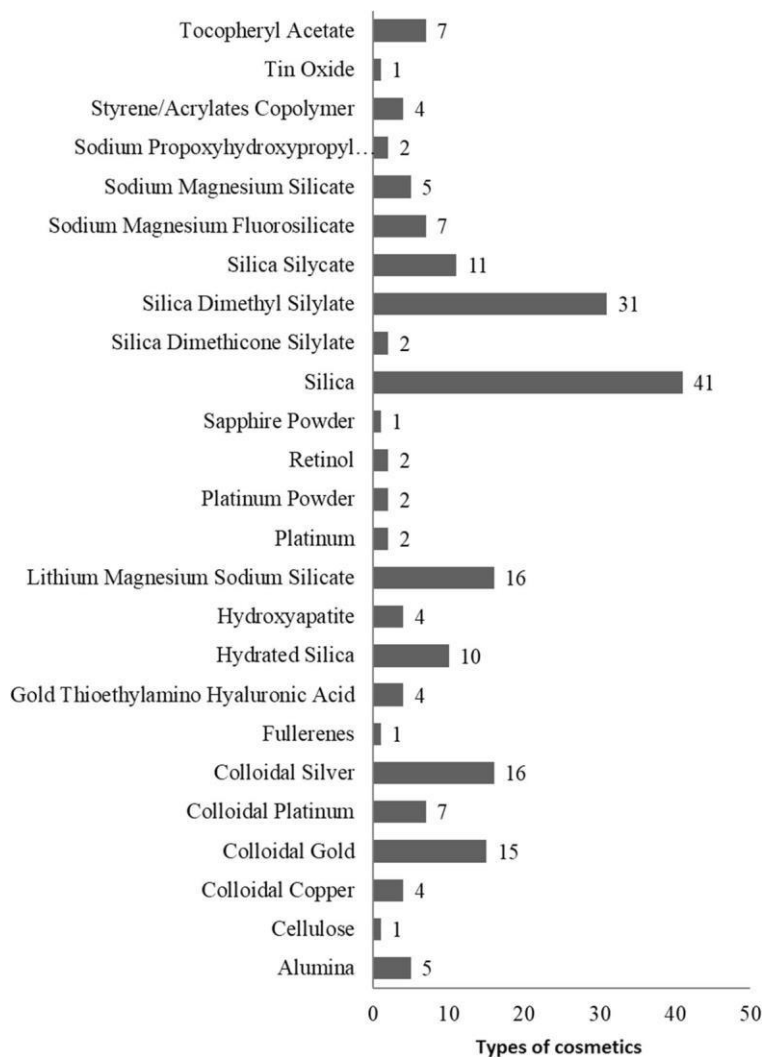
In this case, the responsible person must notify the substance in the same way to the European Commission, which should ask for an SCCS study in case there are doubts about its toxicity. These nanomaterials are not regulated. Although there is not much knowledge about the toxicity of nanomaterials, whether applied as colorants, preservatives, UV filters, or for any other function, they have been registered and published through the catalog of nanomaterials offered by the European Commission, and no other SCCS opinions have been requested to date. Therefore, the notification system created through Regulation 1223 of 2009 has not been applied in its literal sense since a large number of nanomaterials that may be toxic are being incorporated into cosmetic formulas.

After ten years, the control of these materials is not as effective as MEPs probably thought when they passed the regulation on 24 March 2013. On the other hand, it has been a necessary step forward since it has helped greatly to know nanomaterials in the European market and implement toxicity controls while progressing in the state of the art as well as to regulate some of the most commonly commercialized nanomaterials, such as titanium dioxide, zinc oxide, or carbon oxide.

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

Fig. 3

Nanomaterials with other functions used in the European Union in the different types of cosmetic products. Information based on the catalog reported by the EC (version 1 (31.12.2016)) [31]



Proposals

First, since the information on nanomaterials being commercialized in the European Union is available, the SCCS should issue opinions reviewing the toxicity of products containing nanomaterials because each nanomaterial depends on its characteristics, such as its surface area, size, or form, and can express different physicochemical properties and therefore different toxicities. Therefore, at the time of registration, the cosmetic product should be evaluated, to find viable alternatives, depending on the functionality that is supposed to offer and its efficiency. As Hansen Steffen suggests, such information should be taken into account to issue the authorization [40].

The products that the European Commission deem problematic based on previous research carried out in international research centers should be banned until their toxicity profiles are better known following an opinion of the SCCS to evaluate frequently contradictory results from nanomaterial safety studies. In this way, the standard of consumer safety would be greater in terms of being consistent with the principles that inspire the regulation in terms of nanotechnology.

Data on the investigations, methods, and results on toxicity should be incorporated into the catalog of nanomaterials published by the European Commission. Presenting a label with nanoingredients is insufficient to eliminate the asymmetry of information between manufacturer and consumer [41, 42]. Following this proposal would mean that consumers are provided with more information concerning toxicity and thus in a better position to decide whether to buy or not to buy a particular cosmetic product. Since no toxicological aspects are mentioned currently, the consumer does not have all the information necessary to make an evidence-based decision.

In addition to the toxicological information of the nanomaterials in circulation, it is recommended to include further information in the catalog to offer greater transparency for consumers, for example, on the start date of circulation of each nanomaterial in the European market and on which nanomaterials are used most often in the European Union.

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

Concluding Thoughts

Regulation 1223 of 2009 has become a pillar of the particular regulation of nanotechnology in the context of the European Union. Although the possible adverse effects of nanomaterials were noted years ago, the necessary express regulations for cosmetics that incorporate foundational norms in terms of labeling, a definition of nanomaterials, evaluation of toxicity and systems of notification, and control of circulating nanomaterials in European latitudes did not exist. The analysis suggests that adverse effects, such as generating unnecessary differences between European and non-European regulations, have not been observed.

The enforcement of the regulation represented a milestone for nanoregulation, especially by placing the debate in legislative institutions and opting for hard law. In this regard, it paved the way for further regulations in the European Union to incorporate express provisions of nanomaterials.

The analysis of the genesis of the regulation and of the interventions during the debate in the EP shows that nanomaterial regulations were considered to increase the safety of consumers, and research would be promoted to learn more about the toxicity of nanomaterials. Furthermore, according to objectives 3 and 4, the aim was to create a safe space for cosmetic product consumers in Europe as the industrial innovations progress. However, given how difficult it still is to obtain such information, it is conspicuous that meeting the objectives with the current model will be a hard task.

Article 16 (1) of the regulation provides that a high level of protection of human health shall be ensured for any cosmetic product containing nanomaterials. Due to the reasons described in the present paper, especially those associated with the technical difficulties of applying the regulatory precepts of nanomaterials, it can be said that its objectives have only been partially fulfilled. The evidence of this is that certain cosmetic products incorporating nanomaterials are not authorized, which means that there is no clear and precise information on their toxicity or an SCCS opinion. Although the label on the cosmetic product must include the nanotechnological nature of the ingredients, if the consumer does not know the toxicity of this nanomaterial, it does not have the intended

impact. If the objective is to increase consumer safety, then the process is insufficient. In addition, the process compromises the high level of product safety in the cosmetic sector.

It is often emphasized that there is a need to regulate nanotechnology due to the special physicochemical properties of nanomaterials and the potential for toxicological profiles. Unfortunately, information on such profiles is still insufficient. This paper argued that Regulation 1223 of 2009 has encountered technical obstacles. Further research concerning a comprehensive regulation of nanomaterials in cosmetics should be encouraged, and different models of regulation should be developed which are more accurately adjusted to the development of the sector in recent years.

Acknowledgements A special thanks is given to Red NANOCELIA (CYTED) and COLCIENCIAS for the scholarship for the doctorate studies of one of the authors within its competition. In particular, the financial support is given by this institution through the open call: “Convocatoria para Doctorado Nacional 757” from 2017. This original research is part of the project “Investigación en Derecho Internacional y Nanotecnología” registered in the Research Centre of Universidad Pontificia Bolivariana with register number 766B-06/17-37.

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

References

1. Miller G (2008) Contemplating the implications of a nanotechnology “revolution”. In: Fisher E, Selin C, Wetmore JM (eds) Presenting futures. The yearbook of nanotechnology in society, vol 1. Springer, Dordrecht, pp 215–225
2. Organisation for Economic Co-operation and Development (2013) Nanotechnology in the context of technology convergence. OECD, Paris
3. European Commission (2016) Horizon 2020 Work Programme 2016–2017. 5.ii. Nanotechnologies, Advanced Materials, Biotechnology and Advanced Manufacturing and Processing. EC, Brussels
4. Food and Drug Administration (2014) Guidance for industry “considering whether an FDA-regulated product involves the application of nanotechnology”. FDA, Washington, DC
5. World Intellectual Property Organization (2015) Economic growth and breakthrough innovations: a case study of nanotechnology. WIPO, Geneva
6. See: <http://product.statnano.com/>. Accessed 13 Feb 2019
7. Bowman DM (2017) More than a decade on: mapping today’s regulatory and policy landscapes following the publication of nanoscience and nanotechnologies: opportunities and uncertainties. *NanoEthics* 11(2):169–186. <https://doi.org/10.1007/s11569-017-0281-x>
8. Lai RWS, Yeung KWY, Yung MMN, Djurišić AB, Giesy JP, Leung KMY (2017) Regulation of engineered nanomaterials: current challenges, insights and future directions. *Environ Sci Pollut Res* 25:1–18. <https://doi.org/10.1007/s11356-017-9489-0>
9. Food and Drug Administration (2014) Office of the Commissioner, FDA, guidance for industry: considering whether an FDA-regulated product involves the application of nanotechnology. FDA, Washington, DC
10. Burden N, Aschberger K, Chaudhry Q (2017) The 3Rs as a framework to support a 21st century approach for nanosafety assessment. *Nano Today* 12:10–13. <https://doi.org/10.1016/j.nantod.2016.06.007>
11. Reimhult E (2017) Nanoparticle risks and identification in a world where small things do not survive. *NanoEthics* 11(3): 283–290. <https://doi.org/10.1007/s11569-017-0305-6>
12. Maynard AD, Aitken RJ (2016) Safe handling of nanotechnology ten years on. *Nat Nanotechnol* 11:998–1000. <https://doi.org/10.1038/nnano.2016.270>
13. Bowman DM, Chaudhry Q, Gergely A (2015) Evidence-based regulation of food nanotechnologies: a perspective from the European Union and United States. In:

- Sabliov CM, Chen H, Yada RY (eds) Nanotechnology and functional foods: effective delivery of bioactive ingredients. Wiley, NJ, pp 358–374
14. Royal Society and Royal Academy of Engineering (2004) Nanoscience and nanotechnologies: opportunities and uncertainties. RS/RAEng, London
 15. Justo-Hanani R, Dayan T (2015) European risk governance of nanotechnology: explaining the emerging regulatory policy. *Res Policy* 44:1527–1536.
<https://doi.org/10.1016/j.respol.2015.05.001>
 16. Official Journal of the European Union (2009) Regulation (EC) 1223/2009 of the European Parliament and of the Council on cosmetic products (recast). 30 November 2009. EU, Brussels
 17. Future Markets (2012) Nanomaterials in the cosmetics and personal care.
http://www.researchandmarkets.com/reports/2069662/nanomaterials_in_the_cosmetics_and_personal_care
 18. Council of the European Communities (1976) Council Directive on the approximation of the laws of the Member States relating to cosmetic products (No 76/768/EEC, 27 Jul, 1976). CEC, Brussels
 19. European Commission (2008) Proposal for a Regulation of the European Parliament and of the Council on cosmetic products (No COM (2008) 49 final, 05 Feb 2008). EC, Brussels
 20. European Commission (2008) Accompanying document to the communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee (Commission Staff Working Paper No SEC (2008) 117, 05 Feb 2008). EC, Brussels
 21. European Commission (2008) Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee. Regulatory Aspects of Nanomaterials (No COM (2008) 366 final, 17 Jun 2008). EC, Brussels
 22. European Economic and Social Committee (2008) Opinion of the European Economic and Social Committee on the ‘Proposal for a Regulation of the European Parliament and of the Council on cosmetic products’ (No CESE (2008) 1193, 09 Jul 2008). EESC, Brussels
 23. European Parliament (2008) Report on the proposal for a regulation of the European Parliament and of the Council on cosmetic products (No A6 (2008) 484, 02 Dec 2008). EP, Brussels
 24. European Parliament (2009) Debate about cosmetic products (recast version, CRE 23/03/2009–15, 23 Mar 2009). EP, Strasbourg

2) EUROPEAN NANOTECHNOLOGY REGULATION (COSMETIC SECTOR)

25. European Parliament Press Release (2009) MEPs approve new rules on safer cosmetics (24 Mar 2009). EP, Brussels
26. European Commission (2016) Commission Regulation amending Annex VI to Regulation (EC) No 1223/2009 of the European Parliament and of the Council, on cosmetic products (C/2016/4325). EC, Brussels
27. European Parliament (2008) A guide to how the European Parliament co-legislates under the ordinary legislative procedure, Brussels. EP
28. European Commission (2012) Cosmetic products notification portal. Article 13 User Manual. EC, Brussels
29. European Commission (2008) Commission decision setting up an advisory structure of scientific committees and experts in the field of consumer safety, public health and the environment and repealing (No 721/EC (2008)). EC, Brussels
30. European Commission (2012) User manual for the notification of cosmetic products containing nanomaterial according to article 16. EC, Brussels
31. European Commission (2017) Catalogue of nanomaterials used in cosmetic products placed on the EU market. Version 1 (31.12.2016). EC, Brussels
32. Scientific Committee on Consumer Safety (2015) Second revision opinion on carbon black (nano-form) (no SCCS/1515/13). SCCS, Brussels
33. European Commission (2016) Commission Regulation (EU) 2016/1143 amending Annex VI to Regulation (EC) No 1223/2009 of the European Parliament and of the Council on cosmetic products (13 July 2016). EC, Brussels
34. Scientific Committee on Consumer Safety (2015) Second revision opinion on titanium dioxide (nano form) (no SCCS/1516/13). SCCS, Brussels
35. European Commission (2014) Commission Regulation (EU) 2016/1143 amending Annexes III, V and VI to Regulation (EC) No 1223/2009 of the European Parliament and the Council on cosmetic products (08 Aug 2014). EC, Brussels
36. Scientific Committee on Consumer Safety (2015) Second revision opinion on 1,3,5-triazine, 2,4,6-tris[1,1'-biphenyl]- 4-yl (No SCCS/1429/11). SCCS, Brussels
37. European Commission (2016) Commission Regulation (EU) 2016/621 amending Annex VI to Regulation (EC) No 1223/2009 of the European Parliament and of the Council on cosmetic products (21 April 2016). EC, Brussels
38. Scientific Committee on Consumer Safety (2012) Second revision opinion on zinc oxide (nano form). (No SCCS/1489/12). SCCS, Brussels
39. Scientific Committee on Consumer Safety (2013) Second revision opinion on zinc oxide (nano form). (No SCCS/1518/13). SCCS, Brussels

40. Hansen S (2017) React now regarding nanomaterial regulation. *Nat Nanotechnol* 12:714–716. <https://doi.org/10.1038 /nnano.2017.163>
41. European Commission (2012) Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee. Second regulatory review on nanomaterials (no COM(2012) 572, 3 Oct 2012). EC, Brussels
42. JRC (2010) Considerations on a definition of nanomaterial for regulatory purposes. JRC, Brussels

*A man who dares to waste one hour of
time has not discovered the value of
life.*

Charles Darwin

CHAPTER

3

3) European Nanotechnology Regulation (Food sector)

Given the transversal characteristic of the nanotechnology, other sectors such as food sector, has used nanomaterials to improve products in different ways. European regulation includes statements to ensure the safety of these products in the European market. In this chapter, we present an exploration of food regulation on nanomaterials, to show the efficacy of this regulation in the European context and the use of Machine Learning to better apply this regulation.

Cheminformatic methods are able to design and create predictive models with high rate of accuracy saving time, costs and animal sacrifice. This paper makes an analysis of principles of European food law for the appropriateness of applying biological activity Machine Learning prediction models to guarantee

3) EUROPEAN NANOTECHNOLOGY REGULATION (FOOD SECTOR)

public safety, according to **Figure 5**.

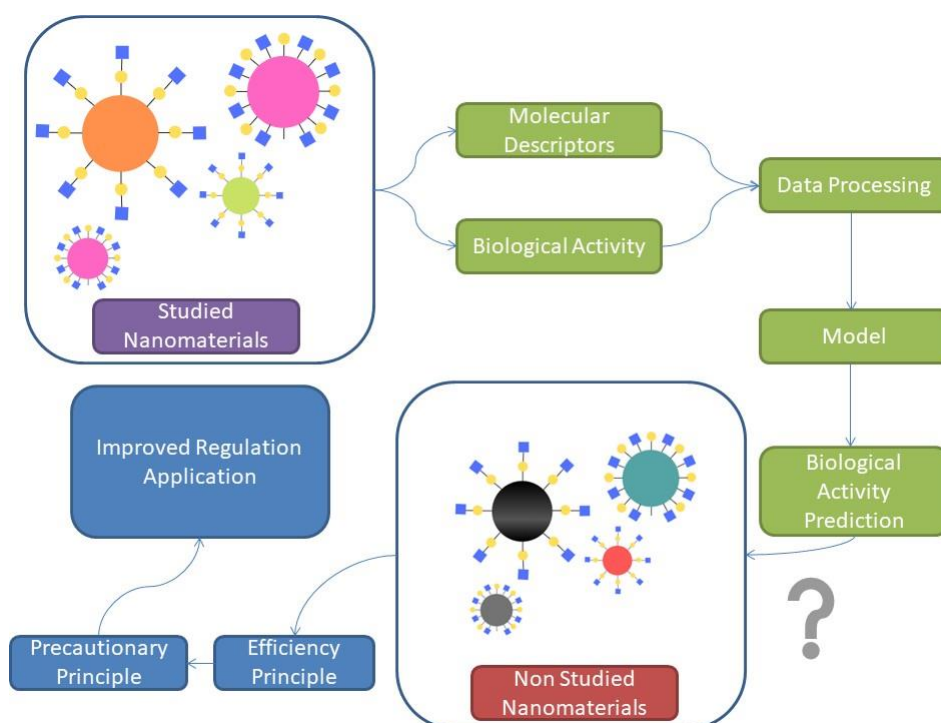


Figure 5. Cheminformatic models workflow to predict biological activity and improve regulation application

Machine Learning as proposal for a better application of food nanotechnology regulation in European Union

Ricardo Santana^{a,b,c}, Enrique Onieva^a, Robin Zuluaga^d,

Aliuska Duardo-Sánchez^e, and Piedad Gañán.^f

^a*DeustoTech-Fundación Deusto, Avda. Universidades, 24, 48007 Bilbao, Spain.*

^b*Faculty of Engineering, University of Deusto, Avda. Universidades, 24, 48007 Bilbao, Spain.*

^c*Grupo de Investigación sobre Nuevos Materiales, Universidad Pontificia Bolivariana, Circular 1° N° 70-01, Medellín, Colombia.*

^d*Facultad de Ingeniería Agroindustrial, Universidad Pontificia Bolivariana UPB, 050031, Medellín, Colombia.*

^e*Department of Public Law, Law and the Human Genome Research Group, University of the Basque Country UPV/EHU, 48940, Leioa, Biscay, Spain.*

^f*Facultad de Ingeniería Química, Universidad Pontificia Bolivariana UPB, 050031, Medellín, Colombia.*

Abstract

Cheminformatic methods are able to design and create predictive models with high rate of accuracy saving time, costs and animal sacrifice. It has been applied on different disciplines including nanotechnology. Given the current gaps of scientific knowledge and the need of efficient application of food law, this paper makes an analysis of principles of European food law for the appropriateness of applying biological activity Machine Learning prediction models to guarantee public safety. For this, a systematic study of the regulation and the incorporation of predictive models of biological activity of

3) EUROPEAN NANOTECHNOLOGY REGULATION (FOOD SECTOR)

nanomaterials was carried out through the analysis of the express nanotechnology regulation on foods, applicable in European Union. It is concluded Machine Learning could improve the application of nanotechnology food regulation, given that it is aligned with principles promoted by the standards of Organization for Economic Co-operation and Development, European Union regulations and European Food Safety Authority. To our best knowledge this is the first study focused on nanotechnology food regulation and it can help to support technical European Food Safety Authority Opinions for complementary information.

Keywords

Nanotechnology, Regulation, Toxicity, Safety, Cheminformatic.

1. Introduction

Nanotechnology regulation has been studied for years by the literature to determine if it was necessary and in what terms^{1,2}. There is scholars that leans towards the favorable position for nanoregulation, given the uncertainty of the biological activity of nanomaterials, the existing number of nanoproducts and the variety of nanomaterials in the market³⁻⁶. This could affect consumer rights, especially safety in particular in food sector^{7,8}. This does not mean that the debate is over, since it is possible to consider that nanotechnology does not need express regulation to be regulated: If a broad interpretation of the existing standards in any jurisdiction that have regulations on the safety of chemical substances, nanomaterials would be possibly included. Therefore, in these contexts, nanotechnology would not be regulated expressly, but implicitly.

Scientific literature of the institutions and universities from the majority of countries has not proposed an analysis from the national legal point of view. In our knowledge, there is no express regulation of nanotechnology in the food sector in another region. This national legal analysis has considerable importance because, by virtue of the principle of territoriality, the limited jurisdiction and the lack of initiative for the generation of international convention, it suggests that nanotechnology regulation, today, is challenging to implement there is not an in depth analysis in each country. This would enable to discover how to incorporate it into their respective legal system. Once analyzed, objectives and strategy for adequate governance can be established. Otherwise, not only

public safety could be compromised, but also an adequate progress of nanotechnology in productive sectors with potential application of nanotechnology.

In the context of the European Union, European Regulations and Directives cover different sectors in terms of nanotechnology application as Regulation (EU) No. 2283/2015 on novel foods or Regulation (EC) No. 1333/2008 on food additives. Therefore, projects with specific objectives related to nanosafety have been carried out to be able to implement a proper governance of this technology⁹⁻¹⁴. For example, European NanoSafety Cluster⁹ makes it possible to exchange information among regulators, researchers, administrators, and industry. Among the different working groups, the Regulations & Risk Governance group¹⁰ is dedicated to monitor and interpret the state of the art in terms of regulation for nanotechnology. In addition, other projects have allowed knowing more technical data about biological activity of nanomaterials such as MembraneNanoPart¹¹ or PreNanoTox¹². On the other hand, risk management tools have been developed in work environments, such as Nanosafer¹³ or Stoffen Manager Nano¹⁴. These projects allow knowing more about the toxicological profile of specific nanomaterials and their legislative impact.

One of the sectors where nanotechnology has many functions is the food industry¹⁵. Nanotechnology in the food sector is also expressly regulated in the European Union through different regulatory instruments¹⁶. However, European Parliament (EP), European Commission (EC) and the rest of European institutions are aware of the progress that can be made in other countries after the approval of these legislative instruments. For instance, Regulation (EU) No. 2283/2015 on novel foods establishes (recital 39 and art. 31) that the regulatory and scientific advances that may occur at the international level will be taken into consideration to modify the list of approved foods including determined specifications.

In spite of the existence of an express nanotechnology regulation, the authorization procedures for the commercialization of nanomaterials depend on opinions issued by the European Food Safety Authority (EFSA), which take into consideration assay results. Predictive models on the biological activity of nanomaterials are a useful way to complement these reports. The method Perturbation Theory Machine Learning (PTML) is proposed for the creation of predictive models given that it has been shown with a considerable level of accuracy and could help to make decisions with more information.

3) EUROPEAN NANOTECHNOLOGY REGULATION (FOOD SECTOR)

Through this research work, to our best knowledge the first of this type, we analyze nanotechnology food regulation in European Union context to assess Machine Learning as alternative technique to improve the application.

2. Regulation of nano foods in the EU.

2.1. European regulations and directives. Scope and authorizations.

Nanotechnology regulation applied to food in the European Union is dispersed in different legislative instruments. Regulation (EC) No. 178/2002 constitutes a horizontal regulation for the sector. In addition, there are regulations such as Regulation (EC) No. 1333/2008, Regulation (EC) No. 1333/2008 or Regulation (EC) No. 1333/2008, which regulate food additives, food enzymes, and flavors. The express regulation of nanomaterials, therefore, is not collected uniformly for all substances; the authorization procedure may vary depending not only on the specific substance but also on the use that is given in the production process. In this section, an analysis of the nanofood regulations is carried out in terms of the authorization process to be commercialized. Given the objective of this study, comments on the regulation of nanotechnology in feed, plant protection products, food contact materials, biocides, or others that do not regulate nanofoods are not included.

Regulation (EC) No. 178/2002 applies to the stages of production, transformation and distribution of food, establishing a general horizontal framework for food legislation (art. 4). Through this Regulation, EFSA is created. This is the authority in charge of scientifically advising the European institutions, and Member States with legislative purposes (art. 23). This function covers the assessment in the authorization process for food commercialization. In addition, this Regulation is notable for setting out, in section 1, the principles that should govern food legislation. Among them, two principles of particular interest in the regulation of nanofoods are incorporated: 1) The food legislation is based on the risk analysis which takes into consideration the scientific evidence available (art. 6) the precautionary principle when evaluating the available information but there is uncertainty of the adverse effects, risk management measures must be applied (art. 7). Although it does not make an explicit regulation of nanotechnology, this principle

must be respected by subsequent legislative measures and commercialization authorizations related to foods in general and nanofoods in particular.

Regulation (EU) No. 2015/2283 entered into force on January 2018, and changes the regime followed until then by Regulation (EC) 258/97. The regulation does not consider in its application field, the place where the novel food is produced but the commercialization in the Union market (recital 1 and 3). The main objective is to protect the health and well-being of citizens. It also aims to promote the free movement of food, legal security and avoid unfair competition. According to the purpose of this investigation, this Regulation is of interest because it includes engineering nanomaterials as new foods (recital 3 and 10), if they have not been widely consumed before May 15, 1997.

All novel foods must be evaluated and approved for commercialization in the European Union market. Therefore, in the case of nanofoods, each nanomaterial must be investigated one by one since the bioavailability can vary depending on the physicochemical properties¹⁷. The Commission grants this authorization by updating the Union list (art. 9). The procedure to authorize the commercialization of the new food begins with the application of an interested party or EC initiative (art. 10). The Member States must be informed in order to promote transparency and, also, they can issue their positions regarding the updating of the list. The Commission will take it into consideration before making the decision. The request for updating the list must refer to the methods applied and the lack of risk of the new food (art. 10.2.e.). Specifically, in the case of nanomaterials, the solicitant must justify the scientific suitability of the methods used for the analysis of risk for the nanomaterial (art. 10.4, recital 27). On the other hand, at the request of the EC, the EFSA will issue an opinion within nine months of receipt (art. 11) if updating the list can influence the health of citizens (art.10). Seven months later the EC will decide on the concession or rejection of the authorization (art. 12).

The Regulation does not apply, in any case, to new genetically modified foods. It also excludes foods when they are used as food additives, enzymes, food flavorings. In these cases, other regulations are applicable if they consist of nanomaterials. Regulation (EC) No. 1333/2008 regulates food additives and includes that a significantly different food additive will need to be reviewed by the EFSA; Therefore, it considers that the use of nanotechnology on an additive turns it into a significantly different additive (recital 13).

3) EUROPEAN NANOTECHNOLOGY REGULATION (FOOD SECTOR)

In this case, a different authorization is needed for the new additive that entails a modification of the Union list.

On the other hand, Regulation (EC) No. 1334/2008, does not refer to particle size change or nanotechnology expressly, but states that if there is a significantly different aromatizer must be reevaluated to be registered in order to have specific authorization (art. 19). If it is interpreted that significantly different includes the modification in the size of the particles, nanotechnology would be included. It is noteworthy that it does not include the interpretation of the expression significantly different, as it is stated in Regulation (EC) No. 1332/2008 and Regulation (EC) No. 1333/2008. It generates inevitable confusion because it could be considered that the European institutions intentionally did not include the change in particle size as a significant change. In any case, it is recommended to revise the redaction in order to not infuse misunderstanding about the application of that recital.

For the granting of the authorization of the substances regulated by Regulation (EC) No. 1332/2008, Regulation (EC) No. 1333/2008 and Regulation (EC) No. 1334/2008, the procedure established in Regulation No. 1331/2008 would be followed. The process is initiated by the Commission or by the interested party who requests it (art. 3). However, in the request of substances with different particle size, a suitability report on the method of analysis of nanomaterials is not required, as the substances regulated by Regulation (EU) No. 2283/2015. Regulation (EU) No. 234/2011, which implements Regulation (EC) No. 1331/2008, establishes that the description of the strategy for risk assessment and selection of test methods must be included in the file. However, it should not necessarily be a strategy that has considered the size of the particles. During the process, additional information may be required. If this is not done, it would be contrary to what was suggested by the Organisation for Economic Co-operation and Development (OECD) Council on safety testing of nanomaterials, which recommends to better understand the specificities of nanomaterials, that the methods be adapted to nanomaterials¹⁸.

Regulation (EC) No. 1331/2008 establishes, unlike Regulation 2283/2012, that the Commission is obliged to request an opinion from EFSA to include a new substance even when there is no risk for human health. However, in case of a request to modify the conditions of permitted substances or to eliminate them, the Commission is only obliged

if it can affect human health (art. 3.2.). In this case, EFSA has nine months to issue the opinion and both the request, and the opinion will be presented to the Member States.

As regards vitamins and minerals, there are two different regulatory frameworks: 1) Food supplements which are included within the scope of Directive 2002/46/EC and Regulation (EC) No. 1170/2009, on vitamins and minerals that can be used in the manufacture of food supplements and 2) Regulation (EC) No. 1925/2006 whose application are vitamins and minerals that can be added to food. In both cases, vitamins and minerals require authorization to be marketed within the European Union market. If they are not in their respective annexes, they are not authorized. Therefore, the vitamins and minerals that are to be introduced in the market after May 15, 1997 and that contain or consist of artificial nanomaterials, are considered new foods, so they must also have the authorization included in Regulation 2283/2015. Subsequently, the respective regulations can be applied.

Other substances that must have the authorization included in Regulation (EU) No. 2283/2015, are those regulated by Regulation (EU) No. 609/2013 on food intended for infants and young children, food for special medical purposes, and total diet replacement for weight control. In this case, an express reference is made to the regulation of new foods (recital 23), alluding to changes in particle size and the application of nanotechnology. After applying Regulation (EU) No. 2283/2015 and being authorized, Regulation (EU) No. 609/2013 will be applied to grant or reject the amendment of the annex and thus be able to be commercialized.

2.2. Authorization of nanotechnology foods under European regulations.

As a result of the application of the commented regulations in section 2.1., EFSA has issued evaluations on nanomaterials applied on foods. Those that refer to food contact materials or feed evaluations are not included. They are ordered by relevance, according to EFSA journal repository:

1) Re-evaluation of silicon dioxide (E 551) as a food additive is analyzed in EFSA J (March 1, 2018) ¹⁹. In this opinion, Panel on Food Additives and Nutrient Sources added to Food (ANS) takes into consideration toxicity studies of nano silicon dioxide given the existence of nanoparticles in the additive E551. The Panel concludes that in the conducted studies no adverse effects were found. Hence, there are no concerns about this

3) EUROPEAN NANOTECHNOLOGY REGULATION (FOOD SECTOR)

additive. However, the lack of long-term studies with nano silicon dioxide, makes the Panel recommends a change of specifications for commercialization since there are gaps on the possible effects of the entire range of particle size that may exist in the E551. These changes refer to a specification in the particle size distribution.

2) Safety and bioavailability of silver hydrosol as a source of silver added for nutritional purposes to food supplements is commented in EFSA J (March 28, 2018)²⁰. Panel on ANS issues this opinion which includes an evaluation of silver hydrosol, consisting of a mixture of positively charged silver ions and silver particles in water. The study that was presented in the request on toxicity showed that it entailed a gastric profusion. The Panel concluded that the study did not provide sufficient information on the characterization of the nanomaterial or on the bioavailability of the source silver, nor the safety for nutritional effects to include it as a food supplement. Some scientific studies are referenced but the Panel considered inadequate for the evaluation of the toxicity of silver hydrosol.

3) Scientific Opinion on the re-evaluation of iron oxides and hydroxides (E 172) as food additives is analyzed in EFSA J (Dec 08, 2015)²¹. Red, yellow, black and brown iron oxides were assessed in this opinion by Panel on ANS. These substances included nanoparticles; however, toxicological databases are remarkably limited. For instance, although there are studies that suggest adverse effects, for instance red and black iron oxide were positive the results of the *in vitro* genotoxicity assays in mammalian cells, *in vivo* oral administration study of red iron oxides did not cause genotoxic effects in rat haemopoietic system and there is not information about the activity in the gastrointestinal tract. EFSA recognized that a proper assessment of the E172 could not be possible. This was given to limitations in the available information in terms of genotoxicity, as well as carcinogenicity and reproductive and developmental toxicity. The Panel recommended that the size of the particles and the size distribution should be included in the specifications given the toxicological potential of the nanoparticles.

4) Scientific Opinion on re-evaluation of calcium carbonate (E 170) as a food additive is analyzed in EFSA J Jul 26, 2011²². In this opinion, Panel on ANS concludes that there is no concern about adverse effects due to traces of nano-calcium carbonate. However, the Panel states that the same conclusion cannot be reached with the information available,

if calcium carbonate is predominantly composed by nanoparticles. It refers to calcium carbonate studies in nanoscale (60-10 nm), but without providing comprehensive information.

5) Re-evaluation of titanium dioxide (E 171) as a food additive is evaluated in EFSA J (Sep 14, 2016)²³. Panel on ANS determined that possible adverse effects of nanomaterials in relation to the reproductive system, being due to the little information available, the same could not be concluded with the additive E171. Adverse effects are identified in the reproductive system but are not considered because the substances under study were either non-food-grade or inadequately characterized nanomaterial. From a carcinogenicity study or the available genotoxicity database there are not considered adverse effects. The new experiments lead EFSA to rethink the previous opinion (July 4, 2018). However, in this opinion, the EFSA has indicated that in the experiments adverse effects are discovered but with uncertainty. Therefore, the previous opinion issued by the EFSA is maintained.

6) Evaluation of di-calcium malate, used as a novel food ingredient and as a source of calcium in foods for the general population, food supplements, total diet replacement for weight control and food for special medical purposes is evaluated in EFSA J (Jun 6, 2018)²⁴. The objective of the opinion of Panel on ANS was not to evaluate the toxicity of the nanoparticles of this compound, but in the information provided by the applicant, there was a study on the particle size distribution. The EFSA determined that vibratory sieve testing with the smallest sieve opening of 45 µm is not an adequate method to determine nano-sized particles.

7) Scientific opinion on the re-evaluation of silver (E 174) as food additive was analyzed in EFSA J (Jan 21, 2016)²⁵. It is noteworthy that the Panel on ANS poses that the strong relationship between the smaller particle size, the release of silver ions from the nanoparticles with the accumulation in organs has been demonstrated. However, it recognizes the ignorance that exists between the metal of the nanoparticle and the silver ions in biological systems due to the large number of variables. It was concluded that the relevance of the available information on toxicological studies of additive E 174 is not possible. It is recommended that in the specifications the particle size distribution, the average particle size and the percentage of particles with at least one dimension below 100 nm.

3) EUROPEAN NANOTECHNOLOGY REGULATION (FOOD SECTOR)

8) Re-evaluation of calcium silicate (E 552), magnesium silicate (553a(i)), magnesium trisilicate (E 553a(ii)) and talc (E 553b) as food additives is analyzed in EFSA J (Aug 02, 2018)²⁶. In this Opinion the Panel could not assess the safety of E 552, 553a(i), 553a(ii) 553b as additive food. The Panel on ANS indicates that there was “no indication of genotoxicity but reliable data on subchronic and chronic toxicity, carcinogenicity and reproductive toxicity of silicates and talc were lacking”²⁶. It is recommended to revise the specifications to include the particle size, distribution and statistical descriptors, following the guidance provided by the EFSA for that purpose.

9) Scientific Opinion on the re-evaluation of gold (E 175) as a food additive is analyzed in EFSA J (Jan 20, 2016)²⁷. The Panel on ANS in this Opinion assess different toxicity studies, for instance it mentions studies *in vitro*, gold nanoparticles over mammalian cells causing DNA strand breaks, micronuclei, chromosomal aberrations, aneuploidy and oxidative stress. However, no conclusive results were found by performing *in vivo* studies. The Panel says: no data on subchronic, chronic toxicity and genotoxicity of elemental gold. There is also limited data about absorption, distribution, metabolism and excretion (ADME), so an assessment of E175 was not possible. The Panel recommended to change the specifications for E 175 by including the mean particle size and particle size distribution, as well as the percentage of nanoparticles.

10) Scientific Opinion on the re-evaluation of vegetable carbon (E 153) as a food additive is analyzed in EFSA J (April 27, 2012)²⁸. Panel on ANS concluded that vegetable carbon (E 153) at the reported uses and use levels is not of safety concern. The Panel concluded that the presence of nanoparticles in vegetable carbon products currently on the market can be excluded. The Panel concluded that EC specifications for plant carbon may need to be amended to include a particle size restriction (<100 nm) in order to exclude the presence of nanoparticles, since the inclusion of particles below 275 nm should be adequate A re-evaluation of plant carbon as a food additive.

All the Opinions point out the need for more studies on the biological activity and especially on the toxicology of the different additives. There are toxicological studies for every additive but there is not integrate study including all possible toxic behaviours. Since, EFSA cannot guarantee the toxicity profile of any nanomaterial assessed. For example the case of iron oxides is characterized by lack of data about carcinogenicity,

reproductive and developmental toxicity. Even when studies confirm that they cause genotoxicity through *in vitro* assays, later *in vivo* studies reconsider the toxicological profile. Therefore, it is concluded that the Opinions are characterized by the uncertainty given the current state of art of foods consisting or including nanotechnology and the use of inadequate methods of analysis that generate heterogeneous information.

3. Machine Learning for Law application.

The Opinions issued and commented in the previous section, do not include results of predictive models to complement the assessment of the biological activity of the nanomaterials under study. Should Machine Learning (ML) be applied for complement the EFSA Opinions? These predictive models are studied by a discipline called Cheminformatics²⁹. They consist of models capable of predicting biological activity, physicochemical properties or biokinetic processes by processing data from simulations or experimental assays³⁰. These models could be applied to “i) Supporting priority setting of chemicals; ii) Guiding experimental design of regulatory tests or testing strategies; iii) Providing mechanistic information; iv) Grouping chemicals into categories based on similarity; v) Filling in a data gap needed for classification and labeling; vi) Filling in a data gap needed for risk assessment”³¹. In **figure 1**, we can see the workflow of the construction of a predictive model and the potential repercussion in regulatory processes, such as the authorization processes described above.

3) EUROPEAN NANOTECHNOLOGY REGULATION (FOOD SECTOR)

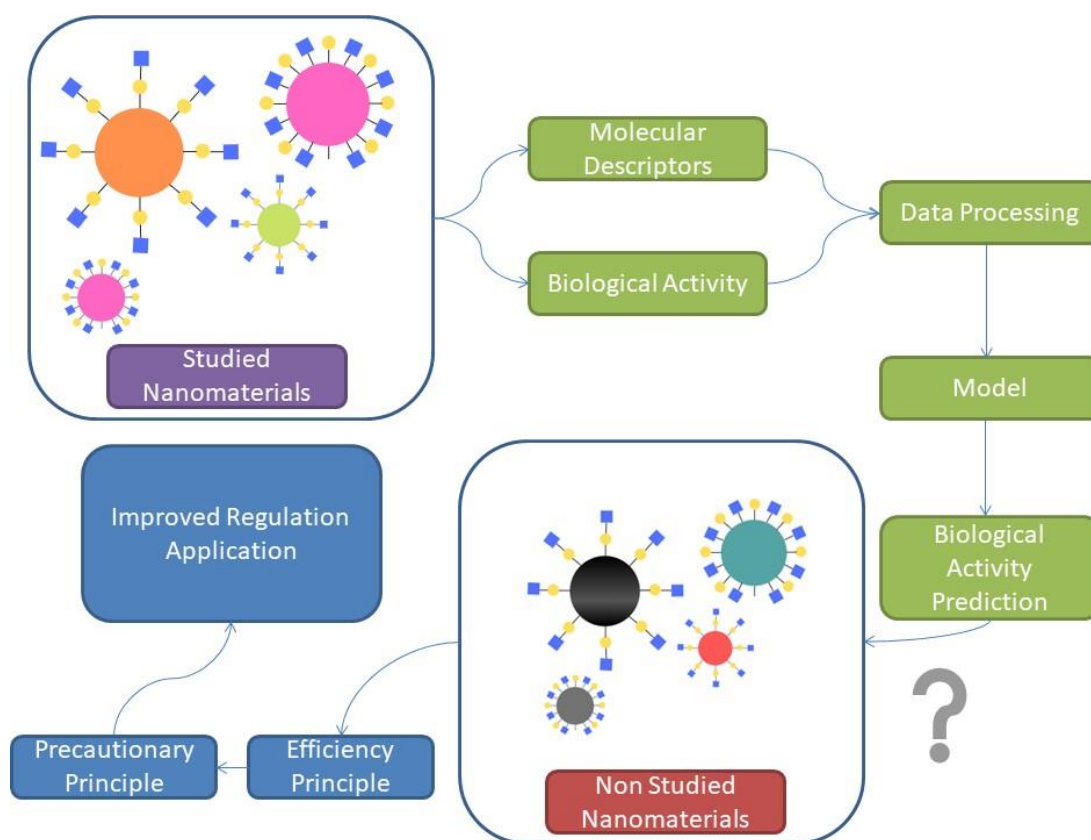


Figure 1. Cheminformatic models workflow to predict biological activity and improve regulation application

Although there are available models capable of generate knowledge about biological activity of nanomaterials, it is debated whether they should apply on legislative field³², bearing in mind the processes discussed above. From the legal point of view, predicting the biological activity of a compound with high probability allows taking measures in accordance with compliance with the precautionary principle. This is useful since this principle consists of taking preventive measures before a possible danger, to diminish risks not scientifically proved. The probability of danger should not be based on mere hypothesis³³. However, there are justified doubts to think that if they are not applied properly, it would affect public health. Despite the difficulties in generating this type of models for nanomaterials, European Union projects have been carried out such as PreNanoTox, MembraneNanoPart, NanoPuzzles³⁴ or ModENPTox³⁵ in order to understand through the development of computational tools, among other aspects, the toxicity of selected classes of engineered nanoparticles.

In the European Union, regarding food, Regulation 178/2002 establishes the precautionary principle. The possible negative consequences and a risk analysis should be identified with the most reliable available data³⁶⁻³⁹. This can be approached through a model, which covers the results of the tests carried out effectively. In this regard, the precautionary principle can be applied in a more efficient way to identify risks, prioritize additional studies, decree urgent measures or discard hypothesis without scientific basis. In addition, the application of ML is aligned with Directive 2010/63/EU on the protection of animals used for scientific purposes⁴⁰. This Directive protects the value of animal welfare, which is established by the Treaty on the Functioning of the European Union (art. 13). In this Regulation the rule of the 3 rs: Replacement, Reduction and Refinement is incorporated as a principle. This principle consists in animals tests should be carried out only when it is not possible to obtain this information without an animal. Therefore, when an animal is tested, the method must be as human as possible and the number of animals as small as possible, without interfering in the achievement of the scientific objective pursued.

On the other hand, these models, product of ML application, must comply with requirements of precision, transparency and robustness, since they would not only have a scholar application. If they do not apply properly, negative repercussions on public safety are possible and must be avoid according to Treaty on the Functioning of the European Union (art. 168) specially in terms of consumer protection (art. 169). To promote international standardization in terms of transparency and robustness, the OECD through the 37th Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology (Joint Meeting) agreed on the OECD Principles for the Validation for Regulatory Purposes⁴¹. It must be specified that the 5 principles were not proposed to be accepted for a regulatory purpose, but to provide a conceptual framework with which to validate them. In addition, the OECD mentions the flexibility of these principles so that each regulatory agency can adapt it according to their needs⁴¹.

The 5 principles that are referred by OECD are: 1) A defined endpoint; Homogeneous information although in the majority of cases this is “rarely feasible in practice”. Therefore, the defined endpoint, any physicochemical, biological or environmental effect that can be measured and used in the training set of the algorithm; 2) an unambiguous algorithm, in order to ensure transparency in the description of the model algorithm,

3) EUROPEAN NANOTECHNOLOGY REGULATION (FOOD SECTOR)

given that numerous commercialized algorithms do not have great transparency; 3) a defined domain of applicability in order to ensure reliable predictions; 4) appropriate measures of goodness-of-fit, robustness and predictivity, depending on the statistical method: The internal performance in terms of predictivity of the training set and the external performance, in terms of predictivity of the test set and 5) a mechanistic interpretation, if possible, for the relation of the descriptors used in order to build the model and the predicted endpoints. The model, to be aligned with principles 3 and 4 proposed by the OECD, must be constantly updated, so it is advisable to regulate the process to include new cases in the database with which the model is generated. This means that using the same model, the biological behavior can be categorized depending on the level of desirability and cutoff for each variable of biological activity. This may be relevant information during the authorization process since tests have to be presented on the interaction with the food, nutrient composition and its bioavailability, both in the procedure contemplated by Regulation (EU) No. 2283/2015 and 1331/2008 on interactions with other nutrients in the matrix.

Although many models have been proposed, they are characterized by lack of abundant good quality data. However great advances have been made by applying different methods to predict certain endpoints as Cellular uptake⁴²⁻⁴⁴, Hemolysis⁴⁵, EC₅₀³⁰ or LDH⁴⁶. A considerable number of algorithms and techniques have been applied⁴⁷. Among them, we point out the method of Perturbation Theory Machine Learning (PTML)⁴⁸, which takes into consideration the heterogeneity of the information; since it considers all the cases and predictions can be more precise. Furthermore this method is able to build multi target predictions for different combination of experimental conditions⁴⁹⁻⁵³ and it would accomplish legal requirements for its validation and acceptance⁵⁴. The novel application is about the method to build algorithms and adapt them to scientific needs. This allows an advantage for subsequent design of the algorithms with for regulation purposes.

3.1. Machine Learning and Regulation (EU) No 2283/2015 Authorization

In accordance with Regulation (EU) No 2283/2015, the EFSA must evaluate, among other factors, all the characteristics of new food that may pose a risk to human health, and consider its possible repercussions on vulnerable groups of the population. Food authorization must be granted through a process, as detailed above, that complies with

the principle of effectiveness, connected with the efficiency in the operation of the internal market (recitals 1 and 2) and expressed in recital 22. ML methods can contribute to the achievement of this principle, as mentioned above.

In addition, Regulation (EU) No. 2283/2015 establishes that EFSA must verify new food consisting of artificial nanomaterials, through the application of the most advanced test methods are used to assess their safety (recital 23). Furthermore it states that the solicitant "must provide an explanation of their scientific suitability for nanomaterials" (Recital 27). This obligation is also included as follows: "Applicants will provide an explanation of their scientific suitability for nanomaterials" (art. 10.4). In addition, it establishes that scientific evidence must be submitted that demonstrates that it does not pose a risk to health (art. 10.2.e.). As we see, flexibility is provided to the applicant, both in the tests presented and in the methods of obtaining them. The justification is that each nanomaterial is different. Therefore, an evaluation will be made on a case-by-case basis. With this, on the one hand, there is evidence of the need for different complementary test methods and, on the other hand, it includes advanced methods as we could interpret them as models capable of predicting the biological activity of nanomaterials.

In the event that the Commission needs more information and time during the process, Regulation 2283/2015 recognizes the authority to extend the terms of the procedure (art. 22). This can involve considerable time and aggregated costs, translated into transaction costs for the applicant. Therefore, we propose, for a more efficient application of the same in line with recitals 1, 2 and 22, the application of complementary methods, as Machine Learning, that allows with the data that the applicant provides, perform a simulation. Once the simulation has been carried out, it could be warned if there is any biological activity that may be out of the expected range, and therefore have uncertainty about the associated risk. In this case, more information would be required of the applicant, under the precautionary principle, mentioned in the previous section.

3.2. Machine Learning and Regulation (EC) No. 1331/2008 Authorization

ML can also promote a more effective application of Regulation (EC) No. 1331/2008. This regulatory instrument includes the principle of effectiveness which is present in different aspects. It can be inferred from recital 26, referring to the framework of the marketing authorization procedure: "For reasons of effectiveness, the terms normally applicable in the framework of the regulatory procedure with control should be abbreviated" (recital 26). It also refers to the need for the process to be effective and

3) EUROPEAN NANOTECHNOLOGY REGULATION (FOOD SECTOR)

limited in time in recital 7. The idea is repeated in recital 11, stating that the beginning of the authorization procedure should be as soon as possible and in recital 10 when claiming a term adjustment. Even in recital 22 it is recommended that such a procedure be extended to other food authorization processes given its legislative simplification and its effectiveness.

It is necessary to comment that Regulation (EC) No. 1331/2008 contemplates other criteria that the Commission must take into account when scientific information is not sufficient to make a risk management decision (recital 14): criteria of sociological, economic, traditional, ethical and environmental character. These criteria are included by the regulator as the last option to make such decisions once the opinion of the Agency has been received. Therefore this is one more reason why predictive models that have been shown and contrasted by the scientific literature should be considered as a criterion superior in hierarchy and complementary to the assays data presented in the state of the art, which will be evaluated by the Agency through its Opinion.

There are difficulties on building models with the current state of arte. Models are more robust if there is a greater number of quality data from preclinical assays. To date, there is abundant information on compounds, thanks to the preclinical tests carried out and included in the scientific literature. However, on nanomaterials until recently, gaps in knowledge are considerable. The data of the tests carried out up to the date remain heterogeneous to apply conventional methods of machine learning⁵⁰. Additionally, nanomaterials are studied in isolation in preclinical trials. At this time the data is not well structured.

4. Concluding thoughts.

For years the need for the express regulation of nanotechnology has been addressed by literature. In the context of the European Union, nanotechnology in food has been expressly regulated horizontally and vertically, through different European Directives and Regulations. This regulation finds its justification mainly in the principle of food safety and the precautionary principle. A compulsory authorization system has been constructed so that a nanomaterial applied to food can be marketed: 1) process included in Regulation (EU) No. 1131/2008 in the case of additives, flavorings and food enzymes and 2) process of Regulation (EU) No. 2283/2015 for new foods. However, the

application of regulation encounters difficulties given the existing uncertainty in the biological activity of nanomaterials. Due to this limitation to apply these legislative instruments, the application of *in silico* methods is recommended to generate greater knowledge in a rigorous and timely manner that can complement the decision taken by the European Food Safety Agency on the authorization of a nanomaterial.

On the other hand, the contribution of QSAR models contributes a great value for the effective application of the precautionary principle. We can identify risks to be managed that are not mere hypotheses, help prioritize additional studies or decree urgent measures. These models must have a considerable level of precision, robustness and transparency due to the possible negative impact on the rights of citizens if they are not.

For this the OECD exposes the 5 principles to create a theoretical framework to be able to apply QSAR models, and given the characteristics and the limited information of the biological activity of the nanomaterials, it must be flexibilized for the application of nano-QSAR models, since principles are conceived to be adapted to different regulations and agencies and thus be able to apply the regulation in the most efficient way possible. In the context of the European Union, nanomaterials have a different regulation than other materials, so principles should not be applied in the same way given the current knowledge is not that developed. This is of special interest when you want to detect anomalies and prioritize toxicological studies. It can be an excellent complementary tool to help, in this case, European Commission to make a decision. Although these conclusions are taken after analyzing the nanotechnology regulation in European Union, it can be applied in different regions or countries in the future, when different national or international food authorities need to evaluate toxicological profile of nanomaterials.

Perturbation Theory Machine Learning (PTML) is a method well received by the scientific literature capable of predicting biological activities based on the disturbance that is generated in a system to the known material, taking into account the test conditions and the descriptors of the nanomaterials. This makes it possible for a single model to provide for many biological activities as IC_{50} , EC_{50} , potency... etc. Additionally, given the problem of heterogeneous data which are provided by different biological activities studies, by the application of the model the authority would be able to discriminate results that are not aligned with predictions. It is concluded that PTML is a valid option to regulate nanotechnology food regulation.

3) EUROPEAN NANOTECHNOLOGY REGULATION (FOOD SECTOR)

Acknowledgements

R.S.C. thanks COLCIENCIAS scholarship for the doctorate studies; “Convocatoria para Doctorado Nacional 757” from 2017. This original research is part of the project “Investigación en Derecho Internacional y Nanotecnología” registered in the Research Centre of Universidad Pontificia Bolivariana with register number 766B-06/17-37. Special gratitude is extended to CYTED NANOCELIA network.

References

1. Forrest, D. *qRegulating nanotechnology development*. (Palo Alto: Foresight Institute, 1989).
2. Fiedler, F. A., & Reynolds, G. H. Legal problems of nanotechnology: an overview. *South. Calif. Interdiscip. Law J.* **3**, 593–630 (1993).
3. Bowman, D. M. & Hodge, G. A. Nanotechnology: Mapping the wild regulatory frontier. *Futures* **38**, 1060–1073 (2006).
4. Bowman, D. M. & Hodge, G. A. A small matter of regulation: an international review of nanotechnology regulation. *Columbia Sci. Technol. Law Rev.* **8**, 1–36 (2007).
5. Reynolds, G. H. Nanotechnology and Regulatory Policy: Three Futures. *Harv. J. Law Technol.* **17**, 179–208 (2003).
6. Wejnert, J. Regulatory Mechanisms for Molecular Nanotechnology. *Jurimetrics* **44**, 323–350 (2004).
7. Eleftheriadou, M., Pyrgiotakis, G. & Demokritou, P. Nanotechnology to the rescue: using nano-enabled approaches in microbiological food safety and quality. *Curr. Opin. Biotechnol.* **44**, 87–93 (2017).
8. Magnuson, B. A. *et al.* Benefits and Challenges of the Application of Nanotechnology to Food. *Tech. Proc. 2007 Nano Sci. Technol. Inst. Nanotechnol. Conf. Trade Show* **2**, 20–24 (2007).
9. Oomen, A.G., Bos, P.M.J., Fernandes, T.F., Hund-Rinke, K., Boraschi, D., Byrne, H. J., Aschberger, K., Gottardo, S. & Kammer, F.V.D., Kühnel, D., Hristozov, D., Marcomini, A., Migliore, L., Scott-Fordsmand, J., Wick, P., Landsiedel, R. Concern-driven integrated approaches to nanomaterial testing and assessment – report of the NanoSafety Cluster Working Group 10. *Nanotoxicology* **8**, 334–348 (2019).
10. WG G: Regulations & Risk Governance. (2019). Available at: <https://www.nanosafetycluster.eu/working-groups/wg-g-regulations-risk-governance.html>.
11. MembraneNanoPart. Available at: <https://www.nanosafetycluster.eu/eu-nanosafety-cluster-projects/seventh-framework-programme->

3) EUROPEAN NANOTECHNOLOGY REGULATION (FOOD SECTOR)

projects/membranenanoart.html.

12. Dasgupta, S., Auth, T., Gompper, G. Shape and orientation matter for the cellular uptake of nonspherical particles. *Nano Lett.* **14**, 687–693 (2014).
13. Nanosafer. Available at: <http://www.nanosafer.org/>.
14. Stoffen Manager. Available at: <https://nano.stoffenmanager.com/>.
15. Karin Aschberger *et al.* Inventory of Nanotechnology applications in the agricultural, feed and food sector. *EFSA J.* **125** (2014). doi:10.2903/sp.efsa.2014.EN-621
16. Coles, D. & Frewer, L. J. Nanotechnology applied to European food production - A review of ethical and regulatory issues. *Trends Food Sci. Technol.* **34**, 32–43 (2013).
17. Maynard, A. Safe handling of nanotechnology. *Nature* **444**, 7–9 (2006).
18. Organisation for Economic Co-operation and Development. *Recommendation of the Council on the Safety Testing and Assessment of Manufactured Nanomaterials.* (2017).
19. EFSA. Re-evaluation of silicon dioxide (E 551) as a food additive. *EFSA J.* **16**, 1–70 (2018).
20. EFSA. Safety and bioavailability of silver hydrosol as a source of silver added for nutritional purposes to food supplements. *EFSA J.* **16**, 1–9 (2018).
21. EFSA. Scientific Opinion on the re-evaluation of iron oxides and hydroxides (E 172) as food additives. *EFSA J.* **13**, 1–57 (2015).
22. EFSA. Scientific Opinion on re-evaluation of calcium carbonate (E 170) as a food additive. *EFSA J.* **9**, 1–73 (2011).
23. EFSA. Re-evaluation of titanium dioxide (E 171) as a food additive. *EFSA J.* **14**, 1–83 (2016).
24. EFSA. Evaluation of di-calcium malate, used as a novel food ingredient and as a source of calcium in foods for the general population, food supplements, total diet replacement for weight control and food for special medical purposes. *EFSA J.* **16**, 1–16 (2018).

25. EFSA. Scientific opinion on the re-evaluation of silver (E 174) as food additive. *EFSA J.* **14**, 1–64 (2016).
26. EFSA. Re-evaluation of calcium silicate (E 552), magnesium silicate (E 553a(i)), magnesium trisilicate (E 553a(ii)) and talc (E 553b) as food additives. *EFSA J.* **16**, 1–50 (2018).
27. EFSA. Scientific Opinion on the re-evaluation of gold (E 175) as a food additive. *EFSA J.* **14**, 1–43 (2016).
28. EFSA. Scientific Opinion on the re-evaluation of vegetable carbon (E 153) as a food additive. *EFSA J.* **10**, 1–34 (2012).
29. Mitchell, J. Machine learning methods in chemoinformatics. *WIREs Comput Mol Sci* **2014**, **4**, 468–481 (2014).
30. Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X., Dasari, T., Michalkova, A., Hwang, H., Toropov, A., Leszczynska, D., Leszczynski, J. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nat. Nanotechnol.* **6**, 175–178 (2011).
31. Duardo-Sanchez, A., González Díaz, H. Legal Issues for Chem-Bioinformatics Models at Biosciences Frontiers. *Front. Biosci.* **E5**, 361–374 (2012).
32. Villaverde, J. J., Sevilla-Morán, B., López-Goti, C., Alonso-Prados, J. L. & Sandín-España, P. Considerations of nano-QSAR/QSPR models for nanopesticide risk assessment within the European legislative framework. *Sci. Total Environ.* **634**, 1530–1539 (2018).
33. European Court. Case C-111/16. (2017).
34. NanoPuzzles. Available at: nanopuzzles.eu/.
35. Modenptox. Available at: <https://fys.kuleuven.be/apps/modenptox/>.
36. European Court. Case C-58/10. (2011).
37. European Court. Case C-282/15. (2017).
38. European Court. Case C-333/08. (2010).
39. European Court. Case C-236/01. (2003).

3) EUROPEAN NANOTECHNOLOGY REGULATION (FOOD SECTOR)

40. Törnqvist, E. *et al.* Strategic focus on 3R principles reveals major reductions in the use of animals in pharmaceutical toxicity testing. *PLoS One* **9**, 1–11 (2014).
41. OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models.* (2007).
42. Kar, S., Gajewicz, A., Puzyn, T., Roy, K. Nano-quantitative structure–activity relationship modeling using easily computable and interpretable descriptors for uptake of magnetofluorescent engineered nanoparticles in pancreatic cancer cells. *Toxicol Vitro* **28**, 600–606 (2014).
43. Epa, V. C., Burden, F. R., Tassa, C., Weissleder, R., Shaw, S., & Winkler, D. A. Modeling biological activities of nanoparticles. *Nano Lett.* **12**, 5808–5812. (2012).
44. Chau, Y. T., & Yap, C. W. Quantitative nanostructure–activity relationship modelling of nanoparticles. *Rsc Adv.* **2**, 8489–8496 (2012).
45. 132. Wang, X.Z., Yang, Y., Li R.F., Mcguinnes, C., Adamson, J., Megson, IL., Donaldson, K. Principal component and causal analysis of structural and acute in vitro toxicity data for nanoparticles. *Nanotoxicology* **8**, 465–476 (2014).
46. Sayes, C., Ivanov, I. Comparative study of predictive computational models for nanoparticle-induced cytotoxicity. *Risk Anal. An Int. J.* **30**, 1723–1734 (2010).
47. Oksel, C., Ma, C. Y., Liu, J. J., Wilkins, T. & Wang, X. Z. Literature Review of (Q)SAR Modelling of Nanomaterial Toxicity. *Model. Toxic. Nanoparticles* **947**, 103–142 (2017).
48. González-Díaz, H., Arrasate, S., Gómez-San Juan, A., Sotomayor, N., Lete, E., Ruso & Besada-Porto, L. General Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry. *Curr. Top. Med. Chem.* **13**, 1713–1741 (2013).
49. Da Costa, J.F., Silva, D., Caamaño, O., Brea, J.M., Loza, M.I., Munteanu, C.R., Pazo, A., García-Mera, X., & González-Díaz, H. Perturbation Theory/Machine Learning Model of ChEMBL Data for Dopamine Targets: Docking, Synthesis, and Assay of New l-Prolyl-l-leucyl-glycinamide Peptidomimetics. *ACS Chem Neurosci* **9**, 2572–2587

(2018).

50. Speck-Planche, A., Kleandrova, V.V., Luan, F., González-Díaz, H., Ruso, J-M. & Cordeiro, M. N. Computational tool for risk assessment of nanomaterials: Novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. *Environ. Sci. Technol.* **48**, 14686–14694 (2014).

51. Luan, F., Kleandrova, V. V., González-Díaz, H., Ruso, J.M., Melo, A., Speck-Planche, A. & Cordeiro, M. N. Computer-aided nanotoxicology: Assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale* **6**, 10623–10630 (2014).

52. Kleandrova, V.V., Luan, F., González-Díaz, H., Ruso, J. M., Melo, A., Speck-Planche, A. & Cordeiro, N. M. Computational ecotoxicology: Simultaneous prediction of ecotoxic effects of nanoparticles under different experimental conditions. *Environ. Int.* **73**, 288–294 (2014).

53. Torquato, P., Ripa, O., Giusepponi, D., Galarini, R., Bartolini, D., Wallert, M., Pellegrino, R., Cruciani, G. & Lorkowski, S., Birringer, M., Mazzini, F. & Galli, F. Analytical strategies to assess the functional metabolome of vitamin E. *J. Pharm. Biomed. Anal.* **124**, 399–412 (2016).

54. Arrasate, S., Duardo Sánchez, A. Perturbation Theory Machine Learning Models: Theory, Regulatory Issues, and Applications to Organic Synthesis, Medicinal Chemistry, Protein Research, and Technology. *Curr. Top. Med. Chem.* **18**, 1203–1213 (2018).

The logic of validation allows us to move between the two limits of dogmatism and skepticism.

Paul Ricoeur

CHAPTER

4

4) European Nanotechnology Regulation (Pharmaceutic Sector)

Given the transversal characteristic of the nanotechnology, other sectors such as pharmaceutical sector, has used nanomaterials to improve products in different ways. European regulation includes statements to ensure the safety of these products in the European market. In this chapter, we present an exploration of pharmaceutical regulation on nanomaterials, to show analyze this regulation in the European context and the use of Machine Learning to better apply this regulation.

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

All the regulations and directives for pharmaceutical sector are analyzed in terms of centralized process dependent on the EMA and the European Commission that includes a wide variety of drugs.

The Role of Machine Learning in Centralized Authorization Process of Nanomedicines in European Union

Ricardo Santana^{a,b,c}, Enrique Onieva^a, Robin Zuluaga^d,

Aliuska Duardo-Sánchez^e, and Piedad Gañán.^f

^aDeustoTech-Fundación Deusto, Avda. Universidades, 24, 48007 Bilbao, Spain.

^bFaculty of Engineering, University of Deusto, Avda. Universidades, 24, 48007 Bilbao, Spain.

^cGrupo de Investigación sobre Nuevos Materiales, Universidad Pontificia Bolivariana, Circular 1° N° 70-01, Medellín, Colombia.

^dFacultad de Ingeniería Agroindustrial, Universidad Pontificia Bolivariana UPB, 050031, Medellín, Colombia.

^eDepartment of Public Law, Law and the Human Genome Research Group, University of the Basque Country UPV/EHU, 48940, Leioa, Biscay, Spain.

^fFacultad de Ingeniería Química, Universidad Pontificia Bolivariana UPB, 050031, Medellín, Colombia.

Abstract

Machine Learning (ML) has experienced an increasing use given the possibilities to expand the scientific knowledge of different disciplines, such as nanotechnology. This has allowed the creation of Cheminformatic models, capable of predicting biological activity and physicochemical characteristics of new components with high success rates in training and test partitions. Given the current gaps of scientific knowledge and the need of efficient application of medicines products law, this paper analyzes the position of

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

regulators for marketing medicinal nanoproducts in European Union and the role of ML in the authorization process.

In terms of methodology, a dogmatic study of the European regulation and the guidances of the European Medicine Agency on the use of predictive models for nanomaterials was carried out. The study has, as the framework of reference, the European Regulation 726/2004 and has focused on the analysis of how ML processes are contemplated in the regulations.

As result, we present a discussion of the information that must be provided for every case for simulation methods. The results show a favorable and flexible position for the development of the use of predictive models to complement the applicant's information.

It is concluded that Machine Learning has the capacity to help to improve the application of nanotechnology medicine products regulation. Future regulations should promote this kind of information given the advanced state of art in terms of algorithms that are able to build accurate predictive models. This especially applies to methods such as Perturbation Theory Machine Learning (PTML), given that it is aligned with principles promoted by the standards of Organization for Economic Co-operation and Development (OECD), European Union regulations and European Authority Medicine. To our best knowledge this is the first study focused on nanotechnology medicine products and machine learning use to support technical European public assessment report (EPAR) for complementary information.

Keywords

Nanotechnology, Regulation, Safety, Cheminformatic.

1. Introduction

Application of Machine Learning (ML) algorithms is gaining momentum both in Pharmaceutical Sciences and Nanotechnology. For instance, in Pharmaceutical Sciences, Lane et al.¹ curated small molecule *Mycobacterium tuberculosis* (Mtb) data and were able to develop and compare predictive models to discover new active molecules targeting Mtb. Lei et al.² developed cheminformatics models in order to predict urinary

tract toxicity, given that it is an adverse event for medications or natural supplements. Oashi et al.³ provided *in silico* models that are able to predict the transport potential of P-glycoprotein; Fusani et al.⁴ presented a predictive model for early detection and risk mitigation of phospholipidosis in lysosomes of various tissues and Li et al.⁵ proposed a multitask model for concurrent inhibition prediction of five major cytochrome P450 isoforms by training a multitask autoencoder deep neural network. Furthermore, Zhavoronkov⁶ explores the potential of high-performance computing, artificial intelligence and machine learning in terms of reversing the decreasing productivity of pharmaceutical industry.

This interest on application of ML to Nanotechnology is given to the numerous functions that the appropriate design of nanomaterials can fulfill.⁷ These functions are possible thanks to the greater surface area and different physicochemical characteristics of nanomaterials with respect to the same materials with no nanoscale. For instance, Toropova et al.⁸ presented a model that is able to predict dark cytotoxicity and photo-induced cytotoxicity of metal oxide nanoparticles to bacteria *Escherichia coli*. Sizochenko et al.⁹ applied ML techniques for metal oxide nanoparticles toxicity prediction towards *Escherichia coli* and HaCaT cells. Mikolajczyk et al.¹⁰ were able to predict the Zeta Potential of metal oxide nanoparticles by utilizing as descriptors the spherical size of nanoparticles and the energy of the highest occupied molecular orbital per metal atom. In recent dates we must highlight the important advances in models that predict, especially, metal oxide nanoparticles cytotoxicity and genotoxicity¹¹⁻¹⁶ among other researches.¹⁶⁻³⁰

As we mentioned, there are studies expanding the scientific knowledge in nanotechnology. However, the increasing number of types of nanomaterials that can be designed and the generalized lack of information on their characterization provoke a considering uncertainty.³² The difficulty to fully understand the behavior of these materials means that the precautionary principle is applied to safeguard the safety of consumers in the context of the European Union, according to Article 191 of the Treaty on the Functioning of the European Union. Therefore, a company must have an authorization to market, in the European context, products that incorporate nanotechnology. The burden of proof is reversed towards the company that wants to market the product: Applicants must show through *in silico*, *in vitro* and - more

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

importantly - *in vivo* studies that the drug is appropriate from the point of view of its effectiveness.

Given the possibilities of specific functions, the variety of possible nanosystems and physicochemical characteristics, the authorization system for the marketing of drugs with nanomaterials must be effective. This means a doable application to achieve the objectives included in the regulation. On the other hand, it is advisable to take into account not to avoid, through disincentives, the normal development of this technology. For instance, Eisenhardt³³ showed how inflexible regulations stifle innovation, through environmental law example by applying agency theory basis. This is also aligned with the principles of green chemistry related to the reduction, replacement and refinement of animal trials.³⁴ In the European Union, regulator institutions have built a partly decentralized system by the delegation in other national agencies of each member state. On the other hand, centralized process is applied by having an authorization system that depends on the report of European Medicines Agency (EMA) and the decision of European Commission.³⁵ Through this centralized process a wide variety of drugs are authorized currently, as will be analyzed in this paper.

This article analyzes the position of European Medicines Agency (EMA) and discusses the existing options regarding the application of Machine Learning, given the limitations regarding the characterization data of nanomaterials and their biological behavior. It is the first article of its kind, to carry out a dogmatic study of the regulation of nanotechnology in medical products, from the point of view of Machine Learning. This type of study is necessary to explore how the application of these techniques can be improved and how to better present information in terms of governance, to ensure a better application of the precautionary principle. It takes as reference the European system, being able to be replicable in any other country or zone.

2. Regulation of Nanomedicines in European Union

There are different possibilities of nanotechnology governance, referring to the institutional configuration and regulatory options proposed by the literature.³⁶ European Union constitutes a common market with supranational institutions and has opted for

governance based on hard law for the granting of authorization for the commercialization of medicinal products. The justification is to reduce the asymmetry of information between the applicant and the institutions. That is, there are standards for ensuring public health that do not depend on non-governmental institutions and are based primarily on mandatory regulations and directives. This system is not based on recommendations. Thus, if a company wishes to market medicine nanoproducts in the context of the European Union, it must have been authorized by the European Commission.³⁵ The authorization process applied depends on the type of the product; the regulation, therefore, is designed vertically.

For medicine products, Directive 2001/83/EC was an important step for the “mutual recognition procedure” harmonization. Through this procedure the evaluation of the medicinal product is carried out and can be authorized in other member states depending on a report issued by the State where it is applied. However, with Regulation 726/2004, considerable progress was made with the creation of EMA and the centralized process: The approval for the commercialization of the new drug through the centralized process allows marketing in all countries of the European Union, Iceland, Norway and Liechtenstein. Among the different justifications for including this novelty in the governance of pharmaceutical products, the regulation points out the need to maintain a high level of scientific evaluation for high-tech medicines and ensure criteria of efficacy and quality (recital 7). Therefore, this procedure is applied to a wide variety of medicinal products, see **Figure 1**.

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

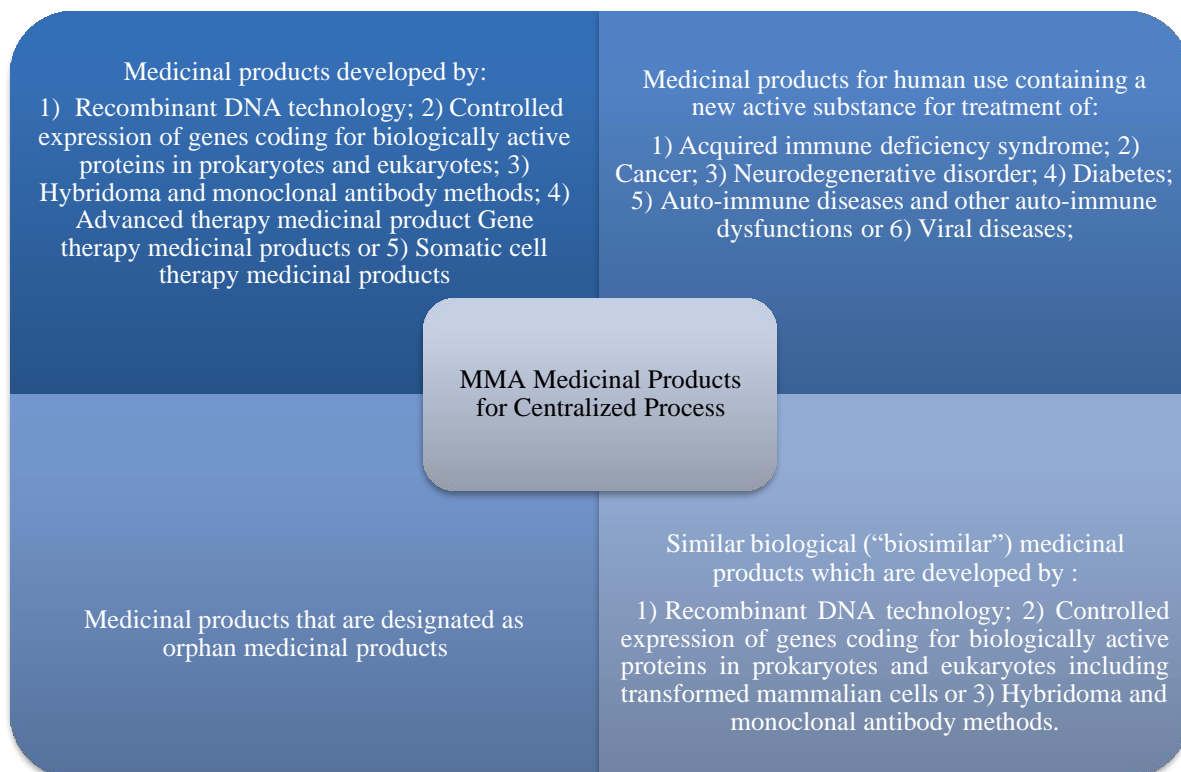


Figure 1. Medicinal products that must be authorized through centralized process in European Union

The complete process lasts around 30 months and begins with a submission of eligibility request, between 18 and 7 months before marketing authorization application (MAA), see **Figure 2**. Subsequently, if it is eligible, the notification of intention to be delivered submit an application, 7 months before the MMA. In the same way, 7 months before, the appointment of rapporteurs is carried out, who will be in charge of the scientific evaluation. Furthermore, meetings are organized, with EMA in order to advise, from a regulatory point of view, the company how to proceed. After 3 months of such meetings, the applicant must confirm that the evaluation process will continue. Subsequently, the request is sent through the “eSubmission” portal. After the scientific evaluation, with a maximum period of 210 days, the Committee for Medicinal Products for Human Use CHMP evaluates the MAA and sends a scientific opinion to the European Commission. Within 67 days, the European Commission must make the decision to authorize or not, through a European Public Assessment Report (EPAR).

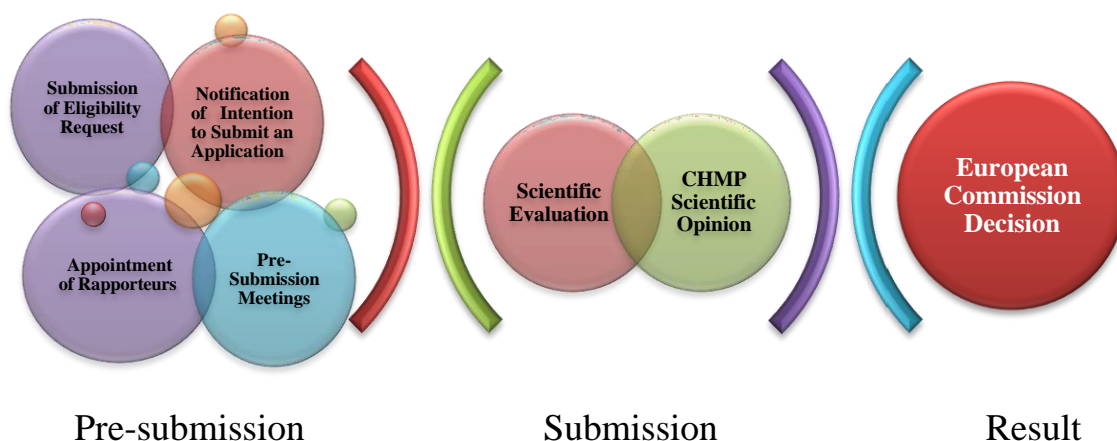


Figure 2. Medicinal products that must be authorized through centralized process.

During the pre-submission stage, the meetings phase allows the applicant to be guided. This implies that the presentation of information and the scientific evaluation phases are more efficient. As we will see in the following sections, this is especially useful for drugs incorporated into nanosystems, given the uncertainty about physicochemical properties, as well as how the information should be characterized and presented. Another aspect detailed in these meetings is how to design, build and use predictive models using Machine Learning techniques for these nanosystems. There are guides proposed by the CHMP, referring to both nanosystems and Machine Learning techniques, which will be discussed in the following sections, on how to build and present these types of models. However, these guides are characterized by being flexible. They are based on the fact that each nanosystem may differ in physicochemical properties, and studies must have been applied one by one.

3. Machine Learning and Nanomedicine Regulation

The application of regulation for nanotechnology medicine products, should consider the protection of certain rights of consumers related to public health. This objective must be achieved without discouraging the development of nanotechnology. The effect must be the opposite: Granting legal certainty to protect consumers, build trust, promote the use of improved safe drugs and encourage the design of new drugs. Given the limitations

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

of knowledge about the behavior of novel nanosystems, prediction using computational techniques is revealed as a convenient tool.³⁷ One of the valuable applications of machine learning is its ability to predict cases (dependent variable) keeping in mind the attributes (independent variables) of previous cases. Similarly, the biological activity of components can be predicted with the reference of attributes of other cases. Therefore, at the legislative level, it is presented as a useful tool for the application of the precautionary principle, for nanostructured drugs.³⁶ The precautionary principle is a legal institution that justifies anticipatory measures to control a given risk. This risk must be justified, although not showed with complete certainty. In this way, the body that authorizes the marketing of drugs will be able to know with certain levels of precision, the behavior of these compounds or the parts that compose them. They can also be used by applicants to complement information on the behavior of new drugs as well as justify the selection of characterization methods and tests performed.

For the construction of the model, a reproducible process must be followed to extract information from the available data at the time. The application of the CRISP-DM³⁸ (Cross Industry Standard Process for Data Mining) methodology is widespread for data mining processes, see **Figure 3**. This methodology presents the advantage of being the most contrasted by data mining experts. It is based on the understanding of the objectives, assess the needs, goals as well as make a plan to specify what is to be predicted. Subsequently, the data is collected and the exploration and preprocessing is carried out to have quality data. The next step is to build and adjust the model. To finalize the model is evaluated and applied. This methodology is also not completely linear since the steps taken feed back the previous steps to improve the prediction. This characteristic permits us to create a model adjusted to reality to face the uncertainty problem of nanosystems in different phases.

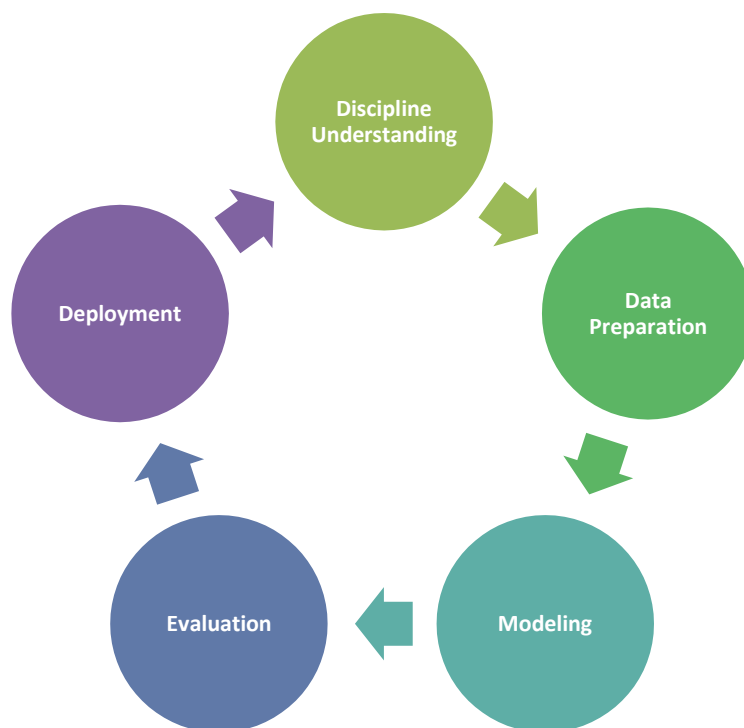


Figure 3. General Machine Learning Process Scheme.

The process described above should result in a predictive model that meets certain international standards such as the OECD. For this, OECD proposed a guidance on how to validate QSAR (Quantitative Structure-Activity Relationships) models.³⁹ The document was generated approved at the Joint Meeting Of The Chemicals Committee And The Working Party On Chemicals, Pesticides And Biotechnology of March 30, 2007. The purpose of this document is to provide a detailed but not prescriptive guidance on how principles should be applied to different QSAR models. These principles consist of model features: 1) A defined endpoint; 2) an unambiguous algorithm, in order to ensure transparency in the description of the algorithm ; 3) a defined domain of applicability in order to ensure reliable predictions; 4) appropriate measures of goodness-of-fit, robustness and predictivity, depending on the statistical method and 5) a mechanistic interpretation.

3.1. EMA and guidances for Machine Learning

EMA provided guidances regarding how modeling and simulation techniques should be applied to generate information in different specific fields. As we will see, many common points are presented, so we could explore the possibility to create one single

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

guidance for modelling any type of nanosystem. The 3 guidances used to gather the principles on which it is based are indicated hereunder:

1) Guidanceline on Reporting the Results of Population Pharmacokinetic Analyses (CHMP/EWP/185990/06).⁴⁰ Document presented by CHMP.

This detailed guidance is the information that must be included for population pharmacokinetic studies. This is applied for regulatory purposes. In this guidance, there is not detailed information about how the process of creating the model should be. The same document establishes that these principles are equally applicable to pharmacodynamic (PD) and pharmacokinetic/pharmacodynamics (PK / PD) studies.

In the presentation of the analysis plan, information such as the objectives of the analysis or justification and source of the data should be included. In addition, the report must reflect information related to the model in order to be evaluated and ensure its usefulness and effectiveness. This information includes: a) general modeling aspects (e.g. software, estimation methods, diagnostics); b) the overall modeling procedure / strategy; c) the structural models to be tested (if this has been decided); d) the variability models to be tested, e) the covariates and covariate models to be tested together with a rationale for testing these covariates based on, for example, biological, pharmacological and / or clinical plausibility; f) the algorithms / methods to be used for covariate model building; g) the criteria to be used for selection of models during model building and inclusion of covariates and h) the model evaluation / qualification procedures to be used.

The guidance does not expressly refer to simulation techniques such as the use of information fusion to generate a model. These methods can be an excellent tool to predict the biological and/or physicochemical behavior of the components in a complex simulated system. If justified within the situation analysis and strategy design section as well as in the preparation of data, the model would comply with the conditions included in this guidance.

The document also includes the recommended sections in the report to be submitted by the applicant, see **table 1**. It should be noted that they are recommendations to provide

detailed information from the regulatory point of view. Therefore, it is not mandatory and should be interpreted and adapted as the scientific knowledge of this area evolves.

Sections	Description	R. Aspects ^a
Summary	Overall summary of the population PK analysis	Context of study, Objectives, Study design, Data, Methods, Results, Main finding and Conclusion
Introduction	To provide a context for the specific population PK analysis	Background information, intent and special features
Objectives	Specific objectives of the population PK analysis	No detailed
Data	Data sets characteristics	Description of studies, method to calculate the endpoints, description and justification of variables transformation, description of missing data process, specifications of datasets used and identification of outliers
Methods	Methods used and same components as analysis plan	Justification of choice of analysis and estimation method. Software used and assumptions Moreover, information about structure of the potential model, covariates, variability and model evaluation
Results	Description of process and important decisions taken to achieve the final model with high level of robustness	Data description, Base model description, covariate selection, final model description and final model evaluation
Discussion	Assessment of how the model describes the data	How the covariates influence and how the results will be used

^aR. Aspects= Recommended Aspects.

Table 1. Recommended information to present for PK/PD models

The guidance also does not include express reference to accepted types of techniques, since there is no machine learning technique that stands out from the rest in a generic

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

way. It will depend on the specific data. Therefore, in a flexible way, the guidance proposes that each case must be justified, in order to grant the appropriate regulatory value. According to this, it also does not comment on the number of cases needed to train the model, or how the information of the different test conditions should be integrated into the data. In this sense, we highlight the Perturbation Theory Machine Learning (PTML) methodology for its versatility by being able to work with heterogeneous data (different endpoints) resulting from different trials with different conditions.

2) Guidanceline on the Investigation of Drug Interactions (CPMP/EWP/560/95/Rev).⁴¹ Document presented by CPMP.

This guidanceline is designed to recommend how information must be presented in terms of drug interaction. That is, the effect produced by medicine on another drug and the drug on medicine. This is due to the problems that may arise from this type of interactions and in some cases it may reduce the effectiveness of the treatment or increase the adverse effects.

In order to know the pharmacokinetic interactions it is recommended to use *in vivo* studies (mainly in humans due to the difficulties of extrapolating results with other methods) and *in vitro*. In addition, the importance of complementary studies *in silico* to inform about the qualitative potential of the interaction as well as estimate its quantitative effect is expressly stated. In this sense, *in silico* studies are conceived as models that are constructed throughout the different phases of drug design. This allows the model to be corrected with the information that is generated in the different stages. Subsequently, it can also be used to simulate effects under different doses.

This guidance refers to the use of simulations to have more information about the potential for interaction between drugs or to improve the design of the live experiment. The indicated guidance, by providing more information once the tests have been carried out, can update the previously constructed model, as a measure to improve its accuracy.

Regarding the report that must be presented on the applied analysis, it must include information on detailed description of the structure of the models, source, justification of parameters, explanation of the assumptions and their physiological and biochemical

plausibility, sensitivity analysis, etc. It does not include the information necessary for its presentation, but it is understood that it will be all the information that helps to understand the model, its design, precision and usefulness, in a transparent way.

3) Guidanceline on the qualification and reporting of physiologically based pharmacokinetic (PBPK) modelling and simulation (EMA/CHMP/458101/2016).⁴² Document presented by CHMP.

This guidance is justified by the high number of drug authorization applications that include Physiologically Based Pharmacokinetic (PBPK) models. It is the first EMA guidance specifies the elements that must include the information that accompanies the presentation of the model. Doing so, it can be evaluated for regulatory terms in an appropriate manner. The guidance does not distinguish between platforms that are operating in the market and platforms created *ad hoc* for a specific application before the EMA. In addition, it expressly states that it may be applied in any area where such a model is presented. The information guidance presented for PBPK modeling and simulation also covers the qualification regime of a platform to be used for regulatory purposes.

Regarding the report, the guidance indicates elements that should be included, see **Table 2**. These elements are the objectives, background information, assumptions, system dependent parameters, drug model, results and discussion. It is expressly established that they are recommendations. So the information should be as complete as possible in order to be evaluated from the regulatory point of view. Therefore, if more information is available that is not included in this guidance, it would be advisable to include it in the report so that there is no information asymmetry between the applicant and the institution.

Sections	Description	R. Aspects ^a
Objective	Objective and the intended regulatory purpose of the PBPK modelling	No detailed
Background information	Information about the investigational drug emphasising in vivo and in vitro ADME	Relevant pharmacokinetic characteristics of the drug. If possible quantitative mass-

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

		balance diagram. Explanatory text and references
Assumptions	Assumptions made in the submitted drug model	Explicit and systematic discussion of the assumptions. Data to support the assumptions and their biological plausibility. Discussion on impact the assumptions have on the model and the outcome
System dependent parameters	Summary of parameters	Any change must be justified
Drug model	Description of the investigational drug model.	Description of model building, Drug dependent parameters, Drug model structure, Sensitivity analysis, Characterizing the level of confidence in PBPK models, Evaluation of the drug model
Results	Results presented in a clear and comprehensive manner	Details of simulation conditions, sensitivity analysis, model files, parameters presented visually with descriptive statistics.
Discussion	Contribution of the model	Contribution for decision making process. Relacionar los conceptos de exposición y eficacia/seguridad.

^aR. Aspects= Recommended Aspects.

Table 2. Recommended information to present for PBPK modelling and simulation

Given that the guidanceline on reporting the results of population pharmacokinetic analyzes guidance, this guidance is characterized by being flexible in terms of model design, provided they are justified and transparent. No restrictions on the preprocessing of data or sources are included. Likewise, no comments are introduced on whether it should be a single endpoint to predict, or if the model can take into account different endpoints with heterogeneous information. This point is of enormous importance since the existing data in public repositories are characterized by being heterogeneous. As we mentioned before, the importance of the PTML methodology is pointed out, since it

allows generating multi input and multi output models, taking into account the information of more cases with different endpoints.

3.2. EMA guidances for the authorization of nanomaterials.

To achieve a proper application of the process described previously, CHMP issued documents that served as a guide for the development and characterization of nanomedicine products. These guidances help companies to identify and present the necessary information to achieve successful marketing authorization applications. As the CHMP points out, they are guidances that must be interpreted together. Next, the 4 reflection papers elaborated on nanomedicine are indicated and commented:

1) Data requirements for intravenous iron-based nano-colloidal products developed with reference to an innovator medicinal product (EMA/CHMP/SWP/620008/2012).⁴³ Document presented by CHMP.

This document was adopted by CHMP on 26 March 2015 and replaced the previous document “Reflection paper on non-clinical studies for generic nanoparticle iron medicinal product applications” (EMA / CHMP / SWP / 100094/2011).

This document aims to identify relevant information to support marketing authorization for an intravenous iron-based nano-colloidal product developed with reference to an innovator product. This product is designed for the treatment of iron deficiency cases. Iron release capacity is related to the size and surface of colloidal iron and matrix. However, in order to comply with the requirements, not only the decision can be taken taking into account the concentration, but other data on the toxicity and performance of this drug.

To do this, they must present a highly similar quality profile. If there is a difference, it must be justified in terms of safety and efficacy. The document refers to 3 aspects that can influence the safety and efficacy of the product: 1) stability of the iron-carbohydrate complex; 2) physicochemical properties of the carbohydrate matrix and 3) physicochemical properties of the iron and iron-carbohydrate complex. To evaluate it and ensure the level of quality, in addition to the physicochemical characterization, both clinical and non-clinical trials are required, in which machine learning fulfills both informative and complementary roles.

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

For the use of Machine Learning techniques in order to create models and simulations of pharmacokinetic performance (PBPK) of the nanoparticles, a biodistribution studies that evaluate distribution, metabolism and excretion of nanoparticles are necessary. In addition, pharmacokinetic information about their *in vivo* degradation or solubilization products should be included. Since the models are created with the intention of predicting the behavior of these nanocolloidal products, the document mentions the need to know the distribution not only in blood/plasma, but also reticular endothelial system (RES) and target tissues / organs. If PBPK modeling or simulation is presented, the information of the model regarding the structure, parameters and discussion of certainty are necessary.

The document includes an endpoint which refers to empirical models. In these cases if parameters are the same, a justification must be presented. Besides, it points out the importance of building a model to estimate a determined endpoint (i.e., the difference in distribution of iron in various tissues). The document does not pronounce on multioutput information models, capable of predicting different outputs with high precision. However, since it is not expressly mentioned and they fulfill the function referred to in the document, we cannot find a reason why it cannot be used. To present it, a graph representing the predicted and observed cases in different tissues or fluids will also be necessary.

2) Data requirements for intravenous liposomal products developed with reference to an innovator liposomal product (CHMP/806058/2009/Rev. 02).⁴⁴ Document presented by CHMP.

This document was adopted by CHMP on 13 March 2013. This document aims to guidance the production of information for marketing authorization for an intravenous liposomal products developed with reference to an innovator product. These products are designed for the encapsulation of active substances in the aqueous phase of the liposome given the different pharmacokinetic properties of this type of systems. The physicochemical properties of these systems such as particle size, membrane fluidity, surface-charge and composition are determinants for biological behavior. However, in

order to be authorized, there must be more information. The reason is that if changes are made in the production process or in the formulation, the interaction with the cell could vary. This may lead to a change in the safety or effectiveness of the new product.

Like the previous section, to be authorized you must present a highly similar quality profile. To do this, a comparison should be made with the reference product to ensure its quality as a first step. To comply with the pharmaceutical comparison, the differences between the applicant's product and the reference product must be discussed and justified in terms of safety and efficacy. Subsequently, clinical and non-clinical studies are conducted to know the biological behavior and ensure equivalence between the products.

The document notes that non-clinical pharmacodynamic studies should include evidences of similarity by using appropriate *in vivo* models and applying different doses levels. Similarly, for pharmacokinetic studies some aspects can be presented using a predictive model, based on both cells and animals. In these cases, in addition to information on the exposure, data should also be provided on the similarity in the distribution and disposal of the product. Besides, the document also refers to the applicant being able to choose the model provided that their suitability to investigate the release of drugs with liposomes is justified and is constructed with appropriate species data for this type of study. As for the quantity of the doses, the document recommends that PBPK modelling or allometric equations should be used.

On the other hand, the document also proposes the use of models for toxicological studies. This type of study, in general, is not necessary in this marketing authorization process. However, depending on the characteristics of the new product, the result of the comparison of quality or the toxicity generated by the product, it may be required: Specifically, the use of models to show toxicological results focused on specific organs. The document also notes that to determine adverse effects such as immune reactogenicity, the use of modeling techniques is appropriate.

3) Joint MHLW/EMA reflection paper on the development of block copolymer micelle medicinal products (EMA/CHMP/13099/2013).⁴⁵ Document presented by CHMP.

This document was adopted by CHMP on 19 December 2013. This document provides information on the production, clinical studies and non-clinical studies of block-

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

copolymer micelle drug products. These products carry an active substance, offering properties such as better stability once created, optimize pharmacokinetics, control the release of the active substance, etc. The document points out the complexity of these systems and the need to foster dialogue with the regulatory organization in order to advise on the critical attributes of the system.

This document was adopted by CHMP on 19 December 2013. This document provides information on the production, clinical studies and non-clinical studies of block-copolymer micelle drug products. These products carry an active substance, offering properties such as better stability once created, optimize pharmacokinetics, control the release of the active substance, etc. The document points out the complexity of these systems and the need to foster dialogue with the regulatory organization in order to advise on the critical attributes of the system. With the information provided by the studies, both *in vivo* and *in vitro*, the use of supervised Machine Learning techniques would allow us to predict the behavior of this type of medicinal products. Once reported, *in silico* studies could provide toxicity studies to obtain results with certain levels of accuracy. These tools are useful for justifying both the design of the assay, as well as predicting the biological behavior of the nanosystem.

4) Reflection paper on surface coatings: general issues for consideration regarding parenteral administration of coated nanomedicine products (EMA/325027/2013).⁴⁶ Document presented by CHMP.

This document was adopted by CHMP on 22 May 2013. It is a document that provides information on the development and lifecycle of coated nanomedicine products. This guidance hosts both covalently and noncovalently bounds coatings. The document refers to the possibilities of this type of product especially for the ability to minimize aggregation and improve stability as well as to modify the critical properties of the product in terms of safety and effectiveness. Although it does not expressly refer to the use of modeling techniques, information regarding *in vivo* and *in vitro* studies, such as determination of the physico-chemical stability of the coating in respect of proposed use,

under conditions relevant to the route of administration or *in vivo* impact of different coating materials / surface coverage on PK and bio-distribution.

The use of modeling and simulation techniques can lead to a considerable advance in the knowledge of the physicochemical and biological behavior of coated nanomedicine products. As we have seen previously, the principle underlying the guidances on the presentation of information related to predictive models and the use of machine learning is the principle of transparency. This principle finds justification to reduce the asymmetry of information between the institution and the applicant. Therefore, we do not identify a reason why the criteria in these guidelines should not apply for coated nanomedicine products. However, under the principle of regulatory economy, it is advisable to include a horizontal guide on the presentation of information for all types of nanosystems.

4. Concluding thoughts.

In the context of the European Union, the regulation of nanotechnology has been applied vertically, that is, depending on the type of product a different regulation is applied. In the case at hand, an express regulation is applied to medicinal products that incorporate nanomaterials. As a result of this regulation, these products must be authorized through a process that shows the effectiveness of the new medicinal product. There is a centralized process dependent on the EMA and the European Commission that includes a wide variety of drugs.

In different guidances, the information that the applicant must present has been indicated. A process with transparency is pursued to reduce the asymmetry of information between the institution and the applicant. The parts included are not taxative, so the recommendation is to provide all the information available, in addition to that indicated in the guidance. Likewise, before carrying out the application process, there is a prior advice process by EMA. In this way, time and regulatory complexity are reduced. This process aims to apply the precautionary principle and thus ensure public safety.

The EMA has ruled on the application of modeling and simulation techniques: It can be done to complement information at different times of the process. These are advanced techniques that predict, with a certain level of precision. The minimum level of precision is not included, such as the percentage of sensitivity, specificity or other statistical parameter. It also does not include which methods are more suitable for this type of study, for example, it is more complicated to give mechanistic interpretation to models

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

constructed with neural networks, comparing to other techniques such as logistic or linear discriminant analysis (LDA). Therefore, they are very flexible guidances that try to open the way to advanced Machine Learning techniques. Taking into account the uncertainty regarding the scientific knowledge of nanotechnology and the need to study products with nanotechnology on a case-by-case basis, it is an adequate approach in terms of governance.

Furthermore, the guidances do not include limitations on the process of data extraction or data preprocessing. This means that different modeling and simulation techniques can be used, if justified, such as the fusion of data with which you can build data sets. In these cases, the Perturbation Theory Machine Learning technique is of particular interest, since it uses tests with different conditions as input measuring different endpoints. This makes it possible to create flexible multioutput models to better understand the behavior of the compound, according to different endpoints, which is useful for both the applicant and the EMA. This method is aligned with the principles recommended by the OECD, on the validation of QSAR models.

As we have seen throughout this study, the tendency of EMA is to create presentation guidances for models and simulations for different types of medicinal products. Likewise, there are also different documents that recommend how to present the information in case of being nanotechnological medicinal products. In the different documents, there is redundant information from the point of view of the information to be presented. The creation of a document with recommendations about the information to be presented for any model or simulation for nanotechnological products is recommended, given the particular characteristics and the current data limitation.


Acknowledgements

R.S.C. thanks COLCIENCIAS scholarship for the doctorate studies; “Convocatoria para Doctorado Nacional 757” from 2017. This original research is part of the project “Investigación en Derecho Internacional y Nanotecnología” registered in the Research Centre of Universidad Pontificia Bolivariana with register number 766B-06/17-37. Special gratitude is extended to CYTED NANOCELIA and USEDAT networks.

Authors information

Corresponding author

*(R.S) E-mail: ricardo.santana@opendeusto.es (R.S.)

 Orcid: 0000-0002-5206-2305

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

References

1. Lane, T., Russo, D.P., Zorn, K.M., Clark, A.M., Korotcov, A., Tkachenko, V., Reynolds, R.C., Perryman, A.L., Freundlich, J.S., Ekins, S. Comparing and Validating Machine Learning Models for Mycobacterium tuberculosis Drug Discovery. *Mol. Pharm.* **15**, 4346–4360 (2018).
2. Lei, T., Sun, H., Kang, Y., Zhu, F., Liu, H., Zhou, W., Wang, Z., Li, D., Li, Y., Hou, T. ADMET Evaluation in Drug Discovery. 18. Reliable Prediction of Chemical-Induced Urinary Tract Toxicity by Boosting Machine Learning Approaches. *Mol. Pharm.* **14**, 3935–3953 (2017).
3. Ohashi, R., Watanabe, R., Esaki, T., Taniguchi, T., Torimoto-Katori, N., Watanabe, T., Ogasawara, Y., Takahashi, T., Tsukimoto, M., Mizuguchi, K. Development of Simplified in Vitro P-Glycoprotein Substrate Assay and in Silico Prediction Models to Evaluate Transport Potential of P-Glycoprotein. *Mol. Pharm.* **16**, 1851–1863 (2019).
4. Fusani, L., Brown, M., Chen, H., Ahlberg, E., Noeske, T. Predicting the Risk of Phospholipidosis with in Silico Models and an Image-Based in Vitro Screen. *Mol. Pharm.* **14**, 4346–4352 (2017).
5. Li, X., Xu, Y., Lai, L., Pei, J. Prediction of Human Cytochrome P450 Inhibition Using a Multitask Deep Autoencoder Neural Network. *Mol. Pharm.* **15**, 4336–4345 (2018).
6. Zhavoronkov, A. Artificial Intelligence for Drug Discovery, Biomarker Development, and Generation of Novel Chemistry. *Mol. Pharm.* **15**, 4311–4313 (2018).
7. Reynolds, G. H. Nanotechnology and Regulatory Policy: Three Futures. *Harv. J. Law Technol.* **17**, 179–208 (2003).
8. Toropova, A. P., Toropov, A. A., Rallo, R., Leszczynska, D. & Leszczynski, J. Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. *Ecotoxicol. Environ. Saf.* **112**, 39–45 (2015).
9. Sizochenko, N., Rasulev, B., Gajewicz, A., Kuzmin, V., Puzyn, T., Leszczynski, J. From basic physics to mechanisms of toxicity: The ‘liquid drop’ approach

- applied to develop predictive classification models for toxicity of metal oxide nanoparticles. *Nanoscale* **6**, 13986–13993 (2014).
10. Mikolajczyk, A., Gajewicz, A., Rasulev, B., Schaeublin, N., Maurer-Gardner, E., Hussain, S., Leszczynski, J., Puzyn, T. Zeta potential for metal oxide nanoparticles: A predictive model developed by a nano-quantitative structure-property relationship approach. *Chem. Mater.* **27**, 2400–2407 (2015).
 11. Ojha, P. K., Kar, S., Roy, K. & Leszczynski, J. Toward comprehension of multiple human cells uptake of engineered nano metal oxides : quantitative inter cell line uptake specificity (QICLUS) modeling. *Nanotoxicology* 1–21 (2018).
 12. Golbamaki, A., Golbamaki, N., Sizochenko, N., Rasulev, B., Cassano, A., Puzyn, T. Leszczynski, J., Benfenati, E. Classification nano-SAR modeling of metal oxides nanoparticles genotoxicity based on comet assay data. *Toxicol. Lett.* **258**, 62–324 (2016).
 13. Natalja, F., Marjana, N., Agnieszka, G. & Bakhtiyor, R. The way to cover prediction for cytotoxicity for all existing nano-sized metal oxides by using neural network method. *Nanotoxicology* **11**, 475–483 (2017).
 14. Sizochenko, N., Mikolajczyk, A., Jagiello, K., Puzyn, T., Leszczynski, J., & Rasulev, B. How the toxicity of nanomaterials towards different species could be simultaneously evaluated: a novel multi-nano-read-across approach. *Nanoscale* **10**, 582–591 (2017).
 15. Mikolajczyk, A., Gajewicz, A., Mulkiwicz, E., Rasulev, B., Marchelek, M., Diak, M., Hirano, S., Medynska A., Puzyn, T. Nano-QSAR modeling for ecosafe design of heterogeneous TiO₂-based nano-photocatalysts. *Environ. Sci. Nano* **5**, 1150–1160 (2018).
 16. Golbamaki, A., Golbamaki, N., Sizochenko, N., Leszczynski, J. & Benfenati, E. Genotoxicity induced by metal oxide nanoparticles : a weight of evidence study and effect of particle surface and electronic properties. *Nanotoxicology* 1–17 (2018).
 17. Singh, K. P. & Gupta, S. Nano-QSAR modeling for predicting biological activity of diverse nanomaterials. *RSC Adv.* **4**, 13215–13230 (2014).
 18. Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X., Dasari, T., Michalkova, A., Hwang, H., Toropov, A., Leszczynska, D., Leszczynski, J. Using nano-QSAR to predict

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

- the cytotoxicity of metal oxide nanoparticles. *Nat. Nanotechnol.* **6**, 175–178 (2011).
19. Toropov, A., Toropova, A., Benfenati, E., Gini, G., Puzyn, T., Leszczynska, D., Leszczynski, J. Novel application of the CORAL software to model cytotoxicity of metal oxide nanoparticles to bacteria *Escherichia coli*. *Chemosphere* **89**, 1098–1102 (2012).
 20. Pathakoti, K., Huang, M.-J., Watts, J. D., He, X. & Hwang, H.-M. Using experimental data of *Escherichia coli* to develop a QSAR model for predicting the photo-induced cytotoxicity of metal oxide nanoparticles. *J. Photochem. Photobiol. B Biol.* **130**, 234–240 (2014).
 21. Gajewicz, A., Schaeublin, N., Rasulev, B., Hussain, S., Leszczynska, D., Puzyn, T. & Leszczynski, J. Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. *Nanotoxicology* **9**, 313–325 (2015).
 22. Sayes, C. & Ivanov, I. Comparative Study of Predictive Computational Models for Nanoparticle-Induced Cytotoxicity. *Risk Anal.* **30**, 1723–1734 (2010).
 23. Toropova, A. P. & Toropov, A. A. Optimal descriptor as a translator of eclectic information into the prediction of membrane damage by means of various TiO₂ nanoparticles. *Chemosphere* **93**, 2650–2655 (2013).
 24. Kar, S., Gajewicz, A., Puzyn, T., Roy, K. & Leszczynski, J. Periodic table-based descriptors to encode cytotoxicity profile of metal oxide nanoparticles: A mechanistic QSTR approach. *Ecotoxicol. Environ. Saf.* **107**, 162–169 (2014).
 25. Liu, R., Rallo, R., George, S., Ji, Z., Nair, S., Nel, A., Cohen, Y. Classification NanoSAR development for cytotoxicity of metal oxide nanoparticles. *Small* **7**, 1118–1126 (2011).
 26. Liu, R., Zhang, H., Ji, Z., Rallo, R., Xia, T., Chang, C., Nel, A. & Cohen, Y. Development of structure-activity relationship for metal oxide nanoparticles. *Nanoscale* **5**, 5644–5653 (2013).
 27. Toropova, A., Toropov, A., Benfenati, E., Korenstein, R., Leszczynska, D., Leszczynski, J. Optimal nano-descriptors as translators of eclectic data into prediction of the cell membrane damage by means of nano metal-oxides. *Environ.*

- Sci. Pollut. Res.* **22**, 745–757 (2015).
28. Patel, T., Telesca, D., Low-Kam, C., Ji, Zx, Zhang, H. Y., Xia, T., Zinc, J. I., Nel, A. E. Relating nano-particle properties to biological outcomes in exposure escalation experiments. *Environmetrics* **25**, 57–68 (2014).
 29. Toropova, A. P., Toropov, A. A., Benfenati, E. & Korenstein, R. QSAR model for cytotoxicity of SiO₂nanoparticles on human lung fibroblasts. *J. Nanoparticle Res.* **16**, (2014).
 30. Burello, E. & Worth, A. P. A theoretical framework for predicting the oxidative stress potential of oxide nanoparticles. *Nanotoxicology* **5**, 228–235 (2011).
 31. Epa, V., Burden, F., Tassa, C., Weissleder, R., Shaw, S., Winkler, D. Modeling biological activities of nanoparticles. *Nano Lett.* **12**, 5808–5812 (2012).
 32. Mitter, N. & Hussey, K. Moving policy and regulation forward for nanotechnology applications in agriculture. *Nature* **14**, 508–510 (2019).
 33. Eisenhardt, K. M. Agency theory: An assessment and review. *Acad. Manag. Rev.* **14**, 57–74 (1989).
 34. Uskokovic, V. Nanotechnologies : What we do not know. *Technol. Soc.* **29**, 43–61 (2007).
 35. Commission, E. Regulation (EC) No 726/2004. (2004).
 36. Stone, V. *et al.* The Essential Elements of a Risk Governance Framework for Current and Future Nanotechnologies. *Risk Anal.* **38**, 1321–1331 (2017).
 37. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. Recent advances and applications of machine learning in solid- state materials science. *Comput. Mater.* **5**, 1–36 (2019).
 38. Wirth, R. & Hipp, J. CRISP-DM : Towards a Standard Process Model for Data Mining. *Proc. 4th Int. Conf. Pract. Appl. Knowl. Discov. data Min.* 29–39 (2000).
 39. OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models.* (2007).
 40. CHMP. Guidanceline on Reporting the Results of Population Pharmacokinetic Analyses. *Eur. Med. Agency* 1–11 (2007).
 41. CHMP. Guidanceline on the Investigation of Drug Interactions. *Eur. Med. Agency* 1–59 (2013).
 42. CHMP. Guidanceline on the qualification and reporting of physiologically based

4) EUROPEAN NANOTECHNOLOGY REGULATION (PHARMACEUTIC SECTOR)

- pharmacokinetic (PBPK) modelling and simulation. *Eur. Med. Agency* 1–18 (2017).
43. CHMP. Data requirements for intravenous iron-based nano-colloidal products developed with reference to an innovator medicinal product. *Eur. Med. Agency* 1–11 (2012).
 44. CHMP. Data requirements for intravenous liposomal products developed with reference to an innovator liposomal product. *Eur. Med. Agency* 1–13 (2009).
 45. CHMP. Joint MHLW/EMA reflection paper on the development of block copolymer micelle medicinal products. *Eur. Med. Agency* 1–18 (2013).
 46. CHMP. Reflection paper on surface coatings: general issues for consideration regarding parenteral administration of coated nanomedicine products. *Eur. Med. Agency* 1–5 (2013).

*If I have seen further it is by standing
on the shoulders of Giants*

Isaac Newton

CHAPTER

5

5) Modelling Vitamin Derivatives

As we will check in the following chapters, vitamin derivatives will be one of the most important compounds in particular nanosystems. This is the case of DVRNs. This compound could give to these systems more desirable biological activities. We are interested in designing even better compounds to integrate them in nanosystems.

To do so, we develop a model able to predict a multi output model able to predict biological activities of new vitamin derivatives. We apply the PTML methodology by following the workflow included in Figure 6.

5) MODELLING VITAMIN DERIVATIVES

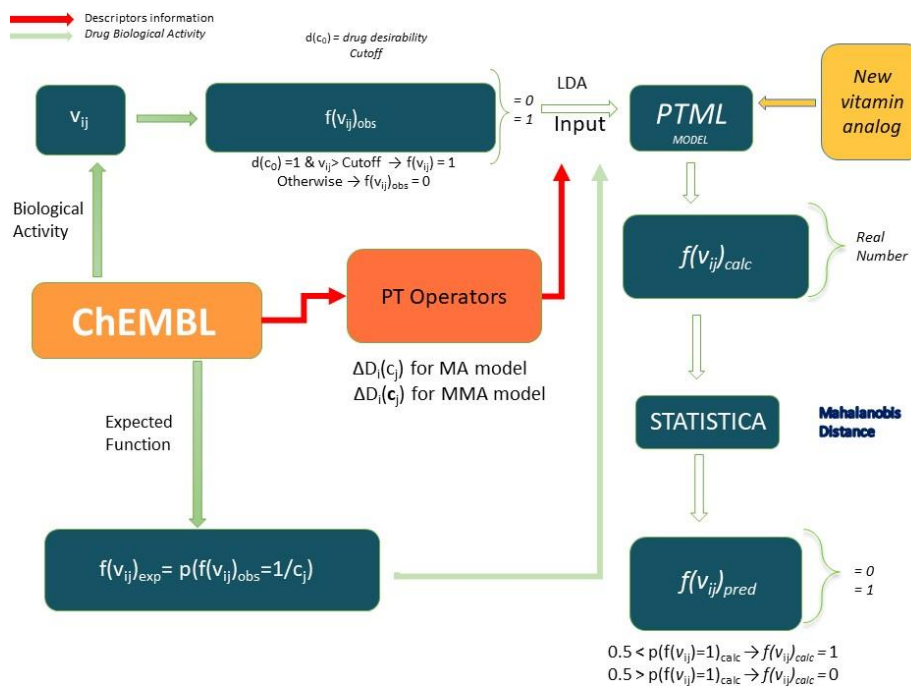


Figure 6. General workflow to build the models for the present study

PTML Model of ChEMBL Compounds Assays for Vitamin Derivatives

Ricardo Santana^{a,b}, Robin Zuluaga^d, Piedad Gañán^c, Sonia Arrasate^e,

Enrique Onieva Caracuel^a, and Humbert González-Díaz^{e,f,}*

^aDeustoTech-Fundación Deusto, Avda. Universidades, 24, 48007 Bilbao, Spain.

^bGrupo de Investigación sobre Nuevos Materiales, Universidad Pontificia Bolivariana UPB, 050031, Medellín, Colombia.

^cFacultad de Ingeniería Química, Universidad Pontificia Bolivariana UPB, 050031, Medellín, Colombia.

^dFacultad de Ingeniería Agroindustrial, Universidad Pontificia Bolivariana UPB, 050031, Medellín, Colombia.

^eDepartment of Organic Chemistry II, University of Basque Country UPV/EHU, 48940, Leioa, Spain.

^fIKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain.

Keywords: ChEMBL; Vitamins; Perturbation Theory; Machine Learning; Big data; Multi-target models.

5) MODELLING VITAMIN DERIVATIVES

ABSTRACT. Determining the biological activity of vitamins derivatives is needed given that organic synthesis of analogs of vitamins is an active field of interest for Medicinal Chemistry, Pharmaceutical and Food Additives. Accordingly, scientists from different disciplines perform preclinical assays (n_{ij}) with a considerable combination of assay conditions (c_j). Indeed, ChEMBL platform contains a database that includes results from 36220 different biological activity bio-assays of 21240 different vitamin and vitamin derivatives. These assays present are heterogeneous in terms of assay combinations of c_j . They are focused on > 500 different biological activity parameters (c_0), > 340 different targets (c_1), > 6200 types of cell (c_2), > 120 organisms of assay (c_3) and > 60 assay strains (c_4). It includes a total of > 1850 niacin assays, > 1580 tretinoin assays, > 1580 retinol assays, 857 ascorbic acid assays, *etc.* Given the complexity of this combinatorial data in terms of being assimilated by researchers, we propose to build a model by combining Perturbation Theory (PT) basis and Machine Learning (ML). Through this study, we propose a PTML (PT + ML) combinatorial model for ChEMBL results on biological activity of vitamins and vitamins derivatives. The Linear Discriminant Analysis (LDA) model presented for training subset a Specificity (%) = 90.38, Sensitivity (%) = 87.51, and Accuracy (%) = 89.89. The model showed for external validation subset Specificity (%) = 90.58, Sensitivity (%) = 87.72, and Accuracy (%) = 90.09. Different types of linear and non-linear PTML models such as Logistic Regression (LR), Classification Tree (CT), Näive Bayes (NB), and Random Forest (RF) were applied in order to contrast the capacity of prediction. The PTML-LDA model predicts with more accuracy by applying combinatorial descriptors. In addition, PCA experiment with chemical structure descriptors allowed to characterize the high structural diversity of the chemical space studied. In any case, PTML models using chemical structure descriptors do not improve the performance of the PTML-LDA model based on ALOGP and PSA. We can conclude that the

three variable PTML-LDA model is a simplified and adaptable tool for the prediction, for different experiment combinations, the biological activity of derivative vitamins.

■ INTRODUCTION

The organic synthesis of analogs of vitamins is an active field of interest for Medicinal Chemistry, Pharmaceutical and Food Additives industry as well¹. For instance, vitamin D₃ analogs synthesis have been promoted given its capacity to modulate signaling pathways with the objective of discover desirable effects in cancer cells.¹⁻⁵ We can consider other examples as vitamin K analogs as possible cancer therapy, due to the inhibition produced in the growth of cancer cells by mechanisms as apoptosis, cell cycle arrest and autophagy.⁶ Also, for their characteristics to develop stem-cell-based therapies for neurodegenerative diseases.⁷ Vitamin E analogs express neuroprotective function due to the anti-apoptotic properties,^{8,9} among other functions.¹⁰ Vitamin B₁₂ analogs for cyanide detection and detoxification¹¹, and more different functions.¹²

Preclinical assays about vitamins and their derivatives are important in order to generate more information about their biological activity.¹³⁻³² In ChEMBL database, 36220 assays results of this type of compound have been found. There is four main types of assays: a) Data measuring binding of vitamin to a molecular target (K_i, IC₅₀, K_d, etc.); b) Data measuring the biological effect of a compound (% cell death in a cell line, rat weight, etc.); c) Data about absorption, distribution, metabolism, and excretion (ADME); d) Data measuring toxicity of a compound (cytotoxicity, hepatotoxicity, *etc.*); e) Assays measuring physicochemical properties of the compounds in the absence of biological material (*e.g.* solubility); and f) Other studies that we cannot consider for any of the previous groups.³³ Given that every compound must be studied for every target for selectivity process, the number of assays would be too high, with high costs involved and animals sacrifice. Consequently multi target computational models that

5) MODELLING VITAMIN DERIVATIVES

predict biological activities and chemical properties are useful.³⁴⁻³⁹ This is aligned to the basis of Replacement, Reduction and Refinement in animal experimentation (three Rs principle)⁴⁰

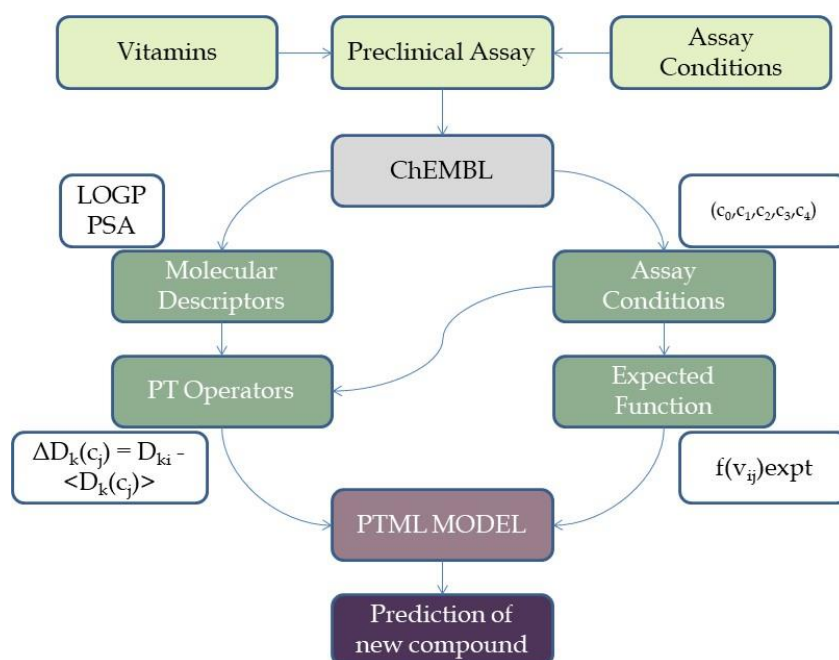
Computational capacities through Machine Learning (ML) give us the opportunity to process information as molecular descriptors; however, traditional techniques to extract metadata from complex databases of preclinical assays are not adequate. This is the case of ChEMBL database, which contains Big Data sets, according to HACE Theorem (starting from Heterogeneous and Autonomous sources that seeks to explore Complex and Evolving relationships among data)⁴¹ of preclinical assays.⁴² In ChEMBL database, a specific biological activity is considered in each assay, such as Activity, EC₅₀, IC₅₀, *etc.*

We must highlight that potential duplicates were a key aspect in preprocessing phase. We detected and deleted duplicated cases. The cases without indispensable information such as biological values, measures or assay conditions were also deleted. Furthermore, the activity comments were not used to build the model. We must point out that we applied PTML method, which has been published and contrasted in literature; even to create previous models with ChEMBL databases. There are not models of this type that take into account techniques such as natural language process of activity comments. We propose to apply this technique for future researches. Regarding the dataset, ChEMBL is managed by the European Bioinformatics Institute (EBI). The data is extracted directly from the literature: There are 7 core journals: Bioorganic & Medicinal Chemistry Letters, Journal of Medicinal Chemistry, Bioorganic & Medicinal Chemistry, Journal of Natural Products, European Journal of Medicinal Chemistry, ACS Medicinal Chemistry Letters and MedChemComm. After extracting the data, a manual curation process is applied. Moreover, the data is updated regularly every 3-4 months. We use this dataset, the version of January 2019. We must say that most of the standard relation data is not "=", which is other cause we use classification techniques instead of regression ones. The

number of every standard relation is: 9 “~”, 152 “<”, 2 “<=”, 11034 “=”, 254 “>”, 13 “>=” and 24756 blanks.

Furthermore, the combination of Perturbation Theory (PT) basis and Machine Learning techniques can be applied to solve this uncertainty of compounds activity. This is due to PTML (Perturbation Theory Machine Learning) models have been applied to different areas like medicinal proteomics, chemistry, and nanotechnology.⁴³⁻⁵⁵ This model is especially adequate for databases with similar Big Data characteristics and combinatorial information. Nevertheless, researches about PTML models for vitamin and vitamin derivatives compounds including multiple biological activity data have not been reported. We advance the first model in order to predict vitamins derivatives biological activity. This is possible thanks to the versatility by PTML method, which, gives the opportunity to decrease costs, reduce, replace and refine animal experimentation.

A general workflow followed for PTML construction is presented In **Figure 1**. This workflow will be taken as reference for this research.



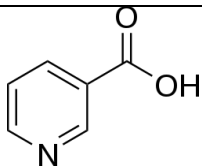
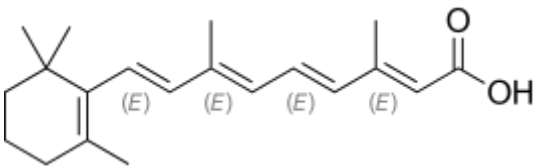
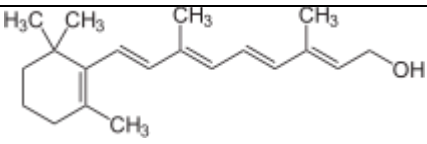
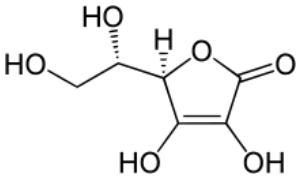
5) MODELLING VITAMIN DERIVATIVES

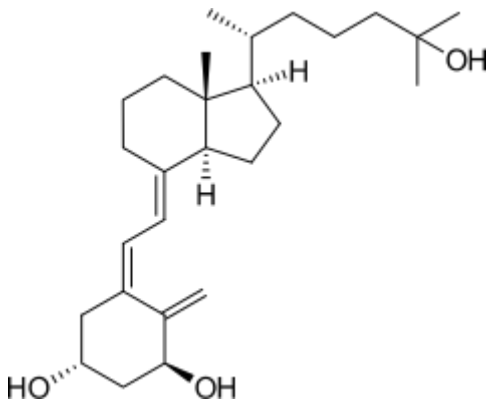
Figure 1. General workflow to develop a PTML model by using ChEMBL database

■ MATERIALS AND METHODS

Vitamins Data pre-processing. The results of the preclinical assays that were used to build the model were extracted from the public database ChEMBL in January 2019. The structure of the dataset is heterogeneous given that there are different compounds. In some cases, a determined compound constitutes different observations given that many assays have been applied. All the compounds have at least 70% of similarity with a vitamin reported in ChEMBL. The compounds with more cases are niacin > 1800, tretinoin > 1500 cases, retinol > 1500 cases and ascorbic acid > 800, calcitriol > 700 see **Table 1**. In supplementary information, we report all the classifiers to identify any compound included in the dataset.

Table 1. Compounds with more observations in dataset

Subset	Cases	Molecular Formula	Structure
Niacin (Vitamin B3)	1863	C ₆ H ₅ NO ₂	
Tretinoin (vitamin A)	1589	C ₂₀ H ₂₈ O ₂	
Retinol (Vitamin A)	1585	C ₂₀ H ₃₀ O	
Ascorbic Acid (Vitamin C)	847	C ₆ H ₈ O ₆	

<p>Calcitriol (Vitamin D)</p>	<p>732</p>	<p>$C_{27}H_{44}O_3$</p>	
-----------------------------------	------------	-------------------------------------	--

Each assay presents an experimental parameter v_{ij} that measures the biological activity of the i^{th} vitamin analog (vit_i) over a given target j^{th} . In all cases, the value of the experimental parameter v_{ij} depends on two elements: 1) The structure of the i^{th} vitamin and 2) conditions $c_j = (c_0, c_1, c_2, \dots, c_n)$ that characterize every preclinical assay. The first considered c_j is $c_0 =$ the biological activity v_{ij} taking into account the units in which the result is presented (IC_{50} (nM), EC_{50} (nM), *etc.*). Besides, other 4 conditions were included: $c_1 =$ target protein, $c_2 =$ name cell, $c_3 =$ assay organism and $c_4 =$ assay strain.

In order to build the model, classification techniques were considered to predict a desirable biological activity. The model gives us the opportunity to predict compounds behavior, in terms of desirable biological values. Below, we propose different combination of cutoffs, in case we need a more restricted model. The aim of this study is to create a useful tool to complement information or event to guide researchers for the development of new compounds. For that, at some point, a decision must be made to continue with the development of the compound. Given the characteristics of the dataset and the literature, PTML technique for classification is a suitable method to extract knowledge. That is the reason why a classification technique is preferred over regression methods. We also present this study to promote other future researches to build models by applying regression techniques. Thus, the discretization of the

5) MODELLING VITAMIN DERIVATIVES

values v_{ij} is as follows: $f(v_{ij})_{obs} = 1$ if $v_{ij} > \text{cutoff}$ and the desirability of the biological activity parameter $d(c_0) = 1$ (**Table 2**). On the other hand, $f(v_{ij})_{obs} = 1$ also if $v_{ij} < \text{cutoff}$ and $d(c_0) = 0$, otherwise $f(v_{ij})_{obs} = 0$. The value $f(v_{ij})_{obs} = 1$ means that there is a desired biological effect of the vitamin over the target. When $d(c_0) = 1$ means that the activity, measured with a determined units, and the biological effect are directly proportional, otherwise $d(c_0) = 0$.

PTML linear model. PTML technique is especially adequate for complex databases with varied preclinical assays registered, as on ChEMBL.⁵⁶ Once the model is prepared, scoring function values $f(v_{ij})_{calc}$ can be calculated for a specific vitamin taking as a reference multiple conditions combination $\mathbf{c}_j = (c_0, c_1, c_2, \dots, c_n)$ of the bio-assay. An important aspect to highlight is that Moving Averages (MA) as input of the model is convenient as they have information of the molecular descriptors and the assay conditions. In this paper, the linear PTML models seek have the following form of equation 1:

$$f(v_{ij})_{calc} = a_0 + a_1 \cdot f(v_{ij})_{expt} + \sum_{k=1, j=0}^{kmax, jmax} a_{kj} \cdot \Delta D_k(c_j) \quad (1)$$

The ML algorithm Linear Discriminant Analysis (LDA) was created with the software STATISTICA.⁵⁷ This algorithm lets us to develop the first PTML classification models (PTML-LDA) with the purpose of predicting different biological activities of a molecule. We also developed other PTML linear models using LDA algorithm and Logistic Regression (LR) algorithm implemented in R studio.⁵⁸

PTML non-linear models. We also explored other PTML non-linear models with the following ML algorithms: Classification Tree (CT), Näive Bayes (NB), and Random Forest (RF) were applied.^{59–62} All these algorithms were run in the platform R studio,⁵⁸ by using MASS, NNET, RPART, E1071 and RANDOMFOREST packages with default parameters. There are other algorithms than are suitable to be applied with PTML like Artificial Neural Networks (ANN), Kernel Support Vector Machine or XGBoost among others. When we apply Perturbation

Operators to a reference function, we can consider PTML is applied, so different algorithms are applicable to explore and better model the data. We propose further studies to explore these other techniques by using the workflow presented in **Figure 2**.

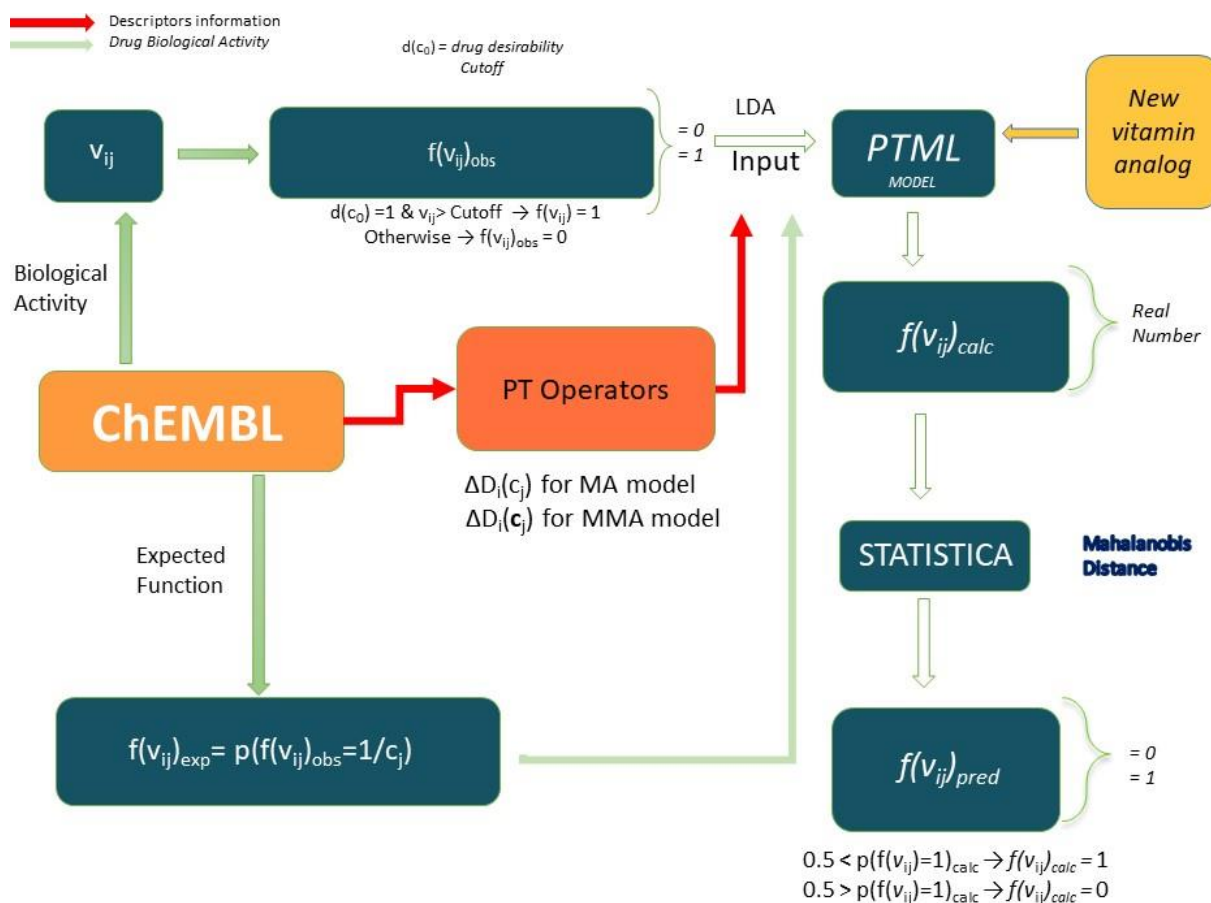


Figure 2. General workflow to build the models for the present study.

■ RESULTS AND DISCUSSION

PTML-LDA MA linear model. PTML model takes as reference the expected value of biological activity $f(v_{ij})_{\text{expt}}$ of a determined vitamin analog and considers the perturbations in the system through the PT operators $\Delta D_k(c_j)$. The best model found in terms of Accuracy, Specificity and Sensitivity is the following:

$$f(v_{ij})_{\text{calc}} = -8.1168817 + 17.540180 \cdot f(v_{ij})_{\text{expt}} \quad (2)$$

5) MODELLING VITAMIN DERIVATIVES

$$\begin{aligned}
 &+0.18278 \cdot \Delta D_1(c_0) \\
 &-0.32862 \cdot \Delta D_1(c_1) \\
 &-0.47524 \cdot \Delta D_1(c_2) \\
 &-0.01883 \cdot \Delta D_2(c_2) \\
 &+0.230552 \cdot \Delta D_1(c_3) \\
 &+0.02133 \cdot \Delta D_2(c_3) \\
 &+0.46245 \cdot \Delta D_1(c_4)
 \end{aligned}$$

$$n = 24146 \quad \chi^2 = 21558.03 \quad p < 0.05$$

In **Table 2**, information of PT operators included in PTML-LDA model is described. Besides, the following are the statistical parameters: n corresponds to the number of bio-assays for training the model, χ^2 is the Chi-square statistics and p is the p-level. In equation 2, the variable $f(v_{ij})_{\text{expt}}$ is the expected value of biological activity for a determined vitamin taking into account an assay conditions combination $\mathbf{c}_j = (c_0, c_1, c_2, \dots, c_j \dots, c_{\text{max}})$. PT operators consist in Moving Averages product of the operation $\Delta D_k(c_j) = D_{ki} - \langle D_k(c_j) \rangle$, for one condition or a combination of different conditions at time. Thus, it takes as reference the value D_{ki} , which is utilized in the model to encode the structure of every vitamin studied. In this present research, the molecular descriptors provided by ChEMBL and taken into consideration were $D_1 = \text{ALOGP}$ (n -Octanol/Water Partition Coefficient) and $D_2 = \text{PSA}$ (Polar Surface Area). PT operators express the perturbation of D_{ki} from the expected value. In this case, the expected value used is the average $\langle D_k(c_j) \rangle$ of every molecular descriptor measured in different conditions or combination of conditions \mathbf{c}_j or \mathbf{c}_j , respectively. As output, the model calculates $f(v_{ij})_{\text{calc}}$ which is a scoring function of v_{ij} of the vitamin derivative in the different conditions or combinations of assay conditions \mathbf{c}_j . After calculating $f(v_{ij})_{\text{calc}}$ the algorithm compute the posterior probabilities $p(f(v_{ij})_{\text{obs}} = 1)_{\text{pred}}$ by applying Mahalanobis's distance metric.⁶³

Table 2. Perturbation theory operators included in the model.

Assay Condition	Condition	Symbol	Operator Formula	Description

c ₀	Activity type	f(v _{ij}) _{expt}	n(f(v _{ij}) _{obs} =1)/n _j	Expected value of probability p(f(v _{ij})=1) _{expt} for c ₀
c ₀	Activity type	ΔD ₁ (c ₀)	$\Delta D_1 = ALOGP_i - \langle ALOGP(c_j) \rangle$ $\Delta D_2 = PSA_i - \langle PSA(c_j) \rangle$	ΔALOGP(c _j) refers to the deviation of the hydrophobicity of the vitamin (ALOGP _i) from the expected value (<ALOGP (c _j)>) for condition j
c ₁	Target	ΔD ₁ (c ₁)		
c ₂	Cell Name	ΔD ₁ (c ₂) ΔD ₂ (c ₂)		
c ₃	Assay	ΔD ₁ (c ₃)		
	Organism	ΔD ₂ (c ₃)		
c ₄	Assay Strain	ΔD ₁ (c ₄)		

Once calculated $p(f(v_{ij}) = 1)_{pred}$ a Boolean function can be built: $f(v_{ij})_{pred} = 1$ if $p(f(v_{ij}) = 1)_{pred} > 0.5$; otherwise, $f(v_{ij})_{pred} = 0$. If $f(v_{ij})_{pred} = f(v_{ij})_{obs}$ the vitamin is considered as properly classified.⁶³ This PTML-LDA model has adequate values of Specificity (%) = 90.6, Sensitivity (%) = 91.27, and Accuracy (%) = 90.75 in training series. The model offered Specificity (%) = 90.69, Sensitivity (%) = 91.67 and Accuracy (%) = 90.85 in external validation series, see **Table 3**. In terms of Medicinal Chemistry, these values suited the requirements regarding the range for classification models⁶⁴. The model was not trained with cases used in validation series.

Residual muy grande o del res muy grande. Leverage muy grande, tiene problemas. Casos con poca dispersión

Table 3. Classification matrix PTML-LDA model with one-condition MA operators

5) MODELLING VITAMIN DERIVATIVES

Obs.	Stat.	Pred.	Pred. sets ^a		
Sets	Param. ^a	Stat. ^a	n_j	$f(v_{ij})_{pred} = 0$	$f(v_{ij})_{pred} = 1$
Training					
$f(v_{ij})_{obs} = 0$	Sp	90.55	18688	18360	328
$f(v_{ij})_{obs} = 1$	Sn	91.52	5458	1914	3544
Total	Ac	90.71	24146	20274	3872
Validation					
$f(v_{ij})_{obs} = 0$	Sp	90.69	9337	9174	163
$f(v_{ij})_{obs} = 1$	Sn	91.67	2737	941	1796
Total	Ac	90.85	12074	10115	1959

^a Obs. Sets = Observed sets, Sp = Specificity, Sn = Sensitivity, Stat. Param. = Statistical parameter, Pred. Stat. = Predicted statistics.

As observed, PTML-LDA model is useful to classify the biological activity of a vitamin analog in assays with different conditions combinations. First we calculate the expected probability of biological activity $p(f(v_{ij})_{obs} = 1)_{expt}$. To that, $p(f(v_{ij})_{obs} = 1)_{expt} = n(f(v_{ij})_{obs} = 1)/n_j$. In this equation, $n(f(v_{ij}) = 1)_{obs}$ refers to the number of vitamins $n(f(v_{ij}) = 1)_{obs}$ with a desired level of a determined activity, see **Table 4**. If $f(v_{ij})_{obs} = 1$, the value of activity $v_{ij} > \text{cutoff}$ for a biological activity type with desirability $d(c_0) = 1$. The value observed of a vitamin can also be classified $f(v_{ij})_{obs} = 1$ when the value of activity $v_{ij} < \text{cutoff}$ for activities with desirability $d(c_0) = 0$. Otherwise, the vitamin is not considered to have a desirable biological activity, $f(v_{ij})_{obs} = 0$. It is important to highlight that the desirability of the vitamins takes into account a cutoff, that in all cases with activity measured in nM will be $\text{cutoff} = 100$. Otherwise, $\text{cutoff} = \langle v_{ij} \rangle$, which is the average of the value of the biological activity v_{ij} .

Table 4. Biological activity parameters (c_0)

Condition c_0	$\langle D1(c_0) \rangle$	$\langle D2(c_0) \rangle$	$n_j(c_0)$	$n_j(f(v_{ij})=1)_{obs}$	$p(f(v_{ij})=1)_{expt}$	Cutoff	$d(c_0)$
Potency(nM)	3.29	74.06	24750	104	0.004	100.00	0

IC ₅₀ (nM)	4.24	63.20	1402	232	0.165	100.00	0
Activity(%)	3.79	79.40	1079	56	0.052	186.79	1
Inhibition(%)	3.25	82.98	415	254	0.612	73.72	0
EC ₅₀ (nM)	4.50	66.09	388	193	0.497	100.00	0
Weight(g)	3.18	35.24	260	192	0.738	4.23	0
Ratio(-)	5.44	63.26	259	253	0.977	46.59	0
GI50(nM)	3.81	38.24	258	2	0.008	100.00	0
Ki(nM)	3.83	77.45	197	106	0.538	100.00	0
Activity(mg/dl)	5.74	58.21	164	95	0.579	5.22	0

For prediction purposes, firstly we introduce the expected value of the molecular descriptors $\langle D_i(c_j) \rangle$ for the condition (or multiple conditions) of a particular assay. In **Table 5**, we see how the expected value $\langle D_i(c_j) \rangle$ vary depending on the subset of data in each condition. This will be important in terms of information for the model and the prediction of the biological activity. The rest of the expected values for every condition can be consulted in supplementary information file SM01.xlsx. The model is able to predict various activity parameters for every no-experimented vitamin. These values change for different activities parameters. Secondly, we introduce the values of new vitamin analog descriptors (in this case, ALOGP and PSA).”

Table 5. One-condition averages $\langle D_i(c_j) \rangle$ and the respective number of cases $n_j(c_j)$

Condition c_1^a	c ₁ Parameters			Condition c_1^a	c ₁ Parameters		
	$n_j(c_1)$	$\langle D_1(c_1) \rangle$	$\langle D_2(c_1) \rangle$		$n_j(c_1)$	$\langle D_1(c_1) \rangle$	$\langle D_2(c_1) \rangle$
088496	71	0.3	224.8	P41231	2.00	-1.86	186.07
P11473	24189	3.3	74.2	Q9NPD5	14.00	2.30	95.94
m.d.	9794	3.6	57.7	Q9Y6L6	13.00	1.88	101.76
P00568	3	-1.9	186.1	P12931	1.00	-1.86	186.07
Condition c_2^a	c ₂ Parameters			Condition c_2^a	c ₂ Parameters		
	$n_j(c_2)$	$\langle D_1(c_2) \rangle$	$\langle D_2(c_2) \rangle$		$n_j(c_2)$	$\langle D_1(c_2) \rangle$	$\langle D_2(c_2) \rangle$

5) MODELLING VITAMIN DERIVATIVES

m.d.	33972	3.33	70.50	K562	28	-0.94	144.87
CHO	45	2.11	77.94	HL-60	527	5.55	66.65
VERO	17	-1.40	168.20	Raji	46	3.38	106.85
HeLa	25	1.84	88.45	U-251	6	3.00	76.95
Condition c_3^a	c ₃ Parameters			Condition c_3^a	c ₃ Parameters		
	$n_j(c_3)$	$\langle D1(c_3) \rangle$	$\langle D2(c_3) \rangle$		$n_j(c_3)$	$\langle D1(c_3) \rangle$	$\langle D2(c_3) \rangle$
m.d.	26435	3.24	74.99	O. cuniculus	22	0.33	134.91
R. norvegicus	5385	3.59	46.36	H. herpesvirus 1	4	-0.13	148.16
Homo sapiens	2756	4.62	60.99	Measles virus	2	-1.86	186.07
C. griseus	74	4.65	63.47	Sindbis virus	2	-1.86	186.07
Condition c_4^a	c ₄ Parameters			Condition c_4^a	c ₄ Parameters		
	$n_j(c_4)$	$\langle D1(c_4) \rangle$	$\langle D2(c_4) \rangle$		$n_j(c_4)$	$\langle D1(c_4) \rangle$	$\langle D2(c_4) \rangle$
m.d.	31513	3.46	73.40	Charles foster	3	-1.69	269.43
Wistar	46	4.64	75.04	BALB/c	12	1.12	176.72
Cultivar 7042S	45	-0.34	106.76	C57BL/6	24	0.65	116.16
KM	5	-1.69	269.43	A/B./1/18	2	-1.69	269.43

^a Full name: Cricetulus griseus, Oryctolagus cuniculus, Human herpesvirus 1, Rattus norvegicus, A/Bervig_Mission/1/18.

PTML-LDA MMA linear model. We present a different model using different PT operators. Previously, we included operators that took into account only one condition at time. For example $\Delta D1 = ALOGP_i - \langle ALOGP(c_j) \rangle$, expresses the deviation of a case regarding all the assays with the same condition c_j ; the average $\langle ALOGP(c_j) \rangle$ included information about a determined condition c_j . In this case the PT operators incorporate information about multiple conditions c_j : a vector \mathbf{c}_j (with \mathbf{c} in boldface). Consequently, the molecular descriptors are maintained $D_1 = ALOGP$ (n -Octanol/Water Partition Coefficient) and $D_2 = PSA$ (Polar Surface Area) but we incorporate $\mathbf{c}_j = (c_0, c_1, c_2, c_3, c_4)$. Consequently, the model has two different PT operators: $D_1(c_0, c_1, c_2, c_3, c_4)$ and $D_2(c_0, c_1, c_2, c_3, c_4)$. Consequently, we do not use one-condition

(MA) but multi-condition combinatorial averages (MMAs),⁶⁵ in this case, combining all the conditions included in **Table 2**. The equation of this model PTML Combinatorial model is the following

$$\begin{aligned}
 f(v_{ij})_{calc} &= -8.13181 + 17.925632 \cdot f(v_{ij})_{expt} \\
 &\quad - 0.03001 \cdot \Delta D_1(c_0, c_1, c_2, c_3, c_4) \\
 &\quad - 0.00103 \cdot \Delta D_2(c_0, c_1, c_2, c_3, c_4)
 \end{aligned} \tag{3}$$

n = 24146 $\chi^2 = 21367.407$ p < 0.05

As we can see in eq. 4, the dimensionality of the model is lower than the previous one. We must highlight that the 2 variables consisting in moving averages actually include information of 12 variables (2 numerical variables treated as descriptors and 8 categorical variables treated as assay conditions). Given the heterogeneity of the dataset and the number of levels in every variable is high (for instance, more than 500 biological activities, represented in c_0), if we apply other conventional method, we would obtain a complex and high-dimension model. By doing that, $\Delta D_1(c_0, c_1, c_2, c_3, c_4)$ and $\Delta D_2(c_0, c_1, c_2, c_3, c_4)$ accumulate all this information and generate the perturbation in the system (taking as reference $f(v_{ij})_{expt}$). That is the main reason we considered PTML is advantageous and completely applicable to this dataset, taking into consideration other cases it has been applied in the state of art, specially to datasets extracted from ChEMBL 49,50,52,55,66,67.

We must point out the meaning of the statistical parameters: n is the number of cases used to train the model, χ^2 is the Chi-square statistics, and p is the p-level, as in the model with one-condition PT operators. The input variable $f(v_{ij})_{expt}$, as in the previous model, represents the expected value of biological activity for different vitamins evaluated in assays with different combinations of experimental conditions $\mathbf{c}_j = (c_0, c_1, c_2, \dots c_j \dots c_{max})$, see **Table 6**. This makes MMA operators very useful given the possibility of including different combinations of assay

5) MODELLING VITAMIN DERIVATIVES

conditions. For example, if we want to predict the result of a preclinical assay, with determined experimental conditions, we take as reference the values of the descriptors D_1 and D_2 for the same combination of experimental conditions.

Table 6. PT multiple condition operators (MMA) included in the equation (3)

Assay Condition	Symbol	Operator Formula	Description
$\mathbf{c}_j = [c_0, c_1, c_2, c_3, c_4]$	$\Delta D_1(\mathbf{c}_j)$	$ALOGP_i - \langle ALOGP(\mathbf{c}_j) \rangle$	Deviation (Δ) of the $D_1 = ALOGP_i$ or $D_2 = PSA_i$ of the i^{th} vitamin from the respective expected value ($\langle ALOGP(\mathbf{c}_j) \rangle$) or ($\langle PSA(\mathbf{c}_j) \rangle$) for cases with the same vector of conditions \mathbf{c}_j
$\mathbf{c}_j = [c_0, c_1, c_2, c_3, c_4]$	$\Delta D_2(\mathbf{c}_j)$	$PSA_i - \langle PSA(\mathbf{c}_j) \rangle$	

In order to highlight the differences presented in MA model and MMA model, see **Table 7**. The common aspect is that $f(v_{ij})_{\text{expt}}$ is an input and give us a reference for the system in which we will apply a perturbation through the perturbation operators, the moving averages. Therefore, we must point out that the main difference to among these models is the information included in the MA vs. MMA operators. On the one hand, MMA model presents as inputs $\Delta D_1(\mathbf{c}_j)$ and $\Delta D_2(\mathbf{c}_j)$. In this case, \mathbf{c}_j contains information of c_0, c_1, c_2, c_3 and c_4 , given that these MA operators correspond to the average of $ALOGP(D_1)$ or $PSA(D_2)$ for all the cases with the same vector of assay conditions c_0, c_1, c_2, c_3, c_4 . On the other hand, the MA model is built with the information of $\Delta D_1(c_0), \Delta D_1(c_1), \Delta D_1(c_2), \Delta D_1(c_3), \Delta D_1(c_4), \Delta D_2(c_0), \Delta D_2(c_1), \Delta D_2(c_2), \Delta D_2(c_3)$ and $\Delta D_2(c_4)$. This means that the average of $ALOGP(D_1)$ or $PSA(D_2)$.

Table 7. Multiple condition PT operators included in the equation (3)

Models	Assay Condition	Symbol	Operator Formula	Description
MMA Model	c_0	$f(v_{ij})_{\text{expt}}$	$n(f(v_{ij})_{\text{obs}}=1)/n_j$	Expected value of probability $p(f(v_{ij})=1)_{\text{expt}}$ for c_0
	$c_j = [c_0, c_1, c_2, c_3, c_4]$	$\Delta D_1(c_j)$	$\text{ALOGP}_i - \langle \text{ALOGP}(c_j) \rangle$	Deviation (Δ) of the $D_1 = \text{ALOGP}_i$ or $D_2 = \text{PSA}_i$ of the i^{th} vitamin from the respective expected value ($\langle \text{ALOGP}(c_j) \rangle$) or ($\langle \text{PSA}(c_j) \rangle$) for cases with the same vector of conditions c_j
	$c_j = [c_0, c_1, c_2, c_3, c_4]$	$\Delta D_2(c_j)$	$\text{PSA}_i - \langle \text{PSA}(c_j) \rangle$	
MA Model	c_0	$f(v_{ij})_{\text{expt}}$	$n(f(v_{ij})_{\text{obs}}=1)/n_j$	Expected value of probability $p(f(v_{ij})=1)_{\text{expt}}$ for c_0
	c_0	$\Delta D_1(c_0)$ $\Delta D_2(c_0)$	$\Delta D_1 = \text{ALOGP}_i - \langle \text{ALOGP}(c_j) \rangle$ $\Delta D_2 = \text{PSA}_i - \langle \text{PSA}(c_j) \rangle$	$\Delta \text{ALOGP}(c_j)$ refers to the deviation of the hydrophobicity of the vitamin (ALOGP_i) from the expected value ($\langle \text{ALOGP}(c_j) \rangle$) for condition j
	c_1	$\Delta D_1(c_1)$ $\Delta D_2(c_1)$		
	c_2	$\Delta D_1(c_2)$ $\Delta D_2(c_2)$		
	c_3	$\Delta D_1(c_3)$ $\Delta D_2(c_3)$		

5) MODELLING VITAMIN DERIVATIVES

c_4	$\Delta D_1(c_4)$ $\Delta D_2(c_4)$
-------	--

In **Table 8**, we summarize the results for the new model. In training series, the model expressed a high Specificity = $Sp(\%) = 90.38$, Sensitivity = $Sn(\%) = 87.51$, and overall Accuracy = $Ac(\%) = 89.89$. Besides, the model in terms of external validation series presented values of $Sp(\%) = 90.58$, $Sn(\%) = 87.72$, and $Ac(\%) = 90.09$. The results for alternative models taking into consideration different cutoffs, are shown in his Table. The method was to apply minus 10%, minus 25%, plus 10%, minus 25% to the initial cutoff. By doing this, we can study the correlation of the cutoff and the performance of the model. By adjusting the cutoffs we see that overall accuracy decreases for lower cutoffs. For instance, with cutoff values plus 25%, the model shows a better performance with 94.57 comparing to the model with reference cutoffs (89.89). This model would be less strict for the cases that measure determined activities. Specially, activities with considerable number of cases such as Potency (%) $IC_{50}(nM)$, Inhibition(%) and $EC_{50}(nM)$. Given the characteristics of the dataset, if we apply minus 25% cutoffs, the specificity is lower but sensitivity increases. Depending on the purpose of the model, we could consider any of them or a version built by adjusting only a few of these cutoffs.

Table 8. Classification matrix PTML-LDA model with MMA operators

Cutoff Change fold (%)	Data Series	Obs. Sets	Stat. Param. ^a	Pred. Stat.	Predicted sets		
					n_j	$f(v_{ij})_{pred} = 0$	$f(v_{ij})_{pred} = 1$
+25	Train	$f(v_{ij})_{obs} = 0$	$Sp(\%)$	95.24	19049	18143	906

		$f(v_{ij})_{obs} = 1$	Sn(%)	92.05	5097	405	4692
		total	Ac(%)	94.57	24146		
	Val	$f(v_{ij})_{obs} = 0$	Sp(%)	95.18	9535	9076	459
		$f(v_{ij})_{obs} = 1$	Sn(%)	92.16	2539	199	2340
		total	Ac(%)	94.55	12074		
+10	Train	$f(v_{ij})_{obs} = 0$	Sp(%)	93.47	19382	18116	1266
		$f(v_{ij})_{obs} = 1$	Sn(%)	90.26	4764	464	4300
		total	Ac(%)	92.83	24146		
	Val	$f(v_{ij})_{obs} = 0$	Sp(%)	93.45	9688	9054	634
		$f(v_{ij})_{obs} = 1$	Sn(%)	90.27	2386	232	2154
		total	Ac(%)	92.83	12074		
Ref	Train	$f(v_{ij})_{obs} = 0$	Sp(%)	90.38	20086	18154	1932
		$f(v_{ij})_{obs} = 1$	Sn(%)	87.51	4060	507	3553
		total	Ac(%)	89.89	24146		
	Val	$f(v_{ij})_{obs} = 0$	Sp(%)	90.58	10021	9077	944
		$f(v_{ij})_{obs} = 1$	Sn(%)	87.72	2053	252	1801
		total	Ac(%)	90.09	12074		
-10	Train	$f(v_{ij})_{obs} = 0$	Sp(%)	93.54	20971	19617	1354
		$f(v_{ij})_{obs} = 1$	Sn(%)	73.41	3175	844	2331
		total	Ac(%)	90.89	24146		
	Val	$f(v_{ij})_{obs} = 0$	Sp(%)	93.58	10456	9785	671

5) MODELLING VITAMIN DERIVATIVES

		$f(v_{ij})_{obs} = 1$	Sn(%)	73.92	1618	422	1196
		total	Ac(%)	90.95	12074		
-25	Train	$f(v_{ij})_{obs} = 0$	Sp(%)	91.06	21397	19484	1913
		$f(v_{ij})_{obs} = 1$	Sn(%)	86.03	2749	384	2365
		total	Ac(%)	90.49	24146		
	Val	$f(v_{ij})_{obs} = 0$	Sp(%)	91.09	10660	9710	950
		$f(v_{ij})_{obs} = 1$	Sn(%)	85.99	1414	198	1216
		total	Ac(%)	90.49	12074		

^aSn(%) = Sensitivity, Sp(%) = Specificity, and Ac(%) = Accuracy

This model can be utilized in order to predict biological activity of a new vitamin analog. Due to the use of MMA PT operators. They also include information of different combinations of conditions; see **Table 9**. The expected probability $p_j(f(v_{ij})=1/c_j)_{\text{expt}}$ also contains information about the experimental conditions: it is the ratio of the number of assays with the same combination of conditions that are desirable.

Table 9. Parameters for combinations of assay conditions

Multi-condition assays ^a					Multi-condition input parameters			
Activity	Protein	Cell	Assay Org.	Assay Strain	Averages		Count & Probs	
(c ₀)	(c ₁)	(c ₂)	(c ₃)	(c ₄)	<D ₁ (c _j)>	<D ₂ (c _j)>	n _j (c ₀)	p(f(v _{ij})=1/c _j) _{expt}
Potency (nM)	P11473	-	-	-	3.27	74.22	23567	0.003
Potency (nM)	-	-	-	-	3.59	73.59	576	0.024
Activity(%)	-	-	<i>Rn</i>	<i>S.D.</i>	2.06	85.27	256	0.000

Weight(g)	-	-	-	-	3.15	35.21	256	0.750
IC ₅₀ (nM)	-	-	-	-	1.39	101.41	202	0.025
Inhibition (%)	-	-	-	-	1.64	87.72	153	0.516
Activity (mg/dl)	-	-	<i>Rn</i>	<i>S.D.</i>	5.83	57.82	142	0.528
Ratio(-)	-	-	<i>Rn</i>	<i>S.D.</i>	5.71	60.22	136	0.993
IC ₅₀ (ug.mL ⁻¹)	-	-	-	-	0.46	97.35	102	0.980
Activity (%)	P11473	HL-60	<i>Hs</i>	-	5.68	77.15	101	0.386

^a *Hs* = *Homo sapiens*, *Mm* = *Mus musculus*, *Sd* = *Sprague Dawley*, *Rn* = *Rattus norvegicus*.

There are different combinations of c_j formed from >500 different biological activity parameters (c_0), >340 different targets (c_1), >6200 types of cell (c_2), > 120 organisms of assay (c_3) and >60 assay strains (c_4) but the number of combinations (c_0, c_1, c_2, c_3, c_4) is > 2280. Every combination has its respective expected (average) values $\langle D_1(c_j) \rangle$ and $\langle D_2(c_j) \rangle$ for D_1 and D_2 . These values encode information of the combination of the conditions (see file SI01.xls). Additionally, if we compare the number of variables that takes into account equation 3 with equation 2, we see that is markedly lower. MMA operators have the information related to explored combinations of an assay conditions c_j , and can be used to predict in a better way. We propose the first PTML model able to predict biological activity of vitamins and vitamins derivatives against different targets.

PTML non-linear models. In this section, we present PTML models other than PTML-LDA to predict the biological activity of vitamin derivatives with ML algorithms implemented on program language R.⁶⁸ The objective is to contrast the predictive capacity. The expected function $f(v_{ij})_{\text{exp}}$ along with all the MAs and the MMAs were introduced as input variables, to

5) MODELLING VITAMIN DERIVATIVES

conform the perturbation of the system. We used the following ML algorithms, implemented in R Studio: LDA, LR, CT, NB, and RF,⁵⁹⁻⁶² see **Table 10**. The packages used were MASS, NNET, RPART, E1071 and RANDOMFOREST respectively. The models were built by using the default parameters of these packages. These ML algorithms have been used for predictive purposes in data mining applied in different disciplines.^{69,70} They also have been implemented in Cheminformatics.⁷¹⁻⁷⁸ The PTML-LR model is the only one linear alternative to the PTML-LDA model tested. The best PTML-LR model found includes all the descriptors and has similar values of Ac(%) and Sp(%) $\approx 90 - 95$ with respect to PTML-LDA. However, PTML-LR has notably lower values of Sp(%) = 72.75 compared to Sp(%) = 90.7.

No other PTML model outperformed the PTML-LDA models developed before in terms of Sp(%), Sn(%), and Ac(%). PTML-NB model shows the highest Sn(%) = 88.88. However, we cannot consider it better in terms of capacity of prediction given that it is not balanced taking into consideration Sp (%) = 87.24%. This is the lowest Sp(%) of all the PTML models. On the other hand, the PMTL-RF model has the highest Sp (%) = 95.21 and Ac (%) = 93.06 of the non-linear models. Anyhow, the Sn (%) = 81.88, that is lower than PTML-LDA. In this case, PTML-RF has more complexity due to it included 13 variables, 3 implemented variables at each split and 50 trees. Regarding the PTML-CT presented has a high Sp (%) = 95.10; which is Sp (%) ≈ 95 like PTML-RF. However the Sn (%) = 75.15. This ratio is better than PTML-LR but lower than any non-linear PTML. Besides, PTML-CT does not include c_1 and c_3 , see **Figure 3**.

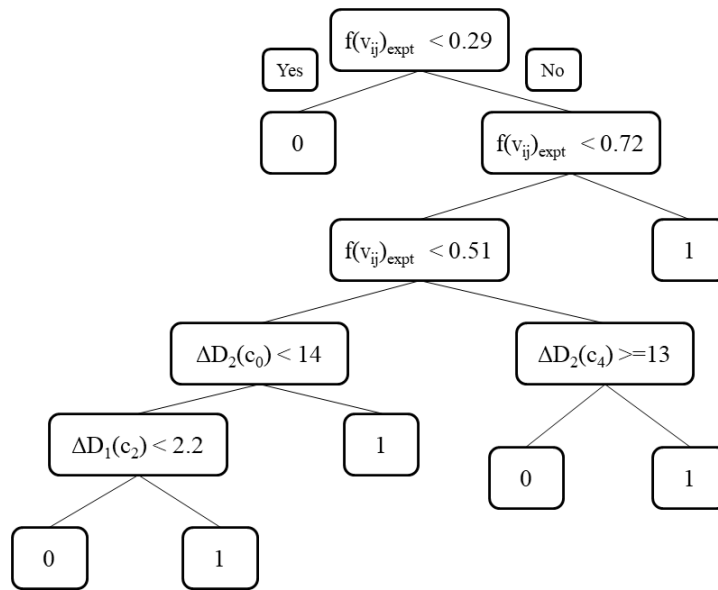


Figure 3. PTML-CT model

Table 10. PTML linear vs. non-linear models results

Soft. ^a	PTML Algorithm ^b	Observed Sets	Statistical Parameter	Predicted Statistics	Observed sets	
					$f(v_{ij})_{obs} = 0$	$f(v_{ij})_{obs} = 1$
S	LDA	$f(v_{ij})_{pred} = 0$	Sp(%)	90.8	9186	169
		$f(v_{ij})_{pred} = 1$	Sn(%)	91.4	929	1790
		total	Ac(%)	90.9	10115	1959
R	LDA	$f(v_{ij})_{pred} = 0$	Sp(%)	91.18	7023	225
		$f(v_{ij})_{pred} = 1$	Sn(%)	84.56	574	1232
		total	Ac(%)	92.44	7597	1457
R	LR	$f(v_{ij})_{pred} = 0$	Sp(%)	94.92	7211	397
		$f(v_{ij})_{pred} = 1$	Sn(%)	72.75	386	1060
		total	Ac(%)	91.35	7597	1457
R	CT	$f(v_{ij})_{pred} = 0$	Sp(%)	95.10	7225	362
		$f(v_{ij})_{pred} = 1$	Sn(%)	75.15	372	1095

5) MODELLING VITAMIN DERIVATIVES

		total	Ac(%)	91.89	7597	1457
R	NB	$f(v_{ij})_{pred} = 0$	Sp(%)	87.24	6628	162
		$f(v_{ij})_{pred} = 1$	Sn(%)	88.88	969	1295
		total	Ac(%)	87.51	7597	1457
R	RF	$f(v_{ij})_{pred} = 0$	Sp(%)	95.21	7233	264
		$f(v_{ij})_{pred} = 1$	Sn(%)	81.88	364	1193
		total	Ac(%)	93.06	7597	1457

^a Software used: S = Statistica, R = R Studio. ^b ML algorithm used: LR = Logistic Regression, CT = Classification Tree, NB = Näive Bayes, RF = Random Forest.

In addition, we carried out a Bootstrap cross validation algorithm to test the robustness of the PTML models found.⁷⁹ Given that, random data is taken to train every model. We repeated the same process 20 times (20-fold Bootstrapping). As a result, we obtained the Ac (%) mean and Ac (%) standard deviation for all training subset. The results do not show significant variations (**Table 11**) comparing to the overall accuracy showed by the first batch trained (**Table 10**). PTML-RF is the model that predict better in terms of overall accuracy.

Table 11. PTML R- non-linear models results for 20-fold bootstrapping

ML Algorithm	Statistics	
	Ac (%) mean	Ac (%) s.d.
LDA	91.6%	0.27
LR	91.6%	0.27
CT	91.9%	0.36
NB	88.2%	0.31
RF	93.2%	0.23

PTML model applicability domain. The applicability domain (AD) of a cheminformatics model is the physico-chemical, structural or biological space, in which it is applicable to make predictions for new cases. This helps us to confirm that the assumptions of the model are accomplished and the compounds that is able to predict. We can consider it an extrapolation process. In order to determine the AD, there is not a generally accepted algorithm. There are approaches that have been discussed and compared by the European Centre for the Validation of Alternative Methods (ECVAM).⁸⁰

One of them and widely used is the Williams Plot, for the structural AD of the regression cheminformatic models. It is constituted by the residuals and the deleted residuals from the cross validation process. The plot shows theses residuals and the distribution depending on the leverage. Product of this information, the plot shows six regions. The cases are separated vertically by the leverage threshold value which is the result of $[3(\text{Number of Variables}) + 1]/(\text{Number of cases}]$. The cases with leverage higher than leverage threshold value have more weight for the model fit. The horizontal lines separate the residuals with high values. This delimits regions where residual values are notably high. If there is a case with notably high residual and high weight for the fit, the probability to be a outlier is high.⁸¹ The obtained map shows the training and testing cases, see **Figure 4**. This plot is built using STATISTICA software. The leverage threshold value is 0.00037 and the errors are distributed. So there are many cases that have rather weight to fit the model. The horizontal lines separate the residuals with values 2 or more, and -2 or minus. There are not cases detected with high residuals.

5) MODELLING VITAMIN DERIVATIVES

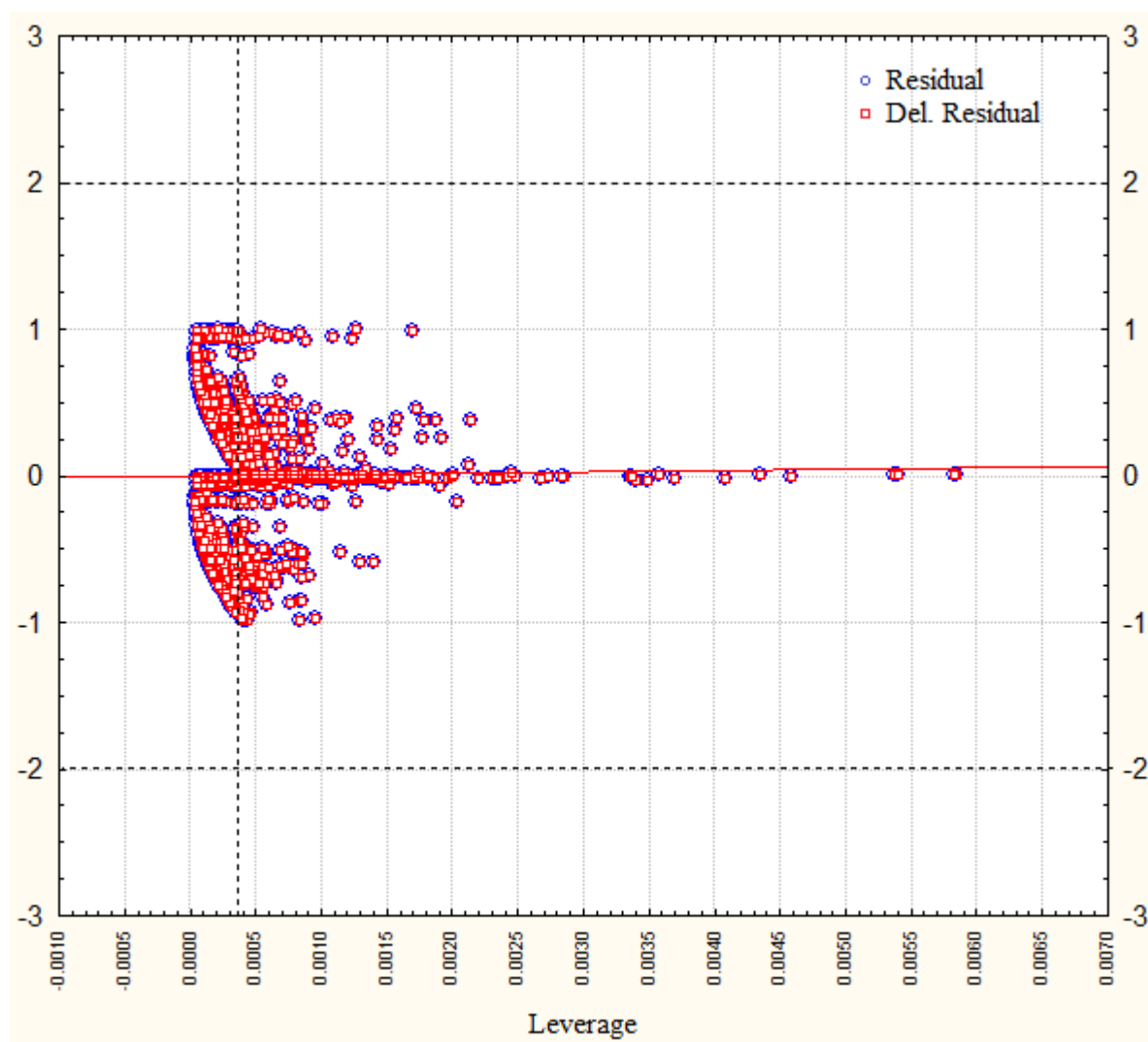
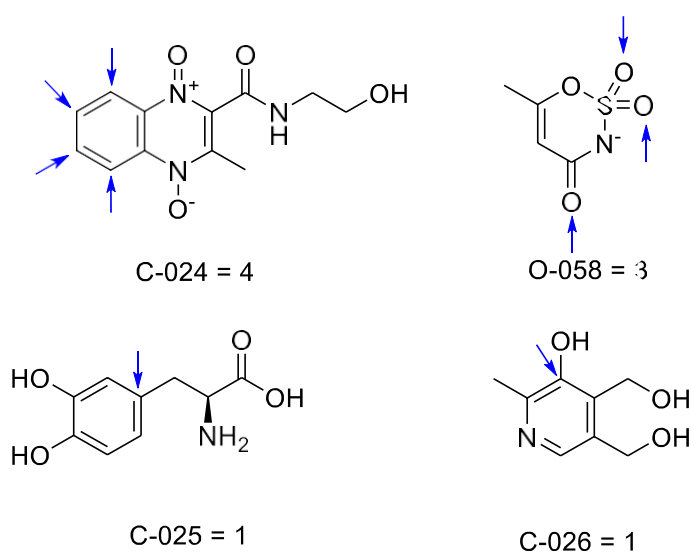


Figure 4. Williams Plot to determine Application Domain.

PTML-PCA chemical space analysis. In the previous sections, we developed PTML models using as structural variables ALOGP and PSA only. These variable correlates highly with the output function $f(v_{ij})_{obs}$. However, these variables are not expected to capture all the structural diversity of our data set. In this section, we calculated 120 new input variables related to structural features of molecules (functional groups, fragments, structural patterns, etc.) in order to characterize the chemical space in structural terms. We used the software DRAGON to calculate these new descriptors.⁸² We were able to calculate 21233 molecules out of 21241 unique molecules in total in the dataset. The other molecules presented problems to be

processed with the software. In addition, we calculated 120 molecular descriptors related to structural patterns in all these 21233 molecules. Only, 104 out of these 120 descriptors (structural patterns) are enough represented in the dataset to be considered (variance >0). In **Table 12**, we depicted selected values of the number of times a chemical pattern appear in the dataset (N) and the number of molecules (Nm) with this pattern. It is notable the high structural diversity of the dataset with >50% of the 104 structural patterns present in >1000 molecules. For instance, almost all groups has a ration $N/N_m > 1$. It means that almost all of the 104 chemical patterns studied appear in the dataset more than one time per molecule. The DRAGON codes for the more represented structural patterns/functional groups in our dataset are: H-047, C-024, O-058, C-025, H-050, and C-026. H-047 refers to the number of H atoms attached to C_{sp^3} or C_{sp^2} . C-024 refers to the number of CH groups inside aromatic rings. O-058 is the code of double-bonded oxygen atoms; e.g., aldehydes, ketones, sulfones, phosphates, *etc.* (O=). The code C-025 refers to the number of quaternary aromatic carbons. The code H-050 refers to H atoms attached to heteroatoms. The code C-026 refers to the number of CH groups of aromatic rings that are attached to one heteroatom. Please, see examples in the **Figure 5**.



5) MODELLING VITAMIN DERIVATIVES

Figure 5. Examples of DRAGON chemical structure descriptors

These chemical patterns/groups appear in more than 15000 different molecules. This coincides, with the results obtained by PCA to analyze these descriptors. The 10 first factors F_{01} - F_{10} with eigenvalue values >1.0 (in range 8.4 – 1.90) were able to explain only 30.7% of variance. Consequently, this PCA analysis showed that the present chemical space is notably heterogeneous in structural terms; *i.e.*, we were unable to explain a high percentage of the variance of structural diversity using few components. The first factor F_{01} can be identified with molecules having alkyl groups CH_3R , CH_2R_2 , CHR_3 and CR_4 and or alkenyl groups $=\text{CH}_2$, $=\text{CHR}$, $=\text{CR}_2$. However, other factors do not represent specific groups of molecules. This result was obtained using a Varimax normalized rotation of factors.⁸³ Using other rotations such as Equamax, Quartimax, and Biquartimax we can obtained different representations of the chemical space. For instance, Quartimax normalize rotation make a better representation of chemical diversity. We can identify F_{01} to F_{05} factors with specific groups of molecules, see **Table 12**. F_{01} corresponds to alkyl groups (C-001, C-002, C-003 structural patterns). F_{02} presents mainly 2 patterns C-008 (CHR_2X) and H-050. F_{03} includes information about N-075 (R--N--R or R--N--X). On the other hand, F_{04} and F_{05} include strong presence of H-047 and N-072 ($\text{RCO-NR}_1\text{R}_2$ or $\text{R}_1\text{R}_2\text{N-X=X}$). In any case, the total variance explained is the same with all rotations and almost all of the 104 chemical patterns are not identified, confirming the high heterogeneity of the dataset.

Table 12. PCA analysis results

Pattern ^a	Details ^b	F_{01}	F_{02}	F_{03}	F_{04}	F_{05}	Nm	N	N/Nm
C-002	CH_2R_2	0.7302	0.0013	0.0151	0.0193	0.0036	9754	31546	3.234
C-003	CHR_3	0.6374	0.0035	0.0089	0.0138	0.0024	3922	7129	1.818
C-015	$=\text{CH}_2$	0.4703	0.0016	0.0001	0.0111	0.0028	762	905	1.188
H-052	H''	0.4672	0.0239	0.0015	0.0329	0.0082	7787	34502	4.431

C-011	CR3X	0.4406	0.0167	0.0084	0.0004	0.0022	1497	1666	1.113
C-001	CH3R / CH4	0.3986	0.0344	0.0280	0.0662	0.0205	11792	23817	2.020
O-056	C-OH, alcohol	0.3499	0.3377	0.0011	0.0291	0.0182	1553	3010	1.938
C-004	CR4	0.3373	0.0237	0.0118	0.0790	0.0189	1806	2552	1.413
C-017	=CR2	0.2603	0.0531	0.0233	0.1360	0.0179	3887	6070	1.562
C-008	CHR2X	0.1944	0.6052	0.0013	0.0003	0.0003	4972	7490	1.506
C-024	R--CH--R	0.1923	0.0004	0.0032	0.3358	0.0006	19364	123122	6.358
C-027	R--CH--X	0.1106	0.0003	0.0411	0.0532	0.0033	2550	3670	1.439
C-025	R--CR--R	0.0897	0.0066	0.0219	0.1073	0.0420	16641	33046	1.986
C-016	=CHR	0.0774	0.0804	0.0275	0.1237	0.0238	5171	9302	1.799
C-026	R--CX--R	0.0717	0.0481	0.0118	0.2702	0.0003	15730	34049	2.165
O-057	Phenol/enol/carbox./ OH	0.0564	0.3457	0.0267	0.0664	0.0181	3936	5242	1.332
O-058	O=	0.0510	0.0130	0.0892	0.1144	0.1908	16867	33545	1.989
H-050	H-heteroatom	0.0451	0.5982	0.0158	0.0202	0.0359	15755	28388	1.802
C-019	=CRX	0.0408	0.2388	0.0263	0.0154	0.0063	2462	3128	1.271
C-039	Ar-C(=X)-R	0.0397	0.0284	0.0548	0.0074	0.1081	2670	2992	1.121
C-040	C'	0.0383	0.0032	0.0573	0.0139	0.3837	14740	23593	1.601
N-075	R--N--R / R--N--X	0.0365	0.0068	0.5445	0.0001	0.0170	7783	13306	1.710
C-022	#CR / R=C=R	0.0261	0.0001	0.0000	0.0003	0.0009	138	236	1.710
Cl-089	Cl attached to C1(sp2)	0.0205	0.0004	0.0002	0.0144	0.0020	3054	3759	1.231
O-060	O'	0.0202	0.0539	0.0015	0.0966	0.0018	11097	17429	1.571
N-074	R#N / R=N-	0.0172	0.0008	0.0018	0.0003	0.1399	4457	5393	1.210
C-012	CR2X2	0.0169	0.0062	0.0015	0.0027	0.0000	292	310	1.062
C-005	CH3X	0.0164	0.0034	0.0013	0.1907	0.0056	6930	10344	1.493
Br-094	Br attached to C1(sp2)	0.0124	0.0001	0.0052	0.0021	0.0003	1079	1184	1.097
F-082	F attached to C2(sp3)	0.0115	0.0015	0.0003	0.0000	0.0002	31	91	2.935
N-072	RCO-N< / >N-X=X	0.0088	0.0001	0.0147	0.1362	0.4970	12829	19301	1.504
F-081	F attached to C1(sp3)	0.0075	0.0001	0.0001	0.0000	0.0003	27	28	1.037
C-007	CH2X2	0.0074	0.0011	0.0001	0.0018	0.0000	534	633	1.185
C-028	R--CR--X	0.0072	0.0037	0.1654	0.0057	0.0111	4699	6052	1.288
C-037	Ar-CH=X	0.0063	0.0001	0.0012	0.0008	0.0015	704	717	1.018
O-059	Al-O-Al	0.0061	0.3818	0.0146	0.0524	0.0053	2108	2540	1.205
H-047	H'	0.0059	0.0004	0.0000	0.7556	0.0003	21088	236506	11.215

^a O-060=> O', Al-O-Ar or Ar-O-Ar; H-047 => H', H attached to C_{sp3} or C_{sp2}; C-040 => C', R-C(=X)-X / R-C#X / X=C=X. ^b DRAGON software code.

PTML-LDA Chemical structure patterns model. In this section, we present a model in order to contrast the capacity of prediction of the moving averages for the new structural descriptors calculated, comparing to ALOGP and PSA. This is a PTML model trained with LDA algorithm applying forward stepwise feature selection with prior probabilities $\pi_1 = 0.5$ and programmed for 10 maximum steps, see the result in equation 4.

5) MODELLING VITAMIN DERIVATIVES

$$\begin{aligned}
 f(v_{ij})_{calc} = & -7.08819 + 15.79509 \cdot f(v_{ij})_{expt} \\
 & -0.33992 \cdot \Delta D_3(c_0, c_1, c_2, c_3, c_4) \\
 & -0.48952 \cdot \Delta D_4(c_0, c_1, c_2, c_3, c_4) \\
 & -0.01669 \cdot \Delta D_5(c_0, c_1, c_2, c_3, c_4) \\
 & +0.16235 \cdot \Delta D_6(c_0, c_1, c_2, c_3, c_4)
 \end{aligned}
 \tag{4}$$

$$n = 24127 \quad \chi^2 = 19490.174 \quad p < 0.05$$

We see that The PTML-PCA comparing to the previous model, presents a more complex structure with the double of variables. In fact, the model includes the variables $\Delta D_3(c_0, c_1, c_2, c_3, c_4)$, $\Delta D_4(c_0, c_1, c_2, c_3, c_4)$, $\Delta D_5(c_0, c_1, c_2, c_3, c_4)$, $\Delta D_6(c_0, c_1, c_2, c_3, c_4)$, which refer to the moving average of the descriptors C-004 (CR4), C-012 (CR2X2), H-047 (H attached to C_{sp3}/C_{sp2}) and O-056 (C-OH, alcohol) respectively, for the combination of assay conditions $c_j = (c_0, c_1, c_2, c_3, c_4)$. This model also presents a high χ^2 and low value of p, giving information about the proper statistical significance. The capacity of prediction is depicted in **Table 13**. As it can be observed, this model presents a high accuracy, specificity and sensitivity (range of 87.91%). However, it is slightly lower the sensitivity of the test set than the previous model built with PSA and ALOGP (87.72 vs. 87.67).

Table 13. PTML-PCA-LDA classification matrix

Obs.	Stat.	Pred.	Pred. sets ^a		
Sets	Param. ^a	Stat. ^a	n_j	$f(v_{ij})_{pred} = 0$	$f(v_{ij})_{pred} = 1$
Training					
$f(v_{ij})_{obs} = 0$	Sp	90.39	20067	18139	1928
$f(v_{ij})_{obs} = 1$	Sn	87.43	4060	510	3550
Total	Ac	89.89	24127	18649	5478
Validation					

$f(v_{ij})_{obs} = 0$	Sp	90.58	10020	9077	943
$f(v_{ij})_{obs} = 1$	Sn	87.67	2052	253	1799
Total	Ac	90.09	12072	9330	2742

^a Obs. Sets = Observed sets, Sp = Specificity, Sn = Sensitivity, Stat. Param. = Statistical parameter, Pred. Stat. = Predicted statistics.

The result is lower comparing to the previous model only with ALOGP and PSA. This is given to the quality of these descriptors. Specially, ALOGP is not a bulky descriptor (simple sum of the atoms present) but fragment-based descriptor including information about many different structural patterns.⁸⁴ Thus, these descriptors are the best we found for the given dataset to accumulate information of assay conditions through MA and create a high accuracy PTML model.

■ CONCLUSIONS

In this paper, we showed there is a high number of preclinical assays to discover different biological activities and chemical properties. Computational capacity is able to generate prediction in order to extract knowledge from that database and predict compounds. This is appropriate not for only decreasing costs but also for reducing, replacing and refining animal experimentation. In this sense, on the one hand, PTML method is adequate to model complex ChEMBL datasets characterized by heterogeneous information of the different compounds and assay conditions. Specifically, PTML-LDA is able to predict biological activity and chemical properties of different vitamins and their derivatives in different assay conditions with high accuracy. On the other hand, for dataset taken into account, the PTML-LDA model including MMA presented better results than the model with MA operators. The PTML-LDA model with MMA presented a Sp(%) = 90.7, Sn(%) = 91.3, and overall Ac(%) = 90.8 in training series. The model also have Sp(%) = 90.8, Sn(%) = 91.4, and overall Ac(%) = 90.9 in external validation series. The PTML-LDA with MMA used less variables in the equation than MA model. On the other hand, we applied LR, CT, NB and RF in order to compare to LDA

5) MODELLING VITAMIN DERIVATIVES

performance. PTML-LDA showed the best Sn. In terms of Sp and overall Ac, RF model presented better results but the Sn decreases ten points comparing to PTML-LDA. Thus, PTML-LDA model with MMA and MA is a useful tool for prediction of biological activity and chemical properties of vitamins and their derivatives. In addition, PCA experiment with chemical structure descriptors allowed to characterize the high structural diversity of the chemical space studied. In any case, PTML models using chemical structure descriptors do not improve the performance of the PTML-LDA model based on ALOGP and PSA.

AUTHOR INFORMATION

Corresponding Author

*E-mail: humberto.gonzalezdiaz@ehu.es (H.G.-D.)

*E-mail: ricardo.santana@opendeusto.es (R.S.)

ASSOCIATED CONTENT

Supporting Information

The full lists of the values v_{ij} , cutoff, $f(v_{ij})_{obs}$, $f(v_{ij})_{pred}$, assay conditions, molecular descriptors, MA, MMA, desirability of each activity, and the rest of data mentioned in this research appear in the Supplementary Information file SI01.xlsx

■ ACKNOWLEDGMENTS

R.S.C. thanks COLCIENCIAS scholarship for the doctorate studies; “Convocatoria para Doctorado Nacional 757” from 2017. This original research is part of the project “Investigación en Derecho Internacional y Nanotecnología” registered in the Research Centre of Universidad Pontificia Bolivariana with register number 766B-06/17-37. Special gratitude is extended to CYTED NANOCELIA network. The authors acknowledge research grants from Ministry of Economy and Competitiveness, MINECO, Spain (FEDER CTQ2016-74881-P) and Basque government (IT1045-16). The authors also acknowledge the support of Ikerbasque, Basque Foundation for Science. The authors also acknowledge the support of Ikerbasque, Basque Foundation for Science.

5) MODELLING VITAMIN DERIVATIVES

■ REFERENCES

- (1) Hadden, J.W. & Kyle, M. Structure–Activity Relationship Studies of Vitamin D3 Analogues Containing an Ether or Thioether Linker as Hedgehog Pathway Inhibitors. *Chem. Med. Chem.* **2018**, *13*, 748–753.
- (2) DeBerardinis, A. M., Raccuia, D. S., Thompson, E. N., Maschinot, C. A. & Hadden, M. K. Vitamin D3 Analogues That Contain Modified A-and Seco-B-Rings as Hedgehog Pathway Inhibitors. *Eur. J. Med. Chem.* **2015**, *93*, 156–171.
- (3) Banerjee, U., Deberardinis, A. M.; Hadden, M. Design, Synthesis, and Evaluation of Hybrid Vitamin D3 Side Chain Analogues as Hedgehog Pathway Inhibitors. *Bioorg. Med. Chem.* **2015**, *23* (3), 548–555. <https://doi.org/10.1016/j.bmc.2014.12.005>.
- (4) Maschinot, C. A.; Hadden, M. K. Synthesis and Evaluation of Vitamin D3 Analogues with C-11 Modifications as Inhibitors of Hedgehog Signaling. *Bioorg. Med. Chem. Lett.* **2017**, *27* (17), 4011–4014.
- (5) Maschinot, C. A.; Chau, L. Q.; Wechsler-Reya, R. J.; Hadden, M. K. Synthesis and Evaluation of Third Generation Vitamin D3 Analogues as Inhibitors of Hedgehog Signaling. *Eur. J. Med. Chem.* **2018**, *162*, 495–506. <https://doi.org/10.1016/j.ejmech.2018.11.028>.
- (6) Dasari, S., Ali, SM., Zheng, G., Chen, A., Dontaraju, S., Bosland, MC. & Kajdacsy-Balla, A. Vitamin K and Its Analogs: Potential Avenues for Prostate Cancer Management. *Oncotarget* **2017**, *8* (34), 57782–57799.
- (7) Kimura, K.; Hirota, Y.; Kuwahara, S.; Takeuchi, A.; Tode, C. Synthesis of Novel

- Synthetic Vitamin K Analogues Prepared by Introduction of a Heteroatom and a Phenyl Group That Induce Highly Selective Neuronal Differentiation of Neuronal Progenitor Cells. *J. Med. Chem.* **2017**, 4–9.
- (8) Osakada, F.; Hashino, A.; Kume, T.; Katsuki, H. Alfa-Tocotrienol Provides the Most Potent Neuroprotection among Vitamin E Analogs on Cultured Striatal Neurons. *Neuropharmacology* **2004**, *47*, 904–915. <https://doi.org/10.1016/j.neuropharm.2004.06.029>.
- (9) Numakawa, Y., Numakawa, T., Matsumoto, T., Yagasaki, Y., Kumamaru, E., Kunugi, H., Taguchi, T. & Niki, E. Vitamin E Protected Cultured Cortical Neurons from Oxidative Stress-Induced Cell Death through the Activation of Mitogen- Activated Protein Kinase and Phosphatidylinositol 3-Kinase. *J. Neurochem.* **2006**, *97*, 1191–1202. <https://doi.org/10.1111/j.1471-4159.2006.03827.x>.
- (10) Torquato, P., Ripa, O., Giusepponi, D., Galarini, R., Bartolini, D., Wallert, M., Pellegrino, R., Cruciani, G.; Lorkowski, S., Birringer, M., Mazzini, F. & Galli, F. Analytical Strategies to Assess the Functional Metabolome of Vitamin E. *J. Pharm. Biomed. Anal.* **2016**, *124*, 399–412. <https://doi.org/10.1016/j.jpba.2016.01.056>.
- (11) Shepherd, G.; Velez, L. I. Role of Hydroxocobalamin in Acute Cyanide Poisoning. *Ann. Pharmacother.* **2008**, *42*, 661–669. <https://doi.org/10.1345/aph.1K559>.
- (12) Zelder, F. Recent Trends in the Development of Vitamin B12 Derivatives for Medicinal Applications. Chemical Communications. *Chem. Commun.* **2015**, *51* (74), 14004–14017. <https://doi.org/10.1039/C5CC04843E>.
- (13) Bruyn, T. De; Westen, G. J. P. Van; Ijzerman, A. P.; Stieger, B.; Witte, P. De;

5) MODELLING VITAMIN DERIVATIVES

- Augustijns, P. F.; Annaert, P. P. Structure-Based Identification of OATP1B1/3 Inhibitors. *Mol. Pharmacol.* **2013**, *Mol*, 1–39.
- (14) Guo, M.; Lu, W.; Li, M.; Wang, W. Study on the Binding Interaction between Carnitine Optical Isomer and Bovine Serum Albumin. *Eur. J. Med. Chem.* **2008**, *43* (10), 2140–2148. <https://doi.org/10.1016/j.ejmech.2007.11.006>.
- (15) Calendula, M., Ukiya, M., Akihisa, T., Yasukawa, K., Tokuda, H., Suzuki, T. & Kimura, Y. Anti-Inflammatory, Anti-Tumor-Promoting, and Cytotoxic Activities of Constituents of Marigold (*Calendula Officinalis*) Flowers. *J. Nat. Prod.* **2006**, *69* (12), 1692–1696.
- (16) Rakers, C., Schwerdtfeger, S-M., Mortier, J., Duwe, S., Wolff, T.; Wolber, G., Melzig & Matthias, F. Inhibitory Potency of Flavonoid Derivatives on Influenza Virus Neuraminidase. *Bioorg. Med. Chem. Lett.* **2014**, *24* (17), 4312–4317. <https://doi.org/10.1016/j.bmcl.2014.07.010>.
- (17) Hamama, W. S.; Gouda, M. A.; Badr, M. H.; Zoorob, H. H. Synthesis, Antioxidant, and Antitumor Evaluation of Certain New N-Substituted-2-Amino-1, 3, 4-Thiadiazoles. *Med. Chem. Res.* **2013**, *22* (8), 3556–3565. <https://doi.org/10.1007/s00044-012-0336-z>.
- (18) Guerra, A.; Campillo, N. E.; Pa, J. A. Neural Computational Prediction of Oral Drug Absorption Based on CODES 2D Descriptors. *Eur. J. Med. Chem.* **2010**, *45* (3), 930–940. <https://doi.org/10.1016/j.ejmech.2009.11.034>.
- (19) Kaci, M., Uttaro, J-P., Lefort, V.; Mathé, C., El, C. & Périgaud, C. Synthesis of {[5-(Adenin-9-Yl)-2-Furyl]methoxy}methyl Phosphonic Acid and Evaluations against Human Adenylate Kinases. *Bioorg. Med. Chem. Lett.* **2014**, *24* (17), 4227–4230. <https://doi.org/10.1016/j.bmcl.2014.07.036>.

- (20) Cameron, K.O., Kung, D. W.; Kalgutkar, A.S., Kurumbail, R.G., Miller, R., Salatto, C.T., Ward, J., Withka, J.M., Bhattacharya, S.K., Boehm, M., Borzilleri, A., Brown, J. A.; Calabrese, M., Caspers, N. L.; Cokorinos, E., Conn, L., Dowling, M.S., Edmonds, D. J.; Eng, H., Fernando, D.P., Hepworth, D., Landro, J., Mao, Y., Rajamohan, F., Reyes, A.R., Colin, R., Ryder, T., Shavnya, A., Smith, A. C.; Tu, M., Wolford, A.C. & Xiao, J. Discovery and Preclinical Characterization of 6-Chloro-5-[4-(1-Hydroxycyclobutyl) Phenyl]-1 H-Indole-3-Carboxylic Acid (PF-06409577), a Direct Activator of Adenosine Monophosphate-Activated Protein Kinase (AMPK), for the Potential Treatment of Diabetic Ne. *J. Med. Chem.* **2016**, *59* (17), 8068–8081. <https://doi.org/10.1021/acs.jmedchem.6b00866>.
- (21) Hong, J.A., Bhave, D.P. & Carroll, K. S. Identification of Critical Ligand Binding Determinants in Mycobacterium Tuberculosis Adenosine-50-Phosphosulfate Reductase. *J. Med. Chem.* **2009**, *52*, 5485–5495.
- (22) El-tayeb, A., Iqbal, J., Behrenswerth, A., Romio, M.; Schneider, M., Zimmermann, H. & Christa, E. M. Nucleoside-5'-Monophosphates as Prodrugs of Adenosine A2A Receptor Agonists Activated by Ecto-5'-Nucleotidase. *J. Med. Chem.* **2009**, *52* (23), 7669–7677. <https://doi.org/10.1021/jm900538v>.
- (23) Amiable, C., Paoletti, J., Haouz, A., Padilla, A., Labesse, G., Kaminski, P. A., & Pochet, S. 6-(Hetero)Arylpurine Nucleotides as Inhibitors of the Oncogenic Target DNPH1: Synthesis, Structural Studies and Cytotoxic Activities Claire. *Eur. J. Med. Chem.* **2014**, *85*, 418–437. <https://doi.org/10.1016/j.ejmech.2014.07.110>.
- (24) Baldisserotto, A., Vertuani, S.; Bino, A., Lucia, D., Milani, R., Gambari, R. & Manfredini, S. Design , Synthesis and Biological Activity of a Novel Rutin Analogue

5) MODELLING VITAMIN DERIVATIVES

- with Improved Lipid Soluble Properties. *Bioorg. Med. Chem.* **2014**, *23* (1), 264–271. <https://doi.org/10.1016/j.bmc.2014.10.023>.
- (25) Kamogawa, E.; Sueishi, Y. A Multiple Free-Radical Scavenging (MULTIS) Study on the Antioxidant Capacity of a Neuroprotective Drug, Edaravone as Compared with Uric Acid, Glutathione, and Trolox. *Bioorg. Med. Chem. Lett.* **2014**, *24* (5), 1376–1379. <https://doi.org/10.1016/j.bmcl.2014.01.045>.
- (26) Bhatia, G. Khanna, A.K., Sonkar, R., Mishra, S.K., Srivastava, S.; Lakshmi, V. Lipid Lowering and Antioxidant Activity of Flavones in Triton Treated Hyperlipidemic Rats. *Med. Chem. Res.* **2011**, *20* (9), 1622–1626. <https://doi.org/10.1007/s00044-010-9444-9>.
- (27) Pandurangan, N.; Bose, C.; Banerji, A. Synthesis and Antioxygenic Activities of Seabuckthorn Flavone-3-Ols and Analogs. *Bioorg. Med. Chem. Lett.* **2011**, *21* (18), 5328–5330. <https://doi.org/10.1016/j.bmcl.2011.07.008>.
- (28) Zhang, F., Yang, Y-N., Song, X-Y., Shao, S-Y., Feng, Zi-M., Jiang, J-S. & Li, L. Forsythoneosides A – D, Neuroprotective Phenethanoid and Flavone Glycoside Heterodimers from the Fruits of Forsythia Suspensa. *J. Nat. Prod.* **2015**, *10*, 4–11. <https://doi.org/10.1021/acs.jnatprod.5b00372>.
- (29) Stark, T.D., Salger, M., Frank, O.; Balemba, O.B., Wakamatsu, J. & Hofmann, T. Antioxidative Compounds from Garcinia Buchananii Stem Bark. *J. Nat. Prod.* **2014**, *78* (2), 234–240. <https://doi.org/10.1021/np5007873>.
- (30) Xu, G-B., He, G., Bai, H-H., Yang, T., Zhang, G-L., Wu, Li-W. & Li, G.-Y. Indole Alkaloids from Chaetomium Globosum. *J. Nat. Prod.* **2014**, *78* (7), 1479–1485. <https://doi.org/10.1021/np5007235>.

- (31) Ramachandrappa, K., Renuka, N., Kameshwar, V.H., Srinivasan, B., Ajay, K. & Shashikanth, S. Synthesis of Lignan Conjugates via Cyclopropanation : Antimicrobial and Antioxidant Studies. *Bioorg. Med. Chem. Lett.* **2016**, 26 (15), 3621–3625. <https://doi.org/10.1016/j.bmcl.2016.06.005>.
- (32) Hofmann, E., Webster, J., Kidd, T., Kline, R., Jayasinghe, M., & Paula, S. Hydroxylated Chalcones with Dual Properties: Xanthine Oxidase Inhibitors and Radical Scavengers. *Int. J. Biosci. Biochem. Bioinforma.* **2014**, 4 (4), 234. <https://doi.org/10.1016/j.bmc.2015.12.024>.
- (33) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Motow, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2016**, 45 (1), 945–954.
- (34) Zimmermann, G. R.; Lehar, J.; Keith, C. T. Multi-Target Therapeutics: When the Whole Is Greater than the Sum of the Parts. *Drug Discov. Today* **2007**, 12 (1–2), 34–42.
- (35) Hopkins, A. L. Network Pharmacology: The next Paradigm in Drug Discovery. *Nat. Chem. Biol.* **2008**, 4 (11), 682.
- (36) Jia, J., Zhu, F., Ma, X., Cao, Z-W., Li, Y-X. & Chen, Y.-Z. Mechanisms of Drug Combinations: Interaction and Network Perspectives. *Nat. Rev. Drug Discov.* **2009**, 8 (2), 111.
- (37) Pujol, A.; Mosca, R.; Farrés, J.; Aloy, P. Unveiling the Role of Network and Systems Biology in Drug Discovery. *Trends Pharmacol. Sci.* **2010**, 31 (3), 115–123.
- (38) Zheng, W., Zhao, Y., Luo, Q.; Zhang, Y., Wu, K. & Wang, F. Multi-Targeted Anticancer Agents. *Curr. Top. Med. Chem.* **2017**, 17 (28), 3084–3098.

5) MODELLING VITAMIN DERIVATIVES

- (39) Ramsay, R. R.; Popovic-Nikolic, M. R.; Nikolic, K.; Uliassi, E.; Bolognesi, M. L. A Perspective on Multi-Target Drug Discovery and Design for Complex Diseases. *Clin. Transl. Med.* **2018**, *7* (1), 3.
- (40) Flecknell, P. Replacement, Reduction and Refinement. *Altex- Altern. to Anim. Exp.* **2002**, *19* (2), 73–78.
- (41) Wu, X.; Zhu, X.; Member, S. Data Mining with Big Data. *IEEE Trans. Knowl. Data Eng.* **2014**, *26* (1), 97–107. <https://doi.org/10.1109/TKDE.2013.109>.
- (42) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2011**, *40* (1), 1100–1107.
- (43) Simón-Vidal, L.; García-Calvo, O.; Oteo, U.; Arrasate, S.; Lete, E.; Sotomayor, N.; González-Díaz, H. Perturbation-Theory and Machine Learning (PTML) Model for High-Throughput Screening of Parham Reactions: Experimental and Theoretical Studies. *J. Chem. Inf. Model.* **2018**, *58* (7), 1384–1396. <https://doi.org/10.1021/acs.jcim.8b00286>.
- (44) Blázquez-Barbadillo, C., Aranzamendi, E., Coya, E., Lete, E., Sotomayor, N. & González-Díaz, H. Perturbation Theory Model of Reactivity and Enantioselectivity of Palladium-Catalyzed Heck-Heck Cascade Reactions. *RSC Adv.* **2016**, *6* (45), 38602–38610. <https://doi.org/10.1039/c6ra08751e>.
- (45) Speck-Planche, A., Kleandrova, V.V., Luan, F., González-Díaz, H., Ruso, J-M. & Cordeiro, M. N. Computational Tool for Risk Assessment of Nanomaterials: Novel QSTR-Perturbation Model for Simultaneous Prediction of Ecotoxicity and Cytotoxicity

- of Uncoated and Coated Nanoparticles under Multiple Experimental Conditions. *Environ. Sci. Technol.* **2014**, *48* (24), 14686–14694. <https://doi.org/10.1021/es503861x>.
- (46) Luan, F.; Kleandrova, V. V.; González-Díaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, N. Computer-Aided Nanotoxicology: Assessing Cytotoxicity of Nanoparticles under Diverse Experimental Conditions by Using a Novel QSTR-Perturbation Approach. *Nanoscale* **2014**, *6* (18), 10623–10630. <https://doi.org/10.1039/c4nr01285b>.
- (47) Blay, V.; Yokoi, T.; González-Díaz, H. Perturbation Theory-Machine Learning Study of Zeolite Materials Desilication. *J. Chem. Inf. Model.* **2018**, *58* (12), 2414–2419.
- (48) Arrasate, S.; Duardo-Sanchez, A. Perturbation Theory Machine Learning Models: Theory, Regulatory Issues, and Applications to Organic Synthesis, Medicinal Chemistry, Protein Research, and Technology. *Curr. Top. Med. Chem.* **2018**, *18* (14), 1203–1213.
- (49) González-Díaz, H.; Arrasate, S.; Gómez-San Juan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. General Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry. *Curr. Top. Med. Chem.* **2013**, *13* (14), 1713–1741. <https://doi.org/10.2174/1568026611313140011>.
- (50) Da Costa, J. F.; Silva, D.; Caamaño, O.; Brea, J. M.; Loza, M. I.; Munteanu, C. R.; Pazo, A.; García-Mera, X.; González-Díaz, H. Perturbation Theory/Machine Learning Model of ChEMBL Data for Dopamine Targets: Docking, Synthesis, and Assay of New L-Prolyl-L-Leucyl-Glycinamide Peptidomimetics. *ACS Chem Neurosci* **2018**, *9* (11), 2572–2587. <https://doi.org/10.1021/acschemneuro.8b00083>.

5) MODELLING VITAMIN DERIVATIVES

- (51) Kleandrova, V. V.; Luan, F.; González-Díaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, N. M. Computational Ecotoxicology: Simultaneous Prediction of Ecotoxic Effects of Nanoparticles under Different Experimental Conditions. *Environ. Int.* **2014**, *73*, 288–294. <https://doi.org/10.1016/j.envint.2014.08.009>.
- (52) Da Costa, J. F.; Silva, D.; Caamaño, O.; Brea, J. M.; Loza, M. I.; Munteanu, C. R.; Pazos, A.; García-Mera, X.; González-Díaz, H. PTML Model of ChEMBL Data for Dopamine Targets, Docking, Synthesis, and Assay of New PLG Peptidomimetics. *ACS Chem. Neurosci.* **2018**, *9* (11), 2572–2587.
- (53) Liu, Y.; Tang, S.; Fernandez-Lozano, C.; Munteanu, C. R.; Pazos, A.; Yu, Y.Z.; Tan, Z.; González-Díaz, H. Experimental Study and Random Forest Prediction Model of Microbiome Cell Surface Hydrophobicity. *Expert Syst. Appl.* **2017**, *72*, 306–316. <https://doi.org/10.1016/j.eswa.2016.10.058>.
- (54) Martínez-Arzate, S., Tenorio-Borroto, E., Barbabosa Pliego, A., Díaz-Albiter, H.M., Vázquez-Chagoyán, J.C. & González-Díaz, H. PTML Model for Proteome Mining of B-Cell Epitopes and Theoretical-Experimental Study of Bm86 Protein Sequences from Colima, Mexico. *J. Proteome Res.* **2017**, *16* (11), 4093–4103. <https://doi.org/10.1021/acs.jproteome.7b00477>.
- (55) Bediaga, H.; Arrasate, S.; González-Díaz, H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Comb. Sci.* **2018**, *20* (11), 621–632. <https://doi.org/10.1021/acscombsci.8b00090>.
- (56) Martínez-Arzate, S.G., Tenorio-Borroto, E., Barbabosa Pliego, A., Diaz-Albiter, H.M., Vázquez-Chagoyán, J.C. & González-Díaz, H. PTML Model for Proteome Mining of

- B-Cell Epitopes and Theoretical--Experimental Study of Bm86 Protein Sequences from Colima, Mexico. *J. Proteome Res.* **2017**, *16* (11), 4093–4103.
- (57) Balakrishnama, S.; Ganapathiraju, A. Linear Discriminant Analysis-a Brief Tutorial. *Inst. Signal Inf. Process.* **1998**, *18*, 1–8.
- (58) R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria 2017.
- (59) Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. *Introduction to Logistic Regression Model*; John Wiley & Sons, 2014; Vol. 398.
- (60) Loh, W. Classification and Regression Trees. *WIREs DataMining Knowl Discov* **2011**, *1*, 14–23. <https://doi.org/10.1002/widm.8>.
- (61) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.
- (62) Ng, A., Jordan, M. On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes. *Adv. Neural Inf. Process. Syst.* **2002**, 841–848.
- (63) Hill, T., Lewicki, P., & Lewicki, P. *Statistics: Methods and Applications: A Comprehensive Reference for Science, Industry, and Data Mining.*; StatSoft, Inc., 2006.
- (64) Marrero-Ponce, Y., Siverio-Mota, D., Gálvez-Llompart, M.; Recio, M.C., Giner, R.M., García-Domènech, R., Torrens, F., Arán, V.J., Cordero-Maldonado, M.; Esguera, C.V., et al. Discovery of Novel Anti-Inflammatory Drug-like Compounds by Aligning in Silico and in Vivo Screening: The Nitroindazolinone Chemotype. *Eur. J. Med. Chem.* **2011**, *46* (12), 5736–5753.
- (65) García, I.; Fall, Y.; García-Mera, X.; Prado-Prado, F. Theoretical Study of GSK- 3alpha:

5) MODELLING VITAMIN DERIVATIVES

- Neural Networks QSAR Studies for the Design of New Inhibitors Using 2D Descriptors. *Mol. Divers.* **2011**, *15* (4), 947–955.
- (66) Vásquez-Domínguez, E.; Armijos-Jaramillo, V. D.; Tejera, E.; González-Díaz, H. Multioutput Perturbation-Theory Machine Learning (PTML) Model of ChEMBL Data for Antiretroviral Compounds. *Mol. Pharm.* **2019**, *16*, 4200–4212. <https://doi.org/10.1021/acs.molpharmaceut.9b00538>.
- (67) González-Díaz, H.; Herrera-Ibatá, D. M.; Duardo-Sánchez, A.; Munteanu, C. R.; Orbegozo-Medina, R. A.; Pazos, A. ANN Multiscale Model of Anti-HIV Drugs Activity vs AIDS Prevalence in the US at County Level Based on Information Indices of Molecular Graphs and Social Networks. *J. Chem. Inf. Model.* **2014**, *54* (3), 744–755. <https://doi.org/10.1021/ci400716y>.
- (68) Gentleman, R. *R Programming for Bioinformatics*; Chapman and Hall/CRC: New York, 2008.
- (69) Kotsiantis, S. B. Supervised Machine Learning : A Review of Classification Techniques. *Informatica* **2007**, *31*, 249–268.
- (70) Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. *Data Mining: Practical Machine Learning Tools and Techniques.*; Morgan Kaufmann, 2016.
- (71) Varnek, A.; Baskin, I. Machine Learning Methods for Property Prediction in Chemoinformatics : Quo Vadis ? *J. Chem. Inf. Model.* **2011**, *52* (6), 1413–1437.
- (72) Chen, B.; Sheridan, R. P.; Hornak, V.; Voigt, J. H. Bayes in Comparison of Random Forest and Pipeline Pilot Naive Prospective QSAR Predictions. **2012**, *52* (3), 792–803.

- (73) Svetnik, V., Liaw, A., Tong, C.; Culberson, J.C., Sheridan, R. P.; Feuston, B. P. Random Forest : A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947–1958.
- (74) Schierz, A. C. Virtual Screening of Bioassay Data. *J. Cheminform.* **2009**, *12*, 1–12. <https://doi.org/10.1186/1758-2946-1-21>.
- (75) Marill, K. A. Advanced Statistics : Linear Regression , Part II : Multiple Linear Regression. *Acad. Emerg. Med.* **2004**, *11* (1), 94–102. [https://doi.org/10.1197/S1069-6563\(03\)00601-8](https://doi.org/10.1197/S1069-6563(03)00601-8).
- (76) Willett, P., Wilton, D.J., Bank, W., Sheffield, S., Acklin, P., Azzaoui, K., Jacoby, E. & Schuffenhauer, A. New Methods for Ligand-Based Virtual Screening : Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model.* **2006**, *46*, 462–470.
- (77) Lo, Y.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* **2018**, *0* (0), 1–9. <https://doi.org/10.1016/j.drudis.2018.05.010>.
- (78) Mathea, M.; Klingspohn, W.; Baumann, K. Chemoinformatic Classification Methods and Their Applicability Domain. *Mol. Inform.* **2016**, *35*, 160–180. <https://doi.org/10.1002/minf.201501019>.
- (79) Efron, B., & Tibshirani, R. J. *An Introduction to the Bootstrap*; CRC press, 1994.
- (80) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Mark, T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Carol, A.; et al. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure – Activity Relationships.

5) MODELLING VITAMIN DERIVATIVES

Atla **2005**, 2, 155–173.

- (81) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, 22 (1), 69–77.
- (82) R., T.; Consonni, V.; Mauri, V.; Pavan, M. DRAGON Professional Version, 2005.
- (83) Jolliffe, I. T. (1989). Rotation of Iii-Defined Principal Components. *J. R. Stat. Soc. Ser. C (Applied Stat.* **1989**, 38 (1), 139–147.
- (84) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Model.* **1999**, 39, 868–873.

*You cannot teach a man anything; you
can only help him discover it in
himself.*

Galileo Galilei

CHAPTER

6

6) Modelling systems of metal oxide nanoparticles and vitamin derivatives

The prediction of the behavior of nanosystems is a relevant information for saving time, costs and reduce the experimentation with animals. In this case, the metal oxide nanoparticles have been explored to design new nanosystems. If we are able to better desing these nanosystems, we would do significant steps in material science knowledgment.

To do so, we develop a model able to predict a multi output and multi input model able to predict biological activities of the components of nanosystems conformed by metal oxide nanoparticles (with or without coating agents) and

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

vitamin derivatives. We apply the PTML methodology by following the workflow included in Figure 7.

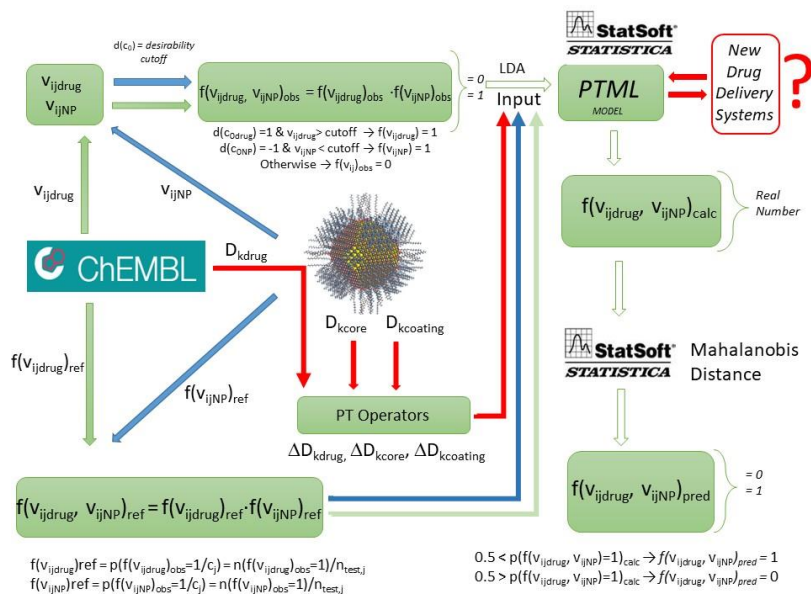


Figure 7. PTML data pre-processing and processing workflow proposed in this work

Predicting Coated-Nanoparticle Drugs Release Systems with Perturbation-Theory Machine Learning (PTML) Models

Ricardo Santana^{*a,b}, Robin Zuluaga^c, Piedad Gañán^b, Sonia Arrasate^c,
Enrique Onieva^a, and Humbert González-Díaz^{*,e,f,g}

^aUniversity of Deusto, Avda. Universidades, 24, 48007 Bilbao, Spain.

^bGrupo de Investigación Sobre Nuevos Materiales, Facultad de Ingeniería Química, Universidad Pontificia Bolivariana, Circular 1° N° 70-01, Medellín, Colombia.

^dFacultad de Ingeniería Agroindustrial, Universidad Pontificia Bolivariana, Circular 1° N° 70-01, Medellín, Colombia.

^eDepartment of Organic Chemistry II, University of Basque Country UPV/EHU, 48940, Leioa, Spain.

^fIKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain.

^gBiofisika Institute CSIC-UPVEHU, University of Basque Country UPV/EHU, 48940, Leioa, Spain

ABSTRACT. Nanoparticles (NPs), decorated with coating agents (polymers, gels, proteins, *etc.*), form Nanoparticle Drug Delivery Systems (DDNS) of high interest in Nanotechnology and Biomaterials science. There is an increasing publication of experimental data sets of biological activity, toxicity, and delivery properties of DDNS. However, these data sets are still disperse and no as large as the datasets of DDNS components (NP and drugs). This prompts researchers train Machine Learning (ML) algorithms able to design new DDNS based on the properties of their components. However, most ML models reported up to date predict specific activities of NP or drugs over determined target or cell line. In this paper, we combine Perturbation Theory and Machine Learning (PTML algorithm) to train a model able to predicting the best components (NP, coating agent, and drug) for DDNS design. In so doing, we downloaded from ChEMBL a dataset of >30000 preclinical assays of drugs. We also downloaded from public sources a NPs data set formed by preclinical assays of coated Metal Oxide Nanoparticles (MONPs). Both, drugs and NPs datasets of preclinical assays cover multiple conditions of assay that can be listed as two arrays \mathbf{c}_{drug} and \mathbf{c}_{NP} , respectively. The

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

$c_{j\text{drug}}$ array includes >504 biological activity parameters ($c_{0\text{drug}}$), >340 target proteins ($c_{1\text{drug}}$), >650 types of cells ($c_{2\text{drug}}$), >120 assay organisms ($c_{3\text{drug}}$), > 60 assay strains ($c_{4\text{drug}}$). On the other side, the $c_{j\text{NP}}$ array includes 3 biological activity parameters ($c_{0\text{NP}}$), 40 types of proteins ($c_{1\text{NP}}$), 10 shapes of nanoparticles ($c_{2\text{NP}}$), 6 assay media ($c_{3\text{NP}}$), and 12 coating agents ($c_{4\text{NP}}$). After downloading, we pre-processed both data sets by separate calculating PT operators able to account for changes (perturbations) in drug, coating agents, and NP chemical structure and/or physicochemical properties as well as for assay conditions. Next, we carry out an information fusion process forming a final dataset of above 500000 DDNS (drug + MONP pairs). We also trained other linear and non-linear PTML models using R studio scripts for comparative purposes. Until the best of our knowledge, this is the first multi-label PTML model useful to select drugs, coating agents, and metal or metal-oxide nanoparticles to be assembled in order to design new DDNS with optimal activity/toxicity profiles.

Keywords: ChEMBL; Nanoparticle; Drug Release; Machine Learning; Big data; Multi-output models.

■ INTRODUCTION

Nanoparticles (NP), decorated with coating agents (polymers, gels, proteins, etc.) form nano-systems and/or biomaterials with desirable properties in nanotechnology and biomaterials research. This can be especially useful in terms of improving the capacity releasing determined drugs with Drug Delivery Nanoparticles (DDNS).¹⁻³ In **Table 1**, we have listed some studies of this increasing research area in recent dates.⁴⁻⁹ We must say that this is not a state-of-art review but a table where we highlight the diversity of the data. More specifically, there is a strong attention of researchers about the potential use of Metal Oxide NPs (MONPs) as drug carrier in DDNS. Specially, the performance as anticancer drug carriers has been explored with adequate results in terms of cytotoxicity and drug release. Other nanosystems without MONPs, such as poly(lactic-co-glycolic acid) nanoparticles (PLGANPs), have been designed for cancer co-therapy purpose. There are multiple combinations of core NP, coating agent, drugs and other compounds to be tested if we want to design new to design new DDNS. Furthermore, as we can see, these researches are characterized by a heterogeneous data between them, in terms of the composition of the materials, the source, the target cell and the biological activity among

other assay conditions. There are different applications so desirable biological activity can be different and the exposition may differ, which gives us a more complex context.

Table 1. Nanoparticle and Nanoparticle-Drug systems experimental and computational studies

Experimental Researches						
Author	Meth. ^a	Syst. ^b	Appl. ^c	Drug ^d	Output ^e	Ref.
Farboudi <i>et al.</i>	TEM, XRD, MTT, etc.	MOF	DD.	DOX/FA	Cytotox.	4
Vlassi <i>et al.</i>	LS	PEO Fe ₂ O ₃	DD.	IND	Size and mass	5
Zheng <i>et al.</i>	TEM, SEM, XRD, MTT, etc.	MONP ZnO	DD.	DOX	Cytotox.	6
Yan <i>et al.</i>	TEM, XRD, NMR	MSNPs	DD.	CUR	Release	7
Yin <i>et al.</i>	SEM, TEM, XRD, etc.	MONP TiO ₂	DD.	DOX/HA/Hyal	Cytotox., Release	8
Zhu, <i>et al.</i>	DLS, LDA, MTT, etc.	PLGANPs	DD.	DOX/TPGS	Cytotox., Release	9
Computational Researches						
Author	Meth. ^a	Syst. ^b	Appl. ^c	Drug. ^d	Output ^e	Ref.
Eunkeu <i>et al.</i>	RF	Cd-QD	Med.	-	Cell viability IC ₅₀	10
Novoselska <i>et al.</i>	RF	MONPs	Med.	-	EC ₅₀ , LC ₅₀	11
Toropova <i>et al.</i>	MC	MONPs	Med.	-	pLC ₅₀	12

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

Pathakoti <i>et al.</i>	LR	MONPs	Med.	-	LC ₅₀	13
Singh <i>et al.</i>	GBBA	MONPs	Med.	-	EC ₅₀	14
Fjodorova <i>et al.</i>	CP ANN	MONPs.	Med.	-	EC ₅₀	15
Mikolajczyk <i>et al.</i>	MLR/GA	MONPs.	Med.	-	Zeta Potential	16
Luan <i>et al.</i>	LDA	MONPs	Med.	-	Multiple	17
Kleandrova <i>et al.</i>	LDA	MONPs	Med.	-	Multiple	18
Santana <i>et al.</i>	LDA	nMONPs	DD.	Multiple	Multiple	19
Santana <i>et al.</i>	Multiple	MONPs	DD.	Multiple	Multiple	This work

^a Method = Meth., LS = Light Scattering, RF = Random Forest, MC = Monte Carlo, LR = Linear Regression, GBBA = Gradient Boosting and Bagging Algorithms, MLR = Multiple Linear Regression, GA = Genetic Algorithm, CPANN = Counter Propagation Artificial Neural Network. ^b Syst. = System, PEO/Fe₂O₃ = poly(ethylene oxide-b-phenyl oxazoline) and poly(isoprene-b-ethylene oxide) (PEO-b-PPhOx and PI-b-PEO), MONP ZnO = ZnO-DOX@ZIF-8 with encapsulated iron oxide nanoparticles (γ -Fe₂O₃), Cd-QD = Cadmium Quantum Dots, MONPs = Metal Oxide Nano-Particles, nMONPs = non MONPs, MSNPs = Functional mesoporous silica nanoparticles, PLGANPs = poly(lactic-co-glycolic acid) nanoparticles, TPGS = Vitamin E TPGS, HA = Hyaluronic acid, Hyal = Hyaluronidase, MONP TiO₂ = upconverting nanoparticles with a mesoporous TiO₂, ^c Appl. = Application, Med. = Medicine, DD. = Drug Delivery. ^d DOX = Doxorubicin, FA = Folic Acid, IND = Indomethacin, CUR = Curcumin.

Despite the increasing report of data about DDNS based on coated NPs, there is still a high necessity of useful methods to measure/predict the biological activity and toxicity of the NPs.¹⁰ ML has been applied to extract knowledge of toxicity nanomaterials,¹¹ specifically of MONPs: **Table 1** includes examples of these researches, showing authors, system, application, activity and output. For instance, Eunkeu *et al.*¹² who were able to research cellular toxicity of

cadmium-containing semiconductor quantum dots. In addition, Novoselska *et al.*¹³ applied ML techniques for MONPs toxicity prediction towards *Escherichia coli* and HaCaT cells. For instance, Toropova *et al.*¹⁴ presented a model that is able to predict dark cytotoxicity and photo-induced cytotoxicity of metal oxide nanoparticles to bacteria *Escherichia coli*, among other researches that discover information of the properties of this type of nanomaterials. Pathakoti, *et al.*¹⁵ also developed a model able to predict LC₅₀ for MONPs with high accuracy. Singh *et al.*¹⁶ by applying gradient boosting and bagging algorithms were able to generate a model to predict EC₅₀ with more than 93% of accuracy for validation test. Other models have been built by applying advanced algorithms, such as Counter Propagation Artificial Neural Network (CPANN) or Genetic Algorithms, to improve the performance.^{17,18} On the other hand, in previous works we have found ML models taking into account multiple biological activities of nanomaterials, by applying different algorithms such as Linear Discriminant Analysis (LDA). However, they do not include information about the drug we should include in the DDNS. Santana *et al.* developed a model that includes information of the drug and the vitamin derivatives of the nanosystems. This is the first general purpose model for multiple biological activities of DDNS of nMONPs (non MONPs) and drugs. This model takes into account multiple descriptors of the nMONPs and the drug structure, as well as multiple external conditions of each assay. Nevertheless, this previous model has not been built with MONPs, which is the reason of the present research.^{19–21}

In general, ML methods as used in Cheminformatics take into consideration the structure of the drug but do not take into consideration at the same time other factors mentioned above (conditions of assay). In this context, Perturbation Theory (PT) ideas were introduced to ML techniques coining the term PTML models as a solution to this type of problem. The general idea is to predict the score function $f(v_{ij})_{\text{calc}}$ for multiple properties of the system under study starting from function of reference $f(v_{ij})_{\text{ref}}$ for a group of systems of reference and adding PT Operators to measure the effect of variations in all assay conditions of the system (perturbations) with refer to the group of reference. The PTML models have been used in different disciplines to predict the biological activity of drugs, proteins, materials, and NPs.^{22–25}

However, until the best of our knowledge, there are no reports of general purpose PTML for the design of new DDNS for MONPs. That is why; in this work, we developed the first PTML model for DDNS of MONPs and drugs. In so doing, we have followed a general workflow for

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

PTML models modified for DDNS dataset; see **Figure 1**, going from the base to the top of the pyramid. Firstly, we use a dataset for drugs extracted from ChEMBL with respective molecular descriptors and assay conditions. Besides, we used a MONPs dataset with different descriptors and assay conditions obtained from public literature. Then, we applied information fusion and preprocessing techniques to create a working data set. After that, we trained and validated a linear PTML model. Once we have the PTML model by using STATISTICA software and R language programming, we are able to predict the best compounds to integrate new DDNs. In **Figure 1**, a general scheme applied in this study to construct the PTML model is included.

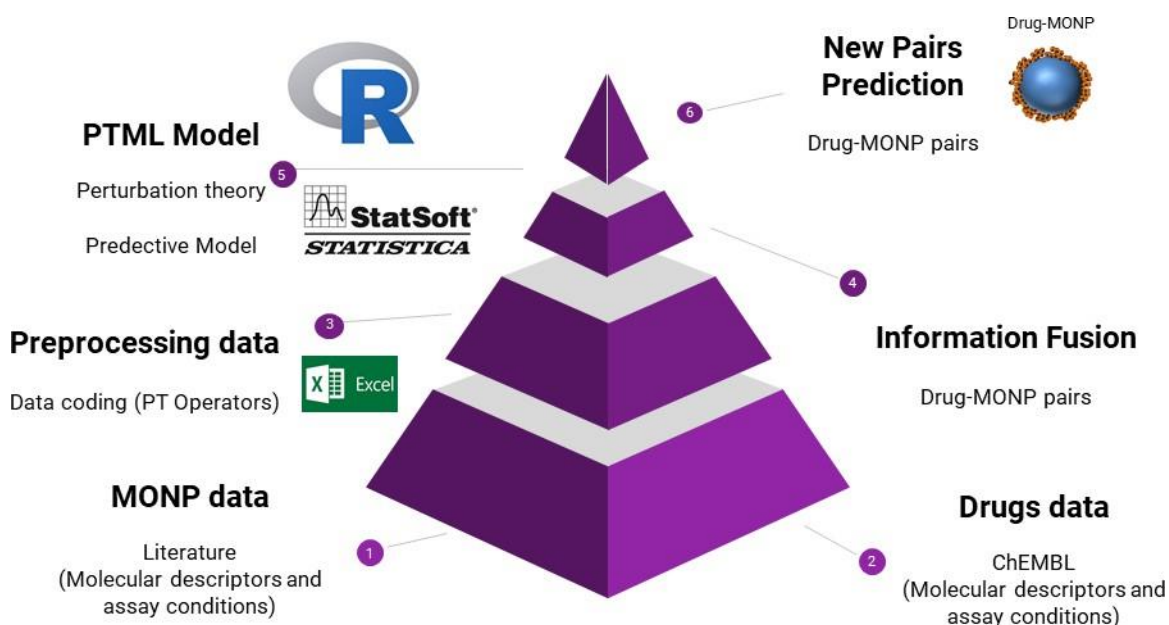


Figure 1. Workflow of the PTML model of DDNS for MONPs and drugs.

■ MATERIALS AND METHODS

Drug data pre-processing. In **Figure 2**, we show the overall data pre-processing and ML data processing workflow proposed in this work. Regarding the drug pre-clinical assays the data points are obtained from preclinical assays registered in ChEMBL free database (February, 2019). Considered drugs are vitamins derivatives or drugs with >70% of similarity of structure. The degree of similarity between the query and the target structures is calculated according to Tanimoto Coefficient.²⁶ Each pre-clinical assay includes a result of the value v_{ijvit} of the

biological activity that the i^{th} drug presents over the j^{th} target. Specifically, $v_{ij\text{drug}}$ varies depending on the structure of each drug and the combination of the assay conditions $c_{j\text{drug}} = (c_{0\text{drug}}, c_{1\text{drug}}, c_{2\text{drug}}, \dots, c_{n\text{vit}})$. Drug assay conditions, $c_{j\text{drug}}$, are $c_{0\text{drug}} =$ the biological activity $v_{ij\text{drug}}$, $c_{1\text{drug}} =$ organism of assay, $c_{2\text{drug}} =$ target protein, *etc* (see **Table 2**). In order to create the PTML model, we discretized $v_{j\text{drug}}$ ²³ as follow: $f(v_{ij\text{drug}})_{\text{obs}} = 1$ if $v_{ij\text{drug}} >$ cutoff and desirability of the biological activity parameter $d(c_{0\text{drug}}) = 1$ (see **Table 5**). The value is also $f(v_{ij\text{drug}})_{\text{obs}} = 1$ when $v_{ij\text{drug}} <$ cutoff and desirability $d(c_{0\text{drug}}) = -1$; otherwise, $f(v_{ij\text{drug}})_{\text{obs}} = 0$. The desirability $d(c_{0\text{drug}}) = 1$ points out for biological activity parameter, if there is a desired effect when it increases. $f(v_{ij\text{drug}})_{\text{obs}} = 1$ means a desirable effect of the drug over the determined target of the bio-assay. Otherwise, if $d(c_{0\text{drug}}) = -1$ it decreases such effect. The cutoff takes the values of 100 for properties with units in nM. If not, $\text{cutoff} = \langle v_{ij\text{drug}} \rangle$; which is the average as expected value (see **Table 5**). The molecular descriptor considered for drugs was $D_{1\text{drug}} = \text{ALOGP}$ (n -Octanol/Water Partition Coefficient).

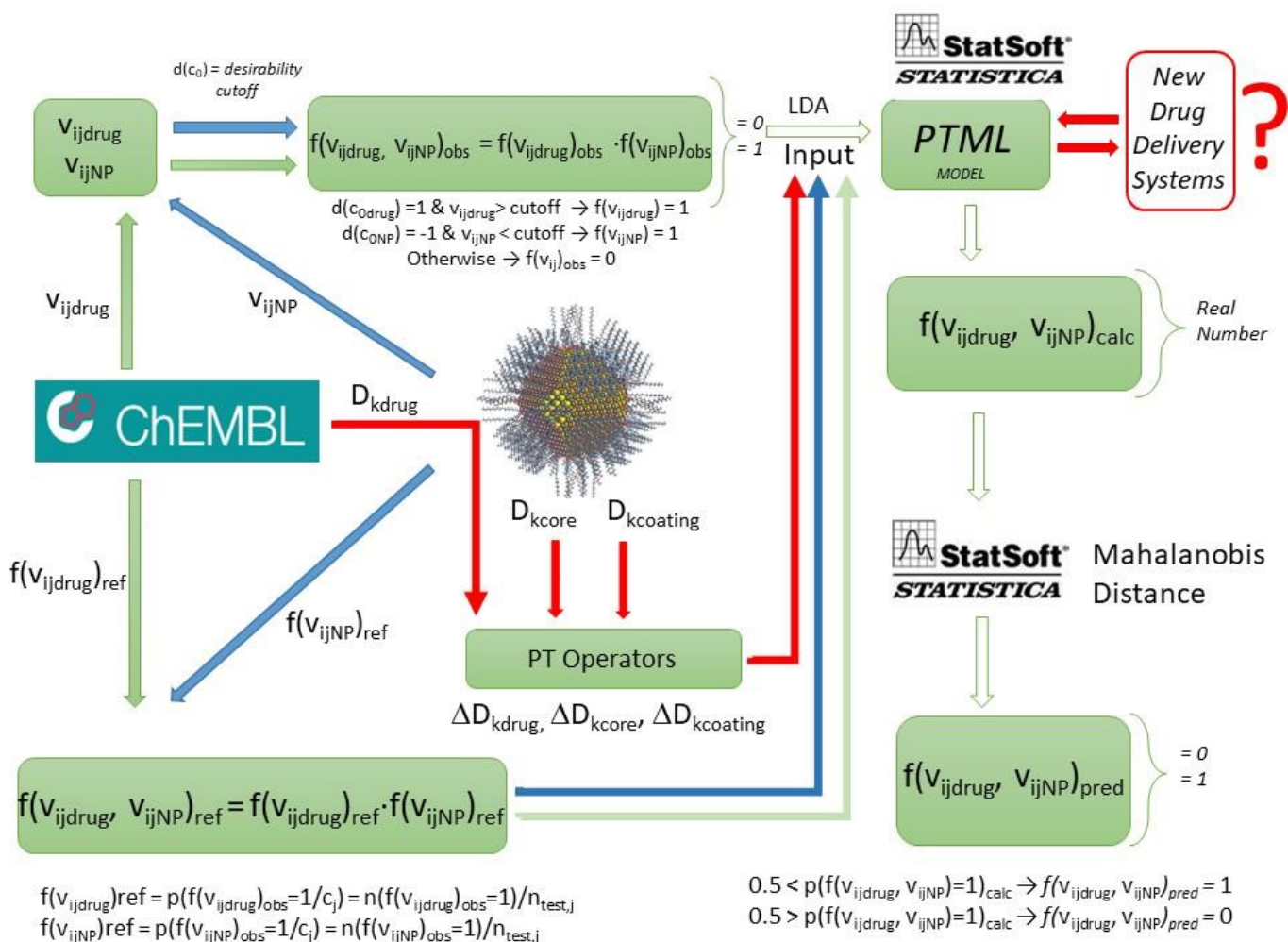


Figure 2. PTML data pre-processing and processing workflow proposed in this work.

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

MONPs data pre-processing. The data for MONPs linked to coating agent assays was obtained from literature.²⁰ In **Figure 2**, we show the overall data processing workflow including also MONPs data. As mentioned, in this work the only type of NP we use are MONPs. As in the case of drugs, each preclinical assay includes a result of the value v_{ijNP} of the biological activity that the i^{th} NP presents used over the j^{th} target. NPs assay conditions c_{jNP} are c_{0NP} = the biological activity v_{ijNP} , c_{1NP} = cell line, c_{2NP} = shape, *etc* (see **Table 3**). By analogy, we also proceeded with the discretization of the v_{ijNP} ²³ as follow: $f(v_{ijNP})_{\text{obs}} = 1$ if $v_{ijNP} > \text{cutoff}$ and biological activity parameter is desirable $d(c_{0NP}) = 1$ (see **Table 6**). If $f(v_{ijNP})_{\text{obs}} = 1$ means that there is a desirable effect of the NP in the nano-toxicity assay. The value for NPs is also $f(v_{ijNP})_{\text{obs}} = 1$ when $v_{ijNP} < \text{cutoff}$ and desirability $d(c_{0NP}) = -1$. Otherwise, $f(v_{ijNP})_{\text{obs}} = 0$. The desirability $d(c_{0NP}) = 1$ indicates that the particular toxicity parameter quantified increases with a non-toxicological effect; otherwise, $d(c_{0NP}) = -1$. Besides, the cutoff = 100 for parameters presented in nM. Otherwise, cutoff = $\langle v_{ijNP} \rangle$; which is the average as expected value, see **Table 6**. Given that the database includes also coated NPs, there are descriptors for the core of the NP D_{icore} and the coating agent D_{icoating} . Thus, the molecular descriptors taken into account for core NPs were: $D_{1\text{core}}$ = Core Monomer Units and $D_{2\text{core}}$ = Core Electronegativity. Regarding the coating agent descriptors, DRAGON software was utilized to calculate them. The model included the following coating agent descriptors: $D_{1\text{coating}}$ = Coating Agent Ghose Crippen Average Molar Refractivity,²⁷ $D_{2\text{coating}}$ = Coating Agent Unsaturation Count, $D_{3\text{coating}}$ = Coating Agent Surface Area of Donor Atoms and $D_{4\text{coating}}$ = Coating Agent Total Surface Area,²⁸ see **Table 3**. $D_{3\text{coating}}$ and $D_{4\text{coating}}$ are calculated by using an estimation of the amount of each atom Van de Waals surface area.²⁹

MONPs-Drug Information Fusion. For this model, an information fusion of the results of NPs tests and drugs tests was carried out, with different conditions for each set. A sample of 500000 drug-NP pairs has been taken to generate the model, see **Table 4** (see **Table S3** in supplementary information for full dataset consultation). We also applied a discretization²³ for the pairs: $f(v_{ij\text{drug}}, v_{ijNP})_{\text{obs}} = 1$ when $f(v_{ij\text{drug}})_{\text{obs}} = 1$ and $f(v_{ijNP})_{\text{obs}} = 1$; $f(v_{ij\text{drug}}, v_{ijNP})_{\text{obs}} = 0$ otherwise. The variable $f(v_{ij\text{drug}}, v_{ijNP})_{\text{ref}}$ is a function that include the expected value of biological activity for a pair (drug-NP), without the perturbation, with vectors of assay

conditions $\mathbf{c}_{j\text{drug}} = (C_{0\text{drug}}, C_{1\text{drug}}, C_{2\text{drug}}, \dots, C_{j\text{drug}}, \dots, C_{\text{maxdrug}})$ and $\mathbf{c}_{j\text{NP}} = (C_{0\text{NP}}, C_{1\text{NP}}, C_{2\text{NP}}, \dots, C_{j\text{NP}}, \dots, C_{\text{maxNP}})$.

PTML linear model. Classification techniques are used given the purpose of the model to predict a desirable biological effect. The model lets us predict $f(v_{ij\text{drug}}, v_{ij\text{NP}})_{\text{calc}}$; which is a scoring function for the drug or drug analog m_i and the NP_i in the combinatorial assay conditions. This PTML model takes into consideration drugs assay conditions ${}^n\mathbf{c}_{j\text{drug}} = (C_{0\text{drug}}, C_{1\text{drug}}, C_{2\text{drug}}, \dots, C_{n\text{drug}})$ and NPs assay conditions $\mathbf{c}_{j\text{NP}} = (C_{0\text{NP}}, C_{1\text{NP}}, C_{2\text{NP}}, \dots, C_{j\text{NP}})$. Note that both ${}^n\mathbf{c}_{j\text{drug}}$ and ${}^n\mathbf{c}_{j\text{NP}}$ can be vectors consisting in combinations of n length. For instance, for the vector $(C_{0\text{drug}}, C_{1\text{drug}})$, $n = 2$. We propose a linear PTML model in order to predict the biological activity and/or classify pairs (drug-NP) as desirable or not desirable. By using Linear Discriminant Analysis (LDA)³⁰ linear classification models can be built having the structure described in equation 1:

$$f(v_{ij\text{drug}}, v_{ij\text{NP}})_{\text{calc}} = a_0 + a_1 \cdot f(v_{ij\text{drug}}, v_{ij\text{NP}})_{\text{expt}} + \sum_{k=1, j=0}^{k_{\text{max}}, j_{\text{max}}} a_{kj} \cdot \Delta D_{k\text{drug}}(\mathbf{c}_{j\text{drug}}) + \sum_{k=1, j=0}^{k_{\text{max}}, j_{\text{max}}} a_{kj} \cdot \Delta D_{k\text{core}}(\mathbf{c}_{j\text{NP}}) + \sum_{k=1, j=0}^{k_{\text{max}}, j_{\text{max}}} a_{kj} \cdot \Delta D_{k\text{coating}}(\mathbf{c}_{j\text{NP}}) \quad (1)$$

PTML-LDA model was built by using STATISTICA Software. The output of the PTML model $f(v_{ij\text{drug}}, v_{ij\text{NP}})_{\text{calc}}$ is an scoring function of the biological activity of the pair drug-NP $v_{ij\text{drug}}$ and $v_{ij\text{NP}}$ for different assay conditions combinations $\mathbf{c}_{j\text{drug}}$ and $\mathbf{c}_{j\text{NP}}$. The first input variable $f(v_{ij\text{drug}}, v_{ij\text{NP}})_{\text{ref}}$ is the function of reference. The PTML model starts with the function $f(v_{ij\text{drug}}, v_{ij\text{NP}})_{\text{ref}}$ as the starting point. Consequently, we defined $f(v_{ij\text{drug}}, v_{ij\text{NP}})_{\text{ref}} = f(v_{ij\text{drug}})_{\text{ref}} \cdot f(v_{ij\text{NP}})_{\text{ref}} = p(f(v_{ij\text{drug}}) = 1) \cdot p(f(v_{ij\text{NP}}) = 1)$. It means that we used as point of reference the probability with which both, the preclinical assay of the drug and the preclinical assay of the NP give a positive result, $f(v_{ij\text{drug}}, v_{ij\text{NP}})_{\text{ref}} = p(f(v_{ij\text{drug}}) = 1, f(v_{ij\text{NP}}) = 1) = p(f(v_{ij\text{drug}}) = 1) \cdot p(f(v_{ij\text{NP}}) = 1)$. After that, the model adds the effect of deviations (perturbations) in all the input variables with respect to their average (expected) values; see **Figure 2**. In order to measure these deviations we used PT operators with the form of one-condition Moving Averages (MA) and multiple-condition Moving Average (MMA) calculated for one condition or combinations of conditions at time, respectively. Thus, we are able to calculate the PT operators $\Delta D_k(c_j) = D_{ki} - \langle D_k(c_j) \rangle$, for descriptors of drugs, coating agents or the core of the NP. The PT operators measure the deviation of D_{ki} ; which is the molecular descriptor of the drug, coating agent or the core of the NP, from the average value $\langle D_k(c_j) \rangle$ of the assays with the same conditions. Furthermore, we built different linear and non-linear models in order to

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

compare the capacity of prediction. These algorithms were: Logistic Regression (LR), Classification Tree (CT), Näive Bayes (NB), AdaBoost (AB) and Random Forest (RF) and Artificial Neural Network (ANN).^{31–34} These different algorithms were implemented on STATISTICA software or program language R,³⁵ with R Studio environment.³⁶ The packages used, with default arguments, were MASS, NNET, RPART, E1071, ADA, and RANDOMFOREST, respectively.

■ RESULTS AND DISCUSSION

PTML linear model. PTML-LDA model considers the expected value of activity $f(v_{ijdrug, v_{ijNP}})_{ref}$ with different added perturbations effects in the system. This article includes an additive model of coated nanoparticles and drug derivatives. In general, for the cases of coated nanoparticle-drug release systems, we can consider four situations: 1) Pristine nanoparticle with drug linked, see **Figure 3A**; 2) Coated nanoparticle and drug linked to nanoparticle, see **Figure 3B**; 3) Coated nanoparticle and drug linked to coating agent, see **Figure 3C**. In **Table 2**, we illustrate the PTML linear equations for these situations. In fact, in a previous work we used linear PTML models, Equation 2, to calculate $f(v_{ijdrug})_{calc}$ which lets us predict biological activity of a free drug (not linked to a nanoparticle delivery system).²⁴ On the other hand, we have used models, like Equation 3 and Equation 4, to predict the biological activity of new nanoparticles without a drug linked to them. This include nanoparticles with coating agent $f(v_{ijNP})_{calc}$ and without it $f(v_{ijNP-pristine})_{calc}$, respectively.³⁷ However, until the best of our knowledge, there are not reports in the literature of PTML models able to predict the activity of the drug linked to a nanoparticle release system.

Table 2. Equations for drug and nanoparticle systems to calculate biological activities.

System	PTML Model	Eq.	Ref.
Drug	$f(v_{ijdrug})_{calc} = a_0 + a_1 \cdot f(v_{ijdrug})_{ref} + \sum_{k=1, j=0}^{kmax, jmax} a_{kj} \cdot \Delta D_{kdrug}(c_{jdrug})$	(2)	36
Pristine NP	$f(v_{ijNP-pristine})_{calc} = a_0 + a_1 \cdot f(v_{ijNP})_{ref} + \sum_{k=1, j=0}^{kmax, jmax} a_{kj} \cdot \Delta D_{kcore}(c_{jNP})$	(3)	15,18

Coated NP (NP)	$f(v_{ijNP})_{calc} = a_0 + a_1 \cdot f(v_{ijNP})_{ref} + \sum_{k=1,j=0}^{kmax,jmax} a_{kj} \cdot \Delta D_{kcore}(c_{jNP}) + \sum_{k=1,j=0}^{kmax,jmax} a_{kj} \cdot \Delta D_{kcoating}(c_{jNP})$	(4)	
Hypothesis	PTML Model	Eq.	Ref.
Additive Without Coating (NP-pristine)	$f(v_{ijdrug}, v_{ijNP-pristine})_{calc} = a_0 + a_1 \cdot f(v_{ijNP})_{ref} + f(v_{ijdrug})_{calc} + f(v_{ijNP-pristine})_{calc}$	(5)	Not Studied
Additive With Coating	$f(v_{ijdrug}, v_{ijNP})_{calc} = a_0 + a_1 \cdot f(v_{ijNP})_{ref} + f(v_{ijdrug})_{calc} + f(v_{ijNP})_{calc}$	(6)	This work
	$f(v_{ijdrug}, v_{ijNP})_{calc} = a_0 + a_1 \cdot f(v_{ijdrug}, v_{ijNP})_{ref} + \sum_{k=1,j=0}^{kmax,jmax} a_{kj} \cdot \Delta D_{kdrug}(c_{jdrug}) + \sum_{k=1,j=0}^{kmax,jmax} a_{kj} \cdot \Delta D_{kcore}(c_{jNP}) + \sum_{k=1,j=0}^{kmax,jmax} a_{kj} \cdot \Delta D_{kcoating}(c_{jNP})$	(7)	

In this work, we present for the first time an additive model, with the capacity to predict drug-nanoparticles pairs with Equation 7. This is the result of adding Equation 2 and Equation 4. That means with $f(v_{ijdrug}, v_{ijNP})_{calc}$ we can predict not only drug-nanoparticle pairs if the drug is linked to the coating but also if the drug is linked to the nanoparticle, with or without a coating (**Figure 3A**, **Figure 3B** and **Figure 3C**). It is necessary to say that the coating agents considered here are organic molecules with chemical structure somehow similar to the drugs. Consequently, the biological activities values are more reliable using v_{ijNP} instead of $v_{ijNP-pristine}$ and the perturbation of the new systems are going to be lower.

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

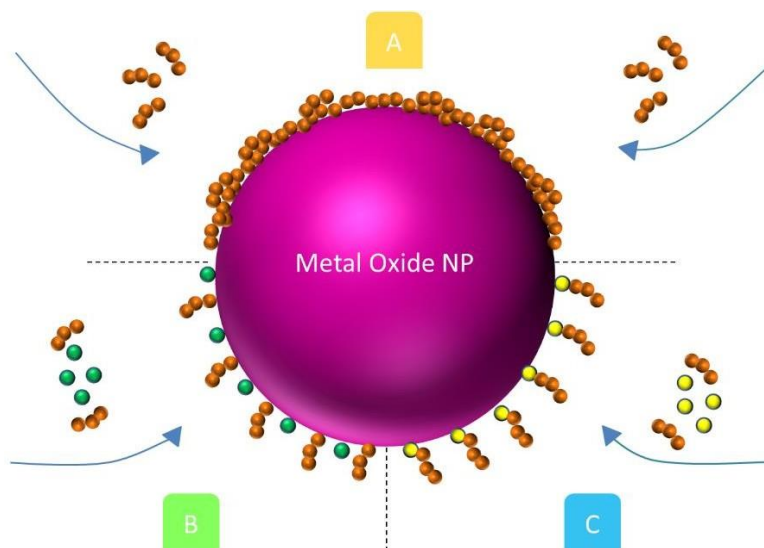


Figure 3. General scheme of drug-nanoparticle systems.

Accordingly, the model developed here include two types of input variables: the expected value function $f(v_{ijdrug}, v_{ijNP})_{ref}$, which is the reference in the system and the variables that add the perturbations for the drug (vit), NP core, and no coating. These are the PT operators $\Delta D_{kdrug}(c_{jdrug})$, $\Delta D_{kcore}(c_{jNP})$ or $\Delta D_{kcoating}(c_{jNP})$. The equation 2 indicates:

$$\begin{aligned}
 f(v_{ijdrug}, v_{ijNP})_{calc} = & -9.4478 + 27.9806 \cdot f(v_{ijdrug}, v_{ijNP})_{ref} \\
 & -0.0055 \cdot \Delta D_{1drug}(c_{jdrug}) + 1.2061 \cdot \\
 \Delta D_{1core}(c_{0NP}) & \\
 & -1.0792 \cdot \Delta D_{1core}(c_{1NP}) + 0.7605 \cdot \\
 \Delta D_{2coating}(c_{1NP}) & \\
 & -0.9859 \cdot \Delta D_{1coating}(c_{1NP}) + 1.2219 \cdot \\
 \Delta D_{3coating}(c_{1NP}) & \tag{2} \\
 & -1.1212 \cdot \Delta D_{4coating}(c_{2NP}) - 0.8670 \cdot \\
 \Delta D_{3coating}(c_{3NP}) & \\
 & -0.2754 \cdot \Delta D_{2core}(c_{4NP}) + 1.6436 \cdot \\
 \Delta D_{1coating}(c_{0NP}) &
 \end{aligned}$$

$$n = 332934 \quad \chi^2 = 173114.59 \quad p < 0.05$$

As we mentioned before, the output of the PTML model $f(v_{ijdrug}, v_{ijNP})_{calc}$ is an scoring function of the biological activity of the pair drug-NP v_{ijdrug} and v_{ijNP} for different assay conditions combinations c_{jdrug} and c_{jNP} .

Table 1 and **Table 2** include information about the input variables. Regarding the statistical parameters of the model, n is the number of cases applied to train the model, χ^2 is the Chi-square statistics, and p is the p-level. The PT operators are one-condition Moving Averages (MA) and multiple-condition Moving Average (MMA) calculated for one condition or combinations of conditions at time, respectively. Thus, we are able to calculate the PT operators $\Delta D_k(c_j) = D_{ki} - \langle D_k(c_j) \rangle$, for descriptors of drugs, coating agents or the core of the NP. The PT operators measure the deviation of D_{ki} ; which is the molecular descriptor of the drug, coating agent or the core of the NP, from the average value $\langle D_k(c_j) \rangle$ of the assays with the same conditions.²³ The output of the PTML model $f(v_{ijdrug}, v_{ijNP})_{calc}$ is an scoring function of the biological activity of the pair drug-NP v_{ijdrug} and v_{ijNP} for different assay conditions combinations c_{jdrug} and c_{jNP} . In this case given that we applied a LDA model $f(v_{ijdrug}, v_{ijNP})_{calc}$ the algorithm must calculate the values of posterior probabilities $p(f(v_{ijdrug}, v_{ijNP})_{obs} = 1)_{pred}$ through the application of the Mahalanobis's distance metric.³⁰ Equal strategy of variable selection was selected to construct the model. This model is characterized by the variety of selected descriptors, five moving average (MA) operators related to the structure of the drug and two multiple MA operators related to the structure of the NP. See details about these operators on **Table 3** (see **Table S1** and **Table S2** respectively in supporting information for full dataset consultation). In addition, **Table S3** we also included all details about each case, observed classification, predicted classification, input variables, experimental conditions, drug derivative and nanoparticle characteristics. This table is freely available online in the public data repository Figshare with doi: <https://doi.org/10.6084/m9.figshare.8143394.v1>, due to volume restrictions.

Table 3. Drugs and nanoparticles (core and coating) operators information

Condition Name	Code	Symbol	Operator Formula	Description
Activity type	c_{0drug}	$f(v_{ijdrug})_{ref}$	$n(f(v_{ijdrug})_{obs}=1)/n_j$	Expected value of probability $p(f(v_{ijdrug})=1)_{ref}$ for the activity v_{ijdrug} of type c_{0drug}
Activity type	c_{0drug}		$D_{1drug\ i} - \langle D_{1drug} (c_{jdrug}) \rangle$	

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

Protein	c_{1drug}	$\Delta D_{1drug}(c_{jdrug})$		Deviation (Δ) of the $D_{1drug} = ALOGP_i$ of the i^{th} drug from the expected value ($\langle ALOGP(c_{jdrug}) \rangle$) for a given subset of multiple assay conditions c_{jdrug} .
Cell Name	c_{2drug}			
Assay Organism	c_{3drug}			
Assay Strain	c_{4drug}			
Activity type	c_{0NP}	$f(v_{ijNP})_{ref}$	$n(f(v_{ijNP})_{obs}=1)/n_j$	Expected value of probability $p(f(v_{ijNP})=1)_{ref}$ for the nanoparticle to have activity v_{ijNP} of type c_{0NP}
Activity type	c_{0NP}, c_{0drug}	$f(v_{ijdrug}, v_{ijNP})_{ref}$	$f(v_{ijdrug})_{ref} \cdot f(v_{ijNP})_{ref}$	Expected value of probability for the activity of the drug-release nano-system
Activity type	c_{0NP}	$\Delta D_{1core}(c_{0NP})$	$D_{1core\ i} - \langle D_{1core}(c_{0NP}) \rangle$	Measures the deviation of the D_{1core} vs. the expected value (average) of all NP_i with the same c_{0NP} = activity type. Same calculus is applied to $D_{1coating}$. Regarding c_{1NP} , the operators used are D_{1core} , $D_{1coating}$, $D_{2coating}$ and $D_{3coating}$.
		$\Delta D_{1coating}(c_{0NP})$	$D_{1coating\ i} - \langle D_{1coating}(c_{0NP}) \rangle$	
Cell line	c_{1NP}	$\Delta D_{1core}(c_{1NP})$	$D_{1core\ i} - \langle D_{1core}(c_{1NP}) \rangle$	
		$\Delta D_{1coating}(c_{1NP})$	$D_{1coating\ i} - \langle D_{1coating}(c_{1NP}) \rangle$	
		$\Delta D_{2coating}(c_{1NP})$	$D_{2coating\ i} - \langle D_{2coating}(c_{1NP}) \rangle$	
		$\Delta D_{3coating}(c_{1NP})$	$D_{3coating\ i} - \langle D_{3coating}(c_{1NP}) \rangle$	
Shape	c_{2NP}	$\Delta D_{4coating}(c_{2NP})$	$D_{4coating\ i} - \langle D_{4coating}(c_{2NP}) \rangle$	Measures the deviation of the $D_{4coating}$ vs. the expected value (average) of all NP_i with the same c_{2NP} = Shape. Same
Medium	c_{3NP}	$\Delta D_{3coating}(c_{3NP})$	$D_{3coating\ i} - \langle D_{3coating}(c_{3NP}) \rangle$	

Assay time	c_{4NP}	$\Delta D_{2core}(c_{4NP})$	$D_{2core} - \langle D_{2core}(c_{4NP}) \rangle$	calculus is applied to $D_{3coating}$ and D_{2core} and with $c_{3NP} = \text{Medium}$ and $c_{4NP} = \text{Assay time}$, respectively.
------------	-----------	-----------------------------	--	--

Once calculated $p(f(v_{ijdrug}, v_{ijNP}) = 1)_{pred}$, it is possible to create a Boolean function: $f(v_{ijdrug}, v_{ijNP})_{pred} = 1$ if $p(f(v_{ijdrug}, v_{ijNP}) = 1)_{calc} > 0.5$; otherwise, $f(v_{ijdrug}, v_{ijNP})_{pred} = 0$. This function is needed in comparison terms: If $f(v_{ij}, v_{ijNP})_{pred} = 1$ and $f(v_{ijdrug}, v_{ijNP})_{obs} = 1$ the case is properly classified; otherwise, it is not.²³ With this comparison we are able to measure the S_n , S_p , and A_c of the generated PTML-LDA model. In this work, the model showed adequate values of $S_p = 95.75$, $S_n = 75.09$, and $A_c = 94.43$ in training. Similar values were presented for external validation, see **Table 4**.

Table 4. Results of the model and input variables analyzed

Obs. Sets ^a	Stat. Param. ^b	Pred. Stat. ^c	Predicted sets		
			n_j	$f(v_{ijdrug}, v_{ijNP})_{pred} = 1$	$f(v_{ijdrug}, v_{ijNP})_{pred} = 0$
Training					
$f(v_{ijdrug}, v_{ijNP})_{obs} = 1$	S_p	75.0	21362	16021	5341
$f(v_{ijdrug}, v_{ijNP})_{obs} = 0$	S_n	95.8	311572	13201	298371
Total	A_c	94.4	332934		
Validation					
$f(v_{ijdrug}, v_{ijNP})_{obs} = 1$	S_p	74.6	10743	8017	2726
$f(v_{ijdrug}, v_{ijNP})_{obs} = 0$	S_n	95.9	156322	6453	149869
Total	A_c	94.5	167065		

^a Obs. Sets = Observed sets, ^b Stat. Param. = Statistical parameter, ^c Pred. Stat. = Predicted statistics

If we want to apply the model in order to select a pair (drug-NP) to assemble a nano-system, a substitution in the model is needed regarding the expected values of the descriptors $\langle D_{idrug}(c_{jdrug}) \rangle$, $\langle D_{icore}(c_{jNP}) \rangle$ and $\langle D_{icoating}(c_{jNP}) \rangle$ for different conditions or combination of

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

conditions. In **Table 5**, considered parameters for drugs and nanoparticles for $d(c_{0drug})$ and $d(c_{0NP})$ are included.

Table 5. Considered parameters for drugs (c_{0drug}) and nanoparticles (c_{0NP})

Condition c_{0drug}^a	$\langle D_1(c_{0drug}) \rangle$	$\langle D_2(c_{0drug}) \rangle$	$n_j(c_{0drug})$	$n_j(f(v_{ijdrug})=1)_{obs}$	$p(f(v_{ijdrug})=1)_{ref}$	cutoff	$d(c_{0drug})$
Potency(nM)	3.29	74.06	24750	104	0.004	100.00	-1
IC ₅₀ (nM)	4.24	63.20	1402	232	0.165	100.00	-1
Activity(%)	3.79	79.40	1079	56	0.052	186.79	1
Inhibition(%)	3.25	82.98	415	254	0.612	73.72	-1
EC ₅₀ (nM)	4.50	66.09	388	193	0.497	100.00	-1
Weight(g)	3.18	35.24	260	192	0.738	4.23	-1
Ratio(-)	5.44	63.26	259	253	0.977	46.59	-1
GI ₅₀ (nM)	3.81	38.24	258	2	0.008	100.00	-1
Ki(nM)	3.83	77.45	197	106	0.538	100.00	-1
Activity(mg/dl)	5.74	58.21	164	95	0.579	5.22	-1
Condition c_{0NP}^b	Input parameters used to specify c_{0NP}						
Activity	$n_j(c_{0NP})$	$n_j(f(v_{ijNP})=1)_{obs}$	$p(f(v_{ijNP})=1)_{ref}$	cutoff	$d(c_{0NP})$	$\langle D_{1core}(c_{0NP}) \rangle$	
EC ₅₀ (μM)	30	27	0.9	25422	-1	51.13	
IC ₅₀ (μM)	29	21	0.72	18714	-1	0.28	
CC ₅₀ (μM)	113	21	0.19	3099	1	21.44	
Condition c_{0NP}	Input parameters used to specify c_{0NP}						
Activity	$\langle D_{2core}(c_{0NP}) \rangle$	$\langle D_{1coating}(c_{0NP}) \rangle$	$\langle D_{2coating}(c_{0NP}) \rangle$	$\langle D_{3coating}(c_{0NP}) \rangle$	$\langle D_{4coating}(c_{0NP}) \rangle$		
EC ₅₀ (μM)	2.48	6.54	0.18	17.57	53.16		
IC ₅₀ (μM)	2.36	7.94	0.55	11.77	74.56		
CC ₅₀ (μM)	2.54	17.17	0.30	6.07	72.26		

^a Condition c_{0drug} = the type of activity parameter measured for drugs. ^b Condition c_{0NP} = the type of activity parameter measured for NPs

In **Table 6**, it is shown the values of the averages $\langle D_i(c_{jNP}) \rangle$ for NPs. We can see examples of how the expected values change depending on the 3 considered conditions c_{1NP} , c_{2NP} and c_{3NP} . This happens if we change conditions not only for NPs but also for drugs. Consequently, if the conditions change affect the result of the model for every pair drug-NP. The list of all values of MA an MMA for drugs and NPs can be consulted on supporting information, **Table S1** and **Table S2**. With regard the expected values of probability we consider $p(f(v_{ijdrug}, v_{ijNP})_{obs}=1)_{ref} = p(f(v_{ijdrug})_{obs}=1)_{ref} * p(f(v_{ijNP})_{obs}=1)_{ref}$. This probability is the result of the multiplication of the probability of drug to be desired $p(f(v_{ijdrug})_{obs}=1)_{ref}$ and the probability of a NP to be non-toxic $p(f(v_{ijNP})_{obs}=1)_{ref}$, see also **Figure 2**.

Table 6. One-condition averages and number of cases for selected NP conditions of assay

c_{1NP}	Parameters used to specify c_{1NP}^c						
Cell Line ^a	Avg ₁	Avg ₂	Avg ₃	Avg ₄	Avg ₅	Avg ₆	$n_j(c_{1NP})$
A549 (H)	0.04	2.72	6.22	0.00	0.00	0.00	23
LE	0.50	2.37	14.40	1.00	21.34	135.1	16
HepG2 (H)	0.47	2.52	27.71	0.60	17.07	129.5	15
3T3 (M)	0.44	2.36	12.80	0.89	18.97	120.1	9
						4	
						1	
						3	
c_{2NP}	Parameters used to specify c_{2NP}^d						
Shape ^b	Avg ₇	Avg ₈	Avg ₉	Avg ₀	Avg ₁	Avg ₁₂	$n_j(c_{2NP})$
Spherical	4.85	2.51	4.85	30.30	0.51	11.47	61
Elliptical	0.00	2.71	0.00	0.00	0.00	0.00	21
Pyramidal	0.00	2.81	0.00	0.00	0.00	0.00	10
PS	0.00	2.51	0.00	0.00	0.00	0.00	8
c_{3NP}	Parameters used to specify c_{3NP}^e						
Medium	Avg ₃	Avg ₄	Avg ₅	Avg ₆	Avg ₇	Avg ₁₈	$n_j(c_{3NP})$
Dry	33.45	2.48	16.23	0.28	8.45	61.31	118

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

H2O	0.41	2.54	10.27	0.50	12.69	106.5 4	44
DMEM	0.00	2.93	0.00	0.00	0.00	0.00	3
RPMI	0.00	2.74	0.00	0.00	0.00	0.00	3

$$\begin{aligned}
 {}^aLE &= Lycopersicon esculentum, {}^bPS = pseudo-spherical. {}^cAvg_1 = \langle D_{1core}(C_{1NP}) \rangle, Avg_2 = \\
 &\langle D_{2core}(C_{1NP}) \rangle, Avg_3 = \langle D_{1coating}(C_{1NP}) \rangle, Avg_4 = \langle D_{2coating}(C_{1NP}) \rangle, Avg_5 = \langle D_{3coating}(C_{1NP}) \rangle, Avg_6 \\
 = &\langle D_{4coating}(C_{1NP}) \rangle, Avg_7 = \langle D_{1core}(C_{2NP}) \rangle, Avg_8 = \langle D_{2core}(C_{2NP}) \rangle, Avg_9 = \langle D_{1coating}(C_{2NP}) \rangle, Avg_{10} = \\
 &\langle D_{2coating}(C_{2NP}) \rangle, Avg_{11} = \langle D_{3coating}(C_{2NP}) \rangle, Avg_{12} = \langle D_{4coating}(C_{2NP}) \rangle, Avg_{13} = \\
 &\langle D_{1core}(C_{3NP}) \rangle, Avg_{14} = \langle D_{2core}(C_{3NP}) \rangle, Avg_{15} = \langle D_{1coating}(C_{3NP}) \rangle, Avg_{16} = \langle D_{2coating}(C_{3NP}) \rangle, Avg_{17} = \\
 &\langle D_{3coating}(C_{3NP}) \rangle, Avg_{18} = \langle D_{4coating}(C_{3NP}) \rangle
 \end{aligned}$$

This model is useful to score the activity of a new pair (drug-NP) in different combinatorial conditions of bio-assays. For that purpose, we must proceed with substitution of the expected probability of activity $p(f(v_{ijdrug}, v_{ijNP})_{obs} = 1)_{ref}$ on the equation, because it contains information depending on the activity is measuring, *e.g.* IC₅₀(μM), CC₅₀(μM), and EC₅₀(μM), *etc.* Consequently, the model is able to predict different activity parameters for each pair. Finally, we must include the values of the new pair descriptors and then we will obtain the prediction of the biological desirability of the new pair MONP-drug.

PTML-LDA DDNS simulation. After training and validating the PTML model we carry out a computational study aimed to show a possible practical use of the model. We selected for the study the output biological properties Inhibition(%) and EC₅₀(nM), two of the more represented in the dataset. Using the linear PTML-LDA model we calculated the posterior probabilities $p(f(v_{ij})=1)$ for >15000 different preclinical assays of drug release nano-systems with different combination of drug, coating agent, and type of metal or MONPs. These are the probabilities calculated with the model with which the selected drug, coating agent, and nanoparticle may be assembled into a useful nano-system. We understand useful here as a system with high probability of having Inhibition(%) of target higher than the cutoff and EC₅₀(nM) lower than the cutoff. We calculated the average value of these probabilities $\langle p(f(v_{ij})=1) \rangle$ for different sub-sets of drug release nano-systems, see **Table 7**. Interestingly, the systems formed by Ag nanoparticles with PSTARCH or CIT as coating agent give higher values of $\langle p(f(v_{ij})=1) \rangle$ for assays of Inhibition(%) of the target and assays of EC₅₀ for drugs like Niacin (>150 assays), Tretinoin (>120 assays), Menadione (>250), *etc.* However the

systems with SiO₂ and Si nanoparticles are predicted to have medium to low values of $\langle p(f(v_{ij})=1) \rangle$ for the set of drugs studied. This kind of computational simulation may be a useful complementary tool to select the components of new drug release nano-systems in future experimental works.

Table 7. PTML simulation of coated nanoparticle drug release systems

Coated Nanoparticle Drug Release Systems				$\langle p(f(v_{ij})=1) \rangle$	
Drug	Coating ^a	Type	n _s	Inhibition(%)	EC ₅₀ (nM)
Niacin	PSTARCH	Ag	159	1	0.997
Tretinoin	PSTARCH	Ag	123	1	0.997
Adenosine Phosphate	CIT	Ag	118	0.998	0.701
Menadione	CIT	Ag	252	0.667	0.980
Rutin	CIT	Ag	223	0.809	0.980
Vitamin E	CIT	Ag	293	0.847	0.979
Niacin	CIT	Ag	1474	0.899	0.696
Cholecalciferol	CIT	Ag	246	0.874	0.654
Calcitriol	CIT	Ag	618	0.856	0.694
Tretinoin	CIT	Ag	1277	0.722	0.706
Niacin	PVP	Ag	597	0.671	0.236
Calcitriol	PEGSi	SiO ₂	479	0.628	0.400
Rutin	PEGSi	SiO ₂	178	0.586	0.317
Niacin	PVA	CoFe ₂ O ₄	139	0.583	0.435
Tretinoin	PVA	CoFe ₂ O ₄	118	0.581	0.435
Niacin	PEGSi	SiO ₂	1177	0.537	0.422
Tretinoin	PEGSi	SiO ₂	1031	0.525	0.367
Vitamin E	PEGSi	SiO ₂	230	0.493	0.375
Adenosine Phosphate	PEGSi	SiO ₂	101	0.47	0.461
Menadione	PEGSi	SiO ₂	226	0.322	0.439
Tretinoin	PVP	Ag	506	0.384	0.160
Vitamin E	PVP	Ag	112	0.369	0.008
Menadione	PVP	Ag	116	0.343	0.008
Calcitriol	PVP	Ag	241	0.014	0.193
Menadione	CIT	Au	209	0.044	0.02

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

Niacin	CIT	Au	997	0.043	0.02
Tretinoin	3NTPA	Ge	241	0.013	0.007
Tretinoin	PAF	Si	487	0.009	0.004
Calcitriol	PAF	Si	201	0.009	0.004
Menadione	PAF	Si	119	0.009	0.004
Niacin	UDAF	Si	319	0.006	0.003
Tretinoin	UDAF	Si	255	0.006	0.003

^a 3NTPA = N,N,N-trimethyl-3(1-propene) ammonium fragment, CIT = Sodium Citrate, PSTARCH = Potato Starch, PAF = Propylamoniun fragment, UDAF = undecylazide fragment, PEGSi = PEG-Si(OMe)₃.

PTML-R Studio linear vs. non-linear models. In any case, the previous PTML-LDA model is a linear model. However, there are other alternative linear and non-linear algorithms useful to seek PTML models as well. In this section, we used different ML algorithms implemented in the software R Studio. These models are built by using program language R. The justification of these models is the necessity to contrast the prediction among them and the PTML-LDA constructed below. The algorithms applied were, Logistic Regression (LR), Classification Tree (CT), Näive Bayes (NB), AdaBoost (AB) and Random Forest (RF), see **Table 8**. These algorithms, along with LDA, have been used in cheminformatics, given their capacity to classify biological activity.³⁸ The best PTML-LR found presented a Sp(%) = 98.54, which is the highest comparing to the rest of models, even non-linear models. However the Sn(%) = 53.14 is the lowest ratio. Thus, we cannot consider it the best linear PTML model given the unbalanced results.

We also developed non-linear PTML models with the same data for comparison purposes: PTML-NB presented the lowest Sp(%) = 90.44 and similar Sn(%) ≈ 53-56 than PTML-RL. Besides, the Ac(%) = 88.18, so we can consider the worst model in terms of prediction. The last two models, PTML-CT and PTML-RF presented higher ratios than PTML-LDA. Both showed Sp(%) ≈ 96-97 and Sn(%) ≈ 85-87. However, if we take into consideration the variables that PTML-CT includes only PT descriptors related to the core of the NP: $\Delta D_{2\text{core}}(C_{0\text{NP}}) = D_{2\text{core } i} - \langle D_{2\text{core}}(C_{0\text{NP}}) \rangle$, $\Delta D_{2\text{core}}(C_{2\text{NP}}) = D_{2\text{core } i} - \langle D_{2\text{core}}(C_{2\text{NP}}) \rangle$, $\Delta D_{3\text{core}}(C_{3\text{NP}}) = D_{3\text{core } i} - \langle D_{3\text{core}}(C_{3\text{NP}}) \rangle$ and $\Delta D_{4\text{core}}(C_{0\text{NP}}) = D_{4\text{core } i} - \langle D_{4\text{core}}(C_{0\text{NP}}) \rangle$, see **Figure 4**. $D_{3\text{core}}$ and $D_{4\text{core}}$ refer to Polarizability and the Size of the NP

respectively. Thus, in this PTML-CT, after applying Classification Tree method, the information of the drug and the coating agent is missed. On the other hand, the PTML-RF is more complex than PTML-CT (includes 50 trees) but includes variables with information of the coating agent and the core of the NP as well as the drug derivatives.

Table 8. PTML Non -LDA models results

PTML Algorithm	Soft. ^a	Predicted Sets ^a	Statistical Parameter ^b	Predicted Statistics	Observed sets	
					$f(V_{ijdrug}, V_{ijNP})_{obs} = 0$	$f(V_{ijdrug}, V_{ijNP})_{obs} = 1$
LDA	S	$f(V_{ijdrug}, V_{ijNP})_{pred} = 0$	Sp(%)	95.75	298360	5321
		$f(V_{ijdrug}, V_{ijNP})_{pred} = 1$	Sn(%)	75.09	13212	16041
		total	Ac(%)	94.43	311572	21362
LDA	R	$f(V_{ijdrug}, V_{ijNP})_{pred} = 0$	Sp(%)	96.93	113389	3035
		$f(V_{ijdrug}, V_{ijNP})_{pred} = 1$	Sn(%)	62.18	3584	4991
		total	Ac(%)	94.7	116973	8026
LR	R	$f(V_{ijdrug}, V_{ijNP})_{pred} = 0$	Sp(%)	98.54	115276	3731
		$f(V_{ijdrug}, V_{ijNP})_{pred} = 1$	Sn(%)	53.14	1697	4295
		total	Ac(%)	95.66	116973	8026
CT	R	$f(V_{ijdrug}, V_{ijNP})_{pred} = 0$	Sp(%)	96.81	113252	1003
		$f(V_{ijdrug}, V_{ijNP})_{pred} = 1$	Sn(%)	87.50	3721	7023
		Total	Ac(%)	96.22	116973	8026
NB	R	$f(V_{ijdrug}, V_{ijNP})_{pred} = 0$	Sp(%)	90.44	105793	3593
		$f(V_{ijdrug}, V_{ijNP})_{pred} = 1$	Sn(%)	55.23	11180	4433
		Total	Ac(%)	88.18	116973	8026
RF	R	$f(V_{ijdrug}, V_{ijNP})_{pred} = 0$	Sp(%)	97.71	114290	1185
		$f(V_{ijdrug}, V_{ijNP})_{pred} = 1$	Sn(%)	85.23	2683	6841
		Total	Ac(%)	96.91	116973	8026
AB	R	$f(V_{ijdrug}, V_{ijNP})_{pred} = 0$	Sp(%)	98.27	114951	2352
		$f(V_{ijdrug}, V_{ijNP})_{pred} = 1$	Sn(%)	70.69	2022	5674
		Total	Ac(%)	96.50	116973	8026

^a Software used: S = Statistica, R = R Studio. ^b ML algorithm used: LR = Logistic Regression,

CT = Classification Tree, NB = Näive Bayes, RF = Random Forest, AB= AdaBoost

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

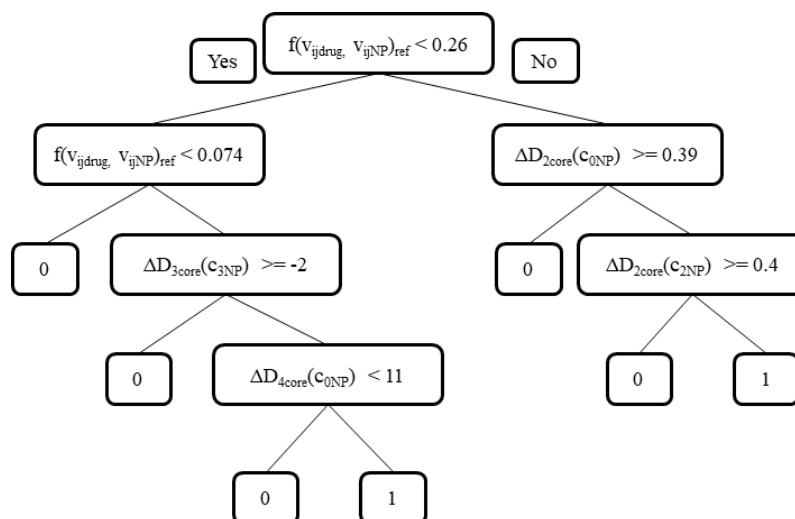


Figure 4. PTML-CT (Classification Tree algorithm with PTML technique) model

Given that random data is taken to train every model, we repeat the same process 20 times in order to have accuracy mean and accuracy standard deviation of all the subsets. This bootstrapping provides us more information about the robustness of the different PTML created.³⁹ The results do not show significant variations (Table 9) comparing to the overall accuracy showed by the first batch trained (**Table 8**). PTML-RF is the model that predict better in terms of overall accuracy.

Table 9. PTML R-Non LDA models results for 20-fold bootstrapping

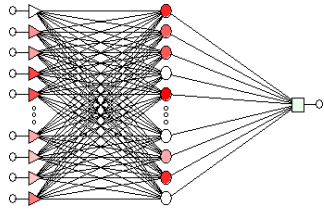
PTML Algorithm	Accuracy mean	Accuracy s.d.
LDA	95.7%	0.00036
LR	95.6%	0.00032
CT	96.1%	0.00056
NB	88.3%	0.00130
RF	96.6%	0.00019
AB	96.5%	0.000659

Cheminformatics models can be transversal, flexible and useful tools for prediction of biological activity of nanomaterials. In this research, we showed that PTML technique is useful to model complex datasets coming from databases like ChEMBL, the drugs in our case, or compiled from literature like NPs toxicological assays. This model has an adequate

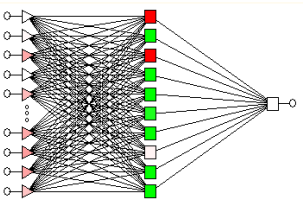
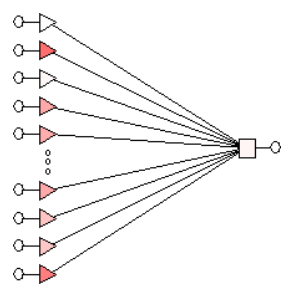
performance, taking into consideration the heterogeneous and combinatorial data with Big Data characteristics. For the present dataset, PTML-LDA model with descriptors including multiple assay conditions are more efficient to predict NP-drug pairs biological activity. In fact, PTML-LDA is a simple and reliable method to predict. The PTML-LDA model presented here is the first multi input and multi output model able to predict biological activity of nanoparticles and drug derivatives pairs. However, RF has showed better results in terms of capacity of prediction. PTML-RF showed higher Specificity, Sensitivity and Accuracy but with higher complexity.

PTML-ANN linear and non-linear models. Using similar criteria than in the previous section, we used here different linear and non-linear ANN algorithms for comparative purposes. We describe these PTML-ANN models in a separated section because they have been obtained with the software STATISTICA and not with R Studio, as in the previous section. The best PTML-ANN models found have more balanced values of Sp and Sn $\approx 95\%$ in training and validation series, **Table 10**. The best PTML-ANN models found presented values of AUROC between 0.95-0.98 for training and external validation series.

Table 10. PTML-ANN models results

Profile ^a Nv:I-H-O:No	AR ^b	Predicted ^c $f(V_{ijdrug}, V_{ijNP})_{pred}$	cPS (%)	Observed test sets $f(V_{ijdrug}, V_{ijNP})_{obs}$	
				0	1
RBF 15:15-25-1:1					
	0.961	0	89.13	139344	1136
		1	89.42	16978	9607
		Total	89.16	156322	10743
MLP 14:14-10-1:1					
	0.982	0	95.13	148710	543
		1	94.94	7612	10200

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

		Total	95.12	156322	10743
LNN 28:28-1:1					
	0.955	0	91.0	142302	972
		1	90.95	14020	9771
		Total	91.02	156322	10743

^a Profile Nv:I-H-O:No = Number of input variables, Inout layers, Hidden layers, Output layers, Number of output variables, MLP = Multi-Layer Perceptron, RBF = Radial Basis Function, LNN = Linear Neural Network. ^bAR = AUROC, ^cPS (%) = Predicted Statistics (Specificity, Sensitivity and Accuracy).

The generation of these PTML-ANN models gives us the information, as well as the other cases, is that PT operators along as reference function present a not-random relation with DDNs elements activity (AUROC = 0.5). Actually they present AUROC values higher than 0.95, see **Figure 5**. All the PTML-ANN present higher complexity, higher Sp(%) and lower Sn(%). We selected the variables using the variable selection algorithm of the ANN module of the software. The best Radial Basis Function (RBF) Network found (AUROC = 0.961) has 15 variables and one hidden layer with 25 nodes. However, it shows an Sp(%) = 89.13, Sn(%) = 89.42 and Ac(%) = 89.16. Regarding the Linear Neural Networks (LNN), the best model we found has 28 variables comparing to the 11 variables that were included in PTML-LDA. The predicted statistics are slightly better than the RBF Neural Network. Regarding the best Multi-Layer Perceptron (MLP) model we found, it presents the best results in terms of Sp(%) = 94.94, Sn(%) = 95.13 and Ac(%) = 95.12. It includes a hidden layer with 10 nodes. Furthermore, we checked the models with y-scrambling to ensure the absence of overtraining. In so doing, we trained 5 extra neural networks with different $f(V_{ijdrug}, V_{ijNP})_{obs}$.⁴⁰ The prediction, in terms of

Sp(%), Sn(%) and Ac(%) decreased to the range of 47%-53%, both for desirable and not desirable cases. This gives us information, together with the percentages of training and test of the neural networks presented, about the adequate performance of the training.

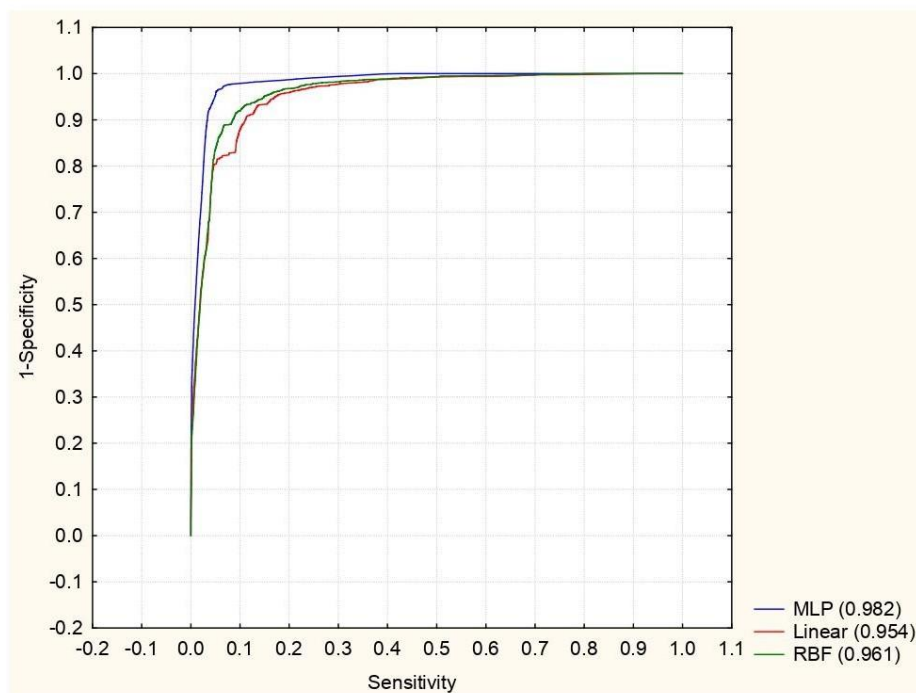


Figure 5. ROC analysis of the PTML-ANN models, Sensitivity (Sn) vs. 1 – Specificity (Sp), random classifier curve (yellow) vs. other models (multiple colors)

■ CONCLUSIONS

Currently, we can find a considerable number of different types of NDDs in literature. Coated MONPs have been studied with special attention given the variety and the showed potential. However, the heterogeneous data of the current bioassays make new challenges raise in terms of cheminformatics models. In this work, we apply PTML method to build a model that is able to predict NDDS of MONPs. The best PTML-LDA linear model found showed values of Sp (%) = 95.75 and Sn(%) = 75.09. The variables of this PTML-LDA include information of the assay conditions, the coating agent, the core of the nanoparticle and the drug of the NDDs. We also trained alternative PTML linear and non-linear models by applying LR, CT, NB, RF, AB, and ANN algorithms. RF and ANN models reached a higher Sp and Sn, although they present a higher complexity in terms of number of variables, neurons layers, number of forests, *etc.* We also illustrated the use of the present PTML-LDA model in a simulation studies to detected

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

promising NDDs. This is a practical use of this model, which can give complementary information in terms of designing new systems. Until the best of our knowledge, this is the first multi-label PTML model useful to select drugs, coating agents, and/or MONPs to be assembled in order to design new DDNS with optimal activity/toxicity profiles.

■ ASSOCIATED CONTENT


Supporting Information

The dataset used, including molecular descriptors, and assay conditions, desirability, cutoff, biological activities etc., was included in tables **Table S1**, **Table S2** and **Table S3** (SI01.xlsx). See details about these Moving Average operators on **Table 3** and **Table 5** (see **Table S1** and **Table S2** respectively in supporting information for full dataset consultation). In addition, **Table S3** we also included all details about each case, observed classification, predicted classification, input variables, experimental conditions, drug derivative and nanoparticle characteristics. This table is freely available online in the public data repository Figshare with doi: <https://doi.org/10.6084/m9.figshare.8143394.v1>, due to volume restrictions.


■ AUTHORS INFORMATION

Corresponding author

*(H.G.D) E-mail: humberto.gonzalezdiaz@ehu.es (H.G.-D.)

 Orcid: 0000-0002-9392-2797

*(R.S) E-mail: ricardo.santana@opendeusto.es (R.S.)

 Orcid: 0000-0002-5206-2305

■ ACKNOWLEDGMENTS

R.S.C. thanks COLCIENCIAS scholarship for the doctorate studies; “Convocatoria para Doctorado Nacional 757” from 2017. This original research is part of the project “Investigación en Derecho Internacional y Nanotecnología” registered in the Research Centre of Universidad Pontificia Bolivariana with register number 766B-06/17-37. Special gratitude is extended to CYTED NANOCELIA network. The authors acknowledge research grants from Ministry of Economy and Competitiveness, MINECO, Spain (FEDER CTQ2016-74881-P) and Basque government (IT1045-16). The authors also acknowledge the support of Ikerbasque, Basque

Foundation for Science. The authors also acknowledge the support of Ikerbasque, Basque Foundation for Science.

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

■ REFERENCES

- 1 B. Rasulev, F. Jabeen, S. Stafslie, B. J. Chisholm, J. Bahr, M. Ossowski and P. Boudjouk, *ACS Appl. Mater. Interfaces*, 2017, **9**, 1781–1792.
- 2 B. Rasulev, M. Quadir, D. C. Webster, S. Stafslie and R. P. Chitemere, *ACS Appl. Bio Mater.*, 2018, **1**, 1830–1841.
- 3 J. A. Hong, D. P. Bhave and K. S. Carroll, *J. Med. Chem.*, 2009, **52**, 5485–5495.
- 4 A. Farboudi, K. Mahboobnia, F. Chogan, M. Karimi, A. Askari, S. Banihashem, S. Davaran and M. Irani, *Int. J. Biol. Macromol.*, 2020, **150**, 178–188.
- 5 E. Vlassi, A. Papagiannopoulos, A. Sergides and S. Pispas, *J. Nanosci. Nanotechnol.*, 2020, **20**, 3981–3988.
- 6 C. Zheng, Y. Wang, S. Z. F. Phua, W. Q. Lim and Y. Zhao, *ACS Biomater. Sci. Eng.*, 2017, **3**, 2223–2229.
- 7 H. Yan, C. Teh, S. Sreejith, L. Zhu, A. Kwok, W. Fang, X. Ma, K. T. Nguyen, V. Korzh and Y. Zhao, *Angew. Chemie - Int. Ed.*, 2012, **51**, 8373–8377.
- 8 M. Yin, E. Ju, Z. Chen, Z. Li, J. Ren and X. Qu, *Chem. - A Eur. J.*, 2014, **20**, 14012–14017.
- 9 H. Zhu, H. Chen, X. Zeng, Z. Wang, X. Zhang, Y. Wu, Y. Gao, J. Zhang, K. Liu, R. Liu, L. Cai, L. Mei and S. S. Feng, *Biomaterials*, 2014, **35**, 2391–2400.
- 10 I. Ali, S. D. Mukhtar, H. S. Ali, M. T. Scotti and L. Scotti, *Curr. Pharm. Des.*
- 11 A. H. Vo, T. R. Van Vleet, R. R. Gupta, M. J. Liguori and M. S. Rao, *Chem. Res. Toxicol.*, 2020, **33**, 20–37.
- 12 R. L. Eunkeu, A. Nel, K. Boeneman Gemill, M. Bilal, Y. Cohen and I. Medintz, *Nat. Nanotechnol.*, 2016, **11**, 479–486.
- 13 N. Novoselska, B. Rasulev, A. Gajewicz, V. Kuzmin, T. Puzyn and J. Leszczynski, *Nanoscale*, 2014, **6**, 13986–13993.
- 14 A. P. Toropova, A. A. Toropov, R. Rallo, D. Leszczynska and J. Leszczynski, *Ecotoxicol. Environ. Saf.*, 2015, **112**, 39–45.
- 15 K. Pathakoti, M. Huang, J. D. Watts, X. He and H. Hwang, *J. Photochem. Photobiol. B Biol.*, 2014, **130**, 234–240.
- 16 K. P. Singh and S. Gupta, *RSC Adv.*, 2014, **4**, 13215–13230.
- 17 N. Fjodorova, M. Novic, A. Gajewicz and B. Rasulev, *Nanotoxicology*, 2017, **11**, 475–

- 483.
- 18 T. Mikolajczyk, A., Gajewicz, A., Rasulev, B., Schaeublin, N., Maurer-Gardner, E., Hussain, S., Leszczynski, J., Puzyn, *Chem. Mater.*, 2015, **27**, 2400–2407.
 - 19 F. Luan, V. V. Kleandrova, H. González-Díaz, J. M. Ruso, A. Melo, A. Speck-Planche and N. Cordeiro, *Nanoscale*, 2014, **6**, 10623–10630.
 - 20 V. V. Kleandrova, F. Luan, H. González Díaz, J. M. Ruso, A. Speck-planche, M. Nata and D. S. Cordeiro, *Environmental Sci. Technol.*, 2014, **48**, 14686–14694.
 - 21 R. Santana, R. Zuluaga, P. Gañán, S. Arrasate, E. Onieva and H. González-Díaz, *Nanoscale*, 2019, **11**, 21811–21823.
 - 22 E. Vásquez-Domínguez, V. D. Armijos-Jaramillo, E. Tejera and H. González-Díaz, *Mol. Pharm.*, 2019, **16**, 4200–4212.
 - 23 L. Simón-Vidal, O. García-Calvo, U. Oteo, S. Arrasate, E. Lete, N. Sotomayor and H. González-Díaz, *J. Chem. Inf. Model.*, 2018, **58**, 1384–1396.
 - 24 H. Bediaga, S. Arrasate and H. González-Díaz, *ACS Comb. Sci.*, 2018, **20**, 621–632.
 - 25 J. F. Da Costa, D. Silva, O. Caamaño, J. M. Brea, M. I. Loza, C. R. Munteanu, A. Pazos, X. García-Mera and H. González-Díaz, *ACS Chem. Neurosci.*, 2018, **9**, 2572–2587.
 - 26 A. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos and J. P. Overington, *Nucleic Acids Res.*, 2014, **42**, 1083–1090.
 - 27 A. K. Ghose and G. M. Crippen, *J. Comput. Chem.*, 1987, **27**, 21–35.
 - 28 Talete, dProperties User’s Manual, http://www.talete.mi.it/help/dproperties_help/index.html?p_vsa_like_descriptors.htm, (accessed 26 February 2019).
 - 29 P. Labute, *J. Mol. Graph. Model.*, 2000, **18**, 464–477.
 - 30 S. Balakrishnama and A. Ganapathiraju, *Inst. Signal Inf. Process.*, 1998, **18**, 1–8.
 - 31 D. W. Hosmer Jr, S. Lemeshow and R. X. Sturdivant, *Introduction to Logistic Regression Model*, John Wiley & Sons, 2014, vol. 398.
 - 32 W. Loh, *WIREs DataMining Knowl Discov*, 2011, **1**, 14–23.
 - 33 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
 - 34 A. Ng and M. Jordan, *Adv. Neural Inf. Process. Syst.*, 2002, 841–848.
 - 35 R. Gentleman, *R Programming for Bioinformatics*, Chapman and Hall/CRC, New

6) MODELLING SYSTEMS OF METAL OXIDE NANOPARTICLES AND VITAMIN DERIVATIVES

- York, 2008.
- 36 R Core Team, 2017.
- 37 A. Speck-Planche, V. V. Kleandrova, F. Luan, H. González-Díaz, J.-M. Ruso and M. N. Cordeiro, *Environ. Sci. Technol.*, 2014, **48**, 14686–14694.
- 38 J. Mitchell, *WIREs Comput Mol Sci* 2014, 2014, **4**, 468–481.
- 39 B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*, CRC press, 1994.
- 40 D. Kang, X. Pang, W. Lian, L. Xu, J. Wang, H. Jia, B. Zhang, A. L. Liu and G. H. Du, *RSC Adv.*, 2018, **8**, 5286–5297.

*You can away from people who try to
belittle your ambitions*

Mark Twain

CHAPTER

7

7) Modelling systems DVRNs (Multiplicative operators)

In this chapter, we explore the design of DVRNs, new nanosystems that are specifically designed for cancer treatment. If we are able to better desing these nanosystems, we also would do significant steps in material science knowledgment.

To do so, we develop a model able to predict a multi output and multi input model able to predict biological activities of the components of nanosystems conformed by DVRNs. We apply the PTML methodology by following the workflow included in Figure 8.

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

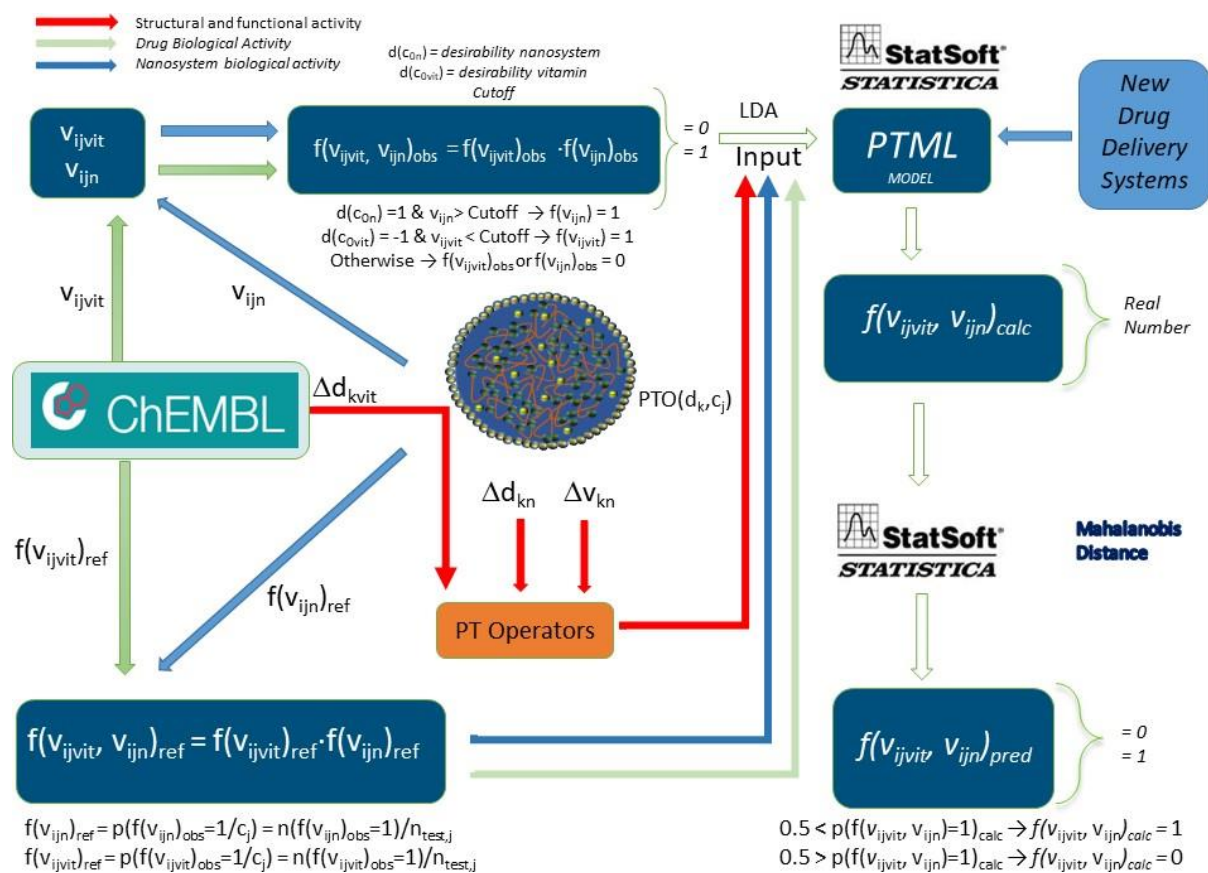


Figure 8. Detailed workflow to build a PTML model used in this work

Designing Nanoparticle Release Systems for Drug-Vitamin Cancer Co-Therapy with Multiplicative Perturbation-Theory Machine Learning (PTML) Models

Ricardo Santana^{*,a,b,c}, Robin Zuluaga^c, Piedad Gañán^d, Sonia Arrasate^e,
Enrique Onieva^a, and Humbert González-Díaz^{*,e,f}

*Corresponding authors

^aDeustoTech-Fundation Deusto, Bilbao, Spain.
E-mail: ricardo.santana@opendeusto.es (R.S.)

^bNew Materials Research Group, Universidad Pontificia Bolivariana UPB, Medellín, Colombia.

^cAgroindustrial Engineering College, Universidad Pontificia Bolivariana, Medellín, Colombia.

^dChemical Engineering College, Universidad Pontificia Bolivariana, Medellín, Colombia.

^eUniversity of Basque Country UPV/EHU, Leioa, Spain.

^fIKERBASQUE, Basque Foundation for Science, Bilbao, Spain.
E-mail: humberto.gonzalezdiaz@ehu.es (H.G.-D.)

Abstract

Nano-systems for cancer co-therapy including vitamins or vitamins derivatives have showed adequate results to continue with further researches to better understand them. However, the number of different combinations of drugs, vitamins, nanoparticle types, coating agents, synthesis conditions, system types (nanocapsules, micelles, *etc.*) to be tested is very large generating a high cost in experimentations. In this context, there are reports of large datasets of preclinical assays of compounds (like in ChEMBL database) and increasing but yet limited reports of experimental measurements of nano-systems *per se*. On the other hand, Machine Learning is gaining momentum in Nanotechnology and Pharmaceutical Sciences as a tool for rational design of new drugs and drug-release nano-systems. In this work, we propose to combine Perturbation Theory principles and Machine Learning to develop a PTML model for rational selection of the components of cancer co-therapy drug-vitamin release nano-systems

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

(DVRNs). In so doing, we apply information fusion techniques with 2 data sets: (1) a large ChEMBL dataset of >36000 preclinical assays of vitamin derivatives and a new dataset of >1000 outcomes of DVRNs, collected herein from literature for the first time. The ChEMBL dataset used covers a considerable number of assay conditions (c_{jvit}) each one with multiple levels. These conditions included >504 biological activity parameters (c_{0vit}), >340 types of proteins (c_{1vit}), >650 types of cells (c_{2vit}), >120 assay organisms (c_{3vit}), > 60 assay strain (c_{4vit}). Regarding the DVRNs, there are 25 different types of nano-systems (n_{jn}), with up to 16 conditions (c_{jn}) including also different levels such as: 8 biological activity parameters (c_{0n}), 9 raw nanomaterials (c_{4n}), 15 assay cells (c_{11n}), *etc.* In a first stage, we used Moving Average operators to quantify the perturbations (deviations) in all input variables with respect to the conditions. After that, we used multiplicative PT operators to carry out data fusion, and dimensions reduction, and Linear Discriminant Analysis (LDA) to seek the PTML model. The best PTML model found showed values of Specificity, Sensitivity, and Accuracy in the range of 83-88% in training and external validation series for >130000 cases (DVRNs *vs.* ChEMBL data pairs) formed after data fusion. Until the best of our knowledge, this is the first general purpose model for the rational design of DVRNs for cancer co-therapy.

Keywords: ChEMBL; Nanoparticle; PTML; Machine Learning; Big data; Multi-target models.

Introduction

Machine Learning (ML) techniques have been widely applied in nanotechnology field given the capacity to accelerate the production of valuable scientific knowledge. Computational methods used in nanoscience give us the opportunity to solve questions and propose new perspectives to be shared in the research community. There have been proposals to tackle difficulties for the construction of *in silico* models, as lack of experimental data. For instance, Yan *et al.*¹ proposed a workflow to profile nanoparticles virtually by constructing a virtual gold nanoparticle library and developing novel universal nanodescriptors or Sizochenko *et al.*² presented a method to predict toxicity towards different species, by solving the problem of missing data. Thanks to these studies among others,^{3,4} we have the opportunity to build multipurpose models with high accuracy for expanding current frontiers of nanosciences. There are models for material design and properties discovery that constitute this kind of advance. For instance, Endo *et al.*⁵ were able to detect molecular behavior of systems by using deep

neural network algorithm, Epa *et al.*⁶ built models of cellular uptake and apoptosis induced by nanoparticles for different cell types, Ekins *et al.*⁷ explained the positive aspects of applying ML techniques in a process assessment in drug discovery and development fields, as a network expressing interactions and perturbations or Sato *et al.*⁸ proposed novel predictive model for the diagnosis of hepatocellular carcinoma and reduced the misclassification rate by about half compared with a single tumor marker.

Furthermore, for the specific area of model development of drug delivery systems, we find significant advances, for instance Hathout *et al.*⁹ presented a model able to predict the mass of loaded drugs in solid lipid nanoparticles; Hashad *et al.*¹⁰ applied artificial neural networks to optimize the process of development of Chitosan-tripolyphosphate nanoparticles to obtain nanocarrier systems given the capacity of good encapsulation; Youshia *et al.*¹¹ built a model by using ML algorithms known as Artificial Neural Networks (ANN) to predict particle size and polydispersity of polymeric nanoparticles, given the biopharmaceutical behavior for different therapeutic functions; Parikh *et al.*¹² proposed a model by applying neural networks to optimize the parameters that affect the size of self-emulsifying drug delivery system. However, there are no reports of models with the novelty of predicting different biological activities, by processing heterogeneous data for this type of nanosystems. Probably, because this data is not public available, it need to be extracted from hundreds of assays with different conditions and also due to the difficulties to process such a complex data with ML algorithms. In this context, this type of study could be of high relevance.

On the other hand, the design of drug-vitamin release nano-systems (DVRNs) with improved drug release ratios, enhanced drug resistance properties, and expressing less toxicity is an emerging challenge for nanosciences. Zhu *et al.*¹³ presented a method to prepare DVRNs of porous PLGA nanoparticles to co-deliver vitamin E TPGS and docetaxel. Othayoth *et al.*¹⁴ shared a method to prepare DVRNs of vitamin–cisplatin-loaded chitosan nano-particles for chemoprevention and cancer fatigue. Wang *et al.*¹⁵ showed the controlled process for designing DVRNs of vitamin E TPGS-functionalized PLGA nanoparticles for delivery of paclitaxel and the antitumor properties, among others.^{16,17}

However, the number of different combinations of drugs, vitamins, nanoparticle types, coating agents, synthesis conditions, system types (nanocapsules, micelles, *etc.*) to be tested in the design of new DVRNs is very large generating a high cost in experimentations. In this context, there are reports of large datasets of preclinical assays of compounds (like in ChEMBL

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

database)^{18,19} and increasing but yet limited reports of experimental measurements of nano-systems *per se*.

In the present work, we propose the combination of the fundamentals of Perturbation Theory (PT) and Multi-Label Machine Learning methods (PTML models) as a solution for this kind of data.²⁰⁻²⁴ The PTML models have been created in different disciplines to be able to predict the biological activity of nps.^{25,26} In any case, there has not been reports of PTML models able to fusion data²⁷ of assay of compounds with nano-systems assays data to predict new DVRNs. In this work, we have created the first benchmark dataset for the study of DVRNs with data collected from literature. Next, we fused this dataset with a large dataset of preclinical assays of compounds downloaded from ChEMBL database. We used Moving Average (MAs) operators to express the perturbations in the assays, and multiplicative PT operators (PTOs) to carry out data fusion, and dimensions reduction. Last, Linear Discriminant Analysis (LDA) algorithm allowed us to seek the PTML model. We must highlight that this study is the first general purpose PTML model for the computational selection of the components of DVRNs. This model is a useful tool to complement information or event to guide researchers for the development of new nanosystems. We must highlight that all experimental tests must be conducted and carefully designed with no exception. All the sources must be checked: *in silico* models along with experimental data. In **Fig. 1**, we summarize the steps we are going to give in this work to train and validate the model.

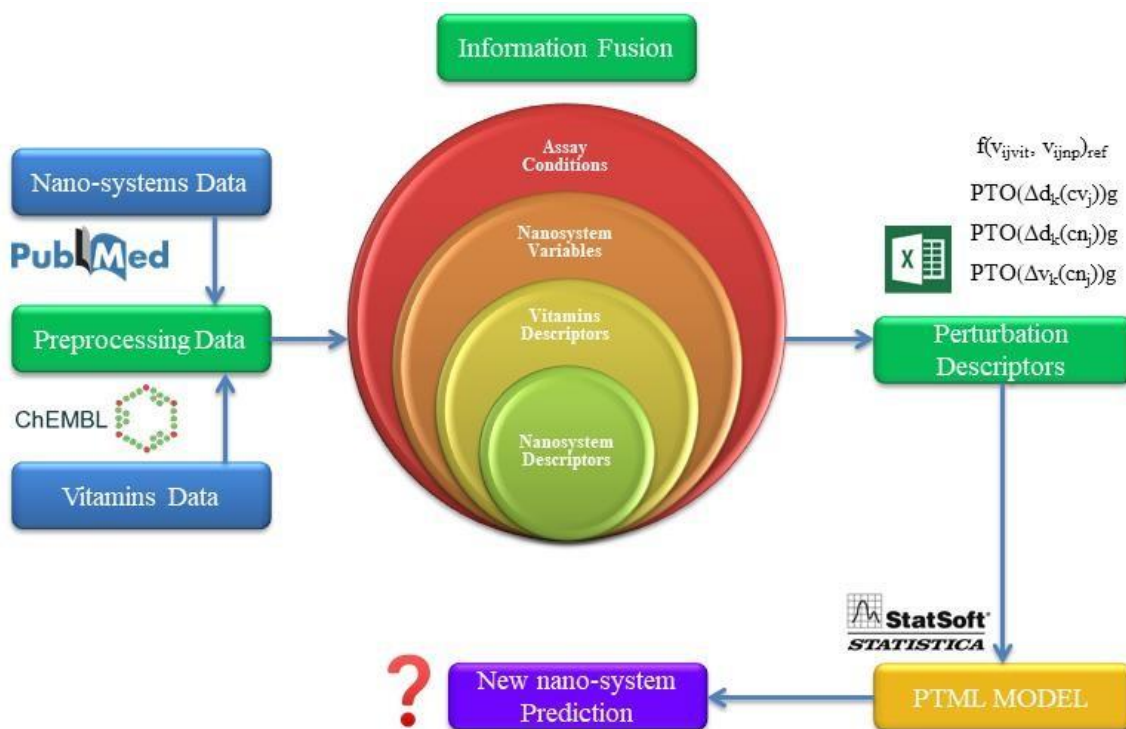


Fig. 1. Workflow to develop a PTML model for DVRNs design.

Materials and methods

Nano-systems dataset

Here we compiled from the first time a data set of 1348 outcomes for assays of DVRNs collected from literature^{13–17,28–32}. We formed these examples after applying different cutoffs to the experimental values reported in the literature (see next section). The search focused on four different types of DVRNs: 1) Emulsions (**Fig. 2A**), 2) Polymer conjugates (**Fig. 2B**), 3) Polymer micelles (**Fig. 2C**) and 4) Polymer particles (**Fig. 2D**). The data included molecular descriptors of the nano-system: d_{1n} = Nanoparticle Size (dimension 1), d_{2n} = Nanoparticle Size (dimension 2), d_{3n} = Zeta Potential, d_{4n} = Polydispersity Index and d_{5n} = Molecular Weight. Besides, we find variables that could affect to DVRNs properties values: v_0 = Cutoff of nano-system biological activity value, v_1 = Concentration of vitamin to synthesize the nano-system, v_2 = Concentration of nanoparticle to synthesize the nano-system, v_3 = Concentration of nano-system applied to the assay and v_4 = Assay time. The data set also includes 16 different assay conditions (see detailed information in **Table S1**, supporting information file SI00.doc). The main conditions related to the DVRNs are c_{n0} = Biological activity, c_{n1} = Drug included in the nano-system, c_{n2} = Vitamin included in the nano-system, c_{n3} = Nano-system shape, c_{n4} =

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

Nanoparticle core material, c_{n5} = Nano-system type. Other conditions related to the synthesis of the DVRNs are c_{n6} = Nanomaterial synthesis method, c_{n7} = Nano-system synthesis method, c_{n8} = Nano-system synthesis solvent, c_{n9} = Nanomaterial synthesis solvent. Last, conditions related to the assay of the DVRNs are c_{n10} = Assay organism, c_{n11} = Assay cell, c_{n12} = Assay protein (only albumin included in the data set), c_{n13} = Assay solution/solvent, c_{n14} = Assay pH, c_{n15} = Type of assay.

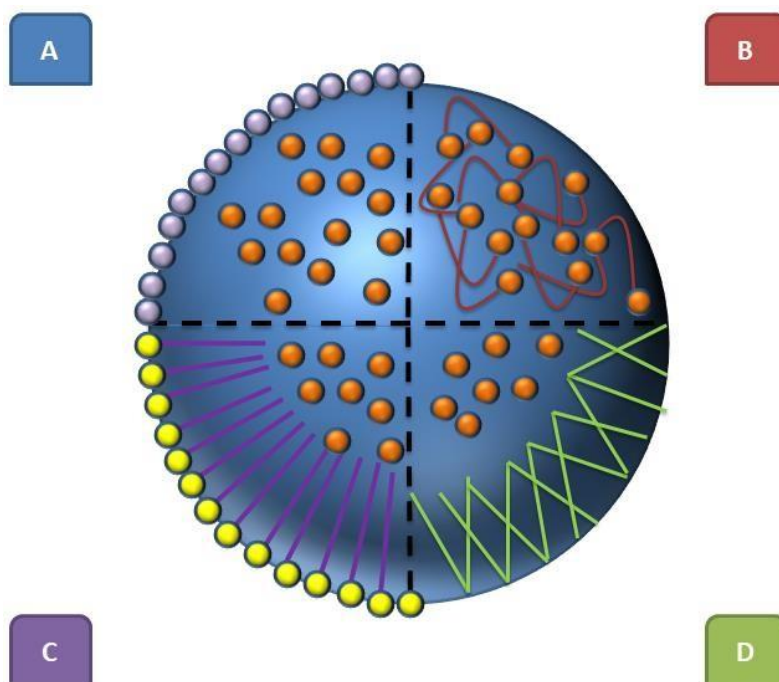


Fig. 2. General scheme of DVRNs studied in this work

Nano-systems data pre-processing

Each nano-system preclinical assay presents a result of the value v_{ijn} of the biological activity that the i^{th} nano-system presents against a j^{th} target. The MAs of the input variables have been calculated for all descriptors and variables with all the conditions c_{jn} , obtaining $\Delta d_k(c_{jn}) = d_{ki} - \langle d_k(c_{jn}) \rangle$ for descriptors of the nano-systems, and $\Delta v_k(c_{jn}) = v_{ki} - \langle v_k(c_{jn}) \rangle$ for variables of nano-systems. Same process of discretization is followed for nano-systems: construction of $f(v_{ijn})_{\text{obs}}$, depending on the desirability of the biological activity parameter considered in the nano-system assay $d(c_{0n})$ and cutoff, that in this case is also considered a variable as mentioned above (v_0). Consequently $f(v_{ijn})_{\text{obs}} = 1$ if desirability of the biological activity parameter $d(c_{0n}) = 1$ and $v_{ijn} > \text{cutoff}$. Besides, $f(v_{ijn})_{\text{obs}} = 1$ also when $v_{ijn} < \text{cutoff}$ and $d(c_{0n}) = 0$; otherwise, $f(v_{ijn})_{\text{obs}} = 0$. Each experimental value was confronted on average to 1-3 (top, medium, bottom) cutoff levels generating 1-3 different values of the discrete variable $f(v_{ijn})_{\text{obs}}$.

ChEMBL drug dataset

The data set for vitamins derivatives has been extracted from public database ChEMBL (February, 2019). This dataset is maintained by the European Bioinformatics Institute (EBI). The data is extracted directly from the literature: There are 7 core journals: Bioorganic & Medicinal Chemistry Letters, Journal of Medicinal Chemistry, Bioorganic & Medicinal Chemistry, Journal of Natural Products, European Journal of Medicinal Chemistry, ACS Medicinal Chemistry Letters and MedChemComm. After extracting the data, a manual curation process is applied. Moreover, the data is updated regularly every 3-4 months. This data set includes >36000 preclinical assays of drugs like vitamins, vitamins derivatives and molecules with at least > 80% of structural similarity. This is the reason we will refer to drugs as vit in formulations. Molecular descriptors for the different drugs are: d_{1vit} = LOPG (n-Octanol/Water Partition Coefficient) and d_{2vit} = Polar Surface Area (PSA). Besides, every case presents assay conditions c_j . These conditions have multiple levels including >504 biological activity parameters (c_{0vit}), >340 types of proteins (c_{1vit}), >650 types of cells (c_{2vit}), >120 assay organisms (c_{3vit}), > 60 assay strain (c_{4vit}). See detailed information in **Table S2**, supporting information file SI00.doc.

Drug data pre-processing

The value of biological activity is represented by v_{ijvit} given that it varies according to the descriptors of each drug and the combination of assay conditions $c_{jvit} = (c_{0vit}, c_{1vit}, c_{2vit}, c_{3vit}, c_{4vit})$. Values v_{ijvit} are the quantitative results of every drug assay. For instance, for a determined assay, we have as a quantitative result. This situation presents a challenge, because they are presented in different units. Besides, every case measures a different biological performance such as inhibition or cumulative release, among others. First of all, we must calculate the Perturbation Operators, in this case Moving Averages (MA). These MA are one-condition Moving Averages (MA), in other words, calculated for one condition at time. Thus, we are able to calculate the PT operators for descriptors of vitamins $\Delta d_k(c_{jvit}) = d_{ki} - \langle d_k(c_{jvit}) \rangle$. The PT operators measure the deviation of d_{ki} from the average value $\langle d_k(c_j) \rangle$ of the assays with the same condition.²⁰⁻²⁴ We need to create perturbation operators to develop the PTML model, so we discretized v_{ijvit} ²⁰⁻²⁴ obtaining as result $f(v_{ijvit})_{obs}$, to be able to know if a specific biological value is desirable or not. Since, $f(v_{ijvit})_{obs}$ refers to a function for observed results (obs). To develop $f(v_{ijvit})_{obs}$, we need to take into consideration two parameters: The first necessary parameter is the desirability of the biological activity $d(c_{0vit})$. This is a value associated to the

7) MODELLING SYSTEMS DVNRs (MULTIPLICATIVE OPERATORS)

biological activity and can be 1 or -1: $d(c_{0vit}) = 1$ if we consider biological activity values are more desirable if it increases; otherwise $d(c_{0vit}) = 0$. The second parameter is the cutoff, which is a limit of the biological activity values that separates adequate results. The cutoff = 100 for properties with units in nM. We applied this cutoff in order to be aligned with previous studies that applied PTML method with heterogeneous data.²² If not, cutoff = $\langle v_{ijvit} \rangle$; which is the average as expected value. Once we know $d(c_{0vit})$ and cutoff, we are able to build $f(v_{ijvit})_{obs}$ as follows: $f(v_{ijvit})_{obs} = 1$ if desirability of the biological activity parameter $d(c_{0vit}) = 1$ and $v_{ijvit} > \text{cutoff}$. Besides, $f(v_{ijvit})_{obs} = 1$ also when $v_{ijvit} < \text{cutoff}$ and $d(c_{0vit}) = 0$; otherwise, $f(v_{ijvit})_{obs} = 0$. See graphical schematization of this process in **Fig. 3**.

DVNRs data fusion and PTML model

To develop this PTML model, we generated a working data set resulting from the fusion of the two previous data sets. This data set included 1348 DVNRs from literature and >36000 ChEMBL vitamin derivatives assays. The process of fusion includes not only descriptors (d_{1vit} , d_{2vit} , d_{1n} , d_{2n} , d_{3n} , d_{4n} , d_{5n}) and variables (v_0 , v_1 , v_2 , v_3 , v_4) but also assay conditions (c_{jvit} and c_{jn}). Once we have built the observed function $f(v_{ijvit}, v_{ijn})_{obs}$. So $f(v_{ijvit}, v_{ijn})_{obs} = 1$ if $f(v_{ijn}) = 1$ and $f(v_{ijvit}) = 1$, otherwise $f(v_{ijvit}, v_{ijn})_{obs} = 0$. this will be the class to predict. The inputs will be the reference function $f(v_{ijvit}, v_{ijn})_{ref}$ and the perturbation operators for variables and descriptors of the nanosystem along with the descriptors of the vitamin derivative. We apply Linear Discriminant Analysis to build the model. If we use the model, we will have as result a real number. This real number must be converted by calculating Mahalanobi's distance into 0 or 1. Graphical summary of this process is presented in **Fig. 3**.

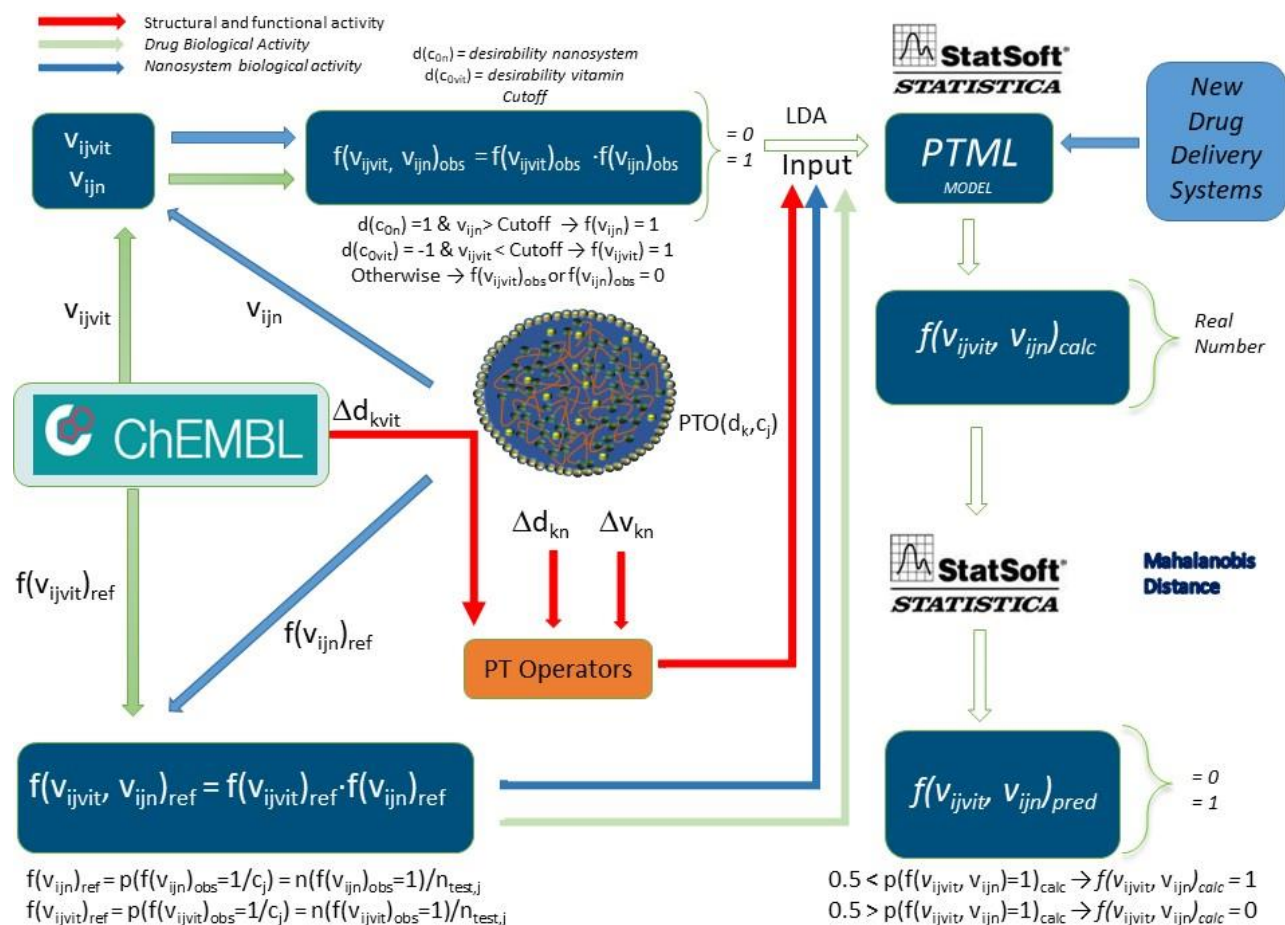


Fig. 3. Detailed workflow to build a PTML model used in this work.

After fusion, the working dataset includes 134901 pairs of DVNRs and vitamin derivatives. This is the result of repeating 100 times the data set of DVNRs but pairing each case with different assays of vitamin derivatives selected at random, see **Table 1**. The generation of possible combinations and the data fusion permit us to have an even larger dataset to deal with. It is possible that in the literature, from all the sources mentioned, some study published unreasonable data. However, the use of Machine Learning minimizes the errors of all the observations, of the whole set. This gives us the opportunity to extract knowledge from the available data nowadays, and to have more information about possible efficient drugs. We also applied a discretization for the pairs generated $f(v_{ijvit}, v_{ijn})_{obs}$.^{20–24} Following same discretization process mentioned above $f(v_{ijvit}, v_{ijn})_{obs} = 1$ when $f(v_{ijvit})_{obs} = 1$ and $f(v_{ijn})_{obs} = 1$; $f(v_{ijvit}, v_{ijn})_{obs} = 0$ otherwise.

Table 1. DVNRs and Vitamins MA operator's information

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

Condition Name	Code	Symbol	Operator Formula	Description
Activity type	c_{0vit}	$f(v_{ijvit})_{ref}$	$n(f(v_{ijvit})_{obs}=1)/n_j$	Expected probability $p(f(v_{ijvit})=1)_{ref}$ for the activity v_{ijvit} of type c_{0vit}
Activity type	c_{0n}	$f(v_{ijn})_{ref}$	$n(f(v_{ijn})_{obs}=1)/n_j$	Expected probability $p(f(v_{ijn})=1)_{ref}$ for the activity v_{ijn} of type c_{0n}
Activity type	c_{0vit}, c_{0n}	$f(v_{ijvit}, v_{ijn})_{ref}$	$f(v_{ijvit})_{ref} \cdot f(v_{ijn})_{ref}$	Reference value of probability for the nano-system
Activity type	c_{0vit}	$\Delta d_{1vit}(c_{0vit})$	$d_{1vit i} - \langle d_{1}(c_{0vit}) \rangle$	Deviation (Δ) of $d_{1vit} = AlogP_i$ and $d_{2vit} = PSA_i$ of the i^{th} vitamin derivative from their reference values $\langle AlogP(c_{jvit}) \rangle$ and $\langle PSA(c_{jvit}) \rangle$ respectively for a given subset of multiple assay conditions c_{jvit}
		$\Delta d_{2vit}(c_{0vit})$	$d_{2vit i} - \langle d_{2}(c_{0vit}) \rangle$	
Protein	c_{1vit}	$\Delta d_{1vit}(c_{1vit})$	$d_{1vit i} - \langle d_{1}(c_{1vit}) \rangle$	
		$\Delta d_{2vit}(c_{2vit})$	$d_{2vit i} - \langle d_{2}(c_{1vit}) \rangle$	
Cell Name	c_{2vit}	$\Delta d_{1vit}(c_{2vit})$	$d_{1vit i} - \langle d_{1}(c_{2vit}) \rangle$	
		$\Delta d_{2vit}(c_{2vit})$	$d_{2vit i} - \langle d_{2}(c_{2vit}) \rangle$	
Assay Organism	c_{3vit}	$\Delta d_{1vit}(c_{3vit})$	$d_{1vit i} - \langle d_{1}(c_{3vit}) \rangle$	
		$\Delta d_{2vit}(c_{3vit})$	$d_{2vit i} - \langle d_{2}(c_{3vit}) \rangle$	
Assay Strain	c_{4vit}	$\Delta d_{1vit}(c_{4vit})$	$d_{1vit i} - \langle d_{1}(c_{4vit}) \rangle$	
		$\Delta d_{2vit}(c_{4vit})$	$d_{2vit i} - \langle d_{2}(c_{4vit}) \rangle$	
Activity type	c_{0n}	$\Delta d_{1n}(c_{0n}), \Delta d_{2n}(c_{0n}) \dots \Delta d_{in}(c_{0n})$	$d_{in i} - \langle d_{in}(c_{0n}) \rangle$	Measures the deviation of the reference d_{in} value (average) of all np_i with the same $c_{0n}, c_{1n}, c_{2n}, c_{3n}, c_{4n}$ and c_{5n} . Furthermore, Δd_{in} refers to values of all considered nano-system variables v_0-v_9 .
Drug/Drug Comb np	c_{1n}	$\Delta d_{1n}(c_{1n}), \Delta d_{2n}(c_{1n}) \dots \Delta d_{in}(c_{1n})$	$d_{in i} - \langle d_{in}(c_{1n}) \rangle$	
DVRNs Vitamin	c_{2n}	$\Delta d_{1n}(c_{2n}), \Delta d_{2n}(c_{2n}) \dots \Delta d_{in}(c_{2n})$	$d_{in i} - \langle d_{in}(c_{2n}) \rangle$	
DVRNs Shape	c_{3n}	$\Delta d_{1n}(c_{3n}), \Delta d_{2n}(c_{3n}) \dots \Delta d_{in}(c_{3n})$	$d_{in i} - \langle d_{in}(c_{3n}) \rangle$	
Core raw material	c_{4n}	$\Delta d_{1n}(c_{4n}), \Delta d_{2n}(c_{4n}) \dots \Delta d_{in}(c_{4n})$	$d_{in i} - \langle d_{in}(c_{4n}) \rangle$	
DVRNs System type	c_{5n}	$\Delta d_{1n}(c_{5n}), \Delta d_{2n}(c_{5n}) \dots \Delta d_{in}(c_{5n})$	$d_{in i} - \langle d_{in}(c_{5n}) \rangle$	
Method Nanomaterial synth	c_{6n}	$\Delta d_{1n}(c_{6n}), \Delta d_{2n}(c_{6n}) \dots \Delta d_{in}(c_{6n})$	$d_{in i} - \langle d_{in}(c_{6n}) \rangle$	
Method Drug-Nano-system	c_{7n}	$\Delta d_{1n}(c_{7n}), \Delta d_{2n}(c_{7n}) \dots \Delta d_{in}(c_{7n})$	$d_{in i} - \langle d_{in}(c_{7n}) \rangle$	
DVRNs Synthesis solvent	c_{8n}	$\Delta d_{1n}(c_{8n}), \Delta d_{2n}(c_{8n}) \dots \Delta d_{in}(c_{8n})$	$d_{in i} - \langle d_{in}(c_{8n}) \rangle$	
Nanoparticle synthesis solvent	c_{9n}	$\Delta d_{1n}(c_{9n}), \Delta d_{2n}(c_{9n}) \dots \Delta d_{in}(c_{9n})$	$d_{in i} - \langle d_{in}(c_{9n}) \rangle$	
DVRNs Assay Organism	c_{10n}	$\Delta d_{1n}(c_{10n}), \Delta d_{2n}(c_{10n}) \dots \Delta d_{in}(c_{10n})$	$d_{in i} - \langle d_{in}(c_{10n}) \rangle$	Measures the deviation of the reference D_{in} value (average) of all np_i with the same c_{6n}, c_{7n}, c_{8n} and c_{9n} . Furthermore, Δd_{in} refers to values of all considered nano-system variables v_0-v_9 .
DVRNs Assay Cell	c_{11n}	$\Delta d_{1n}(c_{11n}), \Delta d_{2n}(c_{11n}) \dots \Delta d_{in}(c_{11n})$	$d_{in i} - \langle d_{in}(c_{11n}) \rangle$	
Albumin	c_{12n}	$\Delta d_{1n}(c_{12n}), \Delta d_{2n}(c_{12n}) \dots \Delta d_{in}(c_{12n})$	$d_{in i} - \langle d_{in}(c_{12n}) \rangle$	
DVRNs Media Assay	c_{13n}	$\Delta d_{1n}(c_{13n}), \Delta d_{2n}(c_{13n}) \dots \Delta d_{in}(c_{13n})$	$d_{in i} - \langle d_{in}(c_{13n}) \rangle$	
DVRNs Assay pH	c_{14n}	$\Delta d_{1n}(c_{14n}), \Delta d_{2n}(c_{14n}) \dots \Delta d_{in}(c_{14n})$	$d_{in i} - \langle d_{in}(c_{14n}) \rangle$	
Type of Assay	c_{15n}	$\Delta d_{1n}(c_{15n}), \Delta d_{2n}(c_{15n}) \dots \Delta d_{in}(c_{15n})$	$d_{in i} - \langle d_{in}(c_{15n}) \rangle$	

^aThe input variables for vitamin derivatives structure are $d_{1vit}=AlogP$, $d_{2vit}=PSA$. ^bThe input variables for nano-system are d_{1n} = Nanoparticle Size (dimension 1), d_{2n} = Nanoparticle Size (dimension 2), d_{3n} = Zeta Potential, d_{4n} = Polydispersity Index and d_{5n} = Molecular Weight. Nano-system properties values: v_0 = cutoff of nano-system biological activity value, v_1 = Concentration of vitamin to synthesize the nano-system, v_2 = Concentration of nanoparticle to synthesize the nano-system, v_3 = Concentration of nano-system applied to the assay and v_4 = assay time. Nano-system assay conditions: c_{0n} = Biological activity, c_{1n} = Drug included in the nano-system, c_{2n} = Vitamin included in the nano-system, c_{3n} = Nano-system shape, c_{4n} = nanoparticle core material, c_{5n} = Nano-system type, c_{6n} = Nanomaterial synthesis method, c_{7n} = Nano-system synthesis method, c_{8n} = Nano-system synthesis solvent, c_{9n} = Nanomaterial synthesis solvent, c_{10n} = Assay organism, c_{11n} = Assay cell, c_{12n} = Assay protein (only albumin included in the data set), c_{13n} = Assay solution/solvent, c_{14n} = Assay pH, c_{15n} = Type of assay

As input variable of the PTML model we include the reference function $f(v_{ijvit}, v_{ijn})_{ref}$. This function accounts for the expected probability of activity of the DVRNs with the new vitamin derivative added. We must highlight that \mathbf{c}_{jvit} and \mathbf{c}_{jn} (with \mathbf{c} in boldface caption) are vectors of combinations of assay conditions unlike c_{jvit} and c_{jn} ; which refers to single assay conditions. We propose a linear PTML model in order to predict the more efficient/safe components (vitamin, drug, nanoparticle, etc.) of DVRNs for cancer co-therapy. By using Linear Discriminant Analysis (LDA) linear classification, PTML-LDA models can be developed with **Equation 1**. As we see the model is able to predict $f(v_{ijvit}, v_{ijn})_{calc}$. This function does not classify the new case because it is a scoring function for the vitamin-np pair in the combinatorial assay conditions. Consequently, the LDA algorithm must calculate the values of posterior probabilities $p(f(v_{ijvit}, v_{ijn})_{obs} = 1)_{pred}$ by applying the Mahalanobis's distance metric.³³

$$\begin{aligned}
 f(v_{ijvit}, v_{ijn})_{calc} = & a_0 + a_1 \cdot f(v_{jvit}, v_{jn})_{ref} + \sum_{g=1}^{gmax} a_{kj} \cdot PTO(D_{kvit}, \mathbf{c}_{jvit}) \\
 & + \sum_{g=1}^{gmax} b_{kj} \cdot PTO(D_{kn}, \mathbf{c}_{jn}) + \sum_{g=1}^{gmax} c_{kj} \cdot PTO(V_k, \mathbf{c}_{jn})
 \end{aligned} \tag{1}$$

Results and discussion

PTML linear model with simple MA

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

Once the algorithm calculates $p(f(v_{ijvit}, v_{ijn}) = 1)_{pred}$, a Boolean function is created depending on the value of $p(f(v_{ijvit}, v_{ijn}) = 1)_{pred}$: If $p(f(v_{ijvit}, v_{ijn}) = 1)_{pred} > 0.5$, $f(v_{ijvit}, v_{ijn})_{pred} = 1$; otherwise, $f(v_{ijvit}, v_{ijn})_{pred} = 0$. This function will be compared to the observed function to identify the S_n , S_p , and A_c : If $f(v_{ij}, v_{ijn})_{pred} = 1$ and $f(v_{ijvit}, v_{ijn})_{obs} = 1$ the case is properly classified.^{20–24} To apply a prediction of biological activity with the proposed model, we need to introduce in the model the new variables taking into consideration the average for the different assay conditions $\langle v_k(c_{jn}) \rangle$, $\langle d_k(c_{jn}) \rangle$ and $\langle d_k(c_{jvit}) \rangle$. In **Table 2** and **Table 3**, the model parameters for vitamins derivatives (c_{0vit}) and DVRNs (c_{0n}) and are included. Besides the cutoff, desirability and reference function.

Table 2. Model parameters for vitamin derivatives for c_{0vit}

Condition c_{0vit} ^a	$\langle d_1(c_{0vit}) \rangle$	$\langle d_2(c_{0vit}) \rangle$	$n_j(c_{0vit})$	$n_j(f(v_{ijvit})=1)_{obs}$	$p(f(v_{ijvit})=1)_{ref}$	cutoff	$d(c_{0vit})$
Potency (nM)	3.29	74.06	24750	104	0.004	100.00	-1
IC ₅₀ (nM)	4.24	63.20	1402	232	0.165	100.00	-1
Activity (%)	3.79	79.40	1079	56	0.052	186.79	1
Inhibition (%)	3.25	82.98	415	254	0.612	73.72	-1
EC ₅₀ (nM)	4.50	66.09	388	193	0.497	100.00	-1
Weight (g)	3.18	35.24	260	192	0.738	4.23	-1
Ratio (-)	5.44	63.26	259	253	0.977	46.59	-1
GI ₅₀ (nM)	3.81	38.24	258	2	0.008	100.00	-1
Ki (nM)	3.83	77.45	197	106	0.538	100.00	-1
Activity(mg/dl)	5.74	58.21	164	95	0.579	5.22	-1

^a Condition c_{0vit} = the type of activity parameter measured for vitamin derivatives

Table 3. Model parameters for nano-systems for c_{0n}

Condition c_{0n} ^a	Input parameters used to specify c_{0n}						
	n_j	$d(c_{0n})$	$\langle \text{cutoff} \rangle$	$p(f(v_{ijn})=1)_{ref}$	$\langle v_1(c_{0n}) \rangle$	$\langle v_2(c_{0n}) \rangle$	$\langle v_3(c_{0n}) \rangle$
Activity type							
Inhibition (%)	90	1	40	0.18	38.5	0.66	520.00
Inhibition Hem. (%)	72	1	3	0.40	38.5	0.66	62.50

Cumulative Rel. (%)	289	-1	57.83	0.55	17.905	15.35	136.56
Survival rate (%)	42	-1	11.66	0.26	36.52	5.23	181.87
Control (%)	24	-1	35.125	0.41	6.39	78.09	6.24
Cell viability (%)	416	-1	34.85	0.35	45.83	5.00	169.68
Tumor volume (mm ³)	308	-1	249.35	0.66	52.01	14.34	181.87
Weight augment. (g)	108	1	1.5	0.48	10	9.92	181.87
Condition c_{0n} ^a	Input parameters used to specify c_{0n}						
Activity type	< $v_4(c_{0n})$ >	< $v_5(c_{0n})$ >	< $v_6(c_{0n})$ >	< $v_7(c_{0n})$ >	< $v_8(c_{0n})$ >	< $v_9(c_{0n})$ >	
Inhibition (%)	159.67	159.67	0.00	24.97	0.23	3375.78	
Inhibition Hem. (%)	159.67	159.67	0.00	24.97	0.23	3375.78	
Cumulative Rel. (%)	102.68	107.87	75.49	1.02	0.17	3302.42	
Survival rate (%)	135.00	135.00	48.00	17.00	0.01	3375.78	
Control (%)	118.33	118.33	48.00	0.50	0.25	3375.78	
Cell viability (%)	77.66	78.14	57.46	-3.07	0.12	3426.74	
Tumor volume (mm ³)	98.51	98.51	9.44	-7.39	0.18	3375.78	
Weight augment. (g)	113.68	113.68	9.44	-7.77	0.19	3375.78	

^a Condition c_{0n} = the type of activity parameter measured for nano-systems; Inhibition Hem. (%) = Inhibition Hemolysis; Cumulative Rel. (%) = Cumulative Release (%); Weight augment. (g) = Body Weight augmentation (g).

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

Table 4 includes the values of the averages $\langle D_i(c_{jn}) \rangle$ for c_{1n} , c_{2n} and c_{3n} . We can see examples of how the reference values of variables and descriptors in this case, according to 3 different conditions c_{1n} , c_{2n} and c_{3n} . For instance, the expected value of Paclitaxel of v_2 is 8.14; however the expected value of v_2 for all the cases with c_{2n} is 12.23. Consequently, if the conditions change affect the result of the model for every nano-system. With regard to the reference we considered $p(f(v_{ijvit}, v_{ijn})_{obs}=1)_{ref} = p(f(v_{ijvit})_{obs}=1)_{ref} * p(f(v_{ijn})_{obs}=1)_{ref}$.

Table 4. Model nano-system parameters for c_{1n} , c_{2n} and c_{3n} .

c_{1n}	Parameters used to specify c_{1n}						
Drug	$n_j(c_{1n})$	$\langle v_0(c_{1n}) \rangle$	$\langle v_1(c_{1n}) \rangle$	$\langle v_2(c_{1n}) \rangle$	$\langle v_3(c_{1n}) \rangle$	$\langle d_4(c_{1n}) \rangle$	$\langle d_5(c_{1n}) \rangle$
Paclitaxel	390	40.19	34.74	8.14	203.70	42.67	42.67
Docetaxel	375	104.07	10.00	9.06	181.87	130.46	130.46
^a PAC.+5-FU							
FU	126	294.29	160.00	25.00	130.35	82.00	82.00
SN-38	120	20.00	36.52	7.36	5.50	175.00	175.00
c_{2n}	Parameters used to specify c_{2n}						
DVRNs Vitamin	$n_j(c_{2n})$	$\langle v_0(c_{2n}) \rangle$	$\langle v_1(c_{2n}) \rangle$	$\langle v_2(c_{2n}) \rangle$	$\langle v_3(c_{2n}) \rangle$	$\langle d_4(c_{2n}) \rangle$	$\langle d_5(c_{2n}) \rangle$
E (TPGS)	557	139.00	46.60	12.23	170.22	119.95	119.95
E	312	42.31	36.52	1.00	213.66	19.73	19.73
Biotin	120	20.00	36.52	7.36	5.50	175.00	175.00
D ₃	87	41.20	4.60	55.25	175.26	113.85	113.85
c_{3n}	Parameters used to specify c_{3n}						
Shape	$n_j(c_{3n})$	$\langle v_0(c_{3n}) \rangle$	$\langle v_1(c_{3n}) \rangle$	$\langle v_2(c_{3n}) \rangle$	$\langle v_3(c_{3n}) \rangle$	$\langle d_4(c_{3n}) \rangle$	$\langle d_5(c_{3n}) \rangle$
Spherical	1315	84.34	37.47	10.50	181.78	104.39	104.39
Rod	34	71.18	0.00	11.90	169.30	50.00	100.00

^aPAC.+5-FU = Paclitaxel+5-FU

This model lets us select compounds in nano-systems for design of nano-systems consisting of DVNRs when adding a new vitamin or vitamin derivative measured with different conditions. To this end, we have to substitute the reference probability of activity $p(f(v_{ijvit}, v_{ijn})_{obs}=1)_{ref}$ on the equation, because it varies according to the measured biological activity, *e.g.* IC₅₀(nM), CC₅₀(nM) and EC₅₀(nM), *etc.* Consequently, the model can predict as desirable the properties

of the compounds of the nano-systems with the new vitamin derivatives. Finally, we need to incorporate the new descriptors and variables of the nano-system.

As we mentioned above, PTML-LDA is a linear model able to predict, in this case biological activities of nano-systems by considering the reference function and the perturbation added to the system by the perturbation operators. Consequently, the **Equation 2** presents the first model proposed. It includes two types of input variables: The reference variable, which is a value function $f(v_{ijvit}, v_{ijn})_{ref}$ and the variables that add the perturbation to the reference value: Vitamins descriptors, nano-systems descriptors and nano-systems variables. The operators that add the perturbation are called Perturbation Theory Operators (PTO) and in these case are: $\Delta v_k(c_{jn})$, $\Delta d_k(c_{jn})$ and $\Delta d_k(c_{jvit})$. See details about these operators on **Table 2** and **Table 3** (see **Table S1** and **Table S2** respectively in supporting information for full data set).

$$\begin{aligned}
 f(v_{ijvit}, v_{ijn})_{calc} = & -6.38631 + 18.74427 \cdot f(v_{ijvit}, v_{ijn})_{ref} \\
 & + 0.00519 \cdot Dv_{0n(c0)} + 0.02117 \cdot Dv_{2n(c0)} \\
 & + 0.09580 \cdot Dv_{0n(c1)} + 0.01583 \cdot Dv_{5n(c1)} \\
 & - 0.00482 \cdot Dv_{6n(c3)} + 0.00190 \cdot Dv_{3n(c4)} \\
 & - 0.06166 \cdot Dv_{2n(c5)} + 9.43873 \cdot Dv_{4n(c6)} \\
 & - 0.07865 \cdot Dv_{0n(c8)} - 0.01589 \cdot Dv_{0n(c11)} \\
 & + 0.07736 \cdot Dv_{2n(c11)} - 0.01380 \cdot Dv_{4n(c11)} \\
 & - 4.97694 \cdot Dd_{4n(c11)} - 0.01034 \cdot Dv_{3n(c12)} \\
 & - 0.07237 \cdot Dv_{2n(c13)} + 0.00077 \cdot Dv_{3n(c13)} \\
 & + 0.00989 \cdot Dv_{3n(c15)} \\
 n = 89934 \quad Chi - sqr = 1 \quad p - level = 1
 \end{aligned} \tag{2}$$

Statistical parameters of the model are: n is the number of cases applied to train the model, χ^2 is the Chi-square statistics, and p is the p-level. In this work, the model showed adequate values of Sp = 88.57, Sn = 84.50, and Ac = 88.27 in with training data set. Similar values were showed for external validation, see **Table 5**. This model used forward stepwise strategy to choose variables with 20 steps, given that there are 20 assay conditions for vitamin assays and nano-systems assays. Given the importance of the assay conditions, we understand that all the conditions should be included in the equation. However no PTO related to vitamins are included in the model so no vitamin conditions are considered; neither c_{14n} , c_{2n} , c_{7n} , c_{9n} or c_{10n} . We increased the number of steps to 30 the model did not include PTO related to vitamins. Also, c_{7n} , c_{9n} and c_{12n} were excluded in this assay. In addition, p-level = 1 indicates that

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

statistically is not significant. Besides, the number of variables is substantial which is not desirable. At this point, we considered to create PTML models with multiplicative operators to be statistically significant and reduce the dimensions of the model. The aim for these new models were to include all the variables, descriptors and assay conditions information without reducing the accuracy (next section).

Table 5. Prediction Results of PTML model

Obs. Sets ^a	Stat. Param. ^b	Pred. Stat. ^c	Predicted sets		
			n_j	$f(v_{ijvit}, v_{ijn})_{pred} = 0$	$f(v_{ijvit}, v_{ijn})_{pred} = 1$
Training					
$f(v_{ijvit}, v_{ijn})_{obs} = 0$	Sp	88.57	83258	73744.00	9514.00
$f(v_{ijvit}, v_{ijn})_{obs} = 1$	Sn	84.50	6676	1035.00	5641.00
Total	Ac	88.27	89934	74779.00	15155.00
Validation					
$f(v_{ijvit}, v_{ijn})_{obs} = 0$	Sp	88.65	41690	36959	4731
$f(v_{ijvit}, v_{ijn})_{obs} = 1$	Sn	84.22	3276	517	2759
Total	Ac	88.33	44966	37476	7490

^a Obs. Sets = Observed sets, ^bStat. Param. = Statistical parameter, ^cPred. Stat. = Predicted statistics (%)

PTML models with multiplicative operators

A notable drawback of the previous model is the necessity of exploring a very high number of MA variables. We can try to sort this using a feature selection technique as we did above. However, another problem is that is mandatory to include at least 21 of these variables in the model. This is because we need to include in the model at least one variable for each one of the boundary conditions c_j defining the biological assay of the vitamin and the synthesis and biological assay of the np-based drug release system. Consequently, feature selection techniques seek very long models and often fail to include all the desired input vars. In this section we propose to carry out a reduction of dimensions for the original input data using different Perturbation-Theory Operators (PTOs), see **Table 6**. We employed multiplicative PTOs using as argument products of MA values. Using these PTOs we reduced the dimensions of the analysis from $N_{ma} = N_v \cdot N_c = 16 \cdot 10 = 160$ input variables (MA) to 5-6 variables (PTOs) depending on the partition of the group of variables (G) used. In the Table 8 we show the general formula for PTML models using the different classes of multiplicative operators. PT

Product (PTP) operator can be calculated as the direct products of different MA values whereas the PT Geometric Mean (PTG) operators are obtained as a further transformation of the PTPs into the respective roots of different orders. We denoted the PTP operators as $PTP(d_k, v_k, c_j)_g = PTP(\Delta d_{vk}(c_j), \Delta d_{nk}(c_j), \Delta v_k(c_j))_g$ and the PTG operators as $PTG(d_k, v_k, c_j)_g = PTG(\Delta d_{vk}(c_j), \Delta d_{nk}(c_j), \Delta v_k(c_j))_g$. This express the fact that the operators have been calculated using the following sequence of transformations: $v_k, d_{vk}, d_{nk} \Rightarrow \Delta v_k(c_j), \Delta d_{vk}(c_j), \text{ or } \Delta d_{nk}(c_j)$ (MA, mean-centered variable) $\Rightarrow PTP(\Delta v_k(c_j))_g = \Delta v_k(c_j) \cdot \Delta d_{vk}(c_j) \cdot \Delta d_{nk}(c_j)$ (product of MA values) $\Rightarrow PTG(\Delta v_k(c_j))_g = [\Delta v_k(c_j) \cdot \Delta d_{vk}(c_j) \cdot \Delta d_{nk}(c_j)]^{1/q}$ (root of higher order q). Take into consideration that q is the number of variables in the operator.

Table 6. PTML models proposed here using different operators

PTO	PTML Model Formula ^a	Eq.
General Model (PTO)	$f(v_{ijvit}, v_{ijn})_{calc} = a_0 + a_1 \cdot f(v_{jvit}, v_{jn})_{ref} + PTO(\Delta d_{kvit}(c_{jvit}))_g$ $+ \sum_{g=1}^{gmax} b_{kj} \cdot PTO(\Delta d_{kn}(c_{jn}))_g + \sum_{g=1}^{gmax} c_{kj} \cdot PTO(\Delta v_k(c_{jn}))_g$	(1)
Product (PTP)	$f(v_{ijvit}, v_{ijn})_{calc} = a_0 + a_1 \cdot f(v_{jvit}, v_{jn})_{ref} + \sum_{g=1}^{gmax} a_{kj} \cdot \{G[\Delta d_{kvit}(c_{jvit})^{\delta_{gj}}]\}$ $+ \sum_{g=1}^{gmax} b_{kj} \cdot \{G[\Delta d_{kn}(c_{jn})^{\delta_{gj}}]\} + \sum_{g=1}^{gmax} c_{kj} \cdot \{G[\Delta v_k(c_{jn})^{\delta_{gj}}]\}$	(2)
Geom. Mean (PTG)	$f(v_{vij}, v_{nij})_{calc} = a_0 + a_1 \cdot f(v_{jvit}, v_{jn})_{ref} + \sum_{g=1}^{gmax} a_{kj} \cdot \{G[\Delta d_{kvit}(c_{jvit})^{\delta_{gj}}]\}^{1/q}$ $+ \sum_{g=1}^{gmax} b_{kj} \cdot \{G[\Delta d_{kn}(c_{jn})^{\delta_{gj}}]\}^{1/r} + \sum_{g=1}^{gmax} c_{kj} \cdot \{G[\Delta v_k(c_{jn})^{\delta_{gj}}]\}^{1/s}$	(3)

^a The parameter g denotes one sub-set of the partition (G) of the group of input variables (MAs) transformed by the operator (PTO), $\delta_{gj} = 1$ when the MA for the condition c_j is included in the group of variables g affected by the operator.

All the PTML models obtained with the $PTP(v_k, c_j)_g$ operators presented high collinearity among the variables and very extreme values of the coefficients. The best model found using

7) MODELLING SYSTEMS DVNRs (MULTIPLICATIVE OPERATORS)

PTML-LDA algorithm was the model based on Geometric mean operators $PTG(v_k, c_j)_g$, see

Equation 7.

$$\begin{aligned}
 f(v_{vij}, v_{nij})_{calc} &= -6.11383 + 18.32699 \cdot f(v_{vj}, v_{nj})_{ref} - 0.01324 \cdot PTG(v_k, c_{vj})_{g1} - 0.00514 \\
 &\cdot PTG(v_k, c_{vj})_{g3} + 0.00132 \cdot PTG(v_k, c_{vj})_{g4} + 0.00183 \cdot PTG(v_k, c_{vj})_{g5} + 0.15420 \\
 &\cdot PTG(v_k, c_{vj})_{g6}
 \end{aligned} \tag{7}$$

$$N = 89934 \quad \text{Chi-sqr} = 27137.04 \quad \text{p-level} < 0.05$$

The results of training and external validation data sets are shown in **Table 7**. Compared to the previous model, this model has higher Chi-sqr and p-level < 0.05 which means that is statistically significant. The training subset consists of 89934 observations: 6676 cases whose values of biological activity are desirable and 83258 that are not (observed data). This model presents high Specificity ratio for training subset given that it is able to predict 87.4% of no desirable values. Furthermore, it provides, equally, a high ratio of Sensitivity, which is aligned with the main aim of the model, which is to discover and develop new efficient and convenient DVNRs. On the other hand, the model was validated by using the testing subset. The performance of the validation showed an excellent robustness of the model. The general accuracy for this subset was 88.33, which is slightly higher comparing to training subsest. In addition, the dimensions of this model are reduced and it presents only 6 input variables. In so doing, despite of the reduction of input variables, ratios of Specificity, Sensitivity, and Accuracy are in the same range than the previous model, of 83-88%. ”. In addition, **Table S3** depicts all details about each case, observed classification, predicted classification, input variables, experimental conditions, vitamin derivative, and nano-systems characteristics (see supporting information file SI01.xlsx).

Table 7. Prediction Results of PTML model

Obs.	Stat.	Pred.	Predicted sets
------	-------	-------	----------------

Sets ^a	Param. ^b	Stat. ^c	n _j	f(v _{ijvit} , v _{ijn}) _{pred} = 0	f(v _{ijvit} , v _{ijn}) _{pred} = 1
Training					
f(v _{ijvit} , v _{ijn}) _{obs} = 0	Sp	87.40	83258	72770	10488
f(v _{ijvit} , v _{ijn}) _{obs} = 1	Sn	83.21	6676	1121	5555
Total	Ac	87.09	89934		
Validation					
f(v _{ijvit} , v _{ijn}) _{obs} = 0	Sp	87.50	41690	36480	5210
f(v _{ijvit} , v _{ijn}) _{obs} = 1	Sn	83.30	3276	547	2729
Total	Ac	87.20	44966		

^a Obs. Sets = Observed sets, ^bStat. Param. = Statistical parameter, ^cPred. Stat. = Predicted statistics

This model was obtained using the $PTG(v_k, c_j)_g$ operators calculated according to partition G_2 . The different MA used here to calculate the $PTG(v_k, c_j)_g$ (as well as all PTOs) were assigned to different operators according to the **Table 8**. This table defines different partitions of the group of variables (G). For instance, in order to calculate the operators $PTP(v_k, c_j)_4$ (not in the model) and $PTG(v_k, c_j)_4$ (in the model) we need to go to the table in the section of the partition G_2 . Next, we need to get the variables in $g_4 = \Delta v_3(cnp_3), \Delta v_3(cnp_4), \Delta v_3(cnp_5)$ and carry out the product and geometric mean operations. After that, we can obtain: $PTP(v_k, c_j)_4 = \Delta v_3(cnp_3) \cdot \Delta v_3(cnp_4) \cdot \Delta v_3(cnp_5)$ and $PTG(v_k, c_j)_4 = [\Delta v_3(cnp_3) \cdot \Delta v_3(cnp_4) \cdot \Delta v_3(cnp_5)]^{1/3}$, respectively. Each partition was created *ad hoc* trying to minimize the number of MA variables used to calculate the operator. We guaranteed it by including MA variables $\Delta v_k(c_j)$ with a certain experimental relationship among the original variable v_k and the boundary condition c_j . For instance, for the partition G_1, G_2 , and G_3 , the sub-set g_1 and g_2 use the molecular only descriptors of the vitamin to calculate MA averages of conditions $c_{v0}, c_{v1}, c_{vit2}, c_{v3}$, and c_{v4} . The sub-set g_1 uses all MA values based on AlogP of the vitamin. The sub-set g_2 (not depicted in the table) uses the MA for the same conditions than g_1 but it is based only on the PSA of the vitamin. The sub-set g_3 includes information about $v_0 =$ cutoff of nanosystem biological activity value, $v_1 =$ Concentration of vitamin to synthesize the nanosystem, $v_2 =$ Concentration of nanoparticle to synthesize the nanosystem, $v_3 =$ Concentration of nanosystem applied to the assay. We related these variables with determined nanosystem characteristics and composition: $c_{0n} =$ Biological activity, $c_{1n} =$ Drug included in the nanosystem, $c_{2n} =$ Vitamin included in the nanosystem, $c_{3n} =$ Nanosystem shape, $c_{4n} =$ Nanoparticle core material, $c_{5n} =$ Nanosystem type. However, g_4 includes information of $d_{1n} =$ Nanoparticle Size (dimension 1), $d_{2n} =$ Nanoparticle Size (dimension 2) and conditions of synthesis of the nanosystem: $c_{6n} =$ Nanomaterial synthesis

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

method, c_{7n} = Nanosystem synthesis method, c_{8n} = Nanosystem synthesis solvent, c_{9n} = Nanomaterial synthesis solvent. Finally, the sub-set g_5 refers to v_4 = assay time and other characteristics of the nanosystem: d_{3n} = Zeta Potential, d_{4n} = Polydispersity Index and d_{5n} = Molecular Weight, taking into consideration c_{10n} = Assay organism, c_{11n} = Assay cell, c_{12n} = Assay protein (only albumin included in the data set), c_{13n} = Assay solution/solvent, c_{14n} = Assay pH, c_{15n} = Type of assay. After this classification, we create 2 variations (G_2 and G_3) of the first group of combinations. G_2 splits g_3 into 2 different sub-sets: On one hand: c_{0n} , c_{1n} , c_{2n} with v_0 , v_1 , v_2 ; on the other hand: c_{3n} , c_{4n} , c_{5n} with v_3 . Finally, G_3 takes into consideration v_0 and c_{0n} , for the last subset g_4 . Note that variables of the np have never been used to calculate MA of the vitamin. Conversely, the variables of the vitamin have never been used to calculate MA of the np system.

Table 8. Partitions of variables in different groups used to calculate the different operators

	Vit.	DVRNs				DVRNs synthesis		DVRNs assay			
G_1	g_1	g_3				g_4		g_5			
	d_{1vit}	v_0	v_1	v_2	v_3	d_{1n}	d_{2n}	v_4	d_{3n}	d_{4n}	d_{5n}
g_1	c_{0vit}	$\Delta d_1(c_{v0})$									
	c_{1vit}	$\Delta d_1(c_{v1})$									
	c_{2vit}	$\Delta d_1(c_{v2})$									
	c_{3vit}	$\Delta d_1(c_{v3})$									
	c_{4vit}	$\Delta d_1(c_{v4})$									
g_3	c_{0n}	$\Delta v_0(c_{0n})$									
	c_{1n}		$\Delta v_2(c_{1n})$								
	c_{2n}		$\Delta v_1(c_{2n})$								
	c_{3n}			$\Delta v_3(c_{3n})$							
	c_{4n}			$\Delta v_3(c_{4n})$							
c_{5n}			$\Delta v_2(c_{5n})$	$\Delta v_3(c_{5n})$							
g_4	c_{6n}					$\Delta d_{1n}(c_{6n})$	$\Delta d_{2n}(c_{6n})$				
	c_{7n}					$\Delta d_{1n}(c_{7n})$	$\Delta d_{2n}(c_{7n})$				
	c_{8n}					$\Delta d_{1n}(c_{8n})$	$\Delta d_{2n}(c_{8n})$				

C9n					$\Delta d_{1n}(c_{9n})$						
C10n										$\Delta d_{4n}(c_{10n})$	
C11n											$\Delta d_{5n}(c_{11n})$
g5 C12n										$\Delta d_{3n}(c_{12n})$	
C13n										$\Delta d_{3n}(c_{13n})$	$\Delta d_{4n}(c_{13n})$
C14n										$\Delta d_{3n}(c_{14n})$	
C15n									$\Delta v_4(c_{15n})$		
G2	g1	g3			g4	g5		g6			
	d _{1vit}	v0	v1	v2	v3	d _{1n}	d _{2n}	v4	d _{3n}	d _{4n}	d _{5n}
C0vit	$\Delta d_1(c_{v0})$										
C1vit	$\Delta d_1(c_{v1})$										
g1 C2vit	$\Delta d_1(c_{v2})$										
C3vit	$\Delta d_1(c_{v3})$										
C4vit	$\Delta d_1(c_{v4})$										
C0n		$\Delta v_0(c_{0n})$									
g3 C1n				$\Delta v_2(c_{1n})$							
C2n			$\Delta v_1(c_{2n})$								
C3n					$\Delta v_3(c_{3n})$						
g4 C4n					$\Delta v_3(c_{4n})$						
C5n					$\Delta v_3(c_{5n})$						
C6n						$\Delta d_5(c_{6n})$	$\Delta d_6(c_{6n})$				
g5 C7n						$\Delta d_5(c_{7n})$	$\Delta d_6(c_{7n})$				
C8n						$\Delta d_5(c_{8n})$	$\Delta d_6(c_{8n})$				
C9n						$\Delta d_5(c_{9n})$					
C10n										$\Delta d_{4n}(c_{10n})$	
C11n											$\Delta d_{5n}(c_{11n})$
g6 C12n										$\Delta d_{3n}(c_{12n})$	
C13n										$\Delta d_{3n}(c_{13n})$	$\Delta d_{4n}(c_{13n})$
C14n										$\Delta d_{3n}(c_{14n})$	
C15n								$\Delta v_4(c_{15n})$			

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

G_3	g_1	g_3				g_4		g_5			
	d_{1vit}	v_0	v_1	v_2	v_3	d_{1n}	d_{2n}	v_4	d_{3n}	d_{4n}	d_{5n}
g_1	c_{0vit}	$\Delta d_1(c_{v0})$									
	c_{1vit}	$\Delta d_1(c_{v1})$									
	c_{2vit}	$\Delta d_1(c_{v2})$									
	c_{3vit}	$\Delta d_1(c_{v3})$									
	c_{4vit}	$\Delta d_1(c_{v4})$									
g_2	c_{1n}			$\Delta v_2(c_{1n})$							
	c_{2n}		$\Delta v_1(c_{2n})$								
	c_{3n}				$\Delta v_3(c_{3n})$						
	c_{4n}				$\Delta v_3(c_{4n})$						
	c_{5n}				$\Delta v_3(c_{5n})$						
g_3	c_{6n}					$\Delta d_{1n}(c_{6n})$	$\Delta d_{2n}(c_{6n})$				
	c_{7n}					$\Delta d_{1n}(c_{7n})$	$\Delta d_{2n}(c_{7n})$				
	c_{8n}					$\Delta d_{1n}(c_{8n})$	$\Delta d_{2n}(c_{8n})$				
	c_{9n}					$\Delta d_{1n}(c_{9n})$					
g_4	c_{0n}	$\Delta v_0(c_{0n})$									
	c_{10n}								$\Delta d_{4n}(c_{10n})$		
	c_{11n}									$\Delta d_{5n}(c_{11n})$	
	c_{12n}								$\Delta d_{3n}(c_{12n})$		
	c_{13n}								$\Delta d_{3n}(c_{13n})$		
	c_{14n}								$\Delta d_{3n}(c_{14n})$		
	c_{15n}									$\Delta v_4(c_{15n})$	

^a $G = [g_1, g_2, g_3, \dots]$, partition of the group of MAs ($\square_{v_k}(c_j)$) of the input variables. ^aThe input variables for vitamin structure are $d_{1vit}=AlogP$, $d_{2vit}=PSA$. ^bThe input variables for nano-system are d_{1n} = Nanoparticle Size (dimension 1), d_{2n} = Nanoparticle Size (dimension 2), d_{3n} = Zeta Potential, d_{4n} = Polydispersity Index and d_{5n} = Molecular Weight. Nano-system properties values: v_0 = cutoff of nano-system biological activity value, v_1 = Concentration of vitamin to synthesize the nano-system, v_2 = Concentration of nanoparticle to synthesize the nano-system, v_3 = Concentration of nano-system applied to the assay and v_4 = assay time . Nano-system assay conditions: c_{0n} = Biological activity, c_{1n} = Drug included in the nano-system, c_{2n} = Vitamin included in the nano-system, c_{3n} = Nano-system shape, c_{4n} = Nanoparticle core material, c_{5n} = Nano-system type, c_{6n} = Nanomaterial synthesis method, c_{7n} = Nano-system synthesis method, c_{8n} = Nano-system synthesis solvent, c_{9n} = Nanomaterial

synthesis solvent, c_{10n} = Assay organism, c_{11n} = Assay cell, c_{12n} = Assay protein (only albumin included in the data set), c_{13n} = Assay solution/solvent, c_{14n} = Assay pH, c_{15n} = Type of assay

PTML simulation of selected DVRNs for cancer co-therapy

To show the potential practical uses of the presented model, we performed a simulation of selected DVRNs for cancer co-therapy. In order to keep it simple, we focused on two relevant parameters for biological activity (Inhibition %) and pharmaceutical function (Cumulative Release %). In fact, these two parameters are included in the dataset with 90 assays for Inhibition % and 289 assays Cumulative Release %. Using the PTML-LDA equation we calculated the values of the $f(v_{ij})_{\text{calc}}$ function using an excel sheet. After that, we transformed these values into a probability-like scale using the following equation: $p(f(v_{ij})=1)_j = [f(v_{ij})_{\text{calc}} - {}^*f(v_{ij})_{\text{max}}] / [{}^*f(v_{ij})_{\text{min}} - {}^*f(v_{ij})_{\text{max}}]$. In this equation, $f(v_{ij})_{\text{calc}}$ is the value of the scoring function for the nano-systems formed by the new drug (vitamin derivative) and a nano-systems (including a classic anticancer drug) measured in conditions c_j . The values ${}^*f(v_{ij})_{\text{max}} = f(v_{ij})_{\text{max}} + 10$ and ${}^*f(v_{ij})_{\text{min}} = f(v_{ij})_{\text{min}} - 10$. We present the values of $p(f(v_{ij})=1)_j$ for all drugs selected (Cisplatin, Paclitaxel, and Doxorubicin) for this study, see **Table 9**. These anticancer drugs are simulated with more or less efficacy, taking into consideration the vitamins or vitamins analogs included in the same nanosystem, the nanosystem material, and the specific measured property. Given that the model is multioutput, it depends on the biological activity we need to predict, the value of the perturbation operators will change. For instance, in terms of cumulative release %, the most adequate nanosystems for Cisplatin are PLGA based nanosystems. Besides, the efficacy will increase if we include vitamin C, the combination of vitamins B₁₂+C+D₃ or in the third place TPGS. We also can consider PLA-TPGS/TPGS-TOOH (PTTT) nanosystems for better cumulative release capacity. However, to improve the efficacy, we should include vitamins D₃ or B₁₂. For inhibition purposes, PLGA nanosystems, according to the model, will be more efficient to deliver cisplatin. As this table shows, the vitamins add a perturbation to the nanosystem taken as reference, to show us the different possibilities of the enriched nanosystems.

Table 9. PTML simulation of selected drug-vit cancer co-therapy nano-systems

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

NP system ^a	Drug ^b	Prop. ^c	B ₁₂	C	D ₃	TPGS	BIOTI N	E	B ₁₂ +C+D 3
Chitosan	CISP	I(%)	0.47 8	0.47 8	0.47 8	0.478	0.478	0.47 8	0.478
PLGA	CISP	I(%)	0.62 0	0.62 5	0.62 0	0.624	0.622	0.62 2	0.625
D ₃ +DGPP	CISP	I(%)	0.49 1	0.47 3	0.48 8	0.475	0.476	0.47 6	0.473
PTTT	CISP	I(%)	0.51 3	0.51 3	0.51 3	0.513	0.513	0.51 3	0.513
PLGA+TPGS	CISP	I(%)	0.48 2	0.48 2	0.48 2	0.482	0.482	0.48 2	0.482
PLGA+PEG	CISP	I(%)	0.48 5	0.48 9	0.48 5	0.488	0.488	0.48 8	0.489
CMMS	CISP	I(%)	0.48 8	0.49 3	0.48 9	0.492	0.491	0.49 1	0.493
Chitosan	CISP	CR(%))	0.48 9	0.47 9	0.47 9	0.479	0.479	0.47 9	0.479
PLGA	CISP	CR(%))	0.88 2	0.88 9	0.88 3	0.888	0.886	0.88 6	0.889
D ₃ +DGPP	CISP	CR(%))	0.49 3	0.47 1	0.49 0	0.474	0.475	0.47 5	0.471
PTTT	CISP	CR(%))	0.56 5	0.55 9	0.56 5	0.559	0.560	0.56 0	0.559
PLGA+TPGS	CISP	CR(%))	0.48 2	0.47 8	0.48 1	0.479	0.479	0.47 9	0.478
PLGA+PEG	CISP	CR(%))	0.48 5	0.48 5	0.48 6	0.487	0.487	0.48 7	0.487
CMMS	CISP	CR(%))	0.48 9	0.49 1	0.48 9	0.490	0.490	0.49 0	0.491

Chitosan	PTX	I(%)	0.479	0.482	0.480	0.481	0.481	0.481	0.482
PLGA	PTX	I(%)	0.618	0.627	0.618	0.625	0.622	0.622	0.627
D ₃ +DGPP	PTX	I(%)	0.491	0.473	0.488	0.475	0.476	0.476	0.473
PTTT	PTX	I(%)	0.513	0.513	0.513	0.513	0.513	0.513	0.513
PLGA+TPGS	PTX	I(%)	0.482	0.482	0.482	0.482	0.482	0.482	0.482
PLGA+PEG	PTX	I(%)	0.484	0.487	0.485	0.486	0.486	0.486	0.487
CMMS	PTX	I(%)	0.488	0.491	0.488	0.490	0.490	0.490	0.491
Chitosan	PTX	CR(%)	0.481	0.485	0.481	0.484	0.484	0.484	0.485
PLGA	PTX	CR(%)	0.881	0.891	0.881	0.889	0.886	0.886	0.891
D ₃ +DGPP	PTX	CR(%)	0.493	0.472	0.490	0.474	0.475	0.475	0.471
PTTT	PTX	CR(%)	0.562	0.565	0.562	0.565	0.565	0.565	0.565
PLGA+TPGS	PTX	CR(%)	0.483	0.481	0.482	0.482	0.482	0.482	0.481
PLGA+PEG	PTX	CR(%)	0.485	0.486	0.485	0.486	0.486	0.486	0.486
CMMS	PTX	CR(%)	0.488	0.490	0.489	0.490	0.490	0.490	0.490
Chitosan	DOX	I(%)	0.480	0.486	0.481	0.485	0.484	0.484	0.486
PLGA	DOX	I(%)	0.614	0.631	0.615	0.628	0.622	0.622	0.631

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

D ₃ +DGPP	DOX	I(%)	0.49 0	0.47 6	0.48 8	0.477	0.478	0.47 8	0.476
PTTT	DOX	I(%)	0.51 3	0.51 3	0.51 3	0.513	0.513	0.51 3	0.513
PLGA+TPGS	DOX	I(%)	0.48 2	0.48 2	0.48 2	0.482	0.482	0.48 2	0.482
PLGA+PEG	DOX	I(%)	0.48 1	0.47 4	0.47 9	0.475	0.476	0.47 6	0.473
CMMS	DOX	I(%)	0.48 4	0.47 7	0.48 3	0.479	0.479	0.47 9	0.477
Chitosan	DOX	CR(%))	0.48 2	0.49 1	0.48 4	0.489	0.488	0.48 8	0.491
PLGA	DOX	CR(%))	0.87 6	0.89 6	0.87 6	0.892	0.886	0.88 6	0.896
D ₃ +DGPP	DOX	CR(%))	0.49 2	0.47 4	0.48 9	0.477	0.477	0.47 7	0.474
PTTT	DOX	CR(%))	0.55 8	0.57 2	0.55 9	0.570	0.570	0.57 0	0.572
PLGA+TPGS	DOX	CR(%))	0.48 5	0.49 3	0.48 7	0.491	0.490	0.49 0	0.493
PLGA+PEG	DOX	CR(%))	0.48 3	0.48 0	0.48 3	0.481	0.481	0.48 1	0.480
CMMS	DOX	CR(%))	0.48 7	0.48 3	0.48 6	0.484	0.484	0.48 4	0.483

^aD₃+DGPP = D₃+1,2-Distearoyl-sn-glycero-3-phosphoethanolamine-N+L-a-phosphatidylcholine, CMMS = Carboxyl-modified mesoporous silica, PTTT: PLA-TPGS/TPGS-TOOH, ^bDOX = Doxorubicin, PTX = Paclitaxel, CISP = Cisplatin ^c Prop. = Property, I(%) = Inhibition (%), CR(%) = Cumulative Release (%).

In this research work, we proposed to combine Perturbation Theory principles and Machine Learning to develop a PTML model for rational selection of the components of cancer co-therapy drug-vitamin release nanosystems (DVRNs). The technique we applied was Linear

Discriminant Analysis given the excellent results in other PTML studies published. However, given the complexity of the new database, there are different assay conditions and characteristics of the nanosystem. We first present a model with so many dimensions and statistically no significant. However, this model presented a high ratio of Sensitivity, Specificity and Accuracy. The challenge with the new collected dataset was to reduce the dimensions selectively, taking into consideration desirable combinations of assay conditions and variables of the nanosystem as well as the vitamins analogs. We created 3 different grills with different combinations of assay conditions and variables. These subsets constituted the new references to generate multiplicative PT operators and Geometric-Mean-based Perturbation Operators. By applying this type of operators, we carry out the data fusion of nanosystems and vitamins. In doing so, the best model found showed high values of Specificity, Sensitivity, and Accuracy in the range of 83-88%. This model included Geometric-Mean-based Perturbation Operators. These operators provoked a remarkable dimension reduction. Until the best of our knowledge, this is the first general purpose model for the rational design of DVRNs for cancer co-therapy. Among the most adequate nanosystems in terms of cumulative release and inhibition for Cisplatin, Paclitaxel and Doxorubicin, the higher probability is presented by nanosystems composed by PLGA and vitamin C or vitamin complex B₁₂+C+D₃.

Supporting Information

The dataset used, including molecular descriptors, and assay conditions, desirability, cutoff, biological activities etc., was included in tables **Table S1**, **Table S2** (SI00.xlsx) and **Table S3** (SI001.xlsx). See details about these Moving Average operators on **Table 2** and **Table 3** (see **Table S1** and **Table S2** respectively in supporting information for full dataset consultation). In addition, **Table S3** we also included all details about each case, observed classification, predicted classification, input variables, experimental conditions, vitamin derivative and nano-system characteristics.

Acknowledgments

R.S.C. thanks COLCIENCIAS scholarship for the doctorate studies; “Convocatoria para Doctorado Nacional 757” from 2017. This original research is part of the project “Investigación en Derecho Internacional y Nanotecnología” registered in the Research Centre of Universidad

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

Pontificia Bolivariana with register number 766B-06/17-37. Special gratitude is extended to CYTED NANOCELIA network. The authors acknowledge research grants from Ministry of Economy and Competitiveness, MINECO, Spain (FEDER CTQ2016-74881-P) and Basque government (IT1045-16). The authors also acknowledge the support of Ikerbasque, Basque Foundation for Science. The authors also acknowledge the support of Ikerbasque, Basque Foundation for Science.

References

- 1 X. Yan, A. Sedykh, W. Wang, X. Zhao, B. Yan and H. Zhu, *Nanoscale*, 2019, **11**, 8352–8362.
- 2 N. Sizochenko, A. Mikolajczyk, K. Jagiello, T. Puzyn, J. Leszczynski and B. Rasulev, *Nanoscale*, 2018, **10**, 582–591.
- 3 L. Li, C. C. Koh, D. Reker, J. B. Brown, H. Wang, N. K. Lee, H. . Liow, H. Dai, H. M. Fan, L. Chen and D. Q. Wei, *Sci. Rep.*, 2019, **9**, 7703.
- 4 J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer and S. Zhao, *Nat. Rev. Drug Discov.*, 2019, **18**, 1474–1784.
- 5 K. Endo, D. Yuhara, K. Tomobeia and K. Yasuoka, *Nanoscale*, 2019, **11**, 10064–10071.
- 6 D. Epa, C., Burden, Tassa, C., Weissleder, R., Shaw, S., Winkler, *Nano Lett.*, 2012, **12**, 5808–5812.
- 7 S. Ekins, A. Puhl, K. Zorn, T. Lane, D. Russo, J. Klein, A. Hickey and A. Clark, *Nat. Mater.*, 2019, **18**, 435–441.
- 8 M. Sato, K. Morimoto, S. Kajihara, R. Tateishi, S. Shiina, K. Koike and Y. Yatomi, *Sci. Rep.*, 2019, **9**, 7704.
- 9 R. M. Hathout and A. A. Metwally, *Eur. J. Pharm. Biopharm.*, 2016, **108**, 262–268.
- 10 R. A. Hashad, R. A. H. Ishak, S. Fahmy and S. Mansour, *Int. J. Biol. Macromol.*, 2017, **86**, 50–58.
- 11 J. Youshia and M. E. Ali, *Eur. J. Pharm. Biopharm.*, 2017, **119**, 333–342.
- 12 K. J. Parikh and K. K. Sawant, *AAPS PharmSciTech*, 2018, **19**, 3311–3321.
- 13 H. Zhu, H. Chen, X. Zeng, Z. Wang, X. Zhang, Y. Wu, Y. Gao, J. Zhang, K. Liu, R. Liu, L. Cai, L. Mei and S. S. Feng, *Biomaterials*, 2014, **35**, 2391–2400.
- 14 R. Othayoth, P. Mathi, K. Bheemanapally, L. Kakarla and M. Botlagunta, *J. Microencapsul.*, 2015, **32**, 578–588.
- 15 G. Wang, B. Yu, Y. Wu, B. Huang, Y. Yuan and C. S. Liu, *Int. J. Pharm.*, 2013, **446**, 24–33.
- 16 S. Patil, S. Gawali, P. S. and S. Basu, *J. Mater. Chem.*, 2013, **1**, 5742–5750.
- 17 J. Zhao and S. S. Feng, *Biomaterials*, 2014, **35**, 3340–3347.
- 18 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S.

7) MODELLING SYSTEMS DVRNS (MULTIPLICATIVE OPERATORS)

- McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2011, **40**, 1100–1107.
- 19 A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit and A. R. Leach, *Nucleic Acids Res.*, 2016, **45**, 945–954.
- 20 H. González-Díaz, S. Arrasate, A. Gómez-San Juan, N. Sotomayor, E. Lete, L. Besada-Porto and J. Ruso, *Curr. Top. Med. Chem.*, 2013, **13**, 1713–1741.
- 21 S. Arrasate and A. Duardo-Sanchez, *Curr. Top. Med. Chem.*, 2018, **18**, 1203–1213.
- 22 L. Simón-Vidal, O. García-Calvo, U. Oteo, S. Arrasate, E. Lete, N. Sotomayor and H. González-Díaz, *J. Chem. Inf. Model.*, 2018, **58**, 1384–1396.
- 23 V. Blay, T. Yokoi and H. González-Díaz, *J. Chem. Inf. Model.*, 2018, **58**, 2414–2419.
- 24 J. F. Da Costa, D. Silva, O. Caamaño, J. M. Brea, M. I. Loza, C. R. Munteanu, A. Pazo, X. García-Mera and H. González-Díaz, *ACS Chem Neurosci*, 2018, **9**, 2572–2587.
- 25 F. Luan, V. V. Kleandrova, H. González-Díaz, J. M. Ruso, A. Melo, A. Speck-Planche and N. Cordeiro, *Nanoscale*, 2014, **6**, 10623–10630.
- 26 V. V. Kleandrova, F. Luan, H. González-Díaz, J. M. Ruso, A. Melo, A. Speck-Planche and N. M. Cordeiro, *Environ. Int.*, 2014, **73**, 288–294.
- 27 P. Willett, *J. Chem. Inf. Model.*, 2013, **53**, 1–10.
- 28 Z. B. Zhao, J. Long, Y. Y. Zhao, J. B. Yang, W. Jiang, Q. Z. Liu, K. Yan, L. Li, Y.-C. Wang and Z. X. Lian, *Biomater. Sci.*, 2018, **6**, 893–900.
- 29 M. Mehdizadeh, H. Rouhani, N. Sepehri, R. Varshochian, M. H. Ghahremani, M. Amini, M. Gharghabi, S. Ostad, F. Atyabi, A. Baharian and R. Dinarvand, *Artif. Cells, Nanomedicine Biotechnol.*, 2017, **45**, 495–504.
- 30 W. Liu, F. Wang, Y. Zhu, X. Li, X. Liu, J. Pang and W. Pan, *Molecules*, 2018, **23**, 3082.
- 31 Y. Ma, D. Liu, D. Wang, Y. Wang, Q. Fu, J. K. Fallon, X. Yang, Z. He and F. Liu, *Mol. Pharm.*, 2014, **11**, 2623–2630.
- 32 J. Lu, Y. Huang, W. Zhao, Y. Chen, J. Li, X. Gao, R. Venkataramanan and S. Li, *Mol. Pharm.*, 2013, **10**, 2880–2890.
- 33 S. Balakrishnama and A. Ganapathiraju, *Inst. Signal Inf. Process.*, 1998, **18**, 1–8.

*The mind that opens up to a new idea
never returns to its original size*

Albert Einstein

CHAPTER

8

8) Modelling DVRNs (Metric operators and enrichment of information)

In this chapter, we present a deeper exploration of the design of DVRNs. We are not using multiplicative PTO, but metric-based

To do so, we develop a model able to predict a multi output and multi input model able to predict biological activities of the components of nanosystems conformed by DVRNs. We apply the PTML methodology by following the workflow included in Figure 9.

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

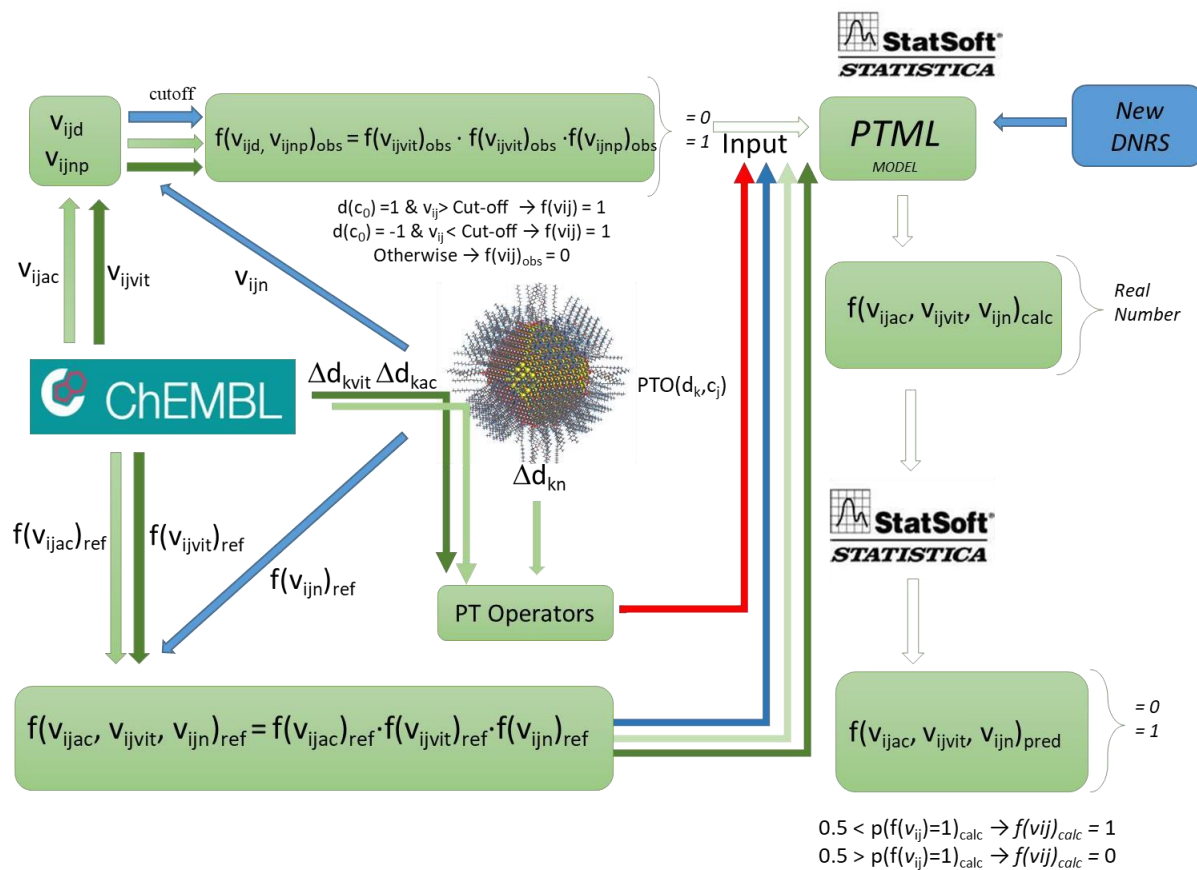


Figure 9. Workflow to build a PTML model used in this work

**PTML Model for Selection of Nanoparticle, Anticancer Drug, and Vitamin in the
Design of Drug-Vitamin Nanoparticle Release Systems for Cancer Co-Therapy**

Ricardo Santana ^{a,b,c}, Robin Zuluaga ^d, Piedad Gañán ^c, Sonia Arrasate ^e,

Enrique Onieva ^b, Matthew M. Montemore ^a, and Humbert González-Díaz ^{e,f,g,*}

^a *Department of Chemical and Biomolecular Engineering, Tulane University, 6823 St Charles Ave, New Orleans, United States.*

^b *University of Deusto, Avda. Universidades, 24, 48007 Bilbao, Spain.*

^c *Grupo de Investigación Sobre Nuevos Materiales, Facultad de Ingeniería Química, Universidad Pontificia Bolivariana, Circular 1° N° 70-01, Medellín, Colombia.*

^d *Facultad de Ingeniería Agroindustrial, Universidad Pontificia Bolivariana, Circular 1° N° 70-01, Medellín, Colombia.*

^e *Department of Organic Chemistry II, University of Basque Country UPV/EHU, 48940, Leioa, Basque Country, Spain.*

^f *Basque Center for Biophysics, Spanish National Research Council (CSIC)-University of Basque Country UPV/EHU, 48940, Leioa, Basque Country, Spain.*

^g *IKERBASQUE, Basque Foundation for Science, Bilbao, Basque Country, Spain.*

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

ABSTRACT. Nano-systems are gaining momentum in pharmaceutical sciences due to the wide variety of possibilities for designing these systems to have specific functions. Specifically, studies of new cancer co-therapy drug-vitamin release nano-systems (DVRNs) including anticancer compounds and vitamins or vitamins derivatives have revealed encouraging results. However, the number of possible combinations of design and synthesis conditions is remarkably high. In addition, there are a high number of anticancer and vitamin derivatives already assayed but a notably less cases of DVRNs assayed as a whole (with the anticancer and the vitamin linked to them). Our approach combine Perturbation Theory and Machine Learning (PTML) to predict the probability of obtaining an interesting DVRN if we change the anticancer compound and/or the vitamin present in a DVRN already tested for other anticancer or vitamin do not tested yet as part of a DVRN. In a previous work, we built a linear PTML model useful for the design of these nano-systems. In so doing, we used Information Fusion (IF) techniques to carry out a data enrichment of DVRNs data compiled from literature with data for preclinical assays of vitamins from ChEMBL database. The design features of DVRNs and the assay conditions nanoparticles and vitamins were included as multiplicative PT Operators (PTOs) to the system, which gives us a measure of the importance of these variables. However, the previous work omitted experiments with non-linear ML techniques and different types of PTOs such as metric-based PTOs. More importantly, the previous work do not considered the structure of the anticancer drug to be included in the new DVRNs. In this work, we are going to accomplish three main objectives (tasks). In the first task, we found a new model, alternative to the published before, for the rational design of DVRNs using metric-based PTOs. The most accurate PTML model was an Artificial Neural Network (ANN) which showed values of specificity, sensitivity, and accuracy in the range of 90-95% in training and external validation

series for more than 130000 cases (DVRNs *vs.* ChEMBL assays). Furthermore, in a second task, we used IF techniques to carry out a data enrichment of our previous dataset. In so doing, we constructed a new working data set of >970000 cases with data of preclinical assays of DVRNs, vitamins, and anticancer compounds from ChEMBL database. All these assays have multiple continue variables or descriptors d_k and categorical variables c_j (conditions of assay) for drug (d_{ack}, c_{acj}), vitamin (d_{vk}, c_{vj}), and nanoparticle (d_{nk}, c_{nj}). It includes, > 20000 potential anticancer with > 270 protein targets (c_{ac1}), > 580 assay cells organisms (c_{ac2}), *etc.* Furthermore, we include > 36000 vitamin derivatives assays in > 6200 types of cell (c_{2vit}), > 120 organisms of assay (c_{3vit}), > 60 assay strains (c_{4vit}) *etc.* The enriched dataset also contains > 20 types of DVRNs (c_{5n}) with 9 nanoparticle core materials (c_{4n}), 8 synthesis methods (c_{7n}), *etc.* We expressed all this information with PTOs and trained a PTML model that is a qualitatively new because it also incorporates information of the anticancer drugs. The new model presents 96-97% of accuracy for training and external validation subsets. Last, in a third task, we carry out a comparative study of ML and/or PTML models published and how the models we are presenting cover a gap of knowledge in terms of drug delivery. In conclusion, we present here by the first time a multipurpose PTML model able to select nanoparticles, anticancer compounds, and vitamins and their conditions of assay for DVRNs design.

Keywords: ChEMBL; Nanoparticle; Anticancer compounds; Perturbation Theory Machine Learning; PTML; Machine Learning; Big data; Multi-target models.

■ INTRODUCTION

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

Machine learning (ML) techniques have gained an important role in inferring new knowledge in pharmaceutical sciences. For instance, Russo *et al.*¹ applied classic algorithms such as random forests, decision trees, and support vector machines to predict compounds for endocrine disrupting capabilities. Lane *et al.*² evaluated different ML methods to efficiently develop new active molecules for those affected by *Mycobacterium tuberculosis*. Korotcov *et al.*³ presented an application of different ML methods including deep learning to pharmaceutical datasets in order to compare their predictive capability. Specifically in nanotechnology, ML has also been used to create efficient predictive models or extract knowledge from data. For instance, Toropova *et al.*⁴ used a Monte Carlo technique to build a model with high accuracy that is able to predict dark cytotoxicity and photo-induced cytotoxicity of metal oxide nanoparticles to the bacteria *Escherichia coli*. Labouta *et al.*⁵ used data mining methods to infer knowledge of nanoparticle cytotoxicity from literature. Yan *et al.*⁶ proposed different ML techniques with possible universal nanodescriptors. Specifically, important steps have been taken in terms of cytotoxicity and genotoxicity by applying ML algorithms.⁷⁻

10

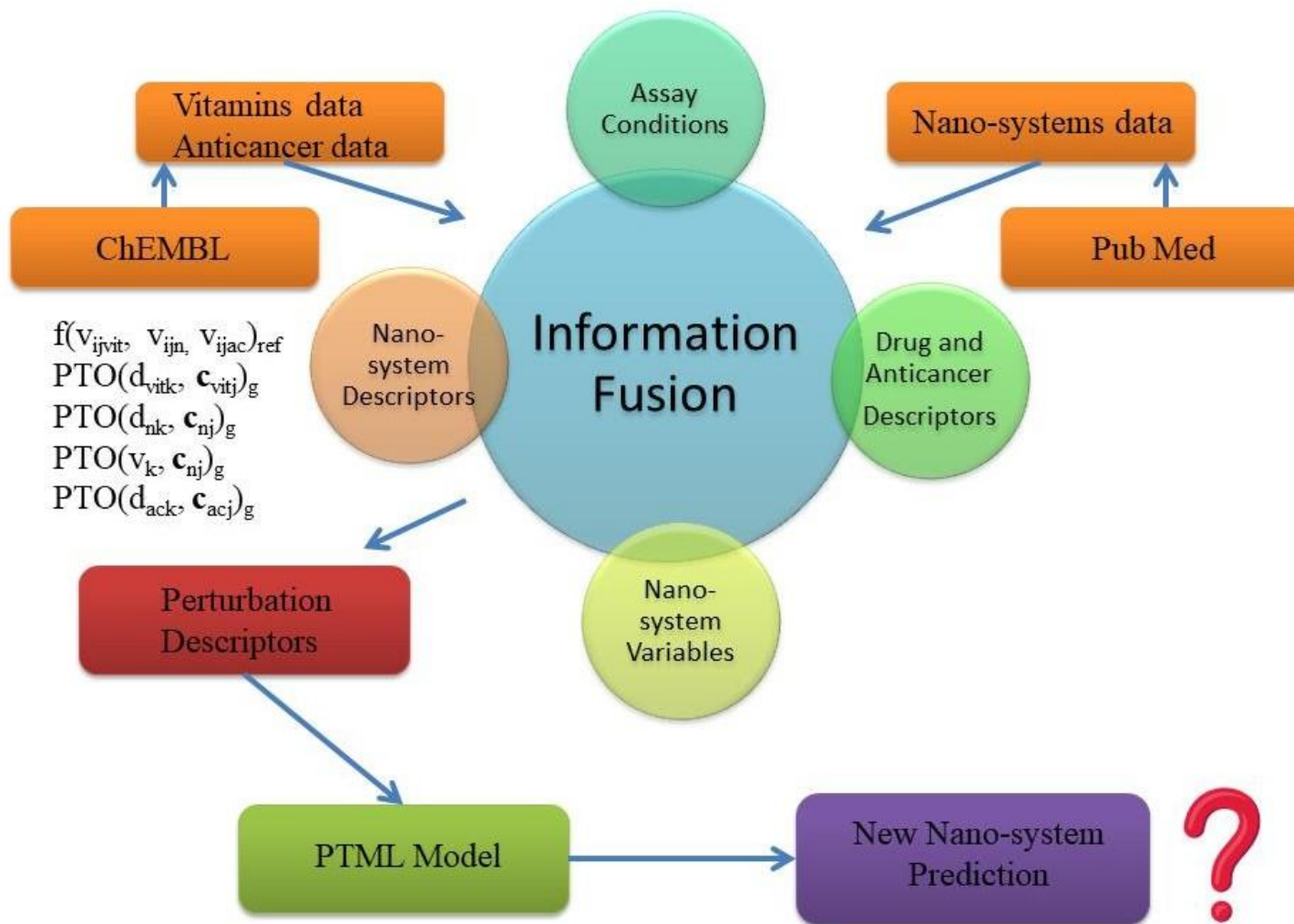
Given the possible impact of nanotechnology on pharmaceutical discoveries, we must highlight the increasing number of studies focused on designing cancer co-therapy drug-vitamin release nano-systems (DVRNs).¹¹ This type of nano-system, which includes vitamins or vitamin derivatives, is designed for chemoprevention and to reduce cancer fatigue.¹² However, the combinatorial space for the design is remarkably high if we take into consideration all the plausible combinations of anticancer compounds, vitamin derivatives, coating agents, nanoparticles, and conditions of assay to be tested. For instance, researchers can vary the vitamins or combination of vitamins, coating agents, polymers, type

(micelles, conjugates, etc.), size, or the drug to be delivered, which are crucial factors in determining the performance of the DVRN. In this context, screening all combinations by using *in vivo* or *in vitro* studies is expensive and time-consuming. Existing ML methods have proven adequate to predict biological activities for specific classes of materials.¹² However, most of the ML models reported up to date in this field does not apply to the design of new DVRNs.

In fact, there are a high number of anticancer and vitamin derivatives already assayed but a notably less cases of DVRNs assayed as a whole (with the anticancer and the vitamin linked to them). These large datasets of bio-assays of compounds are difficult to study given the complexity and heterogeneity of existing datasets¹³, such as the ChEMBL database.^{14,15} Many of these systems are difficult to study with classic ML algorithms. The main difficulty is due to the necessity to include multiple chemical structural descriptors (d_k) and numerical input/output variables (v_k) each one conditioned by other sub-sets of multiple input/output categorical variables (c_j). These categorical variables c_j are used to label multiple experimental conditions (organisms, systems, cells, methods, *etc.*) pre-determining both input/output variables. To address these challenges, our group combined perturbation theory (PT) with ML algorithms to create the PTML method.¹⁶⁻¹⁸ PTML method starts with a function of reference (value for a known system) and adds different PTOs (PTOs). These PTOs measure the variations in the d_k and v_k sets of variables according to the corresponding experimental conditions c_j .¹⁹ The PTML method has been successfully applied to the solution of this type of input/output multi-label problems in order to develop predictive models for biological activity^{16,18,20-24} and the biological performance of nanoparticles.^{9,25}

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

In fact, very recently we developed a PTML multi-output model potentially useful for the design of new DVRNs.¹⁹ However, this work left significant room for improvement in terms of application of suitable algorithms and techniques for better performance. The previous work omitted computational experiments with non-linear ML techniques and different types of PTOs such as metric-based PTOs. More importantly, the previous work do not considered the structure of the anticancer drug to be included in the DVRNs. In this work, we are going to accomplish three main objectives (tasks). In the first objective (**Task 1**), we focus on training new models for rational selection of the compounds of (DVRNs) using metric-based PTOs never used before in PTML modeling. Furthermore, in a second objective (**Task 2**), we are going to use Information Fusion (IF) techniques to carry out a data enrichment of our previous data with data of preclinical assays of anticancer drugs from ChEMBL database. After that, we trained a qualitatively new PTML model that also incorporates information of the anticancer drug and the vitamins inside the DVRNs. The purpose of this new PTML model is to predict the probability of obtaining and interesting DVRN' if we change the anticancer compound and/or the vitamin present in a DVRN already tested for other anticancer' or vitamin' do not tested yet as part of a DVRN. This gives us the possibility to go further in terms of design of DVRNs. Last, in the **Task 3**, we carry out a comparative study for ML and /or PTML models published and how the models we are presenting cover a gap of knowledge in terms of prediction of performance of components of drug delivery. In **Figure 1**, we show the general scheme applied to build the PTML model.



8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

Figure 1. Scheme followed to create the PTML model.

■ MATERIALS AND METHODS

DVRNs and vitamins dataset (Task 1)

For the construction of alternative models without anticancer compounds (Task 1) we use a dataset that we created in previous work.¹⁹ To create this dataset, we collected 1348 data points of biological performance of DVRNs by applying different cutoffs to the reported values. The data set contains four different types of DVRNs: 1) emulsions, 2) polymer conjugates, 3) polymer micelles, and 4) polymer particles. The data includes molecular descriptors of the nano-system: d_{1n} = nanoparticle size (dimension 1), d_{2n} = nanoparticle size (dimension 2), d_{3n} = zeta potential, d_{4n} = polydispersity index and d_{5n} = molecular weight. The subscript n denotes that these variables refer to nano-systems. Additional synthetic and assay variables include: v_0 = cutoff of nano-system biological activity value, v_1 = concentration of vitamin used to synthesize the nano-system, v_2 = concentration of nanoparticle used to synthesize the nano-system, v_3 = concentration of nano-system applied to the assay, and v_4 = assay time. The data set also includes 16 different variables related to the assay conditions (see detailed information in **Table S1**, supporting information file SI00.doc). Of these conditions, the primary variables related to the DVRNs are c_{0n} = biological activity, c_{1n} = drug included in the nano-system, c_{2n} = vitamin included in the nano-system, c_{3n} = nano-system shape, c_{4n} = nanoparticle core material, c_{5n} = nano-system type. Other conditions related to the synthesis of the DVRNs are c_{6n} = nanomaterial synthesis method, c_{7n} = nano-system synthesis method, c_{8n} = nano-system synthesis

solvent, c_{9n} = nanomaterial synthesis solvent. Last, conditions related to the assay of the DVRNs are c_{10n} = assay organism, c_{11n} = assay cell, c_{12n} = assay protein (only albumin is included in the data set), c_{13n} = assay solution/solvent, c_{14n} = assay pH, c_{15n} = type of assay.

PTML models with metric operators without anticancer drugs (Task 1).

In a previous work we proposed a model for the prediction of DVRNs taking into consideration the structure of the vitamin, the nanoparticle, and the conditions of assay. In this previous work we used only multiplicative PTOs based on Geometric means (PTG) and Products of variables (PTP). However, in this previous work we do not tested very interesting types of additive PTOs and ML algorithms. In the present work, with the same dataset collected from literature used in the previous work,¹⁹ we explore how the use of metric-based PTOs in contrast to the multiplicative PTOs. In this study, we also explore linear and non-linear PTML models built with multiple ML techniques to more accurately predict desirable compounds and infer knowledge of DVRNs for cancer co-therapy. To find the best model we applied a variety of classification algorithms: a decision tree, a random forest, quadratic discriminant analysis, naïve Bayes, Ada boost, k neighbors, a Gaussian process, a support vector machine, linear discriminant analysis, and a neural network. These algorithms have been previously used in Chemo-informatics, due to their capacity to classify biological activity.²⁶ The linear form of these PTML models can be seen at follows:

$$\begin{aligned}
 f(\mathbf{v}_{vitij}, \mathbf{v}_{npj})_{calc} = & a_0 + a_1 \cdot f(\mathbf{v}_{ijvit}, \mathbf{v}_{ijn})_{ref} + \sum_{g=1}^{gmax} b_{kj} \cdot PTO(\mathbf{d}_{vitk}, \mathbf{c}_{vitj})_g \\
 & + \sum_{g=1}^{gmax} b_{kj} \cdot PTO(\mathbf{d}_{nk}, \mathbf{c}_{nj})_g + \sum_{g=1}^{gmax} c_{kj} \cdot PTO(\mathbf{v}_k, \mathbf{c}_{nj})_g
 \end{aligned} \tag{1}$$

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

The PTOs can be obtained either from molecular descriptors d_{ks} of different types (k) for different systems (s) or from different non-structural variables v_k . Examples of d_{sk} are $d_{1vit} = \text{Log}P_{vit}$ the logarithm of the partition coefficient of the vitamin or $d_{1n} =$ nanoparticle size. Examples of v_k are $v_1 =$ concentration of vitamin used to synthesize the DVRN or $v_4 =$ assay time. The general notation for these operators is $\text{PTO}(d_{sk}, \mathbf{c}_j)_g = \text{PTO}(\Delta d_{sk}(c_j))_g$ or $\text{PTO}(v_k, \mathbf{c}_j)_g = \text{PTO}(\Delta v_k(c_j))_g$. This indicates that the final value of the operator is obtained in the following sequence: $d_{sk} \Rightarrow \Delta d_{sk}(c_j) \Rightarrow \text{PTO}(\Delta d_{sk}(c_j))_g$ or $v_k \Rightarrow \Delta v_k(c_j) \Rightarrow \text{PTO}(\Delta v_k(c_j))_g$. That is, we transformed the original variables d_{sk} or v_k into Moving Average (MA) variables $\Delta d_{sk}(c_j)$ or $\Delta v_k(c_j)$ (mean-centered variables), see next section. After that, we combine them to calculate the values of the PTOs with the form $\text{PTO}(\Delta d_{sk}(c_j))_g$ or $\text{PTO}(\Delta v_k(c_j))_g$. We employed here the PTOs of Arithmetic mean (PTA), Euclidean distance (PTE), and Manhattan distance (PTM) of each group of variables g is able to predict the observed function. The first PTO depends on d_{vitk} , \mathbf{c}_{vitj} which are the drug descriptors and the combination of drug assay conditions. Both the second and the third PTOs depend on the conditions of the nano-systems assays \mathbf{c}_{nj} . However, one depends on the descriptors d_{nk} and the other one depends on the nano-systems variables v_k . On the other hand, as we see in Eq. 1, each model creates a set of predictions $f(v_{ijvit}, v_{ijn})_{calc}$. This scoring function gives numbers that must be processed to infer knowledge. The algorithm calculates the values of posterior probabilities $p(f(v_{ijvit}, v_{ijn})_{obs} = 1)_{pred}$ by applying the Mahalanobis's distance metric.²⁷

DVRNs, Anticancer Compounds, and Vitamins dataset (Task 2)

In order to train the second type of model including not only DVRNs and vitamins assays but also anticancer compounds assays we created a new dataset.

Our approach to this problem is the following: We could infer probability of obtaining an interesting DVRN' if we start from a DVRN already tested with one anticancer compound and/or the vitamin linked to it and try to infer the effect of causing a perturbation in the system by changing the anticancer or vitamin for a new one already. The approach tries to get advantage of the enrichment of a less common and difficult to obtain datasets of DVRNs with very large datasets of assays of anticancer compounds and vitamin derivatives. This approach is based on the idea of additive perturbations (the compound changed should not introduce a non-linear perturbation on the system). The method should probably fail in the case of synergistic interactions. In any case, there are a very large number of compounds with very close analogues (series of analogues) with similar activity. In all these cases the method is expected to work more accurately and then it could be worthwhile to develop such a model. We constructed this new dataset by data enrichment of the previous dataset created in our earlier work.¹⁹ After IF of the older dataset with the new dataset of anticancer compounds (extracted from ChEMBL database) the new data set contains >970000 data points. Each line entry (data point) includes information about the synthesis and preclinical assay of one DVRNs, preclinical assay of one vitamin, and preclinical assay of one anticancer compound (added now from ChEMBL dataset). However, with the information used, we are not able to predict which anticancer drug and vitamin inside the DVRN will have desirable biological activity. This is given because the structural information of these compounds are not included in the previous models. Thus, we also present in this work, the first multipurpose model able to predict biological activity of DVRNs taking into consideration the information of the anticancer drug and vitamins inside DVRNs apart from the information we already had about the drug and the integral DVRNs. The assay conditions for the anticancer are: c_{0ac} = Biological activity, c_{1ac} = Target protein, c_{2ac} = Assay cell, c_{3ac} = Assay organism and c_{4ac} = Assay type. The descriptors for the anticancer compounds are d_{1ac} = LogP and d_{2ac} = PSA.

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

PTOs used to develop the model (Task 2)

As we mentioned in **Task 1**, there are a large number of PTOs, resulting in a high dimensionality. Therefore, in our previous paper¹⁹ we partitioned the variables into different groups G and calculated PTOs for each partition. After that, the operators in each group $PTG(v_k, c_j)_g$ were used as inputs to different models. In this paper, for the example of model without anticancer drugs, we used only the best partition of variables found in our previous paper. In this case, the parameter g denotes one sub-set of variables of the partition G . This subset of variables includes a sub-group of variables transformed by the PTO. Often, in PTML analysis the variables transformed by the PTO are MAs with notation $\Delta d_{ks}(c_{js})$. These variables have the formula $\Delta d_{ks}(c_{js}) = d_{ksi} - \langle d_k(c_{js}) \rangle$. These are mean-centered variables measuring the deviation of the structural feature or descriptor d_{sk} of type k of the system s with respect to the average value (expected) for all systems measured under experimental condition c_{js} . In fact, the MAs are considered one of the first PTOs used and have notation PTMA in this work. In **Table 1** we summarize the PTMAs used in this work. It includes the PTMAs used in the previous work and the new ones included now (PTMAs for the anticancer drug). Using the previous drug data and the previous formula, we calculated PTMAs for descriptors of anticancer compounds $\Delta d_k(c_{ac}) = d_{ki} - \langle d_k(c_{jac}) \rangle$. The input descriptors or structural features d_{sk} used here to describe the vitamin and anticancer compound are: $\Delta d_{1ac1n}(c_{5n}) = \text{LogP}$ of the first anticancer compound, $\Delta d_{1ac2n}(c_{5n}) = \text{LogP}$ of the second anticancer compound, $\Delta d_{1v1n}(c_{5n}) = \text{LogP}$ of the first vitamin derivative compound, $\Delta d_{1v2n}(c_{5n}) = \text{LogP}$ of the second vitamin derivative compound and $\Delta d_{1v13n}(c_{5n}) = \text{LogP}$ of the third vitamin derivative compound, see **Table 1**. They are added, as we can

observe for c_{5n} which refers to DVRNs system type, according to our earlier work. Please note that PTMAs are the first level PTOs calculated, after that we need to combine PTMAs into more complex PTOs, see previous and next sections.

Table 1. Relevant information for each operator: Conditions, symbols formula and description.

Condition Name		c_j	Symbol	Operator Formula	Description
Anticancer	Actiacy type	c_{0ac}	$\Delta d_{1ac}(c_{0ac})$	$d_{1aci} - \langle d_1(c_{0ac}) \rangle$	Deviation (Δ) of $d_{1ac} = AlogP_i$ and $d_{2ac} = PSA_i$ of the i^{th} acamin derivative from their reference values $\langle AlogP(c_{jac}) \rangle$ and $\langle PSA(c_{jac}) \rangle$ respectively for a given subset of multiple assay conditions c_{jac}
			$\Delta d_{2ac}(c_{0ac})$	$d_{2aci} - \langle d_2(c_{0ac}) \rangle$	
	Protein	c_{1ac}	$\Delta d_{1ac}(c_{1ac})$	$d_{1aci} - \langle d_1(c_{1ac}) \rangle$	
			$\Delta d_{2ac}(c_{1ac})$	$d_{2aci} - \langle d_2(c_{1ac}) \rangle$	
	Cell Name	c_{2ac}	$\Delta d_{1ac}(c_{2ac})$	$d_{1aci} - \langle d_1(c_{2ac}) \rangle$	
			$\Delta d_{2ac}(c_{2ac})$	$d_{2aci} - \langle d_2(c_{2ac}) \rangle$	
	Assay Organism	c_{3ac}	$\Delta d_{1ac}(c_{3ac})$	$d_{1aci} - \langle d_1(c_{3ac}) \rangle$	
			$\Delta d_{2ac}(c_{3ac})$	$d_{2aci} - \langle d_2(c_{3ac}) \rangle$	
	Assay Strain	c_{4ac}	$\Delta d_{1ac}(c_{4ac})$	$d_{1aci} - \langle d_1(c_{4ac}) \rangle$	
			$\Delta d_{2ac}(c_{4ac})$	$d_{2aci} - \langle d_2(c_{4ac}) \rangle$	
Vit. Derivative	Activity type	c_{0vit}	$\Delta d_{1vit}(c_{0vit})$	$d_{1viti} - \langle d_1(c_{0vit}) \rangle$	Deviation (Δ) of $d_{1vit} = AlogP_i$ and $d_{2vit} = PSA_i$ of the i^{th} vitamin derivative from their reference values $\langle AlogP(c_{jvit}) \rangle$ and $\langle PSA(c_{jvit}) \rangle$ respectively for a given subset of multiple assay conditions c_{jvit}
			$\Delta d_{2vit}(c_{0vit})$	$d_{2viti} - \langle d_2(c_{0vit}) \rangle$	
	Protein	c_{1vit}	$\Delta d_{1vit}(c_{1vit})$	$d_{1viti} - \langle d_1(c_{1vit}) \rangle$	
			$\Delta d_{2vit}(c_{2vit})$	$d_{2viti} - \langle d_2(c_{1vit}) \rangle$	
	Cell name	c_{2vit}	$\Delta d_{1vit}(c_{2vit})$	$d_{1viti} - \langle d_1(c_{2vit}) \rangle$	
			$\Delta d_{2vit}(c_{2vit})$	$d_{2viti} - \langle d_2(c_{2vit}) \rangle$	
	Assay organism	c_{3vit}	$\Delta d_{1vit}(c_{3vit})$	$d_{1viti} - \langle d_1(c_{3vit}) \rangle$	

8) MODELLING DVRNs (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

			$\Delta d_{2vit}(c_{3vit})$	$d_{2viti} - \langle d_{2(c_{3vit})} \rangle$	
	Assay strain	c_{4vit}	$\Delta d_{1vit}(c_{4vit})$	$d_{1viti} - \langle d_{1(c_{4vit})} \rangle$	
			$\Delta d_{2vit}(c_{4vit})$	$d_{2viti} - \langle d_{2(c_{4vit})} \rangle$	
DVRNs	Activity type	c_{0n}	$\Delta v_{1n}(c_{0n}), \Delta v_{2n}(c_{0n}) \dots v_{in}(c_{0n})$	$v_{ini} - \langle v_{in}(c_{0n}) \rangle$	<p>Measures the deviation of the reference d_{in} value (average) of all n_i with the same $c_{0n}, c_{1n}, c_{2n}, c_{3n}, c_{4n}, c_{5n}$ and c_{15n}</p> <p>Furthermore, Δv_{in} refers to values of all considered nano-system variables v_0-v_9.</p> <p>It also includes $d_{1ac1n}, d_{1ac2n}, d_{1v1n}, d_{1v2n}$ and d_{1v3n} value (average) of all vitamin and anticancer with the same c_{5n}.</p>
	Drug/Drug comb np	c_{1n}	$\Delta v_{1n}(c_{1n}), \Delta v_{2n}(c_{1n}) \dots \Delta v_{in}(c_{1n})$	$v_{ini} - \langle v_{in}(c_{1n}) \rangle$	
	DVRNs vitamin	c_{2n}	$\Delta v_{1n}(c_{2n}), \Delta v_{2n}(c_{2n}) \dots \Delta v_{in}(c_{2n})$	$v_{ini} - \langle v_{in}(c_{2n}) \rangle$	
	DVRNs shape	c_{3n}	$\Delta v_{1n}(c_{3n}), \Delta v_{2n}(c_{3n}) \dots \Delta v_{in}(c_{3n})$	$v_{ini} - \langle v_{in}(c_{3n}) \rangle$	
	Core raw material	c_{4n}	$\Delta v_{1n}(c_{4n}), \Delta v_{2n}(c_{4n}) \dots \Delta v_{in}(c_{4n})$	$v_{ini} - \langle v_{in}(c_{4n}) \rangle$	
	DVRNs system type	c_{5n}	$\Delta v_{1n}(c_{5n}), \Delta v_{2n}(c_{5n}) \dots \Delta v_{in}(c_{5n})$	$v_{ini} - \langle v_{in}(c_{5n}) \rangle$	
			$\Delta d_{1n}(c_{5n}), \Delta d_{2n}(c_{5n}) \dots \Delta d_{in}(c_{5n})$	$d_{ini} - \langle d_{in}(c_{5n}) \rangle$	
			$\Delta d_{1ac1n}(c_{5n})$	$d_{1ac1n} - \langle d_{1ac1n}(c_{5n}) \rangle$	
			$\Delta d_{1ac2n}(c_{5n})$	$d_{1ac2n} - \langle d_{1ac1n}(c_{5n}) \rangle$	
			$\Delta d_{1v1n}(c_{5n})$	$d_{1v1n} - \langle d_{1v1n}(c_{5n}) \rangle$	
			$\Delta d_{1v2n}(c_{5n})$	$d_{1v2n} - \langle d_{1v2n}(c_{5n}) \rangle$	
	c_{5n}	$\Delta d_{1v3n}(c_{5n})$	$d_{1v3n} - \langle d_{1v3n}(c_{5n}) \rangle$		
Type of assay	c_{15n}	$\Delta v_{1n}(c_{15n}), \Delta v_{2n}(c_{15n}) \dots \Delta v_{in}(c_{15n})$	$v_{ini} - \langle v_{in}(c_{15n}) \rangle$		
Method nanomaterial synth	c_{6n}	$\Delta d_{1n}(c_{6n}), \Delta d_{2n}(c_{6n}) \dots \Delta d_{in}(c_{6n})$	$d_{ini} - \langle d_{in}(c_{6n}) \rangle$	<p>Measures the deviation of the reference D_{in} value (average) of all np_i with the same c_{6n}, c_{7n}, c_{8n} and $c_{9n}, c_{10n}, c_{11n}, c_{12n}, c_{13n}, c_{14n}$. Furthermore, Δd_{in} refers to</p>	
Method drug-nano-system	c_{7n}	$\Delta d_{1n}(c_{7n}), \Delta d_{2n}(c_{7n}) \dots \Delta d_{in}(c_{7n})$	$d_{ini} - \langle d_{in}(c_{7n}) \rangle$		

DVRNs synthesis solvent	c_{8n}	$\Delta d_{1n}(c_{8n}), \Delta d_{2n}(c_{8n}) \dots \Delta d_{in}(c_{8n})$	$d_{ini} - \langle d_{in}(c_{8n}) \rangle$	values of all considered nano-system variables d_1 - d_5 .
Nanoparticle synthesis solvent	c_{9n}	$\Delta d_{1n}(c_{9n}), \Delta d_{2n}(c_{9n}) \dots \Delta d_{in}(c_{9n})$	$d_{ini} - \langle d_{in}(c_{9n}) \rangle$	
DVRNs assay organism	c_{10n}	$\Delta d_{1n}(c_{10n}), \Delta d_{2n}(c_{10n}) \dots \Delta d_{in}(c_{10n})$	$d_{ini} - \langle d_{in}(c_{10n}) \rangle$	
DVRNs assay cell	c_{11n}	$\Delta d_{1n}(c_{11n}), \Delta d_{2n}(c_{11n}) \dots \Delta d_{in}(c_{11n})$	$d_{ini} - \langle d_{in}(c_{11n}) \rangle$	
Albumin	c_{12n}	$\Delta d_{1n}(c_{12n}), \Delta d_{2n}(c_{12n}) \dots \Delta d_{in}(c_{12n})$	$d_{ini} - \langle d_{in}(c_{12n}) \rangle$	
DVRNs media assay	c_{13n}	$\Delta d_{1n}(c_{13n}), \Delta d_{2n}(c_{13n}) \dots \Delta d_{in}(c_{13n})$	$d_{ini} - \langle d_{in}(c_{13n}) \rangle$	
DVRNs assay pH	c_{14n}	$\Delta d_{1n}(c_{14n}), \Delta d_{2n}(c_{14n}) \dots \Delta d_{in}(c_{14n})$	$d_{ini} - \langle d_{in}(c_{14n}) \rangle$	

PTML Information Fusion process (Task 2)

Once, we calculated new PTMAs for the DVRNs, vitamin derivatives, and anticancer compounds we can undergo the Information Fusion (IF) process. The PTOs contain information on the descriptors $\{d_{1vit}, d_{2vit}, d_{1n}, d_{2n}, d_{3n}, d_{4n}, d_{5n}, d_{1ac}, d_{2ac}, d_{1ac1n}, d_{1ac2n}, d_{1v1n}, d_{1v2n}, d_{1v3n}\}$, the non-structural variables $\{v_0, v_1, v_2, v_3, v_4\}$, and assay conditions (c_{jvit} , c_{jac} and c_{jn}). During the IF process we fused the three datasets into a larger one that contains in each row all the variables of one vitamin, one anticancer, and one DVRNs. As part of IF process, we calculated the reference $f(v_{ijvit}, v_{ijn}, v_{ijac})_{ref}$ and new observed $f(v_{ijvit}, v_{ijn}, v_{ijac})_{obs}$ functions used as input/dependent variables to train the new PTML models, see **Figure 2**.

8) MODELLING DVRNs (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

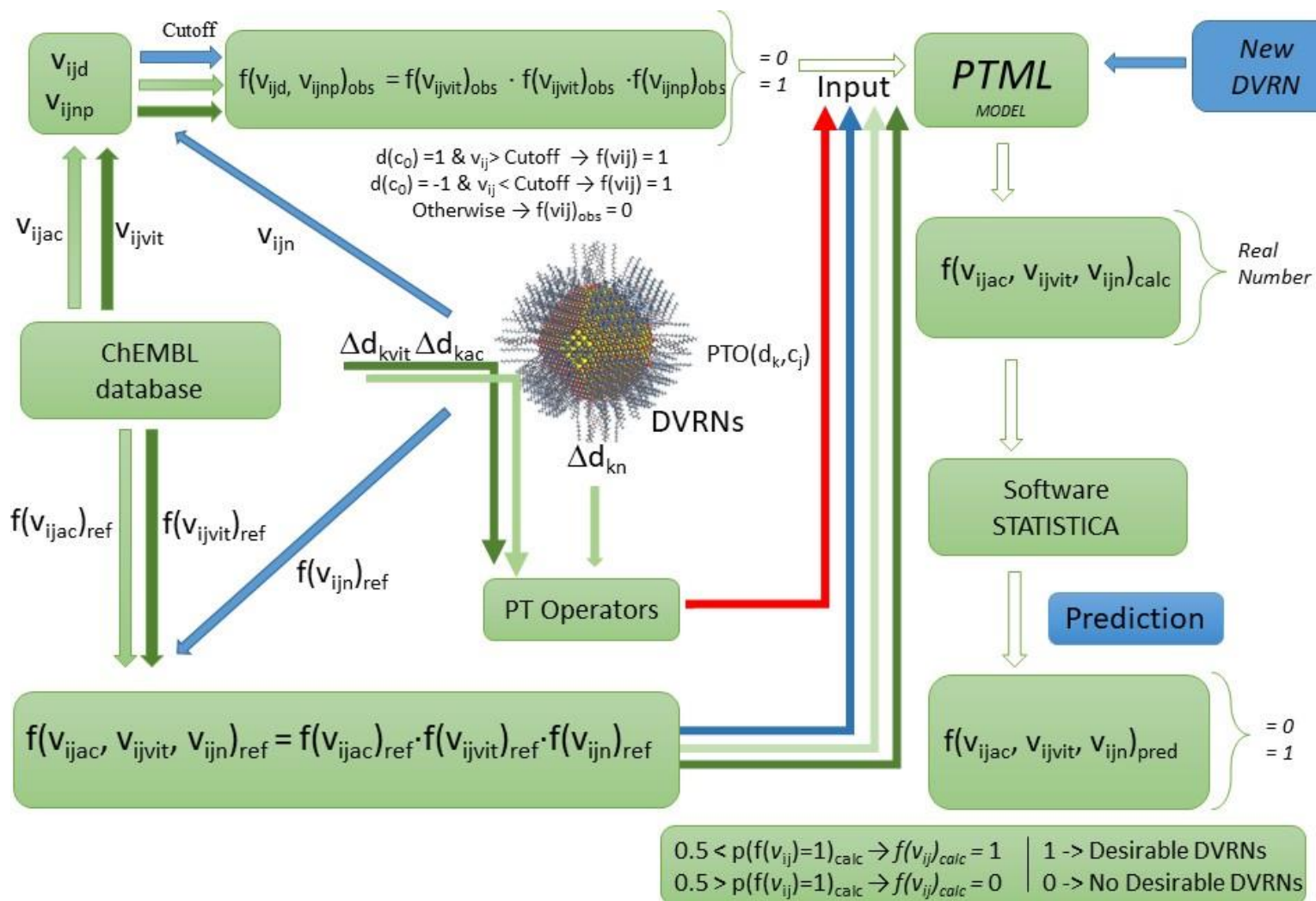


Figure 2. Workflow for IF process

PTML Objective, Reference, and Output Functions (Task 2)

As we mentioned before the aim of our model is to predict the probability of obtaining and interesting DVRN' if we change the anticancer compound and/or the vitamin present in a DVRN already tested for other also tested before. The objective function or observed function $f(v_{ijvit}, v_{ijn}, v_{ijac})_{obs}$, is the function to be predicted by the model. The function $f(v_{ijn}, v_{ijac}, v_{ijvit})_{obs} = 1$ when the DVRNs is potentially interesting; otherwise $f(v_{ijvit}, v_{ijn}, v_{ijac})_{obs} = 0$. To construct this objective function we started by discretizing the experimental parameters of biological activity of the anticancer v_{ijac} compound to obtain the Boolean function $f(v_{ijac})_{obs}$. We followed the same method used to discretize the activity of the vitamin v_{ijvit} and the activity of DVRNs v_{ijn} in order to obtain $f(v_{ijn})_{obs}$ and $f(v_{ijvit})_{obs}$. Consequently, this function is the result taking into considerations also the anticancer compound with respect to the objective function of our previous work.¹⁹ In essence, we compare the value of biological activity v_{ijs} of the property j^{th} for the system s_i with the desirability $d(c_{0j})$ and a cutoff $_j$ of this property . If $[d(c_{0j}) = 1$ and $v_{ijs} > cutoff_j]$ or $[d(c_{0j}) = -1$ and $v_{ijs} < cutoff_j]$ then $f(v_{ijvit}) = 1$, otherwise $f(v_{ijvit}) = 0$. The desirability function indicates if we want to maximize or minimize this property and the cutoff is the level of activity considered interesting .¹⁹ In **Table 2**, we summarize the functions mentioned before. Once we have the values of $f(v_{ijn})$, $f(v_{ijac})$, and $f(v_{ijvit})$ we calculated $f(v_{ijvit}, v_{ijn}, v_{ijac})_{obs} = f(v_{ijn}) \cdot f(v_{ijac}) \cdot f(v_{ijvit})$. Consequently, $f(v_{ijvit}, v_{ijn}, v_{ijac})_{obs} = 1$ if and only if $f(v_{ijn}) = 1$, $f(v_{ijac}) = 1$, and $f(v_{ijvit}) = 1$, otherwise $f(v_{ijvit}, v_{ijn}, v_{ijac})_{obs} = 0$. The $f(v_{ijvit}, v_{ijn}, v_{ijac})_{ref}$ function is the expected probability of the activity of the DVRN, based on the activity of the nano-system, the anticancer compound and the added vitamin derivative, separately. Consequently, we defined $p(f(v_{ijvit}, v_{ijn}, v_{ijac})_{obs} = 1)_{ref} = p(f(v_{ijvit})_{obs} = 1)_{ref} \cdot p(f(v_{ijn})_{obs} = 1)_{ref} \cdot p(f(v_{ijac})_{obs} = 1)_{ref}$. In this formula, $p(f(v_{ijvit})_{obs} = 1)_{ref}$, $p(f(v_{ijac})_{obs} = 1)_{ref}$ and

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

$p(f(v_{ijn})_{obs} = 1)_{ref}$ are the expected probability for the drug's activity v_{ijvit} of type c_{0vit} , anticancer compound's activity v_{ijac} of type c_{0ac} and the expected probability for the DVRN's activity v_{ijn} of type c_{0n} , respectively. Hence, this input function expresses the expected probability of activity of the DVRN with the new vitamin derivative and anticancer compound added.

Table 2. Input/output functions of the IF process

Condition Name	c_j	Symbol	Operator Formula	Description
Objective function (See next sections)	$c_{0ac}, c_{0vit}, c_{0n}$	$f(v_{ijac}, v_{ijvit}, v_{ijn})_{obs}$	$f(v_{ijac})_{obs} \cdot f(v_{ijvit})_{obs} \cdot f(v_{ijn})_{obs}$	Combination of the three observed functions, a DVRN should be assayed only if $f(v_{ijac}, v_{ijvit}, v_{ijn})_{obs} = 1$
Reference Function (See next sections)	$c_{0ac}, c_{0vit}, c_{0n}$	$f(v_{ijac}, v_{ijvit}, v_{ijn})_{ref}$	$f(v_{ijac})_{ref} \cdot f(v_{ijvit})_{ref} \cdot f(v_{ijn})_{ref}$	Expected probability of success for all DVRNs with composed expected probs. $p(f(v_{ijac})=1)_{ref} \cdot p(f(v_{ijvit})=1)_{ref} \cdot p(f(v_{ijn})=1)_{ref}$
Output function	$c_{jac}, c_{jvit}, c_{jn}$	$f(v_{ijac}, v_{ijvit}, v_{ijn})_{calc}$	PTML model	Real-value scoring function obtained as output of the model.
Posterior probability (See next sections)	$c_{jac}, c_{jvit}, c_{jn}$	$p(f(v_{ijac}, v_{ijvit}, v_{ijn})_{calc}=1)$	$1/(1+(\pi_0/\pi_1)\text{Exp}(-f(v_{ijac}, v_{ijvit}, v_{ijn})_{calc}))$	Posterior classification probability for one DVRNs
Predicted function (See next sections)	$c_{jac}, c_{jvit}, c_{jn}$	$f(v_{ijac}, v_{ijvit}, v_{ijn})_{pred}$	= 1 only if $p(f(v_{ijac}, v_{ijvit}, v_{ijn})_{calc}=1) > 0.5$	Predicted value of the objective function
Anticancer reference function	c_{0ac}	$f(v_{ijac})_{ref}$	$n(f(v_{ijac})_{obs}=1)/n_j$	Expected probability $p(f(v_{ijac})=1)_{ref}$ for the activity v_{ijac} of type c_{0ac} of one anticancer
Vitamin reference function	c_{0vit}	$f(v_{ijvit})_{ref}$	$n(f(v_{ijvit})_{obs}=1)/n_j$	Expected probability $p(f(v_{ijvit})=1)_{ref}$ for the activity v_{ijvit} of type c_{0vit} of one vitamin

Nanoparticle reference function	c_{ovit}	$f(v_{ijn})_{ref}$	$n(f(v_{ijvit})_{obs}=1)/n_j$	Expected probability $p(f(v_{ijn})=1)_{ref}$ for the activity v_{ijn} of type c_{on} of one nanoparticle
---------------------------------	------------	--------------------	-------------------------------	--

PTML model with anticancer drug information (Task 2).

After carrying out the IF process to creating the new working dataset we were ready to propose a new PTML model. In this work, we proposed different models built with PTML techniques in order to take into consideration not only variables of the structure of the vitamin but also the structure of the anticancer drug. The linear form of these PTML models is the following:

$$\begin{aligned}
 f(v_{vitij}, v_{acj}, v_{npj})_{calc} &= a_0 + a_1 \cdot f(v_{vitij}, v_{acj}, v_{npj})_{ref} + \sum_{g=1}^{g_{max}} b_{kj} \cdot PTO(d_{acki}, c_{acj})_g \\
 &+ \sum_{g=1}^{g_{max}} b_{kj} \cdot PTO(d_{vitki}, c_{vitj})_g + \sum_{g=1}^{g_{max}} b_{kj} \cdot PTO(d_{nki}, v_{nki}, c_{nj})_g
 \end{aligned} \quad (2)$$

The output of the model here $f(v_{ijvit}, v_{ijn}, v_{ijac})_{calc}$ is an scoring function used score specific DVRNs, see previous **Table 2**. The algorithm transforms the values of this function to posterior probabilities $p(f(v_{ijvit}, v_{ijn}, v_{ijac})_{obs} = 1)_{pred}$ by applying the sigmoidal function. These values of probabilities indicates the probabilities with which a combination of one anticancer drug, vitamin, nanoparticle, and assay conditions should be selected to test a new DVRN.²⁷ On the other hand, the general notation for the PTOs is analogue to the notation describe for the previous model $PTO(d_{sk}, c_j)_g =$

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

$PTO(\Delta d_{sk}(c_j))_g$ or $PTO(v_k, c_j)_g = PTO(\Delta v_k(c_j))_g$. This indicates that the final value of the PTO is obtained in the following sequence: $d_{sk} \Rightarrow \Delta d_{sk}(c_j) \Rightarrow PTO(\Delta d_{sk}(c_j))_g$ or $v_k \Rightarrow \Delta v_{sk}(c_j) \Rightarrow PTO(\Delta v_k(c_j))_g$. The difference in this work is that we include for the first time PTOs depending on d_{ack} and c_{acj} which are the drug descriptors and the combination of drug assay conditions reported in ChEMBL database for the anticancer compound. In addition, as in the previous model, the second PTO depends on d_{vitk} and c_{vitj} which are the drug descriptors and the combination of drug assay conditions reported in ChEMBL for the vitamin derivatives (vit). The remnant PTOs depend on the descriptors of the nanoparticle (d_{nk}) and the conditions of the nano-systems assays c_{nj} . The advantages, disadvantages, and the specific formula used to calculate different PTOs in this and previous papers will be discussed in **Task 3**, next section, as part of the comparative study.

■ RESULTS AND DISCUSSION

PTML models with metric operators without anticancer drugs (Task 1).

We applied the 10 classifiers listed in the Materials and Methods section using the Python package scikit-learn.²⁸ These models were first tested with 5000 data points as a preliminary step to find the most accurate methods. The most accurate models from this initial test—a neural network, Adaboost and a random forest—were then trained on the full data set (see **Table 3**). We found that nonlinear models have a higher prediction capacity compared to our results in our previous study where we applied LDA and obtained 87.09 % and 87.20 % for train and test accuracy respectively.¹⁹ In all cases, we divided the data into two subsets (train and test, 80 %-20 % respectively) and optimized the hyperparameters of

each classifier by performing k-fold cross validation on the training set. The resulting model was then fit on the entire training set and tested on both the training and test subsets.

Table 3. Most accurate PTML model results

^a ML Algorithm	Train Accuracy	Test Accuracy
RF	94.33	94.32
AB	94.00	94.22
ANN	94.46	94.49

^aML algorithm used: RF= Random Forest, AB = Ada Boost, ANN = Artificial Neural Network.

After this process, the most accurate model we found was a neural network classifier (PTML-ANN) with Ac(%) = 94.45 for training subset, and Ac(%) = 94.48 for test subset (see “First Model” in **Table 4**). This ANN includes one hidden layer with 60 neurons, a hyperbolic tangent activation function, and a penalty (regularization term) parameter of 0.02.

Table 4. Results of the models and input variables analyzed

Model	Observed Sets	^a Stat. Param.	^b Pred. stat.		
				$f(v_{ijvit}, v_{ijn})_{pred} = 0$	$f(v_{ijvit}, v_{ijn})_{pred} = 1$
Training					
First Model	$f(v_{ijvit}, v_{ijn})_{obs} = 0$	Sp	98.23	98188	1760

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

	$f(v_{ijvit}, v_{ijn})_{obs} = 1$	Sn	47.01	4224	3748
	Total	Ac	94.45		
^c First Model +	$f(v_{ijvit}, v_{ijn})_{obs} = 0$	Sp	97.45	97405	2543
Δd_{2vit}	$f(v_{ijvit}, v_{ijn})_{obs} = 1$	Sn	48.25	4125	3847
	Total	Ac	93.82		
^d OS Model	$f(v_{ijvit}, v_{ijn})_{obs} = 0$	Sp	89.23	89181	10767
	$f(v_{ijvit}, v_{ijn})_{obs} = 1$	Sn	94.61	4290	75430
	Total	Ac	91.62		
^e OS Model -	$f(v_{ijvit}, v_{ijn})_{obs} = 0$	Sp	90.42	90374	9574
$\Delta d_{2vit} - PTA(g_2)$	$f(v_{ijvit}, v_{ijn})_{obs} = 1$	Sn	94.49	4390	75330
	Total	Ac	92.27		
Test					
	$f(v_{ijvit}, v_{ijn})_{obs} = 0$	Sp	98.22	24556	444
First Model	$f(v_{ijvit}, v_{ijn})_{obs} = 1$	Sn	47.32	1043	937
	Total	Ac	94.48		
First Model +	$f(v_{ijvit}, v_{ijn})_{obs} = 0$	Sp	97.53	24383	617
Δd_{2vit}	$f(v_{ijvit}, v_{ijn})_{obs} = 1$	Sn	49.75	995	985
	Total	Ac	94.03		
OS Model	$f(v_{ijvit}, v_{ijn})_{obs} = 0$	Sp	89.09	22274	2726
	$f(v_{ijvit}, v_{ijn})_{obs} = 1$	Sn	92.63	146	1834
	Total	Ac	89.35		
	$f(v_{ijvit}, v_{ijn})_{obs} = 0$	Sp	90.75	22687	2313

OS Model - Δd_{2vit}	$f(v_{ijvit}, v_{ijn})_{obs} = 1$	Sn	93.23	134	1846
- PTA (g2)	Total	Ac	90.93		

^aStat. Param. = Statistical parameter, ^bPred. Stat. = Predicted statistics, ^cFirst Model + Δd_{2vit} = First model including as new variable $\Delta d_{2vit}(c_{0vit}, c_{1vit}, c_{2vit}, c_{3vit}, c_{4vit})$, ^dOS Model = Model with oversampled training subset, ^eOS Model - Δd_{2vit} - PTA (g2) = Model with oversampled training subset without the variables $\Delta d_{2vit}(c_{0vit}, c_{1vit}, c_{2vit}, c_{3vit}, c_{4vit})$ and PTA (g2).

Using the PTML-ANN allows us to estimate the uncertainty of the model for each prediction. To do so, we created six different PTML-ANN models with the same hyperparameters and dataset, but different initial random states. Each model can potentially have a different prediction for different bioassay results, and the number of the models that agree for a particular case is a measure of uncertainty. We created one model as our primary, reference model and checked whether the other five models agreed with it. The number of cases for which all five models agree is quite high, 25414 cases; the other cases are in **Table 5**. We calculated the percentage of cases that are correctly predicted as a function of the amount of agreement between the five models. As we observe in **Table 5**, the model is especially robust when the five models coincide, where it has 96.93 % accuracy.

Table 5. Agreement across PTML-ANN models with different random states

Number of models that agree	Number of data points	Ac(%)
0	125	32.00

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

1	176	36.93
2	285	50.17
3	438	59.81
4	542	64.02
5	25414	96.93

While this model is accurate, the true positive rate is not high, as seen in the confusion matrix in **Table 4**. The area of the corresponding receiver operating characteristic (ROC) curve is 0.73, which provides us an evaluation of the classifier output quality and thus, the margin to be improved (see First Model in **Figure 3**). The specificity is clearly higher than the sensitivity, allowing prediction of the undesirable cases with high accuracy. However, when designing biocompatible nanomaterials, a model with higher sensitivity would be more useful.

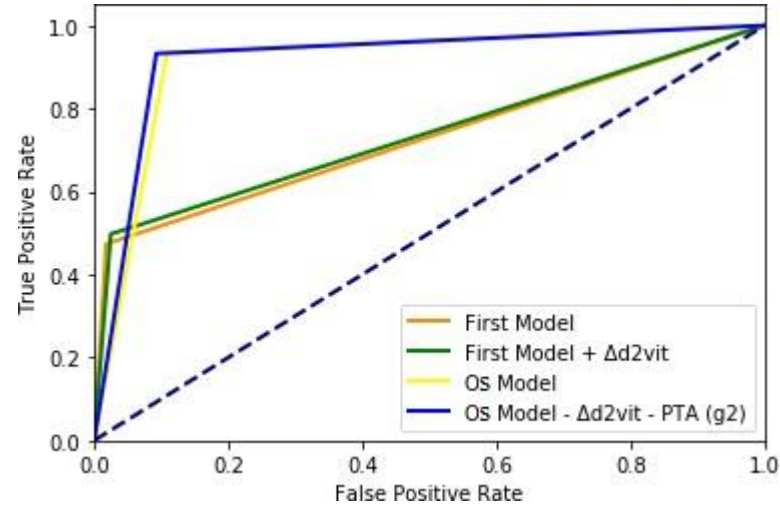


Figure 3. ROC curves for the original PTML-ANN (First Model), the model with an additional feature (First Model + Δd_{2vit}), the oversampled model (OS Model) and the oversampled model with additional features (OS Model - Δd_{2vit} - PTA (g_2)).

In order to improve the sensitivity of the model, we first searched for other variables to be added. We tested all possible variables that resulted from all G partitions and the metrics commented above, to predict the number of agrees. We also tried perturbation operators with multiple assay conditions: $\mathbf{c}_{jvit} = (c_{0vit}, c_{1vit}, c_{2vit}, \dots, c_{nvit})$ of the bio-assay. In order to do this, we created a one-variable PTML-ANN model for every variable and tested the accuracy in each case. This gives insight into the predictive capability of each variable. The best accuracy was obtained with models containing $\Delta d_{1vit}(c_{0vit}, c_{1vit}, c_{2vit}, c_{3vit}, c_{4vit})$ and $\Delta d_{2vit}(c_{0vit}, c_{1vit}, c_{2vit}, c_{3vit}, c_{4vit})$, with 96.22 and 95.96 respectively. We trained new models by adding

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

these to the variables of the first model, both separately and together. The most accurate result was found by adding only $\Delta d_{2vit}(c_{0vit}, c_{1vit}, c_{2vit}, c_{3vit}, c_{4vit})$, see **Table 4**.

Upon adding this variable, the general accuracy decreased somewhat but the sensitivity is slightly higher. The ROC curve in this case is only 0.01 point higher (area = 0.74), see **Figure 3** (First Model + Δd_{2vit}). The percentage of cases where five different models agree is 97.27, which is also higher than the previous model. However, the sensitivity is still fairly low, which suggests that adding more variables is not an effective strategy for achieving high sensitivity. A likely cause for the low sensitivity is the high percentage of undesirable cases in the dataset. Therefore, we tested how oversampling the desirable cases affects the model's sensitivity. Each desirable case was included 10 times in the oversampled dataset. This transforms the original training subset (99948 cases with $f(v_{ijvit}, v_{ijnpc})_{obs} = 0$ and 7972 with $f(v_{ijvit}, v_{ijnpc})_{obs} = 1$) into a new one (99948 cases with $f(v_{ijvit}, v_{ijnpc})_{obs} = 0$ and 79720 with $f(v_{ijvit}, v_{ijnpc})_{obs} = 1$), see **Table 4**. In this case the sensitivity is remarkably higher: 94.61 and 96.63 for training and test subsets respectively. The general accuracy is lower compared to the previous model. However, it is still much higher than models we find in the literature, and is able to predict the positive cases with high precision, see **Figure 3** (OS Model). After applying the oversampling, we observed that removing PTA (g2) and $\Delta d_{2vit}(c_{0vit}, c_{1vit}, c_{2vit}, c_{3vit}, c_{4vit})$ increases the sensitivity. The accuracy is lower compared to the model without oversampling, even for just the cases where all five models agree (92.55 %). However, the ROC curve presents an area of 0.92, as shown in **Figure 3** (OS Model - Δd_{2vit} - PTA (g2)). Given that PTA (G1) includes the information of $c_{0vit}, c_{1vit}, c_{2vit}, c_{3vit}, c_{4vit}$, we can exclude PTA (G2), see **Table 4**.

We attempted to increase the accuracy by applying an ensemble of all the models with different random states mentioned above, and with Ada Boost and random forest models, with the best parameters found in the previous step (random forest with 7 as the maximum depth of the tree, Gini as the function to measure the quality of the split and 5 as the minimum number of samples required to split; Ada Boost with the SAMME.R algorithm). The accuracy with the hard voting method was 90.74 %, compared to 90.93 % of the previous model. Although the accuracy decreases, the sensitivity was slightly higher: 95.15 %, as compared to 93.23 % for the previous model. Overall, the improvement is small over the neural network reference model, despite much higher complexity. Hence, we prefer to use the neural network as the final model.

Once we have an effective model, we can use it learn about uncertainty in the data and feature importance. In terms of prediction, as we have pointed out above, there are cases that are better predicted than others. By applying different random states, we can select the nano-systems where are model has lower uncertainty. For instance, if the property to be predicted is cumulative release for a determined drug, the five models with different random states all agree for 98.51% of cases for nano-systems with Docetaxel. However, if the same property is predicted for other drugs like Cisplatin, Sorafenib or Paclitaxel, the rate decreases to 96.78%, 94.46% and 94.72% respectively. If we want to use the model to predict the system type for the drug Docetaxel, predictions for conjugate systems will present a higher accuracy (97.54 %) than capsule-like systems (95.43 %). The model is especially accurate for conjugate systems that include Docetaxel. This gives insight into where data sets could be improved in order to improve prediction accuracy.

On the other hand, the model has a high accuracy for all the nano-systems explored. This is explained by the distribution of the cases in the study space. The regions with many cases with low agreement among the models also contain a large number of positive cases, where $f(v_{ijvit}, v_{ijn})_{obs} = 1$.

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

For instance, for g1 and g5, we observe in **Figure 4a** the cases for which only one (purple) or two (red) models agree, depending on PTA (G1) and PTA (G5). As we can see in **Figure 4b**, positive cases, where $f(v_{ijvit}, v_{ijn})_{obs} = 1$, are concentrated in the same zones. For instance, for cases in which PTA (g5) is around zero, the uncertainty tends to be higher because that zone includes a high number of cases with $f(v_{ijvit}, v_{ijnpc})_{obs} = 1$.

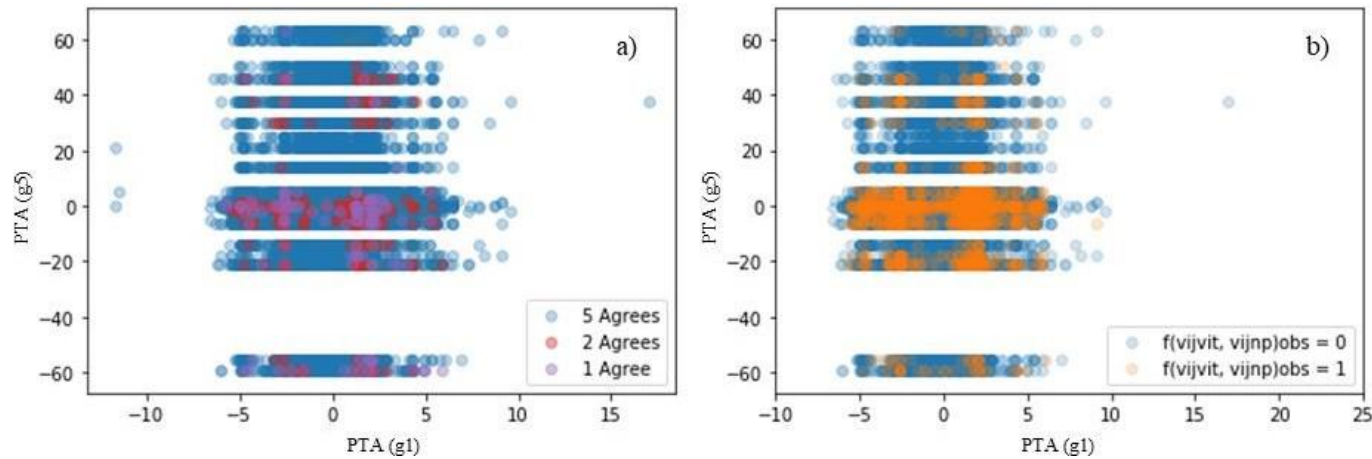


Figure 4. a) Data points where one, two or five models agree as a function of PTA (g1) and PTA (g5). b) $f(v_{ijvit}, v_{ijnpc})_{obs}$ as a function of PTA (g1) and PTA (g5).

Finally, we studied the importance of each variable using the final model (with the oversampled set and without PTA (g2) and Δd_{2vit} (c_{0vit} , c_{1vit} , c_{2vit} , c_{3vit} , c_{4vit})). We performed recursive feature elimination, where the least important feature is recursively removed from the model. That is, we removed each variable from the model, selected the model with the highest sensitivity, and then repeated this process with the new model. This

analysis suggests that the most important feature is $f(v_{ijvit}, v_{ijn})_{ref}$, given that the model with only this variable correctly predicts 93.99 % of the cases with $f(v_{ijvit}, v_{ijn})_{obs} = 1$ (see **Figure 5**). Then, adding the perturbation operators PTA (g4), PTA (g6), PTA (g5) back in increases the sensitivity of the model. These are the groups that provide information about the synthesis, assay conditions and nano-system type of the DVRNs. Once these variables are included, neither PTA (g1) or PTA (g2) give more information for prediction of desirable DVRNs. These groups give information about the biological activity, drug and vitamin incorporated in the DVRNs. By observing the feature importance, we understand why cumulative release is the best biological activity to predict, given that it strongly depends on how the DVRNs are designed. If we exclude PTA (g3) and PTA (g1), we would have an even better model in terms of sensitivity. However, given the purpose of this study, we consider that the information of PTA (g3) and PTA (g1) for rational discovery of DVRNs biological activities is relevant for predictions of biological properties of DVRNs with particular drugs or vitamins.

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

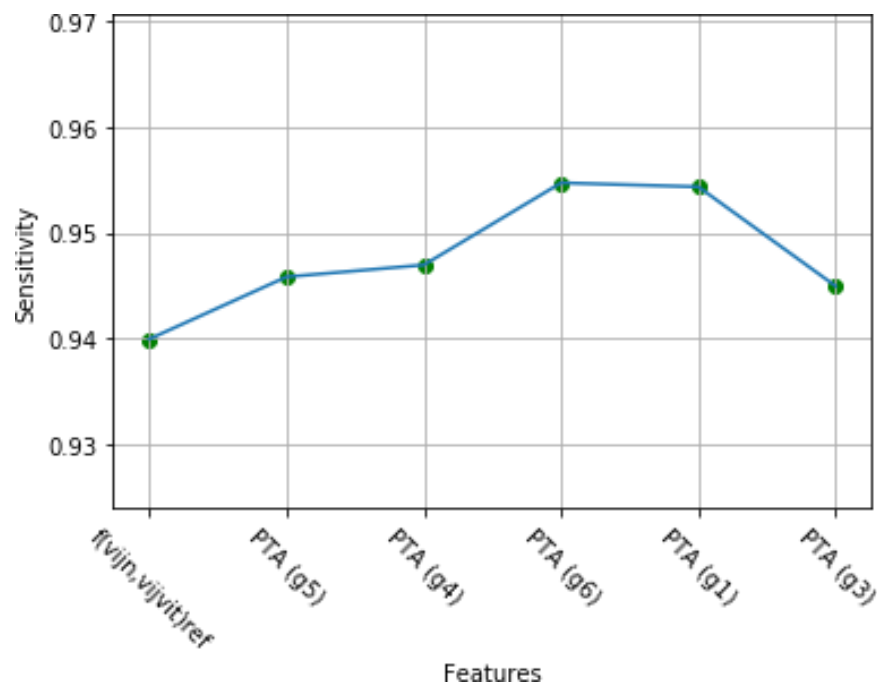


Figure 5. Accumulated sensitivity as we add the variables to the final model in increasing order of importance.

PTML model with anticancer compounds (Task 2). In order to seek the new PTML model the variables have been grouped in a new partition of variables to simplify the information: 1) ChEMBL, given that this information has been provided by this database and 2) DVRNs, which is the information about the Synthesis (g₃), Type (g₄), Size (g₅), Assay (g₆), Anticancer and Vitamins (g₇). After grouping the variables in this new

partition, see **Table 6**, we calculated the new PTOs incorporating information about the anticancer compound and its conditions of assay c_{acj} , see

Figure 6.

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

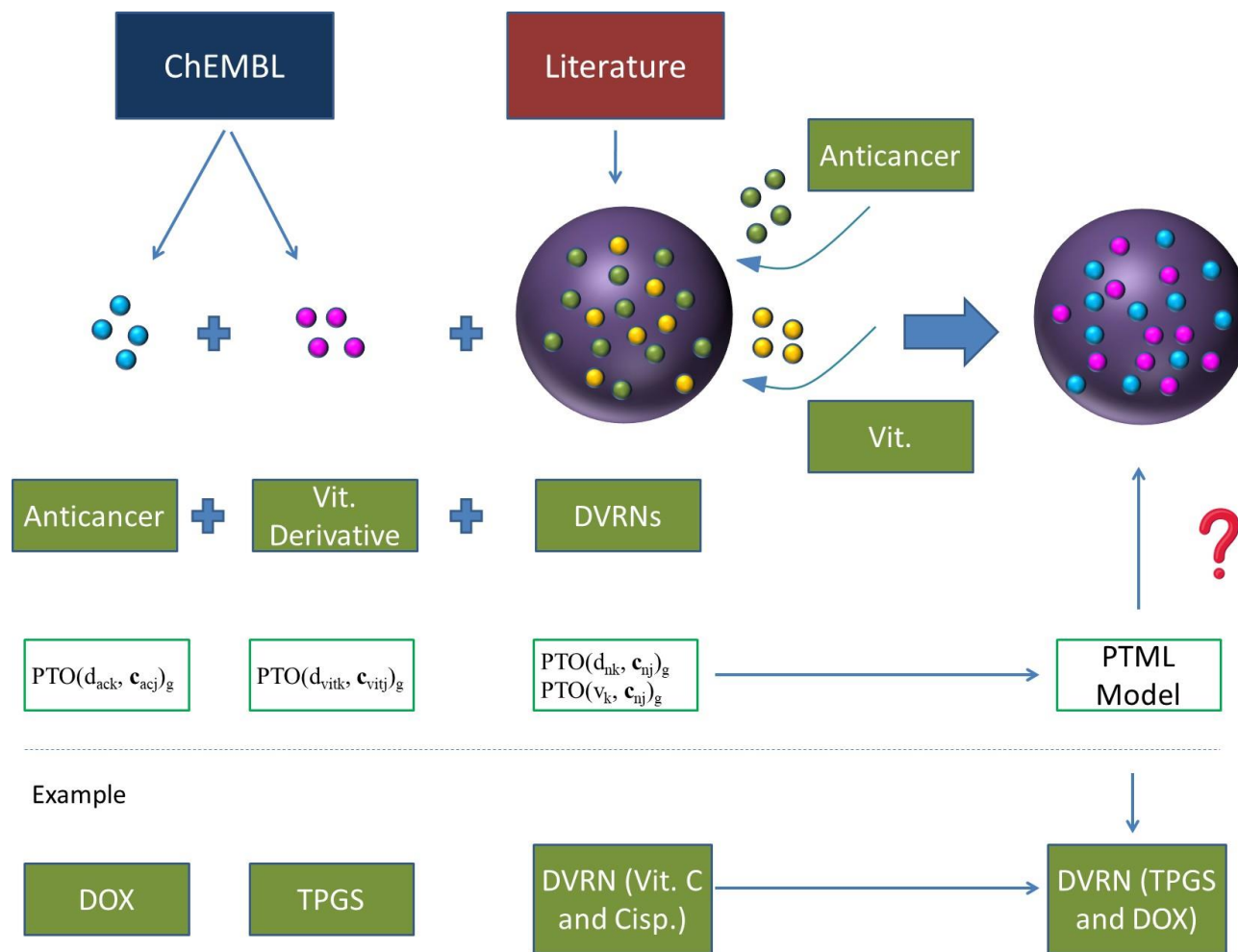


Figure 6. Scheme of compounds of DVRN for the development of PTML model. Original anticancer compounds (green) and vitamin derivatives (yellow) are substituted by new anticancer compound (blue) and vitamin derivative (pink) to improve biological activity and performance.

With all the information provided by these different descriptors as well as assay conditions, we created new a partition in 6 groups (g_i), taking as reference the partition with better results for the previous model (G_2), see **Figure 7**. As result, we can see how the information of the anticancer is accumulated in g_0 ($\Delta d_{1ac}(c_{0ac}), \Delta d_{1ac}(c_{1ac}), \Delta d_{1ac}(c_{2ac}), \Delta d_{1ac}(c_{3ac}), \Delta d_{1ac}(c_{4ac})$). The information regarding the anticancer and vitamins inside DVRNs is included in g_7 ($\Delta d_{1ac1n}(c_{5n}), \Delta d_{1ac2n}(c_{5n}), \Delta d_{1v1n}(c_{5n}), \Delta d_{1v2n}(c_{5n}), \Delta d_{1v3n}(c_{5n})$). The specific formula for calculation of all the PTOs is discussed in the next section, **Task 3**, as part of the comparative purpose.

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)



Figure 7. Partitions of variables in different groups used to calculate the different operators

The resulting model, after the standardization of input variables, is presented in Eq. 8 where we can observe that the reference function $f(v_{ijn}, v_{ijvit}, v_{ijac})_{ref}$ presents an significative weight and the different PTO generate the perturbation. $PTG(g_6)$ and $PTG(g_4)$, present the greater weights in the equation (they incorporate the information of the drug we want to incorporate and the type of the DVRN, respectively). This model shows a high χ^2 and p-value lower than 0.05, which means that it is statistically significative.

$$\begin{aligned}
 f(v_{ijn}, v_{ijvit}, v_{ijac})_{calc} &= -13.7650 + 4.4506 \cdot f(v_{ijn}, v_{ijvit}, v_{ijac})_{ref} \\
 &+ 0.0159 \cdot PTG(g_0) - 0.0202 \cdot PTG(g_1) \\
 &+ 0.0190 \cdot PTG(g_3) + 0.2455 \cdot PTG(g_4) \\
 &- 0.0189 \cdot PTG(g_5) + 0.0797 \cdot PTG(g_6) \\
 &+ 0.0117 \cdot PTG(g_7)
 \end{aligned} \tag{8}$$

$$n = 762637 \quad \chi^2 = 93355.22 \quad p < 0.05$$

The results for training and validations subsets are included in **Table 6**. For both subsets, the Sp and Sn are higher than 70%. The Sp is much higher than Sn in both cases. This result is given mainly that the number of desirable cases ($n = 6676$) is much lower than the number of no desirable cases ($n = 964076$). However, this model gives us the capacity to predict with general Ac of 96.5% for both subsets. We must take into consideration that although the sensitivity of this PTML-LDA model is lower than the previous PTML-ANN model showed in this work, it

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

includes more detailed information about anticancer drug and vitamin inside the DVRNs, which is a relevant step for designing nano-systems with desirable biological activity.

Table 6. Results of the model and input variables analyzed

Obs. Sets ^a	Stat. Param. ^b	Pred. Stat. ^c	Predicted sets		
			n _j	f(v _{ijn} , v _{ijvit} , v _{ijac}) _{pred} = 1	f(v _{ijn} , v _{ijvit} , v _{ijac}) _{pred} = 0
Training					
f(v _{ijn} , v _{ijvit} , v _{ijac}) _{obs} = 1	Sp	70.8	5119	3942	1627
f(v _{ijn} , v _{ijvit} , v _{ijac}) _{obs} = 0	Sn	96.7	757518	25205	732313
Total	Ac	96.5	762637		
Validation					
f(v _{ijn} , v _{ijvit} , v _{ijac}) _{obs} = 1	Sp	71.2	1557	1109	448
f(v _{ijn} , v _{ijvit} , v _{ijac}) _{obs} = 0	Sn	96.7	206558	6845	199713
Total	Ac	96.5	208115		

^a Obs. Sets = Observed sets, ^bStat. Param. = Statistical parameter, ^cPred. Stat. = Predicted statistics

PTML models comparison (Task 3). In previous papers our group and other research groups have published some ML and/or PTML models for the prediction of NP systems. In this section we are going to compare all these models in terms techniques used, applications, sensitivity, specificity, *etc.* In **Table 9**, we depict published works that present PTML models for biological activities of NP systems. We must say that this is not a review of the state of the art, but information to show the heterogeneity of the published studies and how the model we present contributes to cover the

current knowledge gap. At this point, models with Metal Oxide Nanoparticles (MONPs) have attracted significantly the attention of researchers, given the transversal applications in medicine field.^{29,4,30,31,7,32,9,33} Among them we can observe that there are different algorithms such as Linear Regression (LR), LDA or RF. It also exists heterogeneity in terms of the output (such as EC₅₀, LC₅₀, Zeta Potential) and the target cell (such as E. Coli or HaCaT). Other models include multiple levels of target cell and output; they apply Perturbation Operators that constitute PTML models given that include information not only of the descriptors but also the assay conditions. Among these PTML models we must highlight that Santana *et al.*¹⁹ presented in a previous paper a model with Non Metal Oxide Nanoparticles (NMONPs). This previous model, comparing to the recent state of the art, constitutes an important step in terms of design of NP systems in general, and DVRNs in particular. However, as we mentioned before the model published before does not take into consideration the possibility of changing the structure of the anticancer drug released by the DVRNs. This is an important handicap because one of the more important aspects in the design of a DVRNs for anticancer therapy is precisely the possibility of testing different anticancer drugs. In the present work we reached two important goals with respect to the previous model published in Nanoscale.¹⁹ Firstly, see **Task 1**, we improved the previous PTML model significantly by applying all the ML techniques described above for better sensitivity and specificity, see **Table 7**. The resulting PTML model is generated through the application of ANN and oversampling technique, and it presents much better performance. Secondly, see **Task 2**, we highlighted that with the used information, we are not able to change the anticancer drug linked to the DVRN. This is given that the structural information of these compounds are not included in the previous models, Nanoscale paper and **Task 1** of this paper. Please, note in **Table 7** that this is the only model of the table with the items Der. Vit. = **yes**, Vit. Struct. = **yes**, and

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

Antic.Struct. = **yes**. Thus, in this work we developed the first multipurpose model, see **Task 2**, able to design new DVRNs taking into consideration the information of the anticancer drug and vitamins inside DVRNs apart from the information we already had about the drug and the DVRNs.

Table 7. Comparison of ML/PTML models for nanoparticles biological activity

Model	NP Systems ML and/or PTML models *												
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)			
Der. Vit.	no	no	no	no	no	no	no	no	no	yes	yes	yes	yes
Vit. Struct	no	no	no	no	no	no	no	no	no	no	no	no	yes
Antic. Struct.	no	no	no	no	no	no	no	no	no	no	no	no	yes
Meth. ^a	RF	MC	LR	DTB	ANN	MLR/G A	LDA	LDA	LFER	LDA	LDA	ANN	LDA
Syst. ^b	MO	MO	MO	MO	MO	MO	MO	MO	nMO	nMO	MO	nMO	nMO
Appl. ^c	Med.	Med.	Med.	Med.	Med.	Med.	Med.	Med.	DD.	DD.	DD.	DD.	DD.
Drug. ^d	-	-	-	-	-	-	-	-	-	Mult.	Mult.	Mult.	Mult.
Output ^e	EC ₅₀ , LC ₅₀	pLC ₅₀	LC ₅₀	EC ₅₀	EC ₅₀	ZP	Mult.	Mult.	Mult.	Mult.	Mult.	Mult.	Mult.
Cell	HaCaT/ E.Coli	E. Coli	E. Coli	E. coli	E. coli	HaCaT	Mult.	Mult.	-	Mult.	Mult.	Mult.	Mult.

Acc. (Tra) ^f	R ² ≥ 0.93	R ≥ 0.90	R ² ≥ 0.80	Var. 96.9 %	R ² 0.92	R ² 0.82	93.58 %	98 %	93%	87.1 %	94.4 %	89.35 %	96.5 %
Ac. (Test) ^g	R ² 0.92	R 0.73 -0.98	-	Var. 91.3 %	Q ² _{ext} 0.86	Q ² _{ext} 0.87	93.57 %	98 %	-	87.2 %	94.5 %	93.23 %	96.5 %
Vars ^h	LMB, FSD	Quasi Smiles	AE, MHC, LE, <i>etc.</i>	OP, MolRef, PSA	EPS, NMA, NOA, CMC	MIBD, TBD	5 PTM A	8 PTM A	5 PTCO	5 PTP	10 PTMA	5 PTA	7 PTG
Ncases ⁱ	35	34	17	17	17 36 72	15	>1.7K	>36K	25K	>130K	500K	>130K	>970K
Ref.	29	4	30	31	7	32	9	33	34	19	Sub.	This work	This work

* (i) Novoselska, et al.; (ii) Toropova et al.; (iii) Pathakoti, et al.; (iv) Singh et al.; (v) Fjodorova et al. ; (vi) Mikolajczyk et al.; (vii) Luan et al.; (viii) Kleandrova et al.; (ix) Messina et al., and (x)Santana et al. ^a Method = Meth., LS = Light Scattering, RF = Random Forest, MC = Monte Carlo, LR = Linear Regression, DTB = decision treeboost, MLR = Multiple Linear Regression, GA = Genetic Algorithm, LFER = Linear Free Energy Relationship, CPANN = Counter Propagation Artificial Neural Network. ^b Syst. = System, PEO/Fe₂O₃ = poly(ethylene oxide-b-phenyl oxazoline) and poly(isoprene-b-ethylene oxide) (PEO-b-PPhOx and PI-b-PEO), MONP ZnO = ZnO-DOX@ZIF-8 with encapsulated iron oxide nanoparticles (γ -Fe₂O₃), Cd-QD = Cadmium Quantum Dots, MONPs = Metal Oxide Nano-Particles, nMONPs = non MONPs, MSNPs = Functional mesoporous silica nanoparticles, PLGANPs = poly(lactic-co-glycolic acid) nanoparticles, TPGS = Vitamin E TPGS, HA = Hyaluronic acid, Hyal = Hyaluronidase, MONP TiO₂ = upconverting nanoparticles with a mesoporous TiO₂, ^c Appl. = Application, Med. = Medicine, DD. = Drug Delivery. ^d DOX = Doxorubicin, FA = Folic Acid, IND = Indomethacin, CUR = Curcumin, ^e ZP = Zeta Potential ^f Acc. (Tra) = Accuracy Training, DC = Dark Toxicity, PIT = Photo Induced Toxicity, Var = Variance, ^g Q²_{ext} = externally validated regression coefficient (validation set), ^hFSD = Fragmental simplex descriptors, LMBC = ligand-metal binding characteristics, AE = Absolute electronegativity, MHC=molar heat capacity, LE= LUMO energies, OP = oxygen percent, MolRef = Molar refractivity, PSA = Polar surface area, EPS = χ -metal

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

electronegativity by Pauling scale, NMA= number of metal atoms in oxide, NOA = number of oxygen atoms in oxide, CMC = charge of metal cation in oxide, MIBD : microscopic image-based Descriptors, TBD = theory-based (calculated) descriptors, PTO = Perturbation Theory Operator. ⁱ Ncases = Number of cases used to train and/or validate the model, K = 1000.

In **Table 8**, we also give the formula used to calculate the PTOs used here to seek PTML models for VDRNs along with the formula for PTOs used in previous works for comparative purposes. Please, be aware that this last comparison only applies to models using PTML approach for the study of NP systems. The δ_{gj} is the coefficient of the variable in the operator: $\delta_{gj} = 1$ when the moving average for the condition c_j is included in the group of variables g affected by the operator, $\delta_{gj} = 0$ otherwise. Please, be aware, that when c_j is in boldface denotes a vector of multiple conditions of assay as in $PTO(d_{sk}, \mathbf{c}_j)_g$ but when c_j is not bold denotes one single condition of assay. We can conclude from the table that the first PTML models for nanosystems used mainly PTOs with the form of Moving Averages (MA) abbreviated here as PTMA. This limited the application of the operator to only one structural feature (d_k) and one external condition (c_{kj}) of one part of the system at time, *e.g.*; Luan *et al.* and Kleandrova *et al.* In means that, the number of descriptors (d) of type (k) for a sub-system (s) and the number assay conditions (c) of type (j) for a sub-system (s) that can be included in the PTO are $N_{dsk} = 1$ and $N_{csj} = 1$, respectively, see **Table 9**. A more complete PTO introduced by Messina *et al.*³⁴ was the PTO based on Covariance forms (PTCOs). This operator allowed considering up to two structural features ($N_{dsk} = 2$) and up to two experimental conditions ($N_{csj} = 2$) for up to two different parts of the nanosystems.³⁴ It implied the use of many variables in the models when we need to describe systems with multiple

parts and experimental conditions as is the case of the DVRNs. In consequence, in our previous work we explored multiplicative operators like PTPs and PTGs able to zip multiple structural features of different parts of the system (d_{sk}); drug, vitamin, coating, core, etc. and/or multiple external conditions of these systems (c_{sj}) at the same time. However, we can imagine other alternatives to multiplicative PTOs; that is why in this work we tested the metric-based PTOs such as PTA, PTE, and PTM in the **Task 1** for comparative purposes. These new PTOs have $N_{dsk} = q$ and $N_{csj} \leq q$ with q equal to any natural number $q = 1, 2, 3, \dots, q_{max}$ according to the number of variables we want to group in the partition g . The parameter $N_{csj} \leq q$ is because some $\Delta d_{sk}(c_{sj})$ may have the same experimental condition if necessary appearing repeated values of c_{sj} in the PTO, see examples in **Figure 7**.

Table 8. PTOs used in this work and in other papers to describe nanoparticle systems

PTO	PTO formula	N_{dsk}^a	N_{csj}^b	Ref.
Moving Average (PTMA)	$\Delta d_{sk}(c_{sj})$	1	1	33
Co-variance (PTCO)	$\Delta d_{sk}(c_{sj}) \cdot \Delta d_{s'k}(c_{s'j})$	2	2	34

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

Product (PTP)	$PTP(d_{sk}, c_{sj})_g = G \left[\sum_{j=0}^q [\Delta d_{kvit}(c_{jvit})^{\delta_{gj}}] \right]$	q	$\leq q$	19
Geom. Mean (PTG)	$PTG(d_{sk}, c_{sj})_g = \left\{ G \left[\sum_{j=0}^q [\Delta d_{kvit}(c_{jvit})^{\delta_{gj}}] \right]^{1/q} \right\}$	q	$\leq q$	19
Arithm. Mean (PTA)	$PTA(d_{sk}, c_{sj})_g = \frac{1}{q} \sum_{j=0}^q [\delta_{gjk} \cdot \Delta d_{sk}(c_{sj})]$	q	$\leq q$	This work
Euclid. Dist. (PTE)	$PTE(d_{sk}, c_{sj})_g = \left\{ \sum_{j=0}^q [\delta_{gjk} \cdot \Delta d_{sk}(c_{vitj})]^2 \right\}^{1/2}$	q	$\leq q$	This work
Manh. Dist. (PTM)	$PTM(d_{sk}, c_{sj})_g = \sum_{j=0}^q \delta_{gj} \cdot \Delta d_{sk}(c_{sj}) $	q	$\leq q$	This work

^a N_{dsk} = Number of descriptors (d) of type (k) for a sub-system (s) that can be included in the PTO.
 N_{csj} = Number of assay conditions (c) of type (j) for a sub-system (s) that can be included in the PTO.

■ CONCLUSIONS

Cheminformatics models can be useful tools to predict biological activities of compounds. The same methods can be applied to nano-systems that are promising but not well characterized. PTML can be used to create models with heterogeneous datasets and Big Data characteristics. It is useful for rational discovery purposes and has a flexible framework with the capacity to build a multi input and multi output model. In the first task, nonlinear models improve performance significantly over linear models. Among nonlinear models, neural networks (PTML-ANN) performed the best of the models we tried. Combining the predictions of multiple neural networks allows us to estimate the uncertainty in a particular prediction, or for a particular class of materials. We achieved high sensitivity by oversampling the desirable cases in the training subset. This technique allowed us to produce a model to reduce the uncertainty of the prediction for a particular DVRN, with specificity, sensitivity and accuracy in the range of 90 % - 95 % for DVRNs biological activities. In the second task, we used IF techniques to carry out a data enrichment of our previous dataset. In so doing, we constructed a new working data set of >970000 cases with data of preclinical assays of DVRNs, vitamins, and anticancer compounds from ChEMBL database. We expressed the new information with PTOs and trained a PTML model that is a qualitatively new because it also incorporates information of the anticancer drugs. The new model presents 96-97% of accuracy for training and external validation subsets. We present here by the first time a multipurpose PTML model able to select nanoparticles, anticancer compounds, and vitamins and their conditions of assay for DVRNs design. Last, in a third task, we carry out a comparative study of ML and/or PTML models published and how the models we are presenting cover a gap of knowledge in terms of drug delivery.

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

■ ASSOCIATED CONTENT


Supporting Information

The dataset used, including molecular descriptors, and assay conditions, desirability, cutoff, biological activities etc., was included in tables **Table S1**, **Table S2** and **Table S3** (SI01.xlsx). See details about these moving average operators on **Table 2** and **Table 3** (see **Table S1** and **Table S2** respectively in supporting information for full dataset consultation). In addition, **Table S3** we also included all details about each case, observed classification, predicted classification, input variables, experimental conditions, vitamin derivative and nanoparticle characteristics.

■ AUTHORS INFORMATION

Corresponding author

*(H.G.D) E-mail: humberto.gonzalezdiaz@ehu.es (H.G.-D.)

 Orcid: 0000-0002-9392-2797

■ ACKNOWLEDGMENTS

R.S.C. thanks COLCIENCIAS scholarship for the doctorate studies: “Convocatoria para Doctorado Nacional 757” from 2017. This original research is part of the project “Investigación en Derecho Internacional y Nanotecnología” registered in the Research Centre of Universidad

Pontificia Bolivariana with register number 766B-06/17-37. Special gratitude is extended to CYTED NANOCELIA network. The authors acknowledge research grants from Ministry of Economy and Competitiveness, MINECO, Spain (FEDER CTQ2016-74881-P) and Basque government (IT1045-16). The authors also acknowledge the support of Ikerbasque, Basque Foundation for Science.

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

■ REFERENCES

- (1) Russo, D. P.; Zorn, K. M.; Clark, A. M.; Zhu, H.; Ekins, S. Comparing Multiple Machine Learning Algorithms and Metrics for Estrogen Receptor Binding Prediction. *Mol. Pharm.* **2018**, *15* (10), 4361–4370. <https://doi.org/10.1021/acs.molpharmaceut.8b00546>.
- (2) Lane, T.; Russo, D. P.; Zorn, K. M.; Clark, A. M.; Korotcov, A.; Tkachenko, V.; Reynolds, R. C.; Perryman, A. L.; Freundlich, J. S.; Ekins, S. Comparing and Validating Machine Learning Models for Mycobacterium Tuberculosis Drug Discovery. *Mol. Pharm.* **2018**, *15*, 4346–4360. <https://doi.org/10.1021/acs.molpharmaceut.8b00083>.
- (3) Korotcov, A.; Tkachenko, V.; Russo, D. P.; Ekins, S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharm.* **2017**, *14* (12), 4462–4475. <https://doi.org/10.1021/acs.molpharmaceut.7b00578>.
- (4) Toropova, A. P.; Toropov, A. A.; Rallo, R.; Leszczynska, D.; Leszczynski, J. Optimal Descriptor as a Translator of Eclectic Data into Prediction of Cytotoxicity for Metal Oxide Nanoparticles under Different Conditions. *Ecotoxicol. Environ. Saf.* **2015**, *112*, 39–45.
- (5) Labouta, H., Asgarian, N., Rinker, K., Cramb, D. Meta-Analysis of Nanoparticle Cytotoxicity via Data-Mining the Literature. *ACS Nano* **2019**, *13* (2), 1583–1594.
- (6) Yan, X.; Sedykh, A.; Wang, W.; Zhao, X.; Yan, B.; Zhu, H. In Silico Profiling Nanoparticles: Predictive Nanomodeling Using Universal Nanodescriptors and Various Machine Learning Approaches. *Nanoscale* **2019**, *11* (17), 8352–8362.
- (7) Fjodorova, N.; Novic, M.; Gajewicz, A.; Rasulev, B. The Way to Cover Prediction for Cytotoxicity for All Existing Nano-Sized Metal Oxides by Using Neural Network Method. *Nanotoxicology* **2017**, *11* (4), 475–483. <https://doi.org/10.1080/17435390.2017.1310949>.
- (8) Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X., Dasari, T., Michalkova, A., Hwang, H., Toropov, A., Leszczynska, D., Leszczynski, J. Using Nano-QSAR to Predict the Cytotoxicity of Metal Oxide Nanoparticles. *Nat. Nanotechnol.* **2011**, *6* (3), 175–178.
- (9) Luan, F.; Kleandrova, V. V.; González-Díaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, N. Computer-Aided

- Nanotoxicology: Assessing Cytotoxicity of Nanoparticles under Diverse Experimental Conditions by Using a Novel QSTR-Perturbation Approach. *Nanoscale* **2014**, *6* (18), 10623–10630. <https://doi.org/10.1039/c4nr01285b>.
- (10) Gajewicz, A., Schaeublin, N., Rasulev, B., Hussain, S., Leszczynska, D., Puzyn, T.; Leszczynski, J. Towards Understanding Mechanisms Governing Cytotoxicity of Metal Oxides Nanoparticles: Hints from Nano-QSAR Studies. *Nanotoxicology* **2015**, *9* (3), 313–325.
- (11) Zhao, Z. B.; Long, J.; Zhao, Y. Y.; Yang, J. B.; Jiang, W.; Liu, Q. Z.; Yan, K.; Li, L.; Wang, Y.-C.; Lian, Z. X. Adaptive Immune Cells Are Necessary for the Enhanced Therapeutic Effect of Sorafenib-Loaded Nanoparticles. *Biomater. Sci.* **2018**, *6* (4), 893–900. <https://doi.org/10.1039/c8bm00106e>.
- (12) Othayoth, R., Mathi, P.; Bheemanapally, K., Kakarla, L.; Botlagunta, M. Characterization of Vitamin-Cisplatin-Loaded Chitosan Nanoparticles for Chemoprevention and Cancer Fatigue. *J. Microencapsul.* **2015**, *32* (6), 578–588. <https://doi.org/10.3109/02652048.2015.1065921>.
- (13) Bediaga, H.; Arrasate, S.; González-Díaz, H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Comb. Sci.* **2018**, *20* (11), 621–632. <https://doi.org/10.1021/acscombsci.8b00090>.
- (14) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2011**, *40* (1), 1100–1107.
- (15) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2016**, *45* (1), 945–954.
- (16) Da Costa, J. F.; Silva, D.; Caamaño, O.; Brea, J. M.; Loza, M. I.; Munteanu, C. R.; Pazo, A.; García-Mera, X.; González-Díaz, H. Perturbation Theory/Machine Learning Model of ChEMBL Data for Dopamine Targets: Docking, Synthesis, and Assay of New L-Prolyl-L-Leucyl-Glycinamide Peptidomimetics. *ACS Chem Neurosci* **2018**, *9* (11), 2572–2587. <https://doi.org/10.1021/acchemneuro.8b00083>.
- (17) González M., Monserrat, J., Rasulev B., Casañola-Martín, G., Barreiro, J., Paraíso-Medina, S., Maojo, V., González Díaz, H., Pazos, A., Munteanu, C. Carbon Nanotubes ' Effect on Mitochondrial Oxygen Flux Dynamics : Polarography Experimental Study and Machine

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

- Learning Models Using Star Graph Trace Invariants of Raman Spectra. *Nanomaterials* **2017**, *7*, 386–400.
<https://doi.org/10.3390/nano7110386>.
- (18) Simón-Vidal, L.; García-Calvo, O.; Oteo, U.; Arrasate, S.; Lete, E.; Sotomayor, N.; González-Díaz, H. Perturbation-Theory and Machine Learning (PTML) Model for High-Throughput Screening of Parham Reactions: Experimental and Theoretical Studies. *J. Chem. Inf. Model.* **2018**, *58* (7), 1384–1396. <https://doi.org/10.1021/acs.jcim.8b00286>.
- (19) Santana, R.; Zuluaga, R.; Gañán, P.; Arrasate, S.; Onieva, E.; González-Díaz, H. Designing Nanoparticle Release Systems for Drug–Vitamin Cancer Co-Therapy with Multiplicative Perturbation-Theory Machine Learning (PTML) Models. *Nanoscale* **2019**, *11*, 21811–21823. <https://doi.org/10.1039/c9nr05070a>.
- (20) González-Díaz, H.; Arrasate, S.; Gómez-San Juan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. General Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry. *Curr. Top. Med. Chem.* **2013**, *13* (14), 1713–1741. <https://doi.org/10.2174/1568026611313140011>.
- (21) Arrasate, S.; Duardo-Sanchez, A. Perturbation Theory Machine Learning Models: Theory, Regulatory Issues, and Applications to Organic Synthesis, Medicinal Chemistry, Protein Research, and Technology. *Curr. Top. Med. Chem.* **2018**, *18* (14), 1203–1213.
- (22) Da Costa, J. F.; Silva, D.; Caamaño, O.; Brea, J. M.; Loza, M. I.; Munteanu, C. R.; Pazos, A.; García-Mera, X.; González-Díaz, H. PTML Model of ChEMBL Data for Dopamine Targets, Docking, Synthesis, and Assay of New PLG Peptidomimetics. *ACS Chem. Neurosci.* **2018**, *9* (11), 2572–2587.
- (23) Blay, V.; Yokoi, T.; González-Díaz, H. Perturbation Theory-Machine Learning Study of Zeolite Materials Desilication. *J. Chem. Inf. Model.* **2018**, *58* (12), 2414–2419.
- (24) Vásquez-Domínguez, E.; Armijos-Jaramillo, V. D.; Tejera, E.; González-Díaz, H. Multioutput Perturbation-Theory Machine Learning (PTML) Model of ChEMBL Data for Antiretroviral Compounds. *Mol. Pharm.* **2019**, *16*, 4200–4212.

<https://doi.org/10.1021/acs.molpharmaceut.9b00538>.

- (25) Kleandrova, V. V.; Luan, F.; González-Díaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, N. M. Computational Ecotoxicology: Simultaneous Prediction of Ecotoxic Effects of Nanoparticles under Different Experimental Conditions. *Environ. Int.* **2014**, *73*, 288–294. <https://doi.org/10.1016/j.envint.2014.08.009>.
- (26) Mitchell, J. Machine Learning Methods in Chemoinformatics. *WIREs Comput Mol Sci* **2014**, *4*, 468–481. <https://doi.org/10.1002/wcms.1183>.
- (27) Balakrishnama, S.; Ganapathiraju, A. Linear Discriminant Analysis-a Brief Tutorial. *Inst. Signal Inf. Process.* **1998**, *18*, 1–8.
- (28) Learn, S. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.
- (29) Novoselska, N.; Rasulev, B.; Gajewicz, A.; Kuzmin, V.; Puzyn, T.; Leszczynski, J. From Basic Physics to Mechanisms of Toxicity: The “liquid Drop” approach Applied to Develop Predictive Classification Models for Toxicity of Metal Oxide Nanoparticles. *Nanoscale* **2014**, *6* (22), 13986–13993.
- (30) Pathakoti, K.; Huang, M.-J.; Watts, J. D.; He, X.; Hwang, H.-M. Using Experimental Data of Escherichia Coli to Develop a QSAR Model for Predicting the Photo-Induced Cytotoxicity of Metal Oxide Nanoparticles. *J. Photochem. Photobiol. B Biol.* **2014**, *130*, 234–240.
- (31) Singh, K. P.; Gupta, S. Nano-QSAR Modeling for Predicting Biological Activity of Diverse Nanomaterials. *RSC Adv.* **2014**, *4* (26), 13215–13230.
- (32) Mikolajczyk, A., Gajewicz, A., Rasulev, B., Schaeublin, N., Maurer-Gardner, E., Hussain, S., Leszczynski, J., Puzyn, T. Zeta Potential for Metal Oxide Nanoparticles: A Predictive Model Developed by a Nano-Quantitative Structure-Property Relationship Approach. *Chem. Mater.* **2015**, *27* (7), 2400–2407. <https://doi.org/10.1021/cm504406a>.
- (33) Kleandrova, V. V.; Luan, F.; González Díaz, H.; Ruso, J. M.; Speck-planche, A.; Nata, M.; Cordeiro, D. S. Computational Tool for Risk Assessment of Nanomaterials: Novel QSTR-Perturbation Model for Simultaneous Prediction of Ecotoxicity and Cytotoxicity of Uncoated and Coated Nanoparticles under Multiple Experimental Conditions. *Environmental Sci. Technol.* **2014**, *48*, 14686–14694.

8) MODELLING DVRNS (METRIC OPERATORS AND ENRICHMENT OF INFORMATION)

- (34) Messina, P. V.; Besada-Porto, J. M.; González-Díaz, H.; Ruso, J. M. Self-Assembled Binary Nanoscale Systems: Multioutput Model with LFER-Covariance Perturbation Theory and an Experimental-Computational Study of NaGDC-DDAB Micelles. *Langmuir* **2015**, *31* (44), 12009–12018. <https://doi.org/10.1021/acs.langmuir.5b03074>.

*Do what you can with all you have
wherever you are*

Theodore Roosevelt

CHAPTER

9

9) Conclusions and Future

273

Works

Throughout this work the main objective has been achieved: to present applicable models to design nanosystems including vitamin derivatives. To do so, an exhaustive analysis of the current European regulations is presented, for drug, food and cosmetic sections. Then, a vitamin derivative PTML model has been presented in order to better predict their biological activity. Then, we show a model able to predict with high rates of specificity and sensibility nanosystems including metal oxide nanoparticles and vitamin derivatives. Finally we present a PTML model for non metal oxide nanoparticles able to predict notably high, incorporating information about not only the assay conditions, vitamin

9) CONCLUSIONS AND FUTURE WORKS

derivatives and nanoparticles but also anticancer compounds to better design nanosystems with desirable properties. The objective of this chapter is to summarize the main conclusions of each chapter previously exposed.

9.1 Conclusions

The design of nanosystems is gaining momentum given the transversal potential of nanotechnology in different economic sectors. The efficacy and desirable properties of these nanosystems depends on the design. An efficient process for design and production means not only a potential diminution of costs and time but also a justification for authorization process.

Chapter 2 explores the regulation of nanotechnology for cosmetic sector. Regulation 1223/2009 includes the precautionary principle and states the importance of safety for cosmetic including nanomaterials, for public safety purposes. However its application found difficulties given the gaps of knowledge so from a law point of view we recommend to take it into account in future regulations for coherence and legal security.

Chapter 3 presents and evaluates the regulation for food sector in the context of the European Union. The specific directives and regulations, include rules for food including nanotechnology. The authorization process to market these products include criteria in order to ensure public safety according to EFSA technical opinions. No opinions guarantee completely safety of the nanomaterials analyzed. We did not find an opinion that take into consideration ML techniques. Algorithms and statistic methods from ML field should be applied to know more about the behavior of these compounds.

Chapter 4 focuses on drug sector. Specifically, the drugs that include nanotechnology and that must be authorized by the Centralized Authorization Process of Nanomedicines in European Union. As commented in Chapter 2, ML techniques should be applied to infer a greater knowledge from these compounds. The resulting models must accomplish the recommendation from OECD, mainly for transparency purposes. In this sense, we propose PTML, as one of the promising techniques to build models able to improve the referred

9) CONCLUSIONS AND FUTURE WORKS

procedures.

Chapter 5 shows the importance of vitamin derivatives given their functionalities and the important role specially in pharmaceutical sector. It includes the development of the first PTML-LDA model for vitamin derivatives. This model includes information about assay conditions and molecular descriptors. Different algorithms such as Linear Discriminant Analysis (LDA), Näive Bayes (NB), and Random Forest (RF) are applied to find the model with higher sensibility, specificity and accuracy, taking into consideration complexity. These techniques are applied for the models presented in the rest of the chapters.

Chapter 6 includes the discussion about the importance of nanotechnology and specially nanosystems loaded with vitamin derivatives to improve their biological behavior. It shows the first PTML-LDA model that includes information about descriptors of metal oxide nanoparticles with or without coating agent and the descriptors of vitamin derivatives. This model presents high rates of prediction for biological behavior of the different compounds that constitute this kind of nanosystems.

Chapter 7 focuses on a biological behavior of compounds of nanosystems composed by non metal oxide nanoparticles with vitamin derivatives. These nanosystems are specially used for cancer cotherapy. Descriptors of the nanosystem and the vitamin derivatives are used to generate a PTML-LDA model that also incorporates information about the assay conditions.

Chapter 8 presents a summary of the models for nanoparticles in literature and highlight the importance of cover the knowledge gap of drug release. We developed a more accurated PTML-ANN for nanosystems presented in chapter 6. In this case, beside this improvement, we present another PTML-LDA model for non metal oxide nanoparticles that include anticancer drug and vitamin

derivative descriptors. This model is able to generate predictions for different anticancer drug, vitamin derivatives and non metal oxide nanoparticles.

We propose these contributions as a limit for the current gap of knowledge in terms of desing of nanosystems and their influence to better apply nanotechnology regulation. This dissertation has been designed and developed to better ensure efficacy and public safety for products including nanomaterials.

9) CONCLUSIONS AND FUTURE WORKS

9.2 Future work

We are able to affirm that nanotechnology will be one of the most applicable and transversal technologies in different sectors, given the promotion of European projects involving this particular technology.

The regulation must be developed according to the state of art of the technology. The regulation should include in each specific sector, how Machine Learning techniques can be applied to better ensure public safety. As we have more data, we should know how to use it to complement, for instance, process of authorization. The possibilities in nanotechnology field are massive, however we can detect patterns to identify better relevant characteristics that explain the behavior of nanomaterials or their components. European Union should fund this type of research given the potential in innovation, economic and social terms.

The production of data of behavior of nanomaterials will be higher the next years. We are able to produce through Machine Learning techniques models to predict behavior of components of nanosystems. As we have more information we should predict for design purposes in order to respect the three R principles, according to green chemistry fundamentals. Besides the greater information, ML algorithms will be in improvement. We therefore must take it into consideration given the possibilities of understanding these issues will be even higher.

As we have more information about nanosystems behavior we will be able to have more information of the system itself and Machine Learning techniques will let us go further for complex systems with specific functions, not only for cancer cotherapy but also for improvement of the drug delivery and less invasive treatments. Besides pharmaceutical purposes, the knowledge is also applicable in different sectors such as cosmetics and food. The margin for improvement is

notable when nanotechnology is included in manufacturing process.

Bibliography

- 1 K. A. Brown, S. Brittman, N. Maccaferri, D. Jariwala and U. Celano, *Nano Lett.*, 2019, **20**, 2–10.
- 2 Y. Xie, C. Zhang, X. Hu, C. Zhang, S. P. Kelley, J. L. Atwood and J. Lin, *J. Am. Chem. Soc.*, 2020, **142**, 1475–1481.
- 3 B. Sun, M. Fernandez and A. S. Barnard, *J. Chem. Inf. Model.*, 2017, **57**, 2413–2423.
- 4 Z. B. Zhao, J. Long, Y. Y. Zhao, J. B. Yang, W. Jiang, Q. Z. Liu, K. Yan, L. Li, Y.-C. Wang and Z. X. Lian, *Biomater. Sci.*, 2018, **6**, 893–900.
- 5 P. Othayoth, R., Mathi, L. Bheemanapally, K., Kakarla and M. Botlagunta, *J. Microencapsul.*, 2015, **32**, 578–588.
- 6 L. Simón-Vidal, O. García-Calvo, U. Oteo, S. Arrasate, E. Lete, N. Sotomayor and H. González-Díaz, *J. Chem. Inf. Model.*, 2018, **58**, 1384–1396.
- 7 J. F. Da Costa, D. Silva, O. Caamaño, J. M. Brea, M. I. Loza, C. R. Munteanu, A. Pazos, X. García-Mera and H. González-Díaz, *ACS Chem. Neurosci.*, 2018, **9**, 2572–2587.
- 8 H. Martínez-Arzate, S., Tenorio-Borroto, E., Barbabosa Pliego, A., Díaz-Albiter, H.M., Vázquez-Chagoyán, J.C. & González-Díaz, *J. Proteome Res.*, 2017, **16**, 4093–4103.
- 9 R. Santana, R. Zuluaga, P. Gañán, S. Arrasate, E. Onieva and H. González-Díaz, *Nanoscale*, 2019, **11**, 21811–21823.
- 10 H. Martínez-Arzate, S.G., Tenorio-Borroto, E., Barbabosa Pliego, A., Diaz-Albiter, H.M., Vázquez-Chagoyán, J.C. & González-Díaz, *J. Proteome Res.*, 2017, **16**, 4093–4103.
- 11 E. Vásquez-Domínguez, V. D. Armijos-Jaramillo, E. Tejera and H. González-Díaz, *Mol. Pharm.*, 2019, **16**, 4200–4212.
- 12 V. Blay, T. Yokoi and H. González-Díaz, *J. Chem. Inf. Model.*, 2018, **58**,

2414–2419.

- 13 J. F. Da Costa, D. Silva, O. Caamaño, J. M. Brea, M. I. Loza, C. R. Munteanu, A. Pazo, X. García-Mera and H. González-Díaz, *ACS Chem Neurosci*, 2018, **9**, 2572–2587.
- 14 C. González M., Monserrat, J., Rasulev B., Casañola-Martín, G., Barreiro, J., Paraíso-Medina, S., Maojo, V., González Díaz, H., Pazos, A., Munteanu, *Nanomaterials*, 2017, **7**, 386–400.

Declaration

I herewith declare that I have produced this work without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This work has not previously been presented in identical or similar form to any examination board.

The dissertation work was conducted from 2017 to 2020 under the supervision of Dr. Humberto González Díaz at the Basque Country University, Dr. Enrique Onieva Caracuel of University of Deusto and Dr. Piedad Gañán Rojo at Universidad Pontificia Bolivariana.

Bilbao, 17th January 2020.

This dissertation was finished writing in Bilbao on Tuesday 17th Jan, 2021

This page is intentionally left blank