IFPTML Algorithms: From Cheminformatics Models to Software Development, Startup Creation, and Innovation Transference

Author: Harbil Bediaga Bañeres

PhD ThesisUDC / 2023

Director: Prof. Dr. Alejandro Pazos Sierra– UDC

Co-Director: Prof. Dr. Humberto González Díaz – UPV/EHU

DoctorateProgram in Technologies of Information andComunications

UNIVERSIDADE DA CORUÑA

UNIVERSIDADE DA CORUÑA

Prof. Dr. D. Alejandro Pazos Sierra, University Professor of the Department of Computer Science and Information Technology, Faculty of Informatics, CITIC-Information and Communications Technology Research Center, University of A Coruña, Institute of Biomedical Research of ACoruña (INIBIC), A Coruña, Spain.

Prof. Dr. Humberto González Díaz, IKERBASQUE Professor, Faculty of Science and Technology, University of the Basque Country UPV/EHU, Leioa, Basque Country, Spain.

CERTIFY THAT:

The research report **"IFPTML Algorithms: From Cheminformatics Models,toSoftware Development, Startup Creation, and Innovation Transference"** has been carried out by Ms. HARBIL BEDIAGA BAÑERES, under our direction in the Research Program PhD in INFORMATION AND COMMUNICATIONS TECHNOLOGIES, and constitutes the Thesis that she presents to qualify for the Doctorate Degree of the University of A Coruña.

A Coruña, **July 26, 2023.**

Prof. Dr. Alejandro Pazos Sierra
Director de Tesis

Prof. Dr. Humberto González Díaz
Co-Director de Tesis

to my parents,
to my sister Mare-Labrit,
to my boyfriend Jon.

Acknowledgements

Acknowledgements

First of all, I want to express my gratitude to the directors of this thesis, Prof. Dr. Alejandro Pazos University of ACoruña (UDC) and Prof. Dr. Humberto González Díaz, University of A Coruña (UDC), IKERBASQUE, Basque Foundation for Sciences, for all the support and dedication they have given me during this work. Likewise, I want to thank all the collaborators of their respective groups. On one side, all collaborators of Artificial Neuron Networks and Adaptive Systems - Medical Imaging and Radiological Diagnosis (RNASA-IMEDIR) group, University of A Coruña (UDC). On the other hand, I would like to thank the researchers and students of the CHEM.PTML Laboratory, a multi-center group of the Department of Organic and Inorganic Chemistry, University of the Basque Country (UPV/EHU), and the Institute of Biophysics of the University of the Basque Country (UPV/EHU) and the Higher Council for Scientific Research (CSIC). In addition, I would like to especially thank all the IKERDATA S.L. company founding partners and employees for their support along with all the people who helped us in this endeavor. On this regar, I would like too specially thank Prof. Dr. Noemí Peña, Director of Entrepreneurship and Transfer at Campus of Biscay in UPV/EHU, ZITEK bussines incubator programme manager, Prof. Dr Juan B. ArrueMendizabal BEAZ S.A., ZITEK Business Incubator Manager,Profesorof ESI Bilbao UPV/EHU, and Lic. AitorIsasi ZITEK Business Incubator,Entrepreneurship technician.

Last, not the least, I also want to express my deeper gratitude to my parents, my sister, and my boyfriend for their unconditional support throughout this journey.

To my friends and all those people who have support me during the development of this work, for their collaboration and patience.

Thank you very much to you all!

SUMMARY

SUMMARY

The irruption in scence of Drug High Throughput Screening (HTS) technologies has prompted an explosion in the report of pre-clinical assays data for new hit-to-lead compounds with potential as Active Pharmaceutical Ingredients (APIs) in Pharmaceutical Industry. The analysis of all this data with techniques Artificial Intelligence (AI) may lead to the development of new predictive models. These models may be used in turn to predict more specific and safer compounds which the consequent reduction of costs in time and resourcers in APIs development. However, AI analysis of this presents many of the challenges of Big Data problems. It means, shortly, data analysis problems with issues related to Volumen, Velocity, Veracity, Variability, Value, and Complexity(5V + C). The first and second Vs are more or less self-explained and the Variability, Veracity, Value, and Complexity issues refers to data with problems of missing data, not consistent tendencies, errors, contradictory reports, interrelations such as co-llinearity/co-dependent labels forming complex networks, read-across (multi-species, multi-ouput, multi-scale) information, perturbations in multiple input/output variables, multi-labelling problems, *etc*. In this context, our group introduced the Information Fusion, Perturbation Theory, and Machine Learning (IFPTML) algorithm to facilitate the development of read-across multi-output models able to predict multiple outcomes of chemical compounds/drugs in pre-clinical assays. Our group has also reported a software calle SOFT.PTML which is a general purpose platform for IFPTML modeling. However, many aspects are yet to be covered. Many problems such as anti-cancer compounds discovery, allosteric compounds assays studies have not been analyzed with IFPTML algorithms. In addition, specific user-friendly software for these problemas has not

been already reported. Last, despite the potential application in industry no startuptup company was developed (at the beginning of this tesis) for the transference of IFPTML technology to industry. Consequently, the objective of this tesis focus firstly on the development of new IFPTML models of anti-cancer compounds and allosteric compounds assays. Next, we report the development of the user-friendly software LAGA for the read-across prediction of anti-cancer compounds. Last, we describe the planning, creation, structure, services, etc. of IKERDATA S.L a new inter-university startup company focused on the transference of IFPTML technology to Galician and Basque Country companies in the first instance with perspectives of Spain, Europe, and Worldwide proyection.

RESUMO

RESUMO

A irrupción nas tecnoloxías de detección de alto rendemento de drogas (HTS) provocou unha explosión no informe de datos de ensaios preclínicos para novos compostos de acerto a plomo con potencial como ingredientes farmacéuticos activos (API) na industria farmacéutica. A análise de todos estes datos con técnicas de Intelixencia Artificial (IA) pode levar ao desenvolvemento de novos modelos preditivos. Estes modelos poden utilizarse á súa vez para predicir compostos máis específicos e seguros que a consecuente redución de custos en tempo e recursos no desenvolvemento de API. Non obstante, a análise da intelixencia artificial presenta moitos dos desafíos dos problemas de Big Data. Significa, en breve, problemas de análise de datos con problemas relacionados co volume, a velocidade, a veracidade, a variabilidade, o valor e a complexidade (5V + C). A primeira e a segunda V son máis ou menos autoexplicativas e as cuestións de Variabilidade, Veracidade, Valor e Complexidade refírese a datos con problemas de falta de datos, tendencias non consistentes, erros, informes contraditorios, interrelacións como co-linealidade/co-linealidade. etiquetas dependentes que forman redes complexas, información de lectura cruzada (multi-especie, multi-saída, multiescala), perturbacións en varias variables de entrada/saída, problemas de multi-etiquetado, etc. Neste contexto, o noso grupo introduciu a fusión de información, Teoría da perturbación e algoritmo de aprendizaxe automática (IFPTML) para facilitar o desenvolvemento de modelos de lectura cruzada de múltiples saídas capaces de predecir múltiples resultados de compostos/fármacos químicos en ensaios preclínicos. O noso grupo tamén informou dun software calle SOFT.PTML que é unha plataforma de propósito xeral para o modelado IFPTML. Non obstante, moitos aspectos aínda están por cubrir. Moitos

problemas como o descubrimento de compostos anticanceríxenos, os estudos de ensaios de compostos alostéricos non foron analizados con algoritmos IFPTML. Ademais, aínda non se informou de software específico e amigable para estes problemas. Por último, a pesar da potencial aplicación na industria non se desenvolveu ningunha empresa de inicio (ao comezo desta tese) para a transferencia da tecnoloxía IFPTML á industria. En consecuencia, o obxectivo desta tese céntrase en primeiro lugar no desenvolvemento de novos modelos IFPTML de ensaios de compostos anticanceríxenos e compostos alostéricos. A continuación, informamos do desenvolvemento do software LAGA fácil de usar para a predición de lectura cruzada de compostos anticanceríxenos. Por último, describimos a planificación, creación, estrutura, servizos, etc. de IKERDATA S.L unha nova startup interuniversitaria centrada na transferencia de tecnoloxía IFPTML a empresas galegas e do País Vasco en primeira instancia con perspectivas de España, Europa e Europa. Proxección mundial.

RESUMEN

RESUMEN

La irrupción en escena de las tecnologías de detección de alto rendimiento (HTS) de fármacos ha provocado una explosión en el informe de datos de ensayos preclínicos para nuevos compuestos hit-to-lead con potencial como ingredientes farmacéuticos activos (API) en la industria farmacéutica. El análisis de todos estos datos con técnicas de Inteligencia Artificial (IA) puede conducir al desarrollo de nuevos modelos predictivos. Estos modelos pueden utilizarse a su vez para predecir compuestos más específicos y seguros con la consiguiente reducción de costes en tiempo y recursos en el desarrollo de APIs. Sin embargo, el análisis de AI de esto presenta muchos de los desafíos de los problemas de Big Data. Significa, en pocas palabras, problemas de análisis de datos con cuestiones relacionadas con Volumen, Velocidad, Veracidad, Variabilidad, Valor y Complejidad (5V + C). La primera y la segunda V se explican más o menos por sí mismas y los problemas de Variabilidad, Veracidad, Valor y Complejidad se refieren a datos con problemas de falta de datos, tendencias no consistentes, errores, informes contradictorios, interrelaciones como co-linealidad/co-linealidad. etiquetas dependientes que forman redes complejas, extrapolación de información (múltiples especies, múltiples salidas, múltiples escalas), perturbaciones en múltiples variables de entrada/salida, problemas de etiquetado múltiple, etc. En este contexto, nuestro grupo presentó InformationFusion, Algoritmo de Teoría de Perturbación y Aprendizaje Automático (IFPTML) para facilitar el desarrollo de modelos de salida múltiple de lectura transversal capaces de predecir múltiples resultados de compuestos químicos/fármacos en ensayos preclínicos. Nuestro grupo también ha informado sobre un software llamado SOFT.PTML que es una plataforma de propósito general para el modelado

IFPTML. Sin embargo, aún quedan muchos aspectos por cubrir. Muchos problemas, como el descubrimiento de compuestos anticancerígenos y los estudios de ensayos de compuestos alostéricos, no se han analizado con algoritmos IFPTML. Además, aún no se ha informado sobre software fácil de usar específico para estos problemas. Por último, a pesar de la potencial aplicación en la industria, no se ha desarrollado ninguna startup (al inicio de esta tesis) para la transferencia de la tecnología IFPTML a la industria. En consecuencia, el objetivo de esta tesis se centra en primer lugar en el desarrollo de nuevos modelos IFPTML de ensayos de compuestos anticancerígenos y compuestos alostéricos. A continuación, informamos sobre el desarrollo del software LAGA, fácil de usar, para la predicción extrapolada de compuestos anticancerígenos. Por último, se describe la planificación, creación, estructura, servicios, etc. de IKERDATA S.L una nueva startup interuniversitaria enfocada a la transferencia de tecnología IFPTML a empresas gallegas y del País Vasco en primera instancia con perspectivas de proyección hacia España, Europa y a escala global en última instancia. En el Anexo de la tesis se incluye un resumen ampliado de >3000 palabras en idioma castellano (Anexo 3.1).

CONTENS INDEX

CONTENTS

I. Introduction

Chapter 01.

Theoretical Basis

I. PART 01. INTRODUCTION

1. CHAPTER01. THEORETICAL BASIS

Paper 1. Ortega-Tenezaca, B., Quevedo-Tumailli, V., Bediaga, H., Collados, J., Arrasate, S., Madariaga, G., Munteanu, C., Cordeiro, M., & González-Díaz, H. (2020).IFPTML Multi-LabelAlgorithms: Models, Software, and Applications. *Current Topics in Medicinal Chemistry*, *20*(25), 2326–2337. doi: https://doi.org/10.2174/1568026620666200916122616

1.1. CHAPTER 01. ABSTRACT

By combining Machine Learning (ML) methods with Perturbation Theory (PT), it is possible to develop predictive models for a variety of response targets. Such combination often known as Perturbation Theory Machine Learning (PTML) modeling comprises a set of techniques that can handle various physical, and chemical properties of different organisms, complex biological or material systems under multiple input conditions. In so doing, these techniques effectively integrate a manifold of diverse chemical and biological data into a *single* computational framework that can then be applied for screening lead chemicals as well as to find clues for improving the targeted response(s). IFPTML models have thus been extremely helpful in drug or material design efforts and found to be predictive and applicable across a broad space of systems. After a brief outline of the applied methodology, this work reviews the different uses of IFPTML in Medicinal Chemistry, as well as in other applications. Finally, we cover the development of software available nowadays for setting up IFPTML models from large datasets.

Keywords

Drug Discovery, Cheminformatics, Multi-target models, Large data sets, PTML, Perturbation Theory, Machine Learning.

1.2. CHAPTER 01. INTRODUCTION

Quantitative Structure-Activity Relationships (QSAR) modelling is a widely employed computational approach that aims at predicting endpoint response(s) (*e.g.*, activity, property, or toxicity) of chemicals based on their encoding features (*descriptors*), and it is playing an increasingly key role in drug or material design.

Any response value of a chemical compound may vary considerably when determined using different experimental protocols or when applying the same experimental protocol but in different conditions, such as laboratory, environmental, time, and even if different biological measures are employed like IC_{50}, EC_{50},K_i, etc. [1, 2]. Determining the response value of new chemical compounds is a particularly important task in Medicinal Chemistry but simultaneously, highly demanding both in terms of time and resources. Currently, studies are conducted on Cheminformatics models to predict physicochemical properties of small organic molecules, proteins, proteomes, and complex systems. It is useful for reducing time, and research resources in laboratories. Different authors have applied the combination of PT, and ML to obtain IFPTML models on biological systems [3].

Towards such end one should highlight the ChEMBL database, which is nowadays a well-recognized resource in the field of drug discovery and medicinal chemistry research. In fact, this database curates, and stores standardized bioactivity, molecules, targets, and drug data retrieved from multiple sources, as well as from the primary medicinal chemistry literature[4]. It includes in addition multiple conditions of assays, such as different experimental parameters, biological assays, target proteins, cell lines, assay organisms, etc. Other databases that exist and comprise several information are the National Center for Biotechnology Information (NCBI) and the Universal Protein Resource (UnitProt), both allowing merging their information with the one coming from ChEMBL into a dataset for an object of study. UniProt, for instance, is a comprehensive resource for protein sequence and, annotation data that act on drugs [5]. On the other hand, NCBI provides a large suite of online resources for biological information, and data,

including the GenBank nucleic acid sequence database, and the PubMed database of citations, and abstracts for published life science journals [6].

The present review discusses several recent studies that have applied tools as Perturbation Theory (PT) modeling, Machine Learning (ML) techniques, and the Information Fusion (IF) technique. Notice that these tools may be used in an independent or in a combined way to solve a particular combinatorial-like problem. Typically, one resorts to the following combinations: IFPTML (PT + ML) or PTMLIF (PT + ML + IF). This allows making a rational study of complex data to extract useful relationships and predict new chemicals. PT modeling, for example, allows one to predict the endpoint response(s) of a query chemical compound or material system under multiple experimental and/or theoretical conditions based on the endpoint response(s) of a known reference system [7]. To do so, PT is combined with the Box−Jenkins moving average approach, merging unique features of the systems, and simplifying the difficulties of managing whole the information. As to ML tools, these have been used in drug or materials research since at least the ´90s, providing fast, and accurate solutions to a manifold of problems. As to the combination of the latter, that is, IFPTML predictive modeling tools, have been widely applied in medicinal chemistry, proteomics, nanotechnology, etc. for coping with large heterogeneous data sets with numerous features [8-13].

Recently, three software solutions have been launched to automate the process of obtaining models using PT, ML, and IF, namely: QSAR-Co [14], LAGA, and FRAMA [15]. QSAR-Co is a software useful to tackle some of the critical issues that are usually neglected during the development of robust classification-based multi-target models. LAGA is a software developed for drug design by resorting to both perturbation theory and machine learning techniques. FRAMA has been developed to allow calculating descriptors, and setup multi-label conditions to solve several design problems. The latter include, for example, simplifying the analysis of any dataset with a vast number of characteristics, the screening of the activity of new chemical compounds, *etc*. Overall, the scope of this review article is to explore the advances in the last years about the techniques used to setup IFPTML predictive models for medicinal chemistry or

other applications, paying special attention to the availability of user-friendly software for fastening their use.

1.3. IFPTML general methodology.

Mainly the general methodology is developed in two stages. The first stage comprises *data preprocessing*. After retrieving the information of interest from a database that is preprocessed according to valuable criteria. The second stage concerns the application of modeling techniques. This is useful for seek predictive models for complex data set with multiple Big Data features. Finally, the methodology allows to develop linear IFPTML models to predict the biological activity or classify compounds as active or nonnative in terms of biological activity, etc. [16].

$$f(\mathrm{v_{ij}})_{calc} = \mathrm{a_0} + \mathrm{a_1} \cdot f(\mathrm{v_{ij}})_{expt} + \sum_{k=1,j=0}^{k_{max},j_{max}} \mathrm{a}_{kj} \cdot \Delta \mathrm{D_k}(\mathrm{c_j}) \qquad (1)$$

The output of the model $f(\mathrm{v_{ij}})_{calc}$ is a scoring function of the value $\mathrm{v_{ij}}$ of biological activity of the $\mathrm{i^{th}}$ drug in the different combinations of conditions of assay $\mathbf{c}_j$. In the table 1 shows the MA types of operators, detailing the conditions that were used in each one, as well as the symbols, operator formula, and operator information.

1.4. IFPTML models in medicinal chemistry.

In this work we reviewed several models with different applications in Medicinal Chemistry. In the **Table 1** we summarized the main results of for these models. All the results of the statistics parameters such as Sensitivity, Specificity, and Accuracy of the IFPTML models reviewed in this review paper are greater than 70%. The highest average value is 89.0% for training, and validation in the authors' IFPTML model *Nocedo et al.;* with a maximum $\mathrm{n_j}$ equal to 56377. The lowest average value is 75.1% in the authors' IFPTML model *Ferreira et al.;* with a minimum $\mathrm{n_j}$ equal to 18258 that by coincidence are of the same authors in both cases. In next sections we are going to discuss some of these models more in detail.

Table 1. Results of the model, and input variables analyzed.

Authors	Biological Activity	Set[a]	Obs. Sets[a]	Stat. Param.[a]	Pred. Stat.[a]	Pred. Sets[a] $f(v_{ij})_{pred}$		
						n_j	0	1
Nocedo*et al.* [17]	Antibacterial activity	t	$f(v_{ij})_{obs}= 0$	Sp	90.3	30181	27248	2933
			$f(v_{ij})_{obs} = 1$	Sn	88.1	96667	11464	85203
			Total	Ac	88.7	126848		
		v	$f(v_{ij})_{obs}= 0$	Sp	90.3	10030	9062	968
			$f(v_{ij})_{obs} = 1$	Sn	88.1	32252	3842	28410
			Total	Ac	88.6	42282		
Vásquez *et al.* [18]	Antiretroviral activity	t	$f(v_{ij})_{pred}=1$	Sn	73.05	83748	18825	22573
			$f(v_{ij})_{pred}=0$	Sp	86.61	21735	2910	61175
			Total	Ac	75.84	105483		
		v	$f(v_{ij})_{pred}=1$	Sp	73.10	27959	20439	7520
			$f(v_{ij})_{pred}=0$	Sn	87.17	6370	92	6278
			Total	Ac	75.98	34329		
Ferreira *et al.* [7]	Dopamine Pathway Targets	t	$f(v_{ij})_{obs}=0$	Sp	70.1	37080	26005	11075
			$f(v_{ij})_{obs}=1$	Sn	83.9	4001	644	3357
			total	Ac	71.5	41081		
		v	$f(v_{ij})_{obs}=0$	Sp	70.2	12364	8675	3689
			$f(v_{ij})_{obs}=1$	Sn	83.3	1329	222	1107

			total	Ac	71.4	13693		

[a] Set = Training or Validation. SeriesObs. sets = observed sets, stat. param. = statistical parameter, pred. stat. = predicted statistics, Pred. sets = Predicted sets. [b]Sn = Sensitivity (%), Sp = Specificity (%), and Ac = Accuracy (%).

1.5. IFPTML modeling of anti-bacterial activity.

Nocedo-Mena *et al.*reported for the first time the isolation, and characterization of terpenes of the *Cissus incisa* plant. The results of the ChEMBL database are obtained, containing 160,000 results of preclinical antimicrobial activity tests for 55 931 compounds with more than 365 parameters of biological activity, 90 bacteria strains, 25 bacterial species, and the Leong, and Barabási data set includes 40 MRNs of microorganisms. The researchers combined IFPTML with the IF technique to develop the first PTMLIF model. The best linear model found presented values of specificity (%) = 90.31 / 90.40, and sensitivity (%) = 88.14 / 88.07 in training / validation series as shown in table 3. Table 4 shows a comparison between the obtained PT-LDA, and part of the literature, such as the ANN, and BLR model. Finally, the antibacterial activity of terpenes was determined experimentally. The most active compounds were phytol, and α-amirin, with MIC = 100 μg / ml Vancomycin-resistant *Enterococcus faecium,* and *Acinetobacter baumannii* resistant to carbapenems. The model was useful for predicting the activity of these compounds against other microorganisms with different MRNs to find other potential targets [17]. The outcome of PTMLIF model is the next:

$$f(v_{ijs})_{calc} = -5.683 + 14.434 \cdot f(\mathrm{v_{ij}})_{\mathrm{ref}} - 16.426 \cdot Sh_1(\mathrm{Drug}) + 24.818 \cdot \mathrm{D}Sh_1(\mathrm{Assay})_{c1} + 0.211 \cdot \mathrm{D}Sh_2(\mathrm{Assay})_{c1} + 1.882 \cdot \mathrm{D}Sh_1(\mathrm{Assay})_{c0} - 107.050 \cdot Sh_1(\mathrm{MRN})_{c2} + 155.395 \cdot Sh_2(\mathrm{MRN})_{c2} \quad (3)$$

$$n = 126848 x^2 = 122496.8 \; p < 0.05$$

Where, $f(v_{ij})_{calc}$ is the function value that can predict for biological activity of the i_{th} compound assayed in the jth preclinical assay with conditions $c_j = (c_0, c_1)$ against the sth bacteria specie

with MRN_s. The model has four types of input variables. The first type is the expected-value function $f(v_{ij})_{expt}$. The second type are $Sh_k(Drug_i)$ values were used to quantifying the structure of the chemical compounds. And by last, two types of PT operator are the term $\Delta Sh_k(Assay_j)c_j$, and the other type is the term $\Delta Sh_k(MRN_s)c_j$[17]. In Table 3 we show a comparative suty of this model with many other models from the literature including IFPTML and not IFPTML models.

Table 3. Comparison of IFPTML models for anti-bacterial activity with other models

Cmpd. Type[a]	*N*[b]	Var.[b]	Tech.[c]	Acc (%)	Val.[d]	Multi Species[e]	MT[f]	Net.[g]	Ref.
HSC	83,605	6	LDA	88.6	i	MBS	Yes	Yes	
Peptide	3,592	4	LDA	96.0	i	MBS	Yes	No	11
Peptide	2,488	6	LDA	90.0	i	*Gram +bacteria*	Yes	No	20
HSC	30,181	6	LDA	90.0	i	*P. intermedia*	Yes	No	18
HSC	54,000	6	ANN	90.0	i	*Pseudomonas* spp	Yes	No	19
Nano	300	7	LDA	77.7	i	MBS	Yes	No	21
HSC	37,800	5	LDA	95.0	i	No	Yes	No	22
HSC	11,576	4	ANN	97.0	i	*Streptococcus* spp	Yes	No	8
ATD	12,000	4	LDA	90.0	i	*Mycobacterium* spp	Yes	No	9
HSC	667	7	LDA	92.9	i	No	No	No	23
HSC	661	6	LDA	92.6	ii	No	No	No	24
HSC	352	7	LDA	91.0	i	No	No	No	25
HSC	111	7	LDA	94.0	i	No	No	No	26
HSC	-	8	LDA	> 90	i	No	No	No	27
HSC	972	8	LDA	86.8	i	No	No	No	28
HSC	458	2	LDA	~ 85	i	No	No	No	29
HSC	433	6	LDA	~ 85	i	No	No	No	30

[a]Compound type: HSC = Heterogeneous Series of compounds, anti-TB drug = anti-Tuberculosis drugs. [b]Total

number of cases in training and/or validation series, and Var. = Number of variables included in the model. [c]Employed techniques: LDA = Linear Discriminant Analysis, ANN = Artificial Neural Networks, BLR = Binary Logistic Regression. [d]Validation methods: i) external prediction series, ii) leave 30%-out cross-validation, and iii) 100-times-averaged re-substitution technique. Furthermore, notice that methods ii), and iii) are cross-validation methods. [e]Multi Species: Multiple bacterial strain (MBS), *Fusobacteriumnecrophorum, Prevotellaintermedia.* [f]MT = Multi-target: Models that can predict more than one type of biological activity (MIC, IC_{50}, MBC, *etc.*). [g] Net. = MRN_s: Models able to account for changes in the MRN_s of different microorganisms.

1.6. IFPTML modeling of anti-retroviral activity.

Vásquez-Domínguez *et al*. proposed the development of a new predictive model that defines target proteins of new antiretroviral compounds. ChEMBL records more than 140,000 ARV experimental preclinical assays (HIV, HTLV, SIV, HBV, MLV, RSV, FeLV) for 56,105 compounds, covering combinations with 359 biological activity parameters, 55 protein accessions, 83 cell lines, 64 organisms of assay, and 773 subtypes or strains. Also, it has included 5,277 assays for hepatitis B virus. The developed IFPTML model reached considerable values in sensitivity (%) 73.05 / 73.10, specificity (%) 86.61 / 87.17, and accuracy (%) 75.84 / 75.98 in training / validation series as shown in table 3. They compared alternative IFPTML models with different PT operators such as covariance, exponential moments, and terms [18]. The developed model applied to ARVs calculates the probability of interaction of a molecule i with different retrovirus under a set of multiple conditions of assay $\mathbf{c_j}$.HBV has been included as the presence of coinfections with HIV, and HBV in patients are common. HIV prolongs HBV viremia, increases rates of chronicity also the risk of cirrhosis, and liver-related morbidity. For this reason, the treatment of both infections should be coordinated.[19] Some studies have found effective ARV drugs, and have significant activity in the treatment of certain types of resistant HBV in HIV/HBV-coinfected patients.[20] Yang's group suggest that, in case of coinfection, they ARV therapy should include agents with activity against both HIV, and HBV.[21] Multi-condition PT operators were calculated using combinatorial or multiple moving averages (MMA). The IFPTML model equation with LDA is described below:

$$f(\mathrm{v_{ij}})_{calc} = -16.6473 + 10.6828 \cdot f(\mathrm{v_{ij}})_{ref} + 5.4195 \cdot \mathrm{D_1}$$
$$-5.0349 \cdot \Delta\mathrm{D_1}(\mathbf{c_j}) + 0.0512 \cdot \Delta\mathrm{D_2}(\mathbf{c_j}) \quad (4)$$
$$N = 140{,}644 \qquad \chi^2 = 37{,}710.77 \qquad \mathrm{p} < 0.05$$

Where, $f(v_{ij})_{calc}$ is the function value that calculates the probability of interaction of a molecule *i* with different retroviruses under a set of multiple conditions of assay c_j applied to ARV treatments. The model has three types of input variables. The first type is referenced-value function $f(v_{ij})_{ref}$. that represents the biological activity value of the molecule *m* under c_j subset of multiple conditions. The second, and the three type are ΔD_k, and $\Delta D_k(c_j)$ perturbation effects in the molecule structure are added to the equation [18].

1.7. IFPTML models of drug targeting dopamine pathways.

Ferreira da Costa et al. designed a model aimed at predicting drug-protein interactions (DPI) for the target proteins involved in dopamine pathways. The dataset has a total of 50,000 cases. The present work reports the organic synthesis, chemical characterization, and pharmacological assay of a new series of peptidomimetic compounds of *L-prolyl-L-leucyl-glycinamide* (PLG) peptidomimetic compounds [7]. The general results of the training, and validation subsets are shown. In the training series, the model presented high values of Specificity = Sp (%) = 72.8, Sensitivity = Sn (%) = 72.4, and General Accuracy = Ac (%) = 72.7 as shown in table 3. The model was stable in external validation series with values of Sp (%) = 72.7, Sn (%) = 71.4, and Ac (%) = 72.6. The best linear model found was the following:

$$f(v_{ij})_{calc} = -10.780386430140 - 0.000000000020 \cdot f(v_{ij})_{ref}$$
$$+0.440071875560 \cdot \Delta D_1(c_0) + 0.465335484664 \cdot \Delta D_1(c_1)$$
$$-0.541834505781 \cdot \Delta D_1(c_2) - 0.127705300409 \cdot \Delta D_1(c_3)$$
$$-0.114637007349 \cdot \Delta D_1(c_4) - 0.095637330548 \cdot \Delta D_2(c_5)$$
$$-0.054733584740 \cdot \Delta D_2(c_6) - 0.056732915285 \cdot \Delta D_2(c_7) \quad (5)$$
$$n = 41082 \qquad x^2 = 5564.0 \qquad p - level < 0.05$$

Where, $f(v_{ij})_{calc}$ is the function value that predicts DPI for the target proteins involved in the dopamine pathways. The model has nine types of input variables. The first type is referenced-value function $f(v_{ij})_{ref}$. The other types are $\Delta D_k(c_j)$ account for the effects on biological activity of perturbations of the drug hydrophobicity for eight different conditions [7].

1.8. IFPMTL models of chromosome gene orientation inversion networks.

Quevedo-Tumaillie*t al.* defined a new type of complex network called GOIN that encodes short, and long-range inversion patterns of the orientation of gene pairs on the chromosome about *Plasmodium falciparum (Pf)*. These networks have an average of 383 nodes (genes), and 1314 links (pairs of genes with reverse orientation). Some gene communities that encode RIFIN-related proteins were found. The IFPTML model discriminates against the RIFIN type of other proteins. The parameters of the GOINs, and Centralities were used as input values. The model presents values of sensitivity, and specificity 70-80% in the training and external validation series, respectively. In conclusion, a biological relevance of the inversion of gene orientation does not depend directly on the information of the genetic sequence [22]. The best linear model found was the following:

$$f(v_{ij})_{calc} = -68035.949 \cdot f(v_{ij})_{ref} - 4.896 \qquad (6)$$

$$n = 4025 \qquad x^2 = 293.77 \qquad p < 0.05$$

Where, $f(v_{ij})_{calc}$ is the function value that predicts RIFIN in the 5365 proteins in the proteome of *Pf*. The input is a variable of centrality called closeness C_{clo}. It centrality measures the deviation of $gene_i$ in $chromosome_k$ with respect to the expected average value of closeness for all genes in the same $chromosome_k$[22]. This is $f(v_{ij})_{ref}$ is equal to $C_{clo}(Gene_i,Chr_k)$ - $<C_{clo}(Chr_k)>$.

1.9. IFPTML proteome mining of b-cell epitopes.

Martinez-Arzate *et al.* has developed a IFPTML model to discover new B cell epitopes useful for vaccine design, and to predict immunogenic epitope scores in different experimental conditions. The model uses as input the sequence of the peptide q, and the activity of the epitope. The information retrieved contains structural changes in 83683 peptide sequences (Seq) determined in experimental trials reported on IEDB database, and involving 1448 epitope organisms (Org), 323 host organisms (Host), 15 types of in vivo processes (Proc), 28 experimental techniques (Tech), plus 505 adjuvant additives (Adj). The model has accuracy, sensitivity, and specificity between 71, and 80% for training, and external validation series [23]. The equation of this model is presented in the equation 7.

$$S_{qr}(c_i, c_j) = -1.96618 \cdot \varepsilon_r - 0.00368 \cdot^{q} \theta_2(Seq) - 0.01357 \cdot^{q} \theta_0(Org) \quad (7)$$
$$- 0.08383 \cdot^{q} \theta_0(Tech) + 0.00463 \cdot \Delta\theta_5(Seq) + 0.00404 \cdot \Delta\theta_0(Adj)$$
$$+ 0.0025 \cdot \Delta\theta_0(Org) + 0.00089 \cdot \Delta\theta_0(Host) - 0.00095 \cdot \Delta\theta_5(Proc)$$
$$+ 0.00438 \cdot \Delta\theta_0(Tech) + 7.95575$$

Where, Sqr(c_i,c_j) is the function value for the epitope activity of the q-peptide predicted under experimental conditions. The model has nine types of input variables. The first inputs ($^{q}\theta_k$) corresponding to Shannon´s entropy information measures about sequence, epitope organism, and techniques used to determine immunogenicity. The other inputs ($\Delta\theta_k$) corresponding to the quantifies information about the variation or perturbation on the sequence, adjuvant additives, epitope organism, organism exposed to the antigen, In vivo Process type, and techniques[23].

1.10. IFPTML modeling of enzyme subclasses.

Concu*et al.* developed a model to predict a set of enzymes that belong to the Pichia yeast stipe. It has been applied to a data set of 19 187 enzymes that represent the 59 subclasses present in the Protein Data Bank (PDB). In addition, the authors developed IFPTML models based on ANN to predict enzyme-enzyme pairs of template-query sequences with accuracy, specificity, and sensitivity greater than 90% for both the training and validation series [24]. The general form of this IFPTML model is presented in the equation 8.

$$S_{qr}(c_i, c_j) = e_0 + e_1 \cdot \varepsilon_r(c_j) + \sum_{k=0}^{k=5} {}^{k}_{2}e \cdot \Delta TI_k(q,r) + \sum_{k=0}^{k=5} {}^{k}_{3}e \cdot \Delta TI_k(q,j) \quad (8)$$
$$+ \sum_{k=0}^{k=5} {}^{k}_{4}e \cdot \Delta\Delta TI_k(i,j,r) + \sum_{k=0}^{k=5} {}^{k}_{5}e \cdot \nabla\nabla_k(i,j,q,r)$$

Where, the function represents the score value of the query proteins for the enzyme activity of class c_i compared with the enzyme activity $\varepsilon_r(c_j)$ of reference. The first input variable of the model $\varepsilon_r(c_j)$, quantifies the presence or absence of the enzyme activity of subclass c_j. The other types of variables are PT values such as $\Delta TI_k(q, r)$, $\Delta TI_k(q, j)$, $\Delta\Delta TI_k(i, j, q, r)$, and $\nabla\nabla TI_k(i, j, q, r)$ [24].

1.11. IFPTML models of zeolite materials.

The present work applies a IFPTML model for the study in zeolites and represents the effects of disappointment achieving an accuracy of R2 = 0.98 in external validation being useful for the rational design of novel materials [25]. The equation of this model is presented in the equation 9.

$$\varepsilon_k(m_i, c_j)_{new} = -0.22881 + 1.02864 \cdot \varepsilon_k('m_i, 'c_j)_{ref} + 0.02792 \cdot V_1 + 67.29053 \cdot V_{10}$$
$$+ 0.01264 \cdot \Delta\Delta V_1(c_1, c_2, c_3, c_5, c_6, c_7, c_8) + 0.05753$$
$$\cdot \Delta\Delta V_7(c_1, c_2, c_3, c_5, c_6, c_7, c_8) (9)$$
$$n_{tot} = 4975, \quad R^2train = 0.980, \quad R^2val = 0.985, \quad F(1.3730) = 228700,$$
$$p < 0.05$$

Where, the output of IFPTML model represents the effects of disappointment achieving an accuracy. Vk are values of the input variables used to calculate the values of the PT operators such as moving averages, and multi-condition PT operators.27 In a multi-condition PT operator they used the same idea of moving average: $\Delta Vk\ (cj) = Vk - \langle Vk\ (cj)\rangle$[25].

1.12. IFPTMLsoftware available.

The QSAR-Co software version 1.0.0 is a new application useful to tackle some critical issues that are usually neglected during the development of conventional classification-based cheminformatics models. It is a freely available standalone software **QSAR-Co** forcarrying out classification-based studies while considering different experimental conditions as applicable. It is noteworthy that **QSAR-Co** is a short form for "Cheminformatics with conditions", the latter being one of the key features of this software, though one can also develop simple classification-based cheminformatics models with no conditions. Another reason that motivated the development of this software was to provide a distinct platform for deriving classification-based cheminformatics models following all the guidelines recommended by the OECD [26], that is, robust cheminformatics models. The software comprises two modules: 1) the Model development module and 2) the Screen/Predict module. The '**QSAR-Co**' version 1.0.0 software is a standalone tool freely available to download at QSAR-Co webpage. It has two modules ('*model development'* and '*screening/predicting'*) are available in the software, and now we will discuss all the steps, and associated functionalities in each module. In the '*model development*' module, the software provides all the basic steps that are involved in a classification-based Cheminformatics model development, which also includes examining, and treating the input data for several experimental conditions, if applicable.

The software allows to calculate Box-Jenkins moving averages operators for molecular descriptors. The approach was discussed in detail previously. [27-30] In so doing, it calculates the moving average descriptors ($D(D_i)c_j$) for a molecular descriptor D_i of individual compounds '*i*'. The derivative term $D(D_i)c_j$ is called Box-Jenkin's operator[29, 31], and these modified descriptors capture the information about both chemical structures, and specific element of the experimental condition (c_j) under which the samples were assayed. These modified descriptors are calculated by the QSAR-Co software, and further used in subsequent Cheminformatics model development steps. Optionally, one can perform data pre-treatment to remove non-informative descriptors that may not have significant contribution in model building. It can also do dataset division into training, and test sets, so that in later steps the training set is employed for model development, and model selection, while test set is employed for model validation. There is an option to repeat

the same random division to reproduce the model development by using the same seed value in the settings. In rational approaches, two techniques are provided in the software, i.e., Kennard-Stone's algorithm, and Euclidean distance-based division method.[32] The software also remove the less-discriminating descriptors. QSAR-Co also provides '*Genetic Algorithm*'[33] as a variable selection technique for developing 'Linear Discriminant Analysis' (LDA) models. Genetic algorithm (GA) is a well-known technique that is often utilized in regression-based Cheminformatics model development[34-38] as well as for developing classification-based Cheminformatics models[39, 40]. At present, the software provides two ML techniques to develop robust classification-based Cheminformatics models Linear Discriminant Analysis[41] (LDA), and Random Forest[42] is a supervised machine learning algorithm which consists of a collection or ensemble of simple decision tree predictors. In this software, we have used Weka version 3-9-3 java library[43] to perform Random Forest. The validation metrics likeWilk's λ[44] provides a measure of the significance of achieved discrimination. A confusion matrix can be designed using the information of the actual, and predicted response class obtained from the model under evaluation, the software also give parameters like Sensitivity, Specificity, Fisher ratio, etc.[45], and performs Receiver operating characteristics[45] (ROC) curve analysis, and *Y*-randomization test.[46] In addition, the software also carry outs an analysis of the applicability domain (AD)[47] .Last, in the software, module 2 we can carry out predictive analysis of new chemical compounds.

On the other hand, the IFPTML.SOFT Software and his core application FRAMA software version 1.0.0 is a new desktop application in Windows environment. It was developed for the treatment of organized chemical information in grouping, and continuous variables. Use information stored in spreadsheets.[7] The loading time of the information up to FRAMA is directly proportional to the size of the file, and the amount of information. The join variables operation allows to create new grouping variables by joining two or more existing variables preferred by the user. It is done by batch processing on the queue of selected operations. The mentioned processing adds the new variables to the set of variables available in the work context. Once the file has been uploaded, and the join variables, the selection, and classification of grouping variables, and continuous variables is carried out. It is allowed to test if there are

anomalous data as null or empty andmake a decision as to their treatment. In grouping variables, anomalous values can be replaced with the value “MD”. In continuous variables it can be replaced with the average of the column values. It is also possible in both types of variables to eliminate cases.

After pre-treating the data, the selected variables can be subjected to batch operations. Based on the nature of the variables, operations of grouping variables such as Identity, account, probability, Shanon Entropy can be performed. Transformation operations of continuous variables such as: Identity, exponential, absolute value, numerical power, logarithm, minimum maximum probability, z-score. Basic operations for continuous variables such as: Sum, product, difference, division. Parametric operations such as: Maximum, minimum, average, sum, standard deviation, multiplications by a constant. Finally operators between grouping variables, and continuous variables such as Moving Average, Sum by grouping, Standard deviation, Min max probability, Z-Score [48]. It is possible to set the training, and validation values using a character pattern. Letter T is used to represent training, and V for validation. The default pattern is TTTV that expresses 75% of values for training, and 25% for validation values. FRAMA 1.0.0 allows to perform the linear regression analysis operation, based on the selection of independent input variables, and the output dependent variable. Processing results in a CSV file. The resulting file contains the names of the variables, the values of the constants, standard error, statistical T, p-value (T), F Statistic, p-value (F) Upper Confidence Limit, Lower Confidence Limit, Index, Intercept, T-Test, F-Test. Regression information is shown with R2 (r-squared), adjusted r2, F-Test, Z-test, Chi-Squared Test.. It is possible to generate variables of moving average multi labels, and operators that will be used in Perturbation Machine Learning Theory.

1.13. CHAPTER 01. CONCLUSIONS

In conclusion we can show the similarity of the results obtained through IFPTML in the selected models. Similar ranges are presented regarding the study variables. We have reviewed that there is a limited software development aimed at automating IFPTML that contain calculation properties such as dispersion measures applied to the descriptors. The methodology in PTML, allows to establish dispersion measures on descriptors of the physicochemical properties of various organisms. this review shows how the works developed in IFPTML allow to obtain new chemical components, predict biological activities and the development of parallel experimental studies, in multiples that can then be synthesized and develop pharmacological tests among others. Sensitivity, specificity, and accuracy values are greater than 70%. In general, the studies started from the selected information from the miscellaneous sources of data. The IFPTML model with the highest average value of 89% is the study of antibacterial activity about authors *Nocedo-Mena et al.*

1.14. CHAPTER 01. AUTHORS CONTRIBUTIONS.

Wrote the paper: O.T.B., Q.T.V., B.H. (Thesis Author), C.J., A.S., M.G., M.C., C.M., G.D.H. All authors have given approval to the final version of the paper manuscript. The authors declare no competing financial interest.

1.15. CHAPTER 01. REFERENCES.

1. Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P., Comparability of mixed IC50 data–a statistical analysis. *PloS one* **2013**, 8, e61007.
2. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P., Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. *Environmental health perspectives* **2003**, 111, 1361.
3. 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, 1203-1213.
4. Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P., ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res* **2015**, 43, W612-20.
5. Pundir, S.; Martin, M. J.; O'Donovan, C.; UniProt, C., UniProt Tools. *Curr Protoc Bioinformatics* **2016**, 53, 1 29 1-15.
6. Coordinators, N. R. J. N. a. r., Database resources of the national center for biotechnology information. **2017**, 45, D12.
7. Ferreira da Costa, J.; Silva, D.; Caamano, O.; Brea, J. M.; Loza, M. I.; Munteanu, C. R.; Pazos, A.; Garcia-Mera, X.; Gonzalez-Diaz, 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, 2572-2587.
8. Blazquez-Barbadillo, C.; Aranzamendi, E.; Coya, E.; Lete, E.; Sotomayor, N.; Gonzalez-Diaz, H., Perturbation theory model of reactivity and enantioselectivity of palladium-catalyzed Heck-Heck cascade reactions. *Rsc Advances* **2016**, 6, 38602-38610.
9. Casanola-Martin, G. M.; Le-Thi-Thu, H.; Perez-Gimenez, F.; Marrero-Ponce, Y.; Merino-Sanjuan, M.; Abad, C.; Gonzalez-Diaz, H., Multi-output Model with Box-Jenkins Operators of Quadratic Indices for Prediction of Malaria and Cancer Inhibitors Targeting

Ubiquitin-Proteasome Pathway (UPP) Proteins. *Current Protein & Peptide Science* **2016**, 17, 220-227.

10. Romero-Duran, F. J.; Alonso, N.; Yanez, M.; Caamano, O.; Garcia-Mera, X.; Gonzalez-Diaz, H., Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives. *Neuropharmacology* **2016**, 103, 270-278.
11. Kleandrova, V. V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. 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 Science & Technology* **2014**, 48, 14686-14694.
12. Luan, F.; Kleandrova, V. V.; Gonzalez-Diaz, 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* **2014**, 6, 10623-30.
13. Alonso, N.; Caamano, O.; Romero-Duran, F. J.; Luan, F.; Cordeiro, M. N. D. S.; Yanez, M.; Gonzalez-Diaz, H.; Garcia-Mera, X., Model for High-Throughput Screening of Multitarget Drugs in Chemical Neurosciences: Synthesis, Assay, and Theoretic Study of Rasagiline Carbamates. *Acs Chemical Neuroscience* **2013**, 4, 1393-1403.
14. Ambure, P.; Halder, A. K.; Gonzalez Diaz, H.; Cordeiro, M., QSAR-Co: An Open Source Software for Developing Robust Multitasking or Multitarget Classification-Based QSAR Models. *J Chem Inf Model* **2019**, 59, 2538-2544.
15. Bernabe Ortega-Tenezaca, V. Q.-T., Humbert González-Díaz, FRAMA 1.0: Framework for Moving Average Operators Calculation in Data Analysis. *In MOL2NET, International Conference Series on Multidisciplinary Sciences* **2017**, 3.
16. Bediaga, H.; Arrasate, S.; Gonzalez-Diaz, H., IFPTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Comb Sci* **2018**, 20, 621-632.

17. Nocedo-Mena, D.; Cornelio, C.; Camacho-Corona, M. D. R.; Garza-Gonzalez, E.; Waksman de Torres, N.; Arrasate, S.; Sotomayor, N.; Lete, E.; Gonzalez-Diaz, H., Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks. *J Chem Inf Model* **2019**, 59, 1109-1120.
18. Vasquez-Dominguez, E.; Armijos-Jaramillo, V. D.; Tejera, E.; Gonzalez-Diaz, H., Multioutput Perturbation-Theory Machine Learning (PTML) Model of ChEMBL Data for Antiretroviral Compounds. *Mol Pharm* **2019**.
19. Speck-Planche, A.; Cordeiro, M., Erratum to: Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol Divers* **2017**, 21, 525.
20. Speck-Planche, A.; Cordeiro, M., Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol Divers* **2017**, 21, 511-523.
21. Levy, V.; Grant, R. M., Antiretroviral Therapy for Hepatitis B Virus-HIV-Coinfected Patients: Promises and Pitfalls. *Clinical Infectious Diseases* **2006**, 43, 904-910.
22. Benhamou, Y., Antiretroviral therapy and HIV/hepatitis B virus coinfection. *Clinical infectious diseases* **2004**, 38, S98-S103.
23. Yang, R.; Gui, X.; Xiong, Y.; Gao, S.-c.; Yan, Y., Impact of hepatitis B virus infection on HIV response to antiretroviral therapy in a Chinese antiretroviral therapy center. *International Journal of Infectious Diseases* **2014**, 28, 29-34.
24. Ferreira da Costa, J.; Caamano, O.; Fernandez, F.; Garcia-Mera, X.; Sampaio-Dias, I. E.; Brea, J. M.; Cadavid, M. I., Synthesis and allosteric modulation of the dopamine receptor by peptide analogs of L-prolyl-L-leucyl-glycinamide (PLG) modified in the L-proline or L-proline and L-leucine scaffolds. *Eur J Med Chem* **2013**, 69, 146-58.
25. Quevedo-Tumailli, V. F.; Ortega-Tenezaca, B.; Gonzalez-Diaz, H., Chromosome Gene Orientation Inversion Networks (GOINs) of Plasmodium Proteome. *J Proteome Res* **2018**, 17, 1258-1268.
26. Martinez-Arzate, S. G.; Tenorio-Borroto, E.; Barbabosa Pliego, A.; Diaz-Albiter, H. M.; Vazquez-Chagoyan, J. C.; Gonzalez-Diaz, H., IFPTML Model for Proteome Mining of

B-Cell Epitopes and Theoretical-Experimental Study of Bm86 Protein Sequences from Colima, Mexico. *J Proteome Res* **2017**, 16, 4093-4103.

27. Concu, R.; MN, D. S. C.; Munteanu, C. R.; Gonzalez-Diaz, H., IFPTML Model of Enzyme Subclasses for Mining the Proteome of Biofuel Producing Microorganisms. *J Proteome Res* **2019**, 18, 2735-2746.
28. Blay, V.; Yokoi, T.; Gonzalez-Diaz, H., Perturbation Theory-Machine Learning Study of Zeolite Materials Desilication. *J Chem Inf Model* **2018**, 58, 2414-2419.
29. Organization for Economic Co-operation and Development (OECD). Guidance document on the validation of (quantitative) structure-activity relationship ((Q)SAR) models. OECD Series on Testing and Assessment. *69. OECD Document ENV/JM/MONO* **2007**, 55-65.
30. Speck-Planche, A.; Cordeiro, M. N., Simultaneous modeling of antimycobacterial activities and ADMET profiles: a chemoinformatic approach to medicinal chemistry. *Curr Top Med Chem* **2013**, 13, 1656-65.
31. Speck-Planche, A.; Cordeiro, M. N., Chemoinformatics for medicinal chemistry: in silico model to enable the discovery of potent and safer anti-cocci agents. *Future Med Chem* **2014**, 6, 2013-28.
32. Speck-Planche, A.; Cordeiro, M. N. D. S., De novo computational design of compounds virtually displaying potent antibacterial activity and desirable in vitro ADMET profiles. *Medicinal Chemistry Research* **2017**, 26, 2345-2356.
33. Speck-Planche, A.; Kleandrova, V. V.; Ruso, J. M.; Cordeiro, M. N., First Multitarget Chemo-Bioinformatic Model To Enable the Discovery of Antibacterial Peptides against Multiple Gram-Positive Pathogens. *J Chem Inf Model* **2016**, 56, 588-98.
34. Speck-Planche, A.; Cordeiro, M., Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol Divers* **2017**.
35. Kennard, R. W.; Stone, L. A., Computer aided design of experiments. *Technometrics* **1969**, 11, 137-148.

36. Venkatasubramanian, V.; Sundaram, A., Genetic algorithms: introduction and applications. *Encyclopedia of Computational Chemistry* **2002**, 2.
37. Rogers, D.; Hopfinger, A. J., Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *Journal of Chemical Information and Computer Sciences* **1994**, 34, 854-866.
38. Hemmateenejad, B.; Akhond, M.; Miri, R.; Shamsipur, M., Genetic algorithm applied to the selection of factors in principal component-artificial neural networks: application to QSAR study of calcium channel antagonist activity of 1, 4-dihydropyridines (nifedipine analogous). *Journal of Chemical Information and Computer Sciences* **2003**, 43, 1328-1334.
39. Hasegawa, K.; Miyashita, Y.; Funatsu, K., GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *Journal of Chemical Information and Computer Sciences* **1997**, 37, 306-310.
40. Ambure, P.; Roy, K., Understanding the structural requirements of cyclic sulfone hydroxyethylamines as hBACE1 inhibitors against Aβ plaques in Alzheimer's disease: a predictive QSAR approach. *RSC Advances* **2016**, 6, 28171-28186.
41. Gramatica, P.; Chirico, N.; Papa, E.; Cassani, S.; Kovarich, S., QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. *Journal of Computational Chemistry* **2013**, 34, 2121-2132.
42. Gao, H., Application of BCUT metrics and genetic algorithm in binary QSAR analysis. *Journal of Chemical Information and Computer Sciences* **2001**, 41, 402-407.
43. Sutherland, J. J.; O'brien, L. A.; Weaver, D. F., Spline-fitting with a genetic algorithm: A method for developing classification structure− activity relationships. *Journal of Chemical Information and Computer Sciences* **2003**, 43, 1906-1915.
44. Snedecor, G.; Cochran, W., Statistical Methods Oxford and IBH publishing co. *New Delhi* **1967**, 593.
45. Breiman, L., Random forests. *Machine learning* **2001**, 45, 5-32.

46. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H., The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* **2009**, 11, 10-18.
47. Wilks, S. S., Certain generalizations in the analysis of variance. *Biometrika* **1932**, 471-494.
48. Fawcett, T., An introduction to ROC analysis. *Pattern recognition letters* **2006**, 27, 861-874.
49. Fisher, R. A., *The design of experiments*. Oliver And Boyd; Edinburgh; London: 1937.
50. Roy, K.; Kar, S.; Ambure, P., On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems* **2015**, 145, 22-29.
51. Hill, T.; Lewicki, P., *STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining*. StatSoft: Tulsa, 2006 Vol. 1, p 813.

Thesis Hypothesis

1.16. THESIS HYPOTHESIS.

The IFPTML algorithms are useful for the development of new multi-ouput linear and non-linear models for the discovery of new drugs. These new IFPTML models can be implemented in user-friendly software to facilitate the application of the models in real life problems by non experts in computational techniques. In addition, IFPTML models and the software developed can be transferred to society by launching new startup companies like IKERDATA S.L. These companies may offer scientific computing services, consulting, training, and software development focused on solving practical data analys industry problems using IFPTML algorithms.

Thesis Objectives

1.17. THESIS OBJECTIVES.

In this thesis we propose to develop (train/validate) several models based on the IFPTML algorithm for the discovery of new drugs (anticancer, allosteric modulators). In addition, we intend to implement one of the drug design models developed in a beta version of user-friendly software for management by end users. Finally, we intend to introduce a new type of IFPTML algorithm for the development of predictive models using complex networks of complex biological systems as a reference system. Therefore, the specific objectives of this thesis are:

- Develop (train/validate) new linear and non-linear multi-output IFPTML models for the prediction of anti-cancer drugs without taking into account the sequence of the target protein.
- Develop (program) a beta version of the LAGA software, in which IFPTML algorithms for the prediction of anti-cancer compounds are implemented for the first time.
- Develop (train/validate) new linear and non-linear multi-output IFPTML models for the prediction of anti-cancer drugs considering the sequence of the target protein.
- Develop (train/validate) new linear and non-linear multi-output IFPTML models for the study of allosteric modulating drug trials.
- Develop (program) a beta version of the LAGA software, in which IFPTML algorithms for the prediction of anti-cancer compounds are implemented for the first time.
- Outline briefly from the very begenings the genesis, structure, and some projects of the new company IKERDATA S.L. as a case of technological transference of IFPTML modeling algorithms to society.

II. Experimental Work

Chapter 02.
Allosteric Inhibitors

PART 02. EXPERIMENTAL WORK

2. CHAPTER 02. ALLOSTERIC INHIBITORS

Paper 2. Sampaio-Dias IE, Rodríguez-Borges JE, Yáñez-Pérez V, Arrasate S, Llorente J, Brea JM, Bediaga H, Viña D, Loza MI, Caamaño O, García-Mera X, González-Díaz H. Synthesis, Pharmacological, and Biological Evaluation of 2-Furoyl-Based MIF-1 Peptidomimetics and the Development of a General-Purpose Model for Allosteric Modulators (ALLOPTML). *ACS Chem Neurosci*. 2021 Jan 6;12(1):203-215. doi: 10.1021/acschemneuro.0c00687.

2.1. CHAPTER 02. ABSTRACT.

In this study we developed anIFPTML model of the ChEMBL dataset for allosteric modulators. We used a 2D assay conditions array$\mathbf{c}_j$withelements$^a c_j$to denote all the specifications of all the preclinical assays in the data set. We used as input variables different PT operators to train the model. We used LDA to seek a PTML model(ALLOPTML) that predicts the probability of allosteric activity for>20000outcomes of preclinical assayswith specificity Sp = 89.2/89.4% and sensitivity Sn = 71.3/72.2% in training/validation series. To the best of our knowledge, ALLOPTM is the first general-purpose PTML model for the multi-output and multi-condition prediction of allosteric compounds. Using this model, we constructed and studied the topology of a complex network of allosteric modulators. We also illustrated the use of the model with a practical example. In so doing, we reported the synthesis, characterization, and pharmacological assay of eight new Melanostatin (MIF-1) derivatives. Last, we reported a predictive study of the probability of allosteric modulator activity for the new compounds using the model.

Keywords: Allosteric modulators; Artificial Neural Networks; Big data; ChEMBL; Machine Learning; Multi-target models;Perturbation Theory.

1.2. CHAPTER 02. INTRODUCTION.

Any vital biological process for a living organism, *e.g.*, signal transduction, enzyme activity, metabolism, or transport, is susceptible of being modulated in its frequency, rate or extent, relying upon recognition and binding among macromolecules. Some small molecules or ligands (agonists) are able to bind and interact with a given protein and modify that interaction. In this regard, allosteric modulators bind to proteins on a topographically distinct site (allosteric) from the "orthosteric" site of a reference agonist being capable of synergize positively (PAMs) or negatively (NAMs) that agonist activity. Naturally, allostery can both cause diseases and contribute to development of new therapies, holding promise for delivering more selective and less toxic medicines than those that target orthosteric sites[49].In addition, we can talk about allo-network drugs as allosteric compounds that do not act directly over the pharmacological target but over a second protein which in turn interacts directly or indirectly with the target protein. The existence of allo-networks drugs presupposes that the protein interacting with the drug and the final pharmacological target interact each other within a signaling, metabolic, and/or protein-protein interaction network [50-52].

So far, finding allosteric binding sites and lead compounds has been largely serendipitous, accomplished by medicinal chemists through preclinical assays performinghigh-throughput screening with a high number of combinations of experimental conditions ($\mathbf{c}_{a,j}$).In addition to experimental techniques, computational techniques can be used to predict new drugs against different targets [53]. In this sense, many researchers have developed Cheminformatics models for the discovery of new compounds. Nevertheless, almost all of these models are designed for homologous series of compounds, one target, and/or unique cell line.Other models, instead, use heterogeneous series of compounds,although are unable to incorporate information about the target, cell line, organism of assay, etc. [54-69].This drawback can be partially faced by Machine Learning techniques (ML) through which diverse molecular descriptors, encoding the chemical structure of different drugs,can be calculated[70-86]. ML techniques have been already carried out to predict allosteric pockets on proteins[87].Unfortunately, this method fail to account for large

"Big Data (BD)" sets of preclinical assays, such as the ChEMBLdatabase,difficult to study due to the high complexity of the data in addition to the huge volumeofvery heterogeneous preclinical assays [88, 89].

To solve this problem, we have associated Perturbation Theory (PT) ideas and ML modeling [82, 90]. In fact, PTML (PT + ML) models have been used in many fields of knowledge such as Medicinal Chemistry, Proteomics or Nanotechnologyfor modeling large data sets with BD features [8-13, 91-98].Probably due to the mentioned potential benefits of allosteric drugs,a renewed interest in allostery has led to the development of a number of computational approaches to understanding allostery.[99]Nevertheless, to the extent of our knowledge, a general-purpose PTML models for multiple preclinical assays of allosteric compounds have not been reported yet.In**Figure 1** we illustrate the general workflow used in this paper to obtain the PTML model.

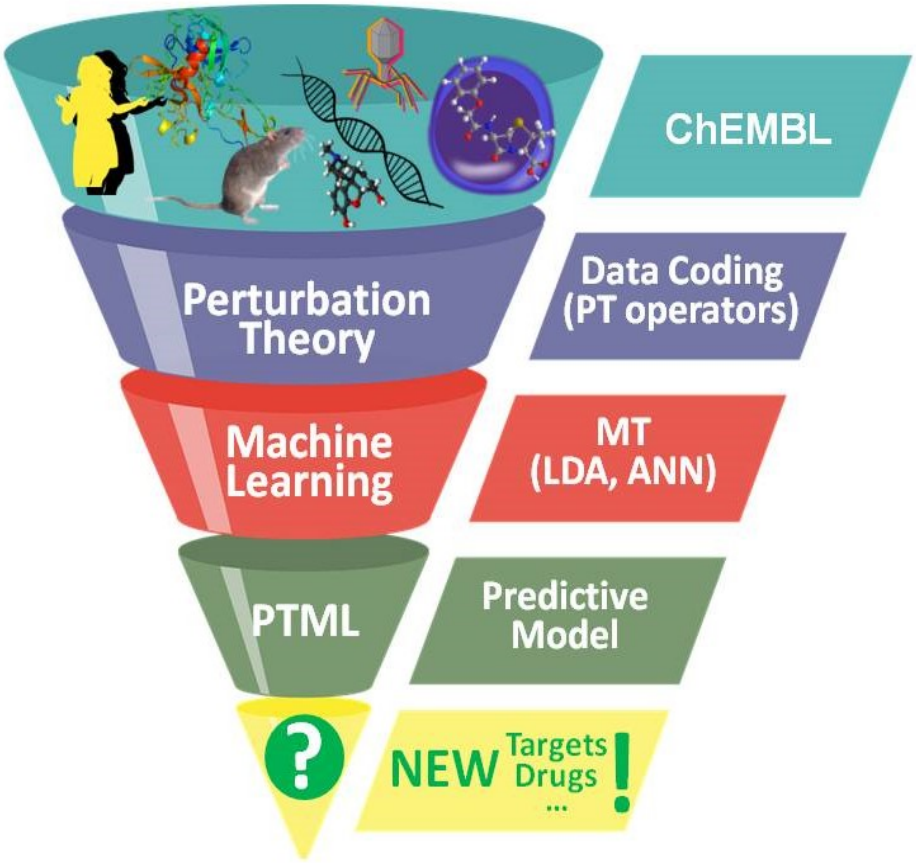


Figure 1. General workflow used in this paper to obtain the PTML model.

In this work, we report a PTML model in which we have proceeded to consider 12 different variables ($\mathbf{c}_{0\text{-}11}$) from the ChEMBL BD sets of compounds covering 38 biological activity parameters ($\mathbf{c}_0$), among them: efficacy, potency, intrinsic activity, IC_{50}, K_i, K_m, etc.); organisms of the protein target ($\mathbf{c}_1$; human, rat, mouse, etc); organism of assay ($\mathbf{c}_2$; rat, swine, herpesvirus 5, etc); different cellular lines ($\mathbf{c}_3$; CHO, HEK293, etc), and so on. Here, we develop the first general-purpose PTML model for the prediction of allosteric compounds for several types of systems and for a hundred different proteins. In fact, the data set used includes 21439 cases from 8984 differentChEMBL assays of 8957 allosteric compounds for 79 different proteins

1.3. IFPTML-LDA linear models with 1-fold moving averages.

The domain of the PTML model is the array $\mathbf{c}_j$ for all sub-sets (assays) with different standard experimental conditions.PTML model starts with the expected value of a given activity and adds the effect of different perturbations in the system. Therefore, the model have two types of input variables: the reference or expected-value function $f(v_{ij})_{ref}$ and the PT operators $\Delta D_k(\mathbf{c}_j)$.The input variable $f(v_{ij})_{ref}$ represents the expected value of a particular biological activity for one compound in a given biological activity c_0.The other PT operators are moving averages (MA) calculated for one sub-set of multiple conditions at time. The PT operators measure the deviation of D_{ki} from the expected value of $\langle D_k(\mathbf{c}_j)\rangle$ (average value) of this descriptor for different sets of drugs c_j[100-102].Hence, we can calculate the PT operators as follow $\Delta D_k(c_j) = D_{ki} - \langle D_k(c_j)\rangle$.They depend on the value of the molecular descriptors D_{ki} of type k used to quantify the structure of the i-[th]molecule. In this work, the specific molecular descriptors used are D_1 = ALOGP (n-Octanol/Water Partition Coefficient) and D_2 = TPSA (Polar Surface Area), taken directly from ChEMBL data set.

Using linear discriminant analysis (LDA)[48]linear classification models can be developed.See details about the form of a PTML linear model-equation on methods section.The equation of the best model found using this kind of PT operators is the following:

$$f(\mathrm{v_{i,j}})_{calc} = -3.62490 + 8{,}82060 \cdot f(\mathrm{v_{i,j}}/c0)_{ref} - 0.02087 \cdot \Delta \mathrm{D_2(c_1)} - 0.21686 \quad (1)$$
$$\cdot \Delta \mathrm{D_1(c_3)} + 0.02817 \cdot \Delta \mathrm{D_2(c_3)} + 0.33521 \cdot \Delta \mathrm{D_1(c_4)} + 0.01765$$
$$\cdot \Delta \mathrm{D_2(c_4)} - 0.01966 \cdot \Delta \mathrm{D_2(c_8)}$$
$$\mathrm{n = 16080} \qquad \chi^2 = 3927.97 \qquad \mathrm{p < 0.05}$$

The statistical parameters of the model are:the number of cases used to train the model, n; the Chi-square test,χ2; and the p-level, p.In**Table 1**it is shown a more detailed explanation about all the input variables analyzed to seek the PTML-LDA linear model.

Table 1. Principal PT operators used as input (c_0-c_8)

MA Type [a]	$\mathbf{c}_{j,l}$	Condition	Symbol	Operator Formula	Operator Information
MA	c_0	Activity types	$f(v_{ij})_{expt}$	$n(f(v_{ij})_{obs}=1)/n_j$	Expected value of probability $p(f(v_{ij})=1)_{expt}$ for the activity v_{ij} of type c_0
	c_0	Activity type	$\Delta D_k(c_0)$	$D_k - \langle D_k(c_j)\rangle$	$\Delta D_k(\mathbf{c}_{j,l})$ operators account for changes (Δ) on the chemical structure of the compound, quantified by the molecular descriptor D_k, with respect to the expected value of the descriptor $\langle D_k(c_{j,l})\rangle$ for molecules assayed under the same single condition of assay c_j.
	c_1	Orgs.	$\Delta D_1(c_1)$		
	c_2	Assay Orgs.	$\Delta D_1(c_2)$		
MMA	$\mathbf{c}_j$	Multiple Condition Array	$\Delta D_k(\mathbf{c}_j)$	$D_k - \langle D_k(\mathbf{c}_j)\rangle$	$\Delta D_k(\mathbf{c}_{j,l})$ account for changes in sub-set of multiple conditions of assay conditions $\mathbf{c}_j$.

[a]MA = Moving Average operators *vs*. MMA = Multiple Moving Average operators

The output of the model $f(v_{ij})_{calc}$ is a scoring function of the value v_{ij} of biological activity of the i^{-th} drug in the conditions of assay c_j. In the particular case of an LDA model $f(v_{ij})_{calc}$is not a probability and therefore it is not in the range 0-1. However, for a given value of $f(v_{ij})_{calc}$,the LDA algorithm, which uses the Mahalanobis's distance metric [48],can calculate the respective values of posterior probabilities $p(f(v_{ij})= 1)_{pred}$.

After calculating $p(f(v_{ij})= 1)_{pred}$it can be easily calculated the Boolean function $f(v_{ij})_{pred} = 1$ when $p(f(v_{ij})= 1)_{pred} > 0.5$ or $f(v_{ij})_{pred} = 0$ otherwise. The values of $f(v_{ij})_{pred} = 1$ or 0 must be compared with the respective observed values $f(v_{ij})_{obs}= 1$ or 0 to calculate the specificity(Sp, (%)), sensitivity (Sn, (%)), and overall accuracy(Ac, (%)) of the model.Hence, when $f(v_{ij})_{pred} = f(v_{ij})_{obs}$ the case is correctly classified [48]. The present one-condition moving averages model resulted in values of Sp = 79.7 %, Sn = 68.6 %, and Ac = 77.8 %in training series andin very similar values in external validation series. It is important to mention that the data points (compound-assay pair) used in validation series have not been used to train the model.

We used the forward-stepwise strategy [48] of variable selection to detect the more important perturbations on different conditions c_j related to pharmacologic property, like assay organism, cell line, among others. See details about these variables on **Table 1**. In this particular model the ML-selected operatorsmeasure perturbations on the values of ALOGP (D_1) andPSA (D_2) compared to other sub-sets c_j of drugs where c_j is: c_1, organisms; c_3, cell line; c_4, ChEMBL targets and c_8, binding domains. The parameters ALOGP and TPSA areextensively used in Medicinal Chemistry.While ALOGP is related to the lipophilicity and TPSA strongly reflects the hydrogen bonding capacity and polarity of a given drug.[103]Accordingly,these parameters mirror their ability to pass through biological membranes or interact with protein hydrophobic pockets.For example, for any drug designed to act on receptors in the central nervous system it is assumed that a TPSA less than 90 $Å^2$is required to be capable of penetrating the blood–brain barrier.[104, 105]

This model can be used to estimate the activity of a supposed new allosteric modulator in different conditions of assay. To begin with, the expected probability of activity $p(f(v_{ij})_{obs}=1)_{ref}$must be substituted on the equation, see **Table 2**.Note that these values change for different activities, like potency(nM), inhibition(%), IC_{50}(nM), activity(%),E_{max} (%), *etc*.Therefore, this modelcan predict different kinds of activity parameters (c_0) for any compounds.For that, the values of ALOGP and TPSA for the new compound (taken from ChEMBL or calculated with software)must be fitted in the model.

Table 2. One-condition parameters for selected biological parameters

Condition c_0[a]	Input parameters used to specify $\mathbf{c}_{0,l}$[b]						
Activity	$<D_1(\mathbf{c}_0)>$	$<D_2(\mathbf{c}_0)>$	$n_j(c_0)$	$n_j(f(v_{ij}) = 1)$	$f(v_{ij}) = 1)_{ref}$	cutoff	$d(c_0)$
Potency (nM)	2.93	83.57	8173	128	0.016	100.0	-1
EC_{50} (nM)	3.51	69.15	4177	649	0.155	100.0	-1
IC_{50} (nM)	3.43	71.60	2256	432	0.191	100.0	-1
Activity (%)	3.63	73.74	1500	593	0.395	48.9	1
$\%_{max}$ (%)	3.50	82.28	1123	487	0.434	70.9	1
E_{max} (%)	3.66	64.44	941	215	0.228	128.3	1
Inhibition (%)	5.03	84.22	777	424	0.546	40.8	1
Efficacy (%)	4.20	80.19	144	85	0.590	55.5	1
Allosteric Enhancer (%)	4.04	75.37	132	47	0.356	22.2	1
K_i (nM)	4.14	130.73	58	24	0.414	100.0	-1

[a]Conditionc_j= c_0 = subscript j refers to the type of activity parameter measured; subscript l are the levels of the j^{th} condition (specific pharmacological parameters).

The expected values of probability, $p(f(v_{ij})_{obs}=1)_{expt}$, must be calculated by the formula: $p(f(v_{ij})_{obs} = 1)_{ref} = n(f(v_{ij}) = 1)_{obs}/n_j$. Where, $n(f(v_{ij}) =1)_{obs}$ is the number of compounds with a desired level of activity for the condition c_j and n_j is the number of compounds assayed for the same condition c_j.The desired level of activity of a compound is $f(v_{ij})_{obs}$ and when that particular c_0 is desirable $f(v_{ij})_{obs}$ is defined to be 1, or else it is set as0. However, depending on the desirability ($d(c_0)$) of a given activity there are two ways to set $f(v_{ij})_{obs}$ = 1. For a property desirability $d(c_0)$ = 1 and when the value of activity is greater than its cutoff (v_{ij}> cutoff) the desired level of activity is $f(v_{ij})_{obs}$= 1. A compound also has a desired level of activity $f(v_{ij})_{obs}$= 1 when the value of activity is smaller than its cutoff (v_{ij}< cutoff) for not desirable properties, i.e., $d(c_0)$ = -1. Otherwise, the compound is not considered to have a desired level of activity so $f(v_{ij})_{obs}$is set as 0. It is, therefore, essential to realize that for properties of the compound that we want to maximize, such as activity (%),the property desirability $d(c_0)$ is set as 1,otherwise $d(c_0)$ is

set as -1 (e.g., EC_{50} (nM)).The cutoff for any particular activity (v_{ij}) was set as the average (cutoff = $<v_{ij}>$) of all the values of thatactivityunless an activityis described with units in nM(e.g., IC_{50}), in which case the cutoff wasset as 100nM.In order to predict a new compound,the expected values of the molecular descriptors $D_k(c_j)$ (ALOGP and PSA) for different conditions ($<D_k(c_j)>$) must be included in the model. In **Table 3**, we depict selected values of the averages $<D_1(c_j)>$. Note that these values change for different conditions, so the model gives a different result for one compound if you change this condition. For instance, $<D_1(c_j)>$ = 3.44 for *H. sapiens* and $<D_1(c_j)>$= 3.51 for *M. musculus*. Consequently, the model is able to predict a different activity in human and mouse for the same drug. The full list of the values of one-condition moving averages appears in the supplementary material file SM00.xlsx.

Table 3. One-condition averages and number of cases for selected conditions of assay

Condition c_1	Parameters used to specify c_1		
Org. of target	$<D_1(c_1)>$	$<D_2(c_1)>$	$n_j(c_1)$
H. sapiens	3.44	76.40	15564
R. norvegicus	3.55	75.90	3075
Condition c_2	Parameters used to specify c_2		
Assay organism	$<D_1(c_2)>$	$<D_2(c_2)>$	$n_j(c_2)$
H. sapiens	3.43	76.99	15046
R. norvegicus	3.35	76.68	3550
Condition c_3[a]	Parameters used to specify c_3		
Cell line	$<D_1(c_3)>$	$<D_2(c_3)>$	$n_j(c_3)$
CHO	4.23	79.13	1924
HEK293	3.54	57.83	1715
Condition c_4[b]	Parameters used to specify c_4		
Target	$<D_1(c_4)>$	$<D_2(c_4)>$	$n_j(c_4)$
$GABA_A$	2.76	83.41	2650
Caspase-1	2.92	85.12	2870

[a]Species full name:Homo sapiens, Mus musculus,Rattus norvegicus and Sus scrofa.[b]$GABA_A$, γ-aminobutyric acid type A receptor; $mGlu_5$, metabotropic glutamate receptor 5; mAChR, M_1, muscarinic acetylcholine receptor M_1.

1.4. IFPTML-LDAmodel with m-fold moving averages.

We need to point out that the PT operators used in the previous model are based on single-condition averages (the simplest case). There is another possibility of calculating PT operators based on multi-condition averages (MMA).[106] It means that the average calculation runs over all cases with the same set of conditions $\mathbf{c}_j$. Remember that, in this context, $\mathbf{c}_j$ (with **c** in boldface) refers to a vector of multiple conditions $\mathbf{c}_j = (c_0, \ldots, c_j)$.A complex case would be the use of all conditions together or all the conditions from the previous model. However, we can calculate the averages using different combinations of conditions.Forinstance,some conditions suggested by the previous approximation (c_1,c_3,andc_4) and some other conditions we estimated essential to be in the model (c_0, biological activity; and c_2, assay organism) can be forced into a new model.Therefore, considering that the simpler the model the better it is,the MMA equation ended up as follows.

$$f(v_{ij})_{calc} = -5.19375867457001000 + 11.75070469825080000 \cdot f(v_{ij}/c_0)_{ref} \quad (2)$$
$$+ 0.12834055618459200 \cdot \Delta D_1(c_0, c_1, c_2, c_3, c_4)$$
$$+ 0.00671023382639323 \cdot \Delta D_2(c_0, c_1, c_2, c_3, c_4)$$
$$n = 16080 \chi^2 = 8556.26 \qquad p < 0.05$$

In training series, the model presented high values of specificity, Sp = 89.2%; sensitivity, Sn = 71.3%; and accuracy, Ac= 86.1%. Likewise, the MMA model was also reliable in external validation series:Sp = 89.4%, Sn = 72.2%, and Ac = 86.4%. These values are in the range considered as useful for classification models with application in Medicinal Chemistry [107].In **Table 4** we compare the results obtained with this model with respect to the results obtained with the 1-fold PTML model with simple MAs (one MA for each experimental condition and descriptor). As can be noted the PTML model with m-fold MMAs (one MMA include all experimental conditions) has higher values of Sp, Sn, and Ac. Specifically, Sp increased in above

10% with respect to the MA model, Sn overpasses the line of 70%, and Ac was over 85% both in training and validation series.This is an important improvement taking into consideration that the number of variables was reduced from 7 variables in the first model to only 3 variables in the second model, reducing the possibilities of chance correlation.[108]

Table 4. Comparison of 1-fold *vs*. 4-fold moving average PTML-LDA models

PTML Model	Data Series	Obs. Sets[a]	Stat. Parm.	Stat. (%)	Predicted sets		
					n_j	$f(v_{ij})_{pred} = 0$	$f(v_{ij})_{pred} = 1$
1-fold (7 vars)	Training series ($\pi_1 = 0.50$)	$f(v_{ij})_{obs} = 0$	Sp	79.7	13272	10578	2694
		$f(v_{ij})_{obs} = 1$	Sn	68.6	2808	883	1925
		Total	Ac	77.8	16080		
	Validation series	$f(v_{ij})_{obs} = 0$	Sp	80.3	4418	3547	871
		$f(v_{ij})_{obs} = 1$	Sn	69.9	941	283	658
		Total	Ac	78.5	5359		
4-fold (3 vars)	Training series ($\pi_1 = 0.55$)	$f(v_{ij})_{obs} = 0$	Sp	89.2	13272	11845	1427
		$f(v_{ij})_{obs} = 1$	Sn	71.3	2808	807	2001
		Total	Ac	86.1	16080		
	Validation series	$f(v_{ij})_{obs} = 0$	Sp	89.4	4418	3951	467
		$f(v_{ij})_{obs} = 1$	Sn	72.2	941	262	679
		Total	Ac	86.4	5359		

Similarly, to previous equation, the input variable $f(v_{ij})_{expt}$ represents the expected value of biological activity for one compound but in different combinations of experimental conditions $c_j = (c_0, c_1, c_2, \ldots c_j \ldots c_{max})$, see analogy with the previous model. The other PT operators (called MMAs) are also MA operators but calculated for multiple conditions at the same time. MMAs operators depend also on the value of the molecular descriptors D_{ki} of type k used to quantify the structure of the i^{-th} drug. However, in this case the average value $\langle D_k(c_j) \rangle$ runs over multiple conditions at the same time. It means that we can calculate the MMAs operators as follows: $\Delta D_k({}^a\mathbf{c}_j) = D_{ki} - \langle D_k(c_j) \rangle$. However, it should be noted that in MMAs the $\mathbf{c}_j$(in boldface) denotes

an array of multiple categorical variables, see experimental conditions in **Table 5**. Conversely, in one-condition MAs c_j denotes a single condition.LDA models are Bayesian methods in the sense that they calculate the posterior probabilities $p(f(v_{ij})= 1)_{pred}$ taking into consideration the prior probability $p(f(v_{ij})= 1)_{prior}$[48].

Table 5. Topmore populated assays in the data set

c_j	c_0	c_1	c_2	c_3	Assay & Conditions Count					
Assay (a)	Actv. (Units)	Target Org	Assay Org.	Cell Line	c_0	c_1	c_2	c_3	c_4	Target [a]
1	Potency (nM)	*Hs*	*Hs*	nd	8173	15550	15033	17519	2870	C1
2	Potency (nM)	*Hs*	*Hs*	nd	8173	15550	15033	17519	2283	C7
3	Potency (nM)	nd	nd	nd	8173	2650	2686	17519	2650	nd
4	Potency (nM)	*Hs*	*Hs*	nd	8173	15550	15033	17519	773	D1R
5	IC_{50} (nM)	*Rn*	*Rn*	nd	2256	3058	3533	17519	1321	mGluR5
6	EC_{50} (nM)	*Hs*	*Hs*	HEK293	4177	15550	15033	1709	1873	mGluR5
7	Activity (%)	*Hs*	*Hs*	CHO	1500	15550	15033	1922	1592	A1R
8	EC_{50} (nM)	*Hs*	*Hs*	nd	4177	15550	15033	17519	812	mGluR1
9	EC_{50} (nM)	*Hs*	*Hs*	nd	4177	15550	15033	17519	807	mGluR4
10	IC_{50} (nM)	*Hs*	*Hs*	HEK293	2256	15550	15033	1709	1873	mGluR5

[a] mGluR1 = Metabotropic glutamate receptor 1, etc., M1 = Muscarinic acetylcholine receptor 1, etc., A1R = Adenosine A1 receptor, TRHR = Thyrotropin-releasing hormone receptor, C1 = Caspase-1, C7 = Caspase-7

This is the probability with which all compounds presented the desired classification $f(v_{ij})_{pred} = 1$.In this paper the *prior* probability was set as $p(f(v_{ij})= 1)_{prior} = 0.8$for the best model found.In fact, the value of $\chi^2$is 8556.3with a p-level < 0.05 indicating that the classifier performs a statistically significant separation of both classes ($f(v_{ij})_{obs} = 1$*vs*. $f(v_{ij})_{obs} = 0$).It can be argued that this kind of multi-condition variable is somehow less revealing since there seems to be an information lost as all conditions of the same assay are merged together in only one input

variable. However, despite this apparently simpler approach, the fact that a lot of information is considered and merged makes this a more powerful model.The activity of a new compound can also be predicted through this new model. Conversely, PT operators, sensible to changes on multiple conditions at the same time, are needed.Thus, the new PT operators are the expected probability $p_j(f(v_{ij}) = 1/\mathbf{c}_j)_{expt}$and the average values $<D_{1,2}(\mathbf{c}_j)>$ for multiple conditions at the same time, see for some examples**Table 6**.

Table 6. Selected values of 5-fold multi-condition averages

Multi-condition assays[a]				Multi-condition input parameters[b]			
Activity	Organism	Cell line	Target	Averages		Count & Probs	
c_0	c_1	c_3	c_4	$<D_1(\mathbf{c}_j)>$	$<D_2(\mathbf{c}_j)>$	$n_j(\mathbf{c}_j)$	$f(v_{ij})_{ref}$
Potency (nM)	*H.sapiens*		D_1	0.49	-51.29	8173	0.140
EC_{50}(nM)	*H.sapiens*	CHO	A_1	2.80	24.85	4177	0.667
Activity (%)	*R. norvegicus*	CHO	$mGlu_4$	1.09	-16.76	1500	0.781
EC_{50} (nM)	HHV-5	HEK293	HHV-5 chemokine	0.60	5.20	4177	0.000
Activity (%)	*H.sapiens*	HEK293	CB_1	-0.18	35.10	1500	1.000
%max (%)	*H.sapiens*	CHO-K1	M_4	-0.65	37.93	1123	0.321
EC_{50} (nM)	*Sus scrofa*		M_2	2.26	-17.48	4177	0.417
IC_{50}(nM)	*Mus musculus*		$mGlu_5$	1.26	1.27	2256	0.000
Activity (%)	*R. norvegicus*	HEK293	NK_2	1.10	-25.98	1500	0.457

[a]*H. sapiens, Homo sapiens; R. norvegicus, Rattus norvegicus;*HHV-5, *Human cytomegalovirus* (HHV-5, *Human herpesvirus 5*); D_1, Dopamine D_1 receptor; A_1, Adenosine A_1 receptor; $mGlu_4$, MGluR4;$mGlu_5$, MGluR 5; HHV-5 chemokine, HHV-5 chemokine receptor; M_4, Muscarinic acetylcholine receptor M_4; M_2, Muscarinic acetylcholine receptor M_2; CB_1, Cannabinoid CB_1 receptor; CB_2, Cannabinoid CB_2 receptor; NK_2, Tachykinin receptors

2.[b]Expected probability of biological activity p = $p(f(v_{ij})= 1/\mathbf{c}_j)_{ref} = n(f(v_{ij})= 1/\mathbf{c}_j)/ n(\mathbf{c}_j)$, where $n(f(v_{ij})= 1/\mathbf{c}_j)$ is the number of cases with $f(v_{ij})= 1$ and measured under the set of conditions $\mathbf{c}_j$. The denominator is the number of cases $n(\mathbf{c}_j)$ measured under the set of conditions $\mathbf{c}_j$ irrespective of the value of $f(v_{ij})= 1$ or 0. Please, note that multi-condition parameters depend on a vector of conditions$\mathbf{c}_j$ (denoted in **boldface**) and not on a single condition c_j.

1.5. IFPTMLandexperimental study mif-1 peptidomimetics

Synthesis of the 2-Furoyl MIF-1 Peptidomimetics.Melanostatin also known asMelanocyte-Inhibiting Factor (MIF-1) is an endogenous peptide fragment of oxytocin with different biological activity. MIF-1 acts by blocking opioid receptor activation effects, as positive allosteric modulator of the D_2/D_4 dopamine receptors and inhibiting the activity of some neuropeptides such as α-melanocyte-stimulating hormone. As a result,MIF-1 shows antidepressant, nootropic, and anti-Parkinsonian activity. Other MIF-1derivatives have been shown to act both as positive or negative allosteric modulators of D_2 receptors.[109-112] In a previous work we have demonstrated that achiral heteroaromatic scaffolds are accepted at the MIF-1 binding site using picolinic acid as proline surrogate.[113]In consonance, herein we decided to carry out the synthesis 2-furoyl-based derivatives of MIF-1, in which proline is replaced by 2-furoic acid (**2-Fu**).

For the synthesis of 2-furoyl-based peptidomimetics it was envisioned a diversity-oriented synthesis (DOS) strategy to create a set of structural-related amino acid combinations at the central and C-terminal positions. Therefore, in some peptidomimetics, L-leucine was replaced by L-valine whereas glycine was substituted by L-alanine. The synthetic routes for preparation of 2-furoyl peptidomimetics is depicted in **Scheme 1**, respectively.Starting from **2-Fu** the first step involves the peptide coupling with either L-leucine or L-valine (**Scheme 1**). For that purpose, TBTU was used as peptide coupling reagent in presence of Hünig's base as tertiary amine. These short reaction periods required for peptide coupling using TBTU are in agreement with our previous experience for peptide coupling in solution-phase using different carboxylic acids and primary amines at room temperature.[114-116]After chromatographic purification, pseudodipeptides **2(a,b)** were obtained in high yields (96%).

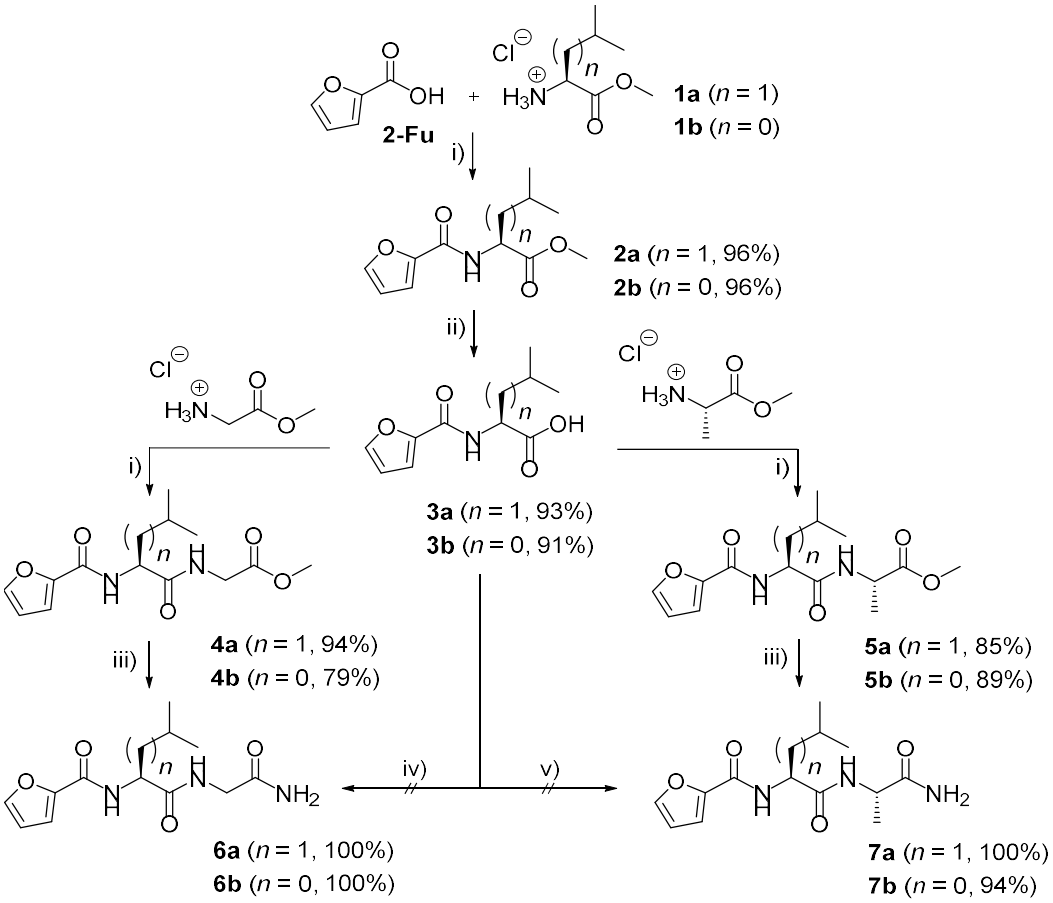
Methyl esterspresent at**2(a,b)**weresaponified using LiOH for subsequent coupling with glycine and L-alanine. After the reactions, the corresponding carboxylates wereconverted into the corresponding carboxylic acids **3(a,b)**upon protonation with H_2SO_4 1M in good to very good yields (91-93%). Special precaution was required during the acidic work-up, since the furan system is sensitive to low pH, which can result in acid-catalyzed ring opening.[117]

Therefore, the pH was carefully adjusted to 4 during the work-up. NMR spectra (^{1}H and ^{13}C-NMR) obtained for peptide products **3(a,b)**corroborate absence of ring opening phenomena(*vide* experimental section). With carboxylic acids **3(a,b)** in hands, pseudotripeptides**4(a,b)**/**5(a,b)** were prepared by condensation with either glycine (a) or L-alanine (b) methyl esters, respectively, under thesame peptide coupling protocol, affording these pseudotripeptides in good to very good yields (79-94%). The final step comprises the conversion of methyl esters **4(a,b)**/**5(a,b)** into the corresponding C-terminal carboxamides **6(a,b)**/**7(a,b)** by treatment witha stream of gaseous ammonia in methanol, prepared *in situ* by reacting $Ca(OH)_2$ and NH_4Cl at 100 ºC using an oil bath. The reactions proceeded cleanly, albeit it took about 48 h for complete conversion (TLC). After the removal of the volatiles,C-terminal carboxamides **6(a,b)**/**7(a,b)**were obtained in practically quantitative yields (94-100%), without the need of further purifications.

We also envisioned an alternative synthetic approach to afford primary carboxamides, peptidomimetics **6(a,b)**/**7(a,b)** by coupling **3(a,b)** with the commercial available C-terminal amino acids in their carboxamide form (glycinamide or L-alaninamide). Although this approach seems reasonable at first glance, these amino acids are insoluble in most of common solvents used in peptide coupling, requiring *N*,*N*-dimethylformamide (DMF) to complete dissolution. The use of DMF results in loss of peptide products during the work-up protocol (liquid-liquid extractions) since these C-terminal carboxamide peptides are highly soluble in water. Furthermore, the presence of residual DMF during the chromatographic purifications make isolation of **6(a,b)**/**7(a,b)** troublesome, making this alternative route disadvantageous in contrast with the simple and efficient conversion of methyl esters into the primary amide using methanolic ammonia. The synthetic route delivered 2-furoyl peptidomimetics in high global yields (69-84%).

Experimental Assay of 2-Furoyl MIF-1 Peptidomimetics. Eight novel MIF-1 peptidomimetics were tested for their ability to potentiate the maximum binding of the radiolabelled agonist *N*-propylnorapomorphine ([^{3}H]-NPA) to cloned human dopamine D_{2S} receptors and their activity was compared to that of MIF-1 as described in the literature.[118] These compounds were tested at eight different concentrations in the range between 1 pM and 10 μM following the protocol previously reported in our research group.[114,115]

Scheme 1. Synthesis of 2-furoyl-based peptidomimetics **4-7(a,b)**.



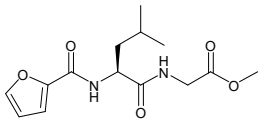
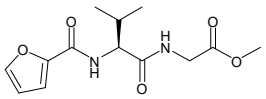
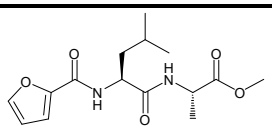
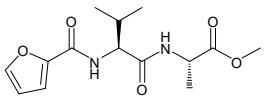
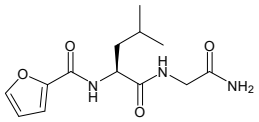
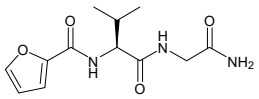
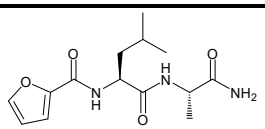
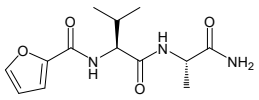
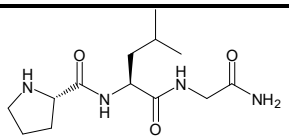
The reagents and conditions of **Scheme 1** are: (i) DIEA, TBTU, CH_2Cl_2; (ii) LiOH, MeOH/H_2O followedby H_2SO_4 1 M; (iii) NH_3 (g), MeOH; (iv) DMF, TBTU, glycinamidehydrochloride,

CH_2Cl_2; (v) DMF, TBTU, L-alaninamidehydrochloride, CH_2Cl_2.The experimental results obtained for the binding assays are shown in **Table 7** and **Figure 2**. A statistically significant enhancement ($P< 0.05$; ANOVA test; post hoc Dunnet T3 test) of the [^{3}H]-NPA response was observed for peptidomimetic **6a**. The maximum effect for **6a** was 11 ± 1% at 10 pM (Figure 2) against 22 ± 9% for MIF-1 at the same concentration. It is worth mentioning that the effect observed with MIF-1 in our work is different from that previously reported by Verma and co-workers,[118] showing increasing [^{3}H]-NPA binding at lower concentrations than those reported. This difference can be explained by different host cells where human D_{2S} receptors were expressed as it has been demonstrated that allosteric modulators are sensitive to environmental changes which may condition different active conformations elicited by the endogenous agonists on GPCRs.[119]

Several studies have shown that the allosteric activity of MIF-1 relies on the C-terminal carboxamide, the key pharmacophore to achieve active peptidometics.[120,121]Peptidomimetic **6a**, like MIF-1, displays the C-terminal dipeptide L-leucylglycinamide, and therefore its activity indicate that the substitution of pyrrolidine moiety by 2-furoyl scaffold is tolerable in the putative MIF-1 binding pocket.

IFPTML prediction of new 2-Furoyl-Based Peptidomimetics of MIF-1. After training and validating the PTML model we decided to illustrate how to use it in a real-life example. In the previous section, we reported the synthesis and biological assay of a new series of MIF-1 peptidomimetic compounds. However, due to resources and time constrains we were unable to test these compounds against many other possible assay conditions $\mathbf{c}_j$ (target proteins, cell lines, assay organisms). In so doing, we have given the following steps. Firstly, we draw the structure of the compounds into ChemDraw software to obtain their SMILE codes. Second step was to upload the file with all SMILE codes to VCCLAB, Virtual Computational Chemistry Laboratory (http://www.vcclab.org, 2005).[122]

Table 7. Experimental results for novel MIF-1 derivatives in the D_2R binding assay.

Compound	Molecular Structure	% [^{3}H]-NPA$_{max}$	AUC	Log (1/C$_{max}$)
4a		11.31	53.72	10
4b		8.75	34.35	12
5a		11.09	27.25	10
5b		6.35	9.36	12
6a		11.01	12.98	11
6b		ND	0	ND
7a		2.96	2.76	12
7b		ND	0	ND
MIF-1		22.45	44.19	10

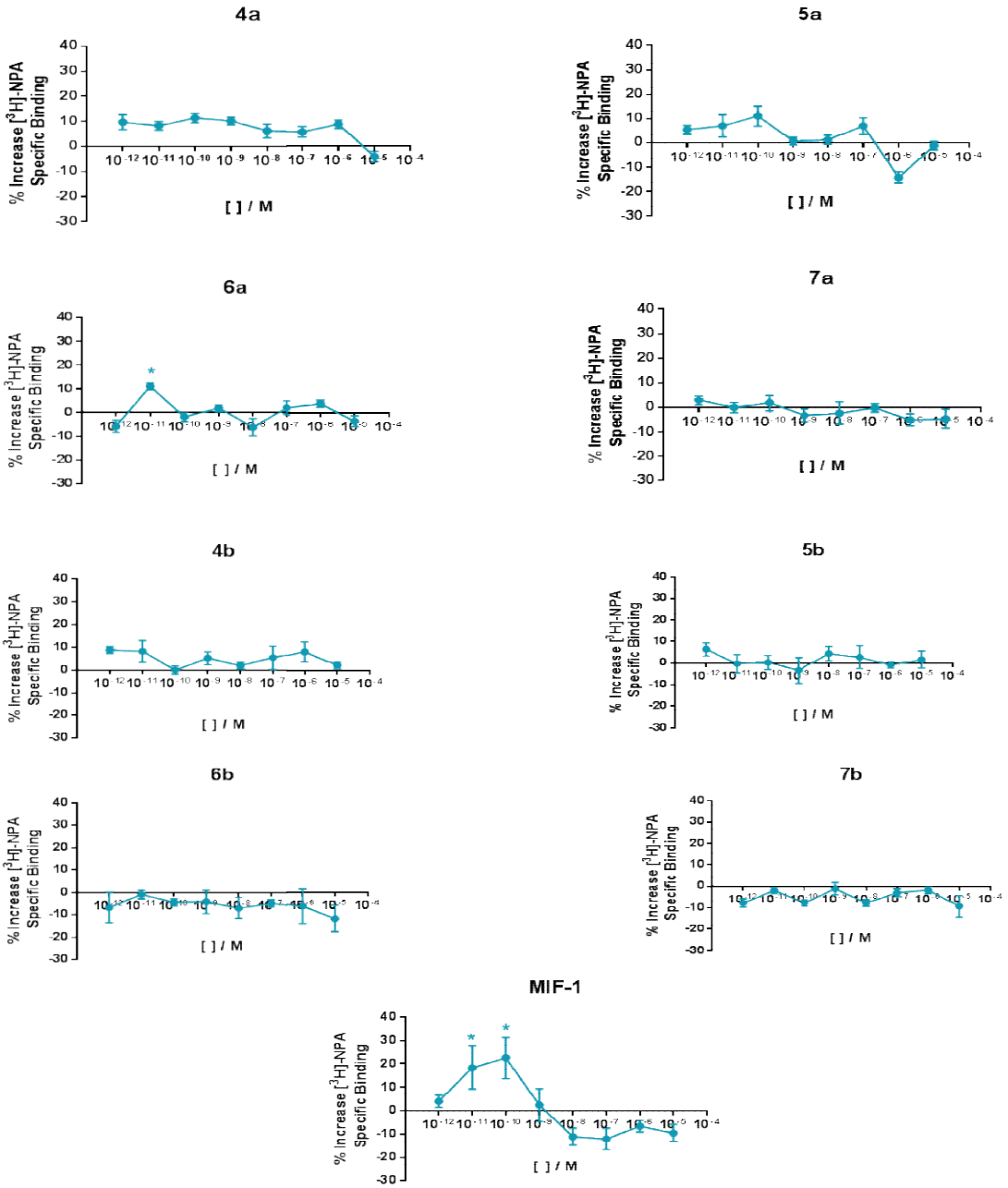


Figure 2. Modulation of [^{3}H]-NPA binding by peptidomimetics 2-furoyl-based peptidomimetics exerted by the different compounds at eight different concentrations. Points represent the mean ± standard deviation (vertical bars) of three independent experiments carried out with duplicates. *P< 0.05 (ANOVA test; *post-hoc*Dunnet T3 test).

In order to calculate the molecular descriptors D_k (LogP, PSA) of each compound. After that, we selected different conditions of assay from our dataset different from the experimental assays carried before. As the idea was to illustrate the practical use of the model, we only selected a few assays. After that we obtained from Supporting Information file (SI.xls) the values of the function of reference $f(v_{ij})_{ref}$ (expected probability of activity) for each assay.We also obtained from this file theexpected or average values $<D_k(\mathbf{c}_j)>$of the molecular descriptors D_k for all compounds already assayed on these conditions. With these values we calculated the MMAs operators $\Delta D_k(\mathbf{c}_j)$ for each compound vs. every set of experimental conditions selected. We substituted $f(v_{ij})_{ref}$ and $\Delta D_k(\mathbf{c}_j)$ values into the equation of the model to obtain the values of the output function $f(v_{ij})_{calc}$. Last, we used these values to calculate the posterior probabilities $p(f(v_{ij})_{pred} = 1)$; with which the compounds may considered as candidates for assay according to our criteria. The equation used to calculate this probabilities was a sigmoid function $p(f(v_{ij})_{pred} = 1) = 1 /(1 + (\pi_0/\pi_1)\cdot Exp(-f(v_{ij})_{calc})$. Remember that $\pi_1 = 1 - \pi_0$are the probabilities (defined before to seek the model) with which a compound may considered as active $f(v_{ij})_{pred} = 1$ or non-active and $f(v_{ij})_{pred} = 0$, *a priori*. In the **Table 8** we show the results obtained for the selected examples including different c_0= Activity parameter (units) such as Effective 50% concentration EC_{50}(nM), Maximun Effect E_{max}(%), and Inhibition(%) = Inhibition(%). The examples also involve two different organisms (Human and Rat) expressing the protein target and organisms of assays ($c_1$and c_2). We also changed the condition c_3 = cell line of assay (HEK293, CHO) and the condition c_4 = target proteins ($mGluR_5$ and$mGluR_2$). It allows us to illustrate with a practical case how we can use the model to obtain predictions of probabilities of activity $p(f(v_{ij})_{pred} = 1)$ for the same compounds in different conditions of assay with minimal computational cost.Interestingly, the model predict a homogeneous behavior for all the series of compounds in the same conditions; which is could be reasonably expected for a homogeneous series of compounds (remember this is only a short example). The model predicts low probabilities of having interesting values of EC_{50}(nM) but high probabilities of Inhibition(%). This may indicate certain probability of activity against mGluR5 in humans but perhaps at higher concentrations.

Table 8. PTML predictive study of new compounds

	c_0	EC_{50}(nM)	E_{max}(%)	$\%_{max}$	Inhibition(%)
	c_1	*H. sapiens*	*H. sapiens*	*R. norvegicus*	*H. sapiens*
Assay[a]	c_2	*H. sapiens*	*H. sapiens*	*R. norvegicus*	*H. sapiens*
	c_3	HEK293	CHO	null	HEK293
	c_4	mGluR5	mGluR2	mGluR2	mGluR5
	4a	0.042	0.087	0.404	0.796
	4b	0.039	0.082	0.388	0.785
	5a	0.043	0.090	0.413	0.802
	5b	0.041	0.085	0.397	0.791
Cmpnd	6a	0.041	0.085	0.397	0.791
	6b	0.038	0.078	0.378	0.778
	7a	0.042	0.087	0.403	0.796
	7b	0.040	0.083	0.392	0.787
	MIF-1	0.036	0.076	0.368	0.770

[a] mGluR5 = Metabotropic glutamate receptor 5 (CHEMBL2564 in Rat and CHEMBL3227 in Human); mGluR2 = Metabotropic glutamate receptor 2(CHEMBL5137 in Human and CHEMBL2581 in Rat).

1.6. CAHPTER 02. CONCLUSIONS.

IFPTML algorithm is able to seek predictive models for very large and heterogeneous data series of preclinical assays reported in ChEMBL for allosteric inhibitors. The linear model with multi-condition MAs showed to be superior to a linear model with simple condition MAs. We also combined the development of the model with the synthesis, experimental assay, and computational study of a new series of analogues of MIF-1 based on2-furoyl scaffold as proline surrogate.They have been tested as peptidomimetics of the neuropeptide MIF-1 for their ability to potentiate the binding of the dopamine receptor agonist [^{3}H]-NPA to cloned human dopamine

D_{2S} receptors. Compound **6a** show significant activity in enhancing the binding of the dopamine D_2 receptor agonist [^{3}H]-NPA. We were able to predict the results of multiple preclinical assays for the new compounds in order to illustrate the practical uses of the model.

1.7. CHAPTER 02. MATERIALS ANDMETHODS

ChEMBL Data pre-processing. Firstly, we downloaded a large and complex data set of pre-clinical assays of allosteric modulators from ChEMBL database. The data include molecular descriptors of the structure of the drugs D_k (input continuous variables) and several discrete variables for assay specifications. We used a 2D arrange $\mathbf{c}_j$of 12 different experimental conditions of the assayto summarize all the specifications of the preclinical assays (discrete variables) in the data set.These conditions include, but are not limited to, 38 biological activity parameters $\mathbf{c}_0$ with levels: $c_{0,1}$ = efficacy, $c_{0,2}$ = potency, $c_{0,3}$ = IC_{50}, *etc*. The arrange also includes different organisms $\mathbf{c}_1$, with levels $c_{1,1}$ = human, $c_{1,2}$ = rat, $c_{1,3}$ = mouse, *etc*. Another experimental condition is cellular lines used in the assy$\mathbf{c}_2$, with levels $c_{2,1}$ = CHO, $c_{2,2}$ = HEK293, *etc*., and so on. Next, we used as input variables different PT operators with form of multi-condition moving averagesdeviations$\Delta D_k(\mathbf{c}_j)$ to train the model.

As we stated above, the first condition is$\mathbf{c}_j$is$\mathbf{c}_0$the typebiological activity measured with valuesv_{ij}levels (IC_{50}, EC_{50}, *etc*.). The values of v_{ij}compiled are not exact numbers in many cases. In fact, many of them appears in ChEMBL with a >, <, and/or ≈ symbol. Eliminating this data could lead to an important reduction in the large of the dataset with the consequent lose of biologically relevant information. In addition, many of the biological activity values v_{ij} are desired to be as high as possible; e.g., Inhibition(%). However, other values of the biological activity values v_{ij} are desired to be as low as possible; e.g., IC_{50}, EC_{50},*etc*. That is why we decided to useclassification techniques instead of regression methods. Using a classification technique we can assign the original values v_{ij} into two one of two possible classes (active or interesting compounds vs. not active or control group). In so doing, we transformed the real numerical values of v_{ij}into a Boolean variable $f(v_{ij})_{obs}$. In order to construct this variable, we defined two parameters called cutoff(c_0) and desirability d(c_0). The cutoff(c_0) is the threshold

value to consider v_{ij} as high or low. The desirability gets the values $d(c_0)$ = 1 or -1 for properties that are desired to be as high or as low as possible, respectively. After that, the function$f(v_{ij})_{obs}$is calculated as $f(v_{ij})_{obs}$ = 1 when v_{ij}>cutoffand $d(c_0)$ = 1 (see **Table 1**).The function is also $f(v_{ij})_{obs}$ = 1 when v_{ij}< cutoff and $d(c_0)$ = -1;$f(v_{ij})_{obs}$ = 0 otherwise. The value $f(v_{ij})_{obs}$ = 1 points to and strong effect of the compound over the target.

PTML linear model. PTML modeling technique is useful to seek predictive models for complex datasets with multiple BD features [123, 124]. We can predict scoring function values $f(v_{ij})_{calc}$ for the i^{-th} compound in the j^{-th} preclinical assay with a specific sub-set of conditions of assay selected out of multiple conditions of assay $\mathbf{c}_j$= (c_0, c_1, c_2, …c_{jmax}). PT operators similar to Box-Jenkins MA operators are used as input[82, 90]. PTML linear models have the following form.

$$f(v_{ij})_{calc} = a_0 + a_1 \cdot f(v_{ij})_{ref} + \sum_{k=1,j=0}^{k_{max},j_{max}} a_{kj} \cdot \Delta D_k(\mathbf{c}_j) \quad (3)$$

Chemistry. All chemicals were of reagent grade and used without further purifications: 2-furoic acid was obtained from Alfa Aesar; *O*-(Benzotriazol-1-yl)-*N,N,N′,N′*-tetramethyluronium tetrafluoroborate (TBTU), H-L-Leu-OMe, H-L-Val-OMe, H-L-Ala-OMe and H-Gly-OMe were obtained from Bachem and *N*,*N*-diisopropylethylamine (DIEA) was obtained from Sigma-Aldrich. All air sensitive reactions were carried out under argon atmosphere. Analytical TLC was carried out on pre-coated silica gel plates (Merck 60 F_{254}, 0.25 mm) using UV light and an ethanolic solution of phosphomolybdic acid (followed by gentle heating) for visualization. Flash chromatography was performed on silica gel (Merck 60, 230-240 mesh).

Apparatus. Mass spectra were recorded on a LTQ OrbitrapTM XL hybrid mass spectrometer (Thermo Fischer Scientific, Bremen, Germany) controlled by LTQ Tune Plus 2.5.5 and Xcalibur 2.1.0 (Centro de Materiais da Universidade do Porto, CEMUP). The capillary voltage of the electrospray ionization source (ESI) was set to 3.1 kV. The capillary temperature was set at 275 °C. The sheath gas was set at 6 (arbitrary unit as provided by the software settings). The capillary voltage was set at 35 V and the tube lens voltage set at 110 V. ^{1}H- and ^{13}C-NMR spectra were recorded at CEMUP with a Bruker Avance III 400 at 400.15 MHz and 100.62 MHz,

respectively. The NMR spectra were calibrated using residual solvents signals ($CDCl_3$: δ_H = 7.26, δ_C = 77.16; CD_3OD: δ_H = 3.31, δ_C = 49.00 and DMSO-d_6: δ_H = 2.50, δ_C = 39.52)[125] and are reported in ppm. The nomenclature used for the assignment of protons and/or carbons for each α-amino acid residue in the peptide chains was made using a single letter system in subscript for the amino acid residue (Leu: L-leucine; Val: L-valine, Ala: L-alanine; Gly: glycine) and indicating the proton (or group of protons) and/or the carbons in the structures by starting the numeration at the carbonyl carbon of the main chain of each α-amino acid residue. Optical rotations were measured on a JASCO P-2000thermostated polarimeter using a sodium lamp and are reported as follows: α_D^θ expressed in (º) (dm^{-1}) (g^{-1}), in which θ is temperature in Celsius and c (g / 100 mL, solvent). Melting points were determined using a STUART Scientific, model SMP1, and are not corrected.Solvents were evaporated in a Buchi rotavapor model R-210.

General protocol for peptide coupling – Synthesis of 2(a,b), 4(a,b), and 5(a,b). The appropriate carboxylic acid (1 equiv.) was charged in round bottom flask and dissolved in anhydrous CH_2Cl_2 (50 mL), under argon atmosphere, followed by iterative addition of DIEA (3 equiv.) and TBTU (1.1 equiv.) and left stirring for 30 min at RT. Finally, the convenient amine (1.2 equiv.) was added, and the reaction was left stirring for about 1h. After completion of the reaction (TLC), CH_2Cl_2 was removed *in vacuo* and the crude oil was dissolved with EtOAc (100 mL) and transferred to a separatory funnel. The organic phase was washed with saturated solutions of $NaHCO_3$ (3 x 100 mL) and NaCl (100 mL) and dried over anhydrous Na_2SO_4. After filtration, the solvent was removed *in vacuo* and the crude oil was chromatographed as specified for each peptide product.

General protocol for saponification – Synthesis of pseudodipeptides 3(a,b). Pseudodipeptide methyl ester (1 equiv.) was charged in round bottom flask and dissolved in MeOH (50 mL) and the solution was cooled at 0 ºC using an ice bath with magnetic stirring. Then, LiOH (3 equiv.) were slowly added during a period of 30 min and left stirring until completion of the reaction (TLC). The solvent was removed under the solvent was removed *in vacuo* (at 40 ºC) and the crude solid was then dissolved in water (20 mL), cooled at 0 ºC using an ice bath with magnetic stirring. After that, 1 M H_2SO_4 was carefully added until pH = 4. The solvent was removed *in*

vacuo (at 40 ºC) and the crude solid was then triturated with hot $CHCl_3$ and filtered. Removal of the volatiles were performed in a rotatory evaporator to afford the desired compound without further purifications.

General protocol for C-terminal carboxamide – Synthesis of pseudotripeptides 6(a,b) and 7(a,b). Pseudotripeptide methyl ester (1 equiv.) was charged in reaction flask and dissolved in a mixture of MeOH p.a. (40 mL). The solution was cooled at 0 ºC using an ice bath with magnetic stirring and gaseous ammonia stream (generated *in situ* by mixing NH_4Cl and $Ca(OH)_2$ 2:1 (w/w) at 100 ºC using an oil bath) was bubbled into the reaction flask. The ammonia stream was maintained by one hour and the reaction was left stirring at room temperature (*ca*. 25 ºC) until completion (*ca*. 48 h). Removal of the volatiles were performed in a rotatory evaporator to afford the desired compound without further purifications.

Methyl 2-furoyl-L-leucinate (2a). Following the general protocol for peptide coupling, it was prepared a solution of **2-Fu** (1.00 g, 8.92 mmol) in anhydrous CH_2Cl_2 (30 mL) followed by addition of DIEA (4.66 mL, 26.8 mmol), TBTU (3.15 g, 9.81 mmol) and methyl L-leucinate hydrochloride (1.94 g, 10.7 mmol). After the typical work-up, the crude oil was chromatographed using EtOAc as eluent, affording 2.05 g of **2a** as a white solid. Yield: 96%. **m.p.**: 83–87 ºC. $\mathbf{R_f}$: 0.84 in EtOAc.$[\alpha]_D^{24}$: +8.3 (*c*1, $CHCl_3$). **^{1}H-NMR** ($CDCl_3$, 400 MHz) δppm: 7.56 – 7.36 (m, 1H, H-5), 7.23 – 7.00 (m, 1H, H-3), 6.77 (d, *J* = 8.0 Hz, 1H, CONH), 6.59 – 6.38 (m, 1H, H-4), 4.82 (dt, *J* = 8.6, 4.9 Hz, 1H, H_{Leu}-2), 3.76 (s, 3H, CO_2CH_3), 1.89 – 1.52 (m, 3H, H_{Leu}-3 + H_{Leu}-4), 1.11 – 0.82 (m, 6H, H_{Leu}-5). **^{13}C-NMR and DEPT** ($CDCl_3$, 101 MHz) δ ppm: 173.4 (C, CO_2CH_3), 158.1 (C, CONH), 147.6 (C, C-2), 144.2 (CH, C-5), 114.8 (CH, C-3), 112.3 (CH, C-4), 52.4 (CH_3, CO_2CH_3), 50.4 (CH, C_{Leu}-2), 41.9 (CH_2, C_{Leu}-3), 25.0 (CH, C_{Leu}-4), 22.9 (CH_3, C_{Leu}-5), 22.0 (CH_3, C_{Leu}-5). **HRMS** (ESI-TOF) *m*/*z*: $[M + H]^+$Calcd for $C_{12}H_{18}NO_4^+$: 240.12303; Found: 240.12312.

Methyl 2-furoyl-L-valinate (2b). Following the general protocol for peptide coupling, it was prepared a solution of **2-Fu** (1.5022 g, 13.403 mmol) in anhydrous CH_2Cl_2 (30 mL) followed by addition of DIEA (7.00 mL, 40.2 mmol), TBTU (4.7340 g, 14.743 mmol) and methyl L-valinatehydrochloride (2.6960 g, 16.084 mmol). After the typical work-up, the crude oil was

chromatographed using EtOAc as eluent, affording 2.8981 g of **2b** as a white solid. Yield: 96%. **m.p.**: 75–78 ºC. **R$_f$**: 0.89 in EtOAc. $[\alpha]_D^{21}$: +9.0 (*c*1.015, $CHCl_3$). **^{1}H-NMR**($CDCl_3$, 400 MHz) δppm: 7.45 (dd, *J* = 1.8, 0.8 Hz, 1H, H-5), 7.10 (dd, *J* = 3.5, 0.8 Hz, 1H, H-3), 6.80 (d, *J* = 8.3 Hz, 1H, CONH), 6.48 (dd, *J* = 3.5, 1.8 Hz, 1H, H-4), 4.70 (dd, *J* = 9.0, 5.0 Hz, 1H, H_{Val}-2), 3.74 (s, 3H, CO_2CH_3), 2.24 (dhept, *J* = 6.9, 5.1 Hz, 1H, H_{Val}-3), [0.98 (d, *J* = 6.9 Hz, 3H), 0.95 (d, *J* = 6.9 Hz, 3H), H_{Val}-4]. **^{13}C-NMR and DEPT** ($CDCl_3$, 101 MHz) δppm: 172.4 (C, CO_2CH_3), 158.2 (C, CONH), 147.7 (C, C-2), 144.2 (CH, C-5), 114.8 (CH, C-3), 112.3 (CH, C-4), 56.8 (CH_3, CO_2CH_3), 52.3 (CH, C_{Val}-2), 31.6 (CH, C_{Val}-3), 19.1 (CH, C_{Val}-4), 17.9 (CH, C_{Val}-4). **HRMS** (ESI-TOF) *m/z*: $[M + H]^+$Calcd for$C_{11}H_{16}NO_4^+$: 226.10738; Found: 226.10738.

2-Furoyl-L-leucine (3a). Following the general protocol for saponification, it was prepared a solution of **2a** (0.8034 g, 3.360 mmol) in MeOH (30 mL), followed by addition of LiOH (0.2414 g, 10.08 mmol). After the typical work-up was obtained 0.7038 g of **3a** as a white solid. Yield: 93%. **m.p.**: 81–84 ºC. **R$_f$**: 0.09 in EtOAc.$[\alpha]_D^{24}$: +8.3 (*c*1, $CHCl_3$). **^{1}H-NMR** ($CDCl_3$, 400 MHz) δppm: 8.95 (br s, 1H, CO_2H), 7.41 (d, *J* = 1.3 Hz, 1H, H-5), 7.10 (d, *J* = 3.4 Hz, 1H, H-3), 7.06 (d, *J* = 7.9 Hz, 1H, CONH), 6.42 (dd, *J* = 3.4, 1.7 Hz, 1H, H-4), 4.76 – 4.61 (m, 1H, H_{Leu}-2), 1.80 – 1.57 (m, 3H, H_{Leu}-3 + H_{Leu}-4), 1.10 – 0.80 (m, 6H, H_{Leu}-5). **^{13}C-NMR and DEPT** ($CDCl_3$, 101 MHz) δ ppm: 177.6 (C, CO_2H), 158.7 (C, CONH), 147.3 (C, C-2), 144.6 (CH, C-5), 115.3 (CH, C-3), 112.3 (CH, C-4), 51.6 (CH, C_{Leu}-2), 41.5 (CH_2, C_{Leu}-3), 25.0 (CH, C_{Leu}-4), 23.1 (CH_3, C_{Leu}-5), 21.8 (CH_3, C_{Leu}-5). **HRMS** (ESI-TOF) *m/z*: $[M - H]^-$Calcd for$C_{11}H_{14}NO_4^-$: 224.09283; Found: 224.09482.

2-Furoyl-L-valine (3b). Following the general protocol for saponification, it was prepared a solution of **2b** (0.8019, 3.560 mmol) in MeOH (30 mL), followed by addition of LiOH (0.2558 g, 10.68 mmol). After the typical work-up was obtained 0.6843 g of **3b** as a white solid. Yield: 91%.**m.p.**: 65–70 ºC. **R$_f$**: 0.05 in EtOAc. $[\alpha]_D^{20}$: +67.5 (*c*1.09, $CHCl_3$). **^{1}H-NMR**($CDCl_3$, 400 MHz) δppm: 10.12 (br s, 1H, CO_2H), 7.42 (d, *J* = 0.9 Hz, 1H, H-5), 7.10 (d, *J* = 3.3 Hz, 1H, H-3), 7.04 (d, *J* = 8.5 Hz, 1H, CONH), 6.42 (dd, *J* = 3.4, 1.7 Hz, 1H, H-4), 4.60 (dd, *J* = 8.5, 5.1 Hz, 1H, H_{Val}-2), 2.37 – 2.13 (m, 1H, H_{Val}-3), 0.95 (d, *J* = 7.1 Hz, 3H, H_{Val}-4), 0.93 (d, *J* = 7.1 Hz, 3H, H_{Val}-4). **^{13}C-NMR and DEPT** ($CDCl_3$, 101 MHz) δppm: 176.2 (C, CO_2H), 158.8 (C,

CONH), 147.3 (C, C-2), 144.6 (CH, C-5), 115.3 (CH, C-3), 112.3 (CH, C-4), 57.9 (CH, C_{Val}-2), 31.3 (CH, C_{Val}-3), 19.3 (CH, C_{Val}-4), 17.9 (CH, C_{Val}-4). **HRMS** (ESI-TOF) *m*/*z*: [M - H]$^-$Calcd for$C_{10}H_{12}NO_4^-$: 210.07718; Found: 210.07777.

Methyl 2-furoyl-L-leucylglycinate (4a). Following the general protocol for peptide coupling, it was prepared a solution of **3a** (0.8767 g, 3.944 mmol) in anhydrous CH_2Cl_2 (30 mL) followed by addition of DIEA (2.06 mL, 11.8 mmol), TBTU (1.3933 g, 4.3393 mmol) and methyl glycinate hydrochloride (0.5942 g, 4.733 mmol). After the typical work-up, the crude oil was chromatographed using EtOAc as eluent, affording 1.1394 g of **4a** as a white solid. Yield: 94%. **m.p.**: 135–136 ºC. $\mathbf{R_f}$: 0.58 in EtOAc.$[\alpha]_D^{20}$: −8.5 (*c*1, $CHCl_3$).^{1}**H-NMR** ($CDCl_3$, 400 MHz) δppm: 7.44 – 7.40 (m, 1H, H-5), 7.26 – 7.12 (m, 1H, CONH), 7.12 – 7.06 (m, 1H, H-3), 7.04 – 6.86 (m, 1H, CONH), 6.46 (dt, *J* = 4.8, 1.7 Hz, 1H, H-4), 4.82 – 4.64 (m, 1H, H_{Leu}-2), 4.13 – 3.87 (m, 2H, H_{Gly}-2), 3.76 – 3.63 [3.70 (s), 3.69 (s), 3H, CO_2CH_3], 1.87 – 1.53 (m, 3H, H_{Leu}-3 + H_{Leu}-4), 1.04 – 0.79 (m, 6H, H_{Leu}-5). ^{13}**C-NMR e DEPT** ($CDCl_3$, 101 MHz) δppm: [172.5 (C), 172.4 (C), 170.2 (C) CONH + CO_2CH_3], 158.5 (C, CONH), 147.4 (C, C-2), 144.5 (CH, C-5), 115.0 (CH, C-3), 112.3 (CH, C-4), 52.4 (CH_3, CO_2CH_3), 51.3 (CH, C_{Leu}-2), 41.3 (2 CH_2, C_{Gly}-2 + C_{Leu}-3), 24.8 (CH, C_{Leu}-4), 23.0 (CH_3, C_{Leu}-5), 22.1 (CH_3, C_{Leu}-5). **HRMS** (ESI-TOF) *m*/*z*: [M + H]$^+$Calcd for$C_{14}H_{21}N_2O_5^+$: 297.14450; Found: 297.14377.

Methyl 2-furoyl-L-valylglycinate (4b). Following the general protocol for peptide coupling, it was prepared a solution of **3b** (1.5004 g, 7.1035 mmol) in anhydrous CH_2Cl_2 (30 mL) followed by addition of DIEA (3.70 mL, 21.3 mmol), TBTU (2.5089 g, 7.8138 mmol) and methyl glycinate hydrochloride (1.0701 g, 8.5242 mmol). After the typical work-up, the crude oil was chromatographed using EtOAc as eluent, affording 1.5842 g of **4b** as a white solid. Yield: 79%. **m.p.**: 133–136 ºC.$\mathbf{R_f}$: 0.60 in EtOAc. $[\alpha]_D^{22}$: −38.4 (*c*1.015, $CHCl_3$). ^{1}**H-NMR**($CDCl_3$, 400 MHz) δppm: 7.50 – 7.40 (m, 1H, H-5), 7.32 (br s, 1H, CONH), 7.20 – 7.00 [7.11 (dd, *J* = 6.9, 4.2 Hz), 7.09 (br s), H-3 + CONH], 6.49 (dd, *J* = 3.4, 1.7 Hz, 1H, H-4), 4.63 – 4.53 (m, 1H, H_{Val}-2), 4.20 – 3.90 (m, 2H, H_{Gly}-2), 3.73 (s, 3H, CO_2CH_3), 2.28 – 2.15 (m, 1H, H_{Val}-3), 1.12 – 0.92 (m, 6H, H_{Val}-4). ^{13}**C-NMR and DEPT** ($CDCl_3$, 101 MHz) δppm: [171.6 (C), 170.2, CO_2CH_3 + CONH], 158.5 (C, CONH), 147.5 (C, C-2), 144.5 (CH, C-5), 114.8 (CH, C-3), 112.2 (CH, C-4),

58.1 (CH, C_{Val}-2), 52.4 (CH_3, CO_2CH_3), 41.2 (CH_2, C_{Gly}-2), 31.5 (CH, C_{Val}-3), 19.3 (CH, C_{Val}-4), 18.3 (CH, C_{Val}-4). **HRMS** (ESI-TOF) *m/z*: $[M + H]^+$Calcd for$C_{13}H_{19}N_2O_5^+$: 283.12885; Found: 283.12878.

Methyl 2-furoyl-L-leucyl-L-alaninate (5a). Following the general protocol for peptide coupling, it was prepared a solution of **3a** (1.6678 g, 7.5045 mmol) in anhydrous CH_2Cl_2 (30 mL) followed by addition of DIEA (3.92 mL, 22.5 mmol), TBTU (2.6506 g, 8.2550 mmol) and methyl L-alaninate hydrochloride (1.2570 g, 9.0054 mmol). After the typical work-up, the crude oil was chromatographed using EtOAc as eluent, affording 1.9797 g of **5a** as a white solid. Yield: 85%. **m.p.**: 135–136 ºC. $\mathbf{R_f}$: 0.58 in EtOAc. $[\alpha]_D^{22}$: +83.7 (*c*1.16, $CHCl_3$).**^{1}H-NMR** ($CDCl_3$, 400 MHz) δ ppm (rotamers present 75:35): [7.43 (dd, *J* = 1.7, 0.8 Hz, minor), 7.41 (dd, *J* = 1.7, 0.8 Hz, major), 1H, H-5], [7.11 (dd, *J* = 3.5, 0.7 Hz, minor), 7.09 (dd, *J* = 3.5, 0.7 Hz, major), 7.07 (br s), 2H, H-3 + CONH], [6.95 (d, *J* = 8.6 Hz, major), 6.91 (d, *J* = 8.6 Hz, minor), 1H, CONH], 6.51 – 6.41 (m, 1H, H-4), 4.78 – 4.66 (m, 1H, H_{Leu}-2), 4.60 – 4.47 (m, 1H, H_{Ala}-2), [3.72 (s, major), 3.66 (s, minor), 3H, CO_2CH_3], 1.78 – 1.59 (m, 3H, H_{Leu}-3 + H_{Leu}-4), [1.40 (d, *J* = 7.2 Hz, minor), 1.36 (d, *J* = 7.2 Hz, major), 3H, H_{Ala}-3], 0.94 – 0.90 (m, 6H, H_{Leu}-5). **^{13}C-NMR and DEPT** ($CDCl_3$, 101 MHz) δ ppm (rotamers present 67:33): [173.3 (C), 173.1 (C), 171.8 (C), 171.6 (C), CONH + CO_2CH_3], [158.4 (C), 158.3 (C), CONH], [147.6 (C), 147.5 (C), C-2], 144.4 (CH, C-5), [114.9 (CH), 114.8 (CH), C-3], [112.3 (CH), 112.2 (CH), C-4], 52.5 (CH_3, CO_2CH_3), [51.3 (CH), 51.2 (CH), C_{Leu}-2], [48.2 (CH), 48.2 (CH), C_{Ala}-2], [41.7 (CH_2), 41.5 (CH_2), C_{Leu}-3], [24.9 (CH), 24.8 (CH), C_{Leu}-4], [23.0 (CH_3), 23.0 (CH_3), C_{Ala}-3], [22.2 (CH_3), 22.2 (CH_3), 18.2 (CH_3), 18.0 (CH_3), C_{Leu}-5]. **HRMS** (ESI-TOF) *m/z*: $[M + H]^+$Calcd for$C_{15}H_{23}N_2O_5^+$: 311.16015; Found: 311.15826.

Methyl 2-furoyl-L-valyl-L-alaninate (5b). Following the general protocol for peptide coupling, it was prepared a solution of **3b** (0.5012 g, 2.373 mmol) in anhydrous CH_2Cl_2 (30 mL) followed by addition of DIEA (1.24 mL, 7.12 mmol), TBTU (0.8381 g, 2.610 mmol) and methyl L-alaninate hydrochloride (0.3975 g, 2.848 mmol). After the typical work-up, the crude oil was chromatographed using EtOAc as eluent, affording 0.6258 g of **5b** as a white solid. Yield: 89%. **m.p.**: 130–132 ºC. $\mathbf{R_f}$: 0.69 in EtOAc. $[\alpha]_D^{20}$: −8.1 (*c*1.01, DMSO). **^{1}H-NMR**($CDCl_3$, 400 MHz)

δppm (rotamerspresent 60:40): 7.44 – 7.40 (m, 1H, H-5), 7.40 – 7.26 [7.38 (d, *J* = 6.6 Hz, major), 7.29 (d, *J* = 6.9 Hz, minor), 1H, CONH], 7.19 – 6.99 [7.13 (d, *J* = 9.0 Hz), H-3 + CONH], 6.50 – 6.39 (m, 1H, H-4), 4.68 – 4.46 (m, 2H, H_{Val}-2 + H_{Ala}-2), [3.71 (s, major), 3.65 (s, minor), 3H, CO_2CH_3], 2.26 – 2.04 (m, 1H, H_{Val}-3), [1.40 (d, *J* = 7.2 Hz, minor), 1.36 (d, *J* = 7.3 Hz, major), 3H, H_{Ala}-3], 1.04 – 0.90 (m, 6H, H_{Val}-4). **^{13}C-NMR and DEPT** ($CDCl_3$, 101 MHz) δppm (rotamers present): [173.3 (C), 173.0 (C), 170.9 (C), 170.7 (C), CO_2CH_3 +CONH], [158.3 (C), 158.3 (C), CONH], [147.6 (C), 147.6 (C), C-2], [144.4 (CH), 144.4 (CH), C-5], [114.7 (CH), 114.7 (CH), C-3], [112.2 CH), 112.2 (CH), C-4], [57.8 (CH), 57.8 (CH), C_{Val}-2], [52.4 (CH_3), 52.4 (CH_3), CO_2CH_3], [48.2 (CH), 48.1 (CH), C_{Ala}-2], [31.9 (CH), 31.7 (CH), C_{Val}-3], [19.3 (CH_3), 19.1 (CH_3), 18.3 (CH_3), 18.2 (CH_3), 18.1 (CH_3), 17.8 (CH_3), C_{Val}-4 + C_{Ala}-3]. **HRMS** (ESI-TOF) *m/z*: $[M + H]^+$Calcd for$C_{14}H_{21}N_2O_5^+$: 297.14450; Found: 297.14463.

Methyl 2-furoyl-L-leucylglycinamide (6a). Following the general protocol for the synthesis of primary amide, it was prepared a solution of **4a** (0.3090 g, 1.043 mmol) in MeOH p.a. (20 mL) and the system was left stirring for 48 h at rt. After the typical work-up it was obtained 0.2921 g of **6a** as a beige solid. Yield: quantitative. **m.p.**: 90–95 ºC. **R_f**: 0.05 in EtOAc. $[\alpha]_D^{20}$: +161.2 (*c*1, $CHCl_3$).**^{1}H-NMR** ($CDCl_3$, 400 MHz) δppm: 8.01 – 7.77 (m, 1H, CONH), 7.49 – 7.32 (m, 2H, H-5 + CONH), 7.12 – 7.03 (m, 1H, H-3), 6.94 (br s, 1H, $CONH_2$), 6.48 (br s, 1H, $CONH_2$), 6.42 (dd, *J* = 3.1, 1.5 Hz, 1H, H-4), 4.74 – 4.50 (m, 1H, H_{Leu}-2), 3.98 (dd, *J* = 16.9, 6.1 Hz, 1H, H_{Gly}-2), 3.87 – 3.66 (m, 1H, H_{Gly}-2), 1.80 – 1.58 (m, 3H, H_{Leu}-3 + H_{Leu}-4), 0.99 – 0.81 (m, 6H, H_{Leu}-5). **^{13}C-NMR and DEPT** ($CDCl_3$, 101 MHz) δppm: [173.3 (C), 173.3 (C), 172.5 (C), 159.1 (C), 2 x CONH+ $CONH_2$], 147.2 (C, C-2), 144.9 (CH, C-5), 115.1 (CH, C-3), 112.3 (CH, C-4), 52.3 (CH, C_{Leu}-2), 42.9 (CH_2, C_{Leu}-3), 41.0 (CH_2, C_{Gly}-2), 24.9 (CH, C_{Leu}-4), 23.0 (CH_3, C_{Leu}-5), 21.9 (CH_3, C_{Leu}-5). **HRMS** (ESI-TOF) *m/z*: $[M + H]^+$Calcd for$C_{13}H_{20}N_3O_4^+$: 282.14483; Found: 282.14417.

Methyl 2-furoyl-L-valylglycinamide (6b). Following the general protocol for the synthesis of primary amide, it was prepared a solution of **4b** (0.3819 g, 1.353 mmol) in MeOH p.a. (20 mL) and the system was left stirring for 48 h at rt. After the typical work-up it was obtained 0.3611 g of **6b** as a white solid. Yield: quantitative. **m.p.**: 180–183 ºC. **R_f**: 0.20 in EtOAc. $[\alpha]_D^{21}$: +14.6

(*c*1.025, DMSO). **^{1}H-NMR** (DMSO-d_6, 400 MHz) δ ppm: 8.30 (t, *J* = 5.8 Hz, 1H, CONH), 8.04 (d, *J* = 8.4 Hz, 1H, CONH), 7.87 – 7.82 (m, 1H, H-5), 7.29 – 7.15 [7.23 (br s), 7.21 (d, *J* = 3.5 Hz), 2H, CONHH + H-3], 7.04 (br s, 1H, CONHH), 6.63 (dd, *J* = 3.5, 1.8 Hz, 1H, H-4), 4.25 (dd, *J* = 8.1, 7.7 Hz, 1H, H_{Val}-2), [3.66 (dd, *J* = 16.7, 6.0 Hz), 3.65 (dd, *J* = 16.7, 5.5 Hz), 2H, H_{Gly}-2], 2.18 – 2.04 (m, 1H, H_{Val}-3), [0.91 (d, *J* = 6.8 Hz), 0.90 (d, *J* = 6.7 Hz), 6H, H_{Val}-4]. **^{13}C-NMR and DEPT** (DMSO-d_6, 101 MHz) δ ppm: [171.1 (C), 170.7 (C), 157.8 (C), 2 x CONH + $CONH_2$], 147.4 (C, C-2), 145.2 (CH, C-5), 114.0 (CH, C-3), 111.9 (CH, C-4), 58.3 (CH, C_{Val}-2), 41.8 (CH_2, C_{Gly}-2), 30.1 (CH, C_{Val}-3), [19.3 (CH), 18.6 (CH), C_{Val}-4]. **HRMS** (ESI-TOF) *m*/*z*: $[M + H]^+$Calcd for$C_{12}H_{18}N_3O_4^+$: 268.12918; Found: 268.12884.

Methyl 2-furoyl-L-leucyl-L-alaninamide (7a). Following the general protocol for the synthesis of primary amide described, it was prepared a solution of **5a** (0.5552 g, 1.788 mmol) in MeOH p.a. (20 mL) and the system was left stirring for 48 h at rt. After the typical work-up it was obtained 0.5279 g of **7a** as a white solid. Yield: quantitative. **m.p.**: 85–90 ºC. **R_f**: 0.27 in EtOAc. $[\alpha]_D^{23}$: −36.6 (*c*1.015, $CHCl_3$).**^{1}H-NMR** ($CDCl_3$, 400 MHz) δ ppm (rotamers present 70:30): [7.84 (d, *J* = 7.8 Hz, minor), 7.76 (d, *J* = 7.5 Hz, major), 1H, CONH], 7.47 – 7.20 [7.44 (dd, *J* = 1.7, 0.7 Hz), 2H, H-5 + CONH], [7.13 (dd, *J* = 3.5, 0.6 Hz, major), 7.10 (dd, *J* = 3.6, 0.6 Hz, minor), 1H, H-3], [7.02 (br s, minor), 6.94 (br s, major), 1H, CONHH], 6.63 – 6.19 [6.42 (br s, minor), 6.38 (br s, major), 2H, H-4 + CONHH], 4.81 – 4.63 (m, 1H, H_{Leu}-2), [4.53 (p, *J* = 7.0 Hz), 4.52 (p, *J* = 7.1 Hz), 1H, H_{Ala}-2], 1.76 – 1.63 (m, 3H, H_{Leu}-3 + H_{Leu}-4), [1.39 (d, *J* = 7.1 Hz, minor), 1.33 (d, *J* = 7.1 Hz, major), 3H, H_{Ala}-3], 0.98 – 0.87 (m, 6H, H_{Leu}-5). **^{13}C-NMR and DEPT** ($CDCl_3$, 101 MHz) δ ppm (rotamers present): [176.3 (C), 176.2 (C), 173.4 (C), 173.2 (C), 159.8 (C), 159.5 (C), 2 x CONH + $CONH_2$], [148.2 (C), 148.0 (C), C-2], [145.5 (CH), 145.4 (CH), C-5], [115.9 (CH), 115.8 (CH), C-3], 113.1 (CH, C-4), [53.0 (CH), 52.6 (CH), C_{Leu}-2], [49.7 (CH), 49.6 (CH), C_{Ala}-2], [42.3 (CH_2), 42.1 (CH_2), C_{Leu}-3], [25.7 (CH), 25.7 (CH), C_{Leu}-4], [23.8 (CH_3), 23.7 (CH_3), 22.9 (CH_3), 22.8 (CH_3), C_{Leu}-5], [18.8 (CH_3), 18.7 (CH_3), C_{Ala}-3]. **HRMS** (ESI-TOF) *m*/*z*: $[M + H]^+$Calcd for$C_{14}H_{22}N_3O_4^+$: 296.16048; Found: 296.16033.

Methyl 2-furoyl-L-valyl-L-alaninamide (7b). Following the general protocol for the synthesis of primary amide, it was prepared a solution of **5b** (0.4012 g, 1.353 mmol) in MeOH p.a. (20

mL) and the system was left stirring for 48 h at rt. After the typical work-up it was obtained 0.3580 g of **7b** as a white solid. Yield: 94%. **m.p.**: 202–204 ºC. $\mathbf{R_f}$: 0.14 in Et_2O. $[\boldsymbol{\alpha}]_{\mathbf{D}}^{\mathbf{21}}$: +4.3 (*c*1.04, DMSO). **^{1}H-NMR**(CD_3OD, 400 MHz) δppm (rotamerspresent): 7.75 – 7.58 (m, 1H, H-5), 7.25 – 7.07 (m, 1H, H-3), 6.66 – 6.50 [6.59 (dd, *J* = 3.5, 1.8 Hz), 6.58 (dd, *J* = 3.5, 1.8 Hz), 1H, H-4], 4.46 – 4.16 (m, 2H, H_{Val}-2 + H_{Ala}-2), 2.25 – 2.10 (m, 1H, H_{Val}-3), 1.42 – 1.33 [m, 3H, H_{Ala}-3], 1.08 – 0.93 (m, 6H, H_{Val}-4). **^{13}C-NMR and DEPT,**(CD_3OD, 101 MHz) δppm (rotamers present): [177.6 (C), 177.3 (C), 173.6 (C), 173.1 (C), 160.9 (C), 160.6 (C), 2 x CONH + $CONH_2$], [148.5 (C), 148.5 (C), C-2], 146.7 (CH, C-5), [115.9 (CH), 115.9 (CH), C-3], [113.1 CH), 113.0 (CH), C-4], [60.8 (CH), 59.9 (CH), C_{Val}-2], [50.1 (CH), 50.0 (CH), C_{Ala}-2], [32.2 (CH), 31.6 (CH), C_{Val}-3], [19.8 (CH_3), 19.6 (CH_3), 19.3 (CH_3), 18.9 (CH_3), 18.3 (CH_3), 18.0 (CH_3), C_{Val}-4 + C_{Ala}-3]. **HRMS** (ESI-TOF) *m/z*: $[M + H]^+$Calcd for$C_{13}H_{20}N_3O_4^+$: 282.14483; Found: 282.14437.

Supporting Information. The full lists of the values of MA and MMA appear in the Supplementary Information filesSM01.xlsxand SM00.xlsx, respectively. The dataset used and the results of the MMA model for each case, including compound code, molecular descriptors, and conditions of assay was delivered in the fileSM03.xlsx. 1D (^{1}H, ^{13}C, DEPT-135) and 2D (COSY, HSQC) NMR spectra for all compounds reported. This material is available free of charge via the Internet at http://pubs.acs.org.

1.8. CHAPTER 02. AUTHORS CONTRIBUTIONS.

Conceived and designed the experiments: IESD, HGD. Performed the experiments: IESD, HGD. Analysed the data: B.H. (Thesis Author), IESD, XGM, JERB, HGD. Wrote the paper: B.H. (Thesis Author), IESD, XGM, HGD. All authors have given approval to the final version of the paper manuscript. The authors declare no competing financial interest.

1.9. CHAPTER 02. REFERENCES.

1. Nussinov, R.; Tsai, C. J., Allostery in disease and in drug discovery. *Cell* **2013**, 153, 293-305.
2. Csermely, P.; Nussinov, R.; Szilagyi, A., From allosteric drugs to allo-network drugs: state of the art and trends of design, synthesis and computational methods. *Curr Top Med Chem* **2013**, 13, 2-4.
3. Szilagyi, A.; Nussinov, R.; Csermely, P., Allo-network drugs: extension of the allosteric drug concept to protein- protein interaction and signaling networks. *Curr Top Med Chem* **2013**, 13, 64-77.
4. Csermely, P.; Korcsmaros, T.; Kiss, H. J.; London, G.; Nussinov, R., Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.* **2013**, 138, 333-408.
5. Taylor, P. J., *Comprehensive Medicinal Chemistry*. 1990 ed.; Pergamon Press: Oxford, 1990; Vol. 4, p 241-294.
6. Wang, J.; Yun, D.; Yao, J.; Fu, W.; Huang, F.; Chen, L.; Wei, T.; Yu, C.; Xu, H.; Zhou, X.; Huang, Y.; Wu, J.; Qiu, P.; Li, W., Design, synthesis and QSAR study of novel isatin analogues inspired Michael acceptor as potential anticancer compounds. *Eur. J. Med. Chem.* **2018**, 144, 493-503.
7. Pogorzelska, A.; Slawinski, J.; Zolnowska, B.; Szafranski, K.; Kawiak, A.; Chojnacki, J.; Ulenberg, S.; Zielinska, J.; Baczek, T., Novel 2-(2-alkylthiobenzenesulfonyl)-3-(phenylprop-2-ynylideneamino)guanidine derivatives as potent anticancer agents - Synthesis, molecular structure, QSAR studies and metabolic stability. *Eur. J. Med. Chem.* **2017**, 138, 357-370.
8. Slawinski, J.; Szafranski, K.; Pogorzelska, A.; Zolnowska, B.; Kawiak, A.; Macur, K.; Belka, M.; Baczek, T., Novel 2-benzylthio-5-(1,3,4-oxadiazol-2-yl)benzenesulfonamides with anticancer activity: Synthesis, QSAR study, and metabolic stability. *Eur. J. Med. Chem.* **2017**, 132, 236-248.

9. Murahari, M.; Kharkar, P. S.; Lonikar, N.; Mayur, Y. C., Design, synthesis, biological evaluation, molecular docking and QSAR studies of 2,4-dimethylacridones as anticancer agents. *Eur. J. Med. Chem.* **2017**, 130, 154-170.
10. Ruddarraju, R. R.; Murugulla, A. C.; Kotla, R.; Chandra Babu Tirumalasetty, M.; Wudayagiri, R.; Donthabakthuni, S.; Maroju, R.; Baburao, K.; Parasa, L. S., Design, synthesis, anticancer, antimicrobial activities and molecular docking studies of theophylline containing acetylenes and theophylline containing 1,2,3-triazoles with variant nucleoside derivatives. *Eur. J. Med. Chem.* **2016**, 123, 379-396.
11. Singh, H.; Kumar, R.; Singh, S.; Chaudhary, K.; Gautam, A.; Raghava, G. P., Prediction of anticancer molecules using hybrid model developed on molecules screened against NCI-60 cancer cell lines. *BMC Cancer* **2016**, 16, 77.
12. Pingaew, R.; Prachayasittikul, V.; Worachartcheewan, A.; Nantasenamat, C.; Prachayasittikul, S.; Ruchirawat, S.; Prachayasittikul, V., Novel 1,4-naphthoquinone-based sulfonamides: Synthesis, QSAR, anticancer and antimalarial studies. *Eur. J. Med. Chem.* **2015**, 103, 446-59.
13. Anwer, Z.; Gupta, S. P., A QSAR study on some series of anticancer tyrosine kinase inhibitors. *Medicinal chemistry (Shariqah, United Arab Emirates)* **2013**, 9, 203-12.
14. Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J., SMILES-based QSAR approaches for carcinogenicity and anticancer activity: comparison of correlation weights for identical SMILES attributes. *Anticancer Agents Med Chem* **2011**, 11, 974-82.
15. Huang, X. Y.; Shan, Z. J.; Zhai, H. L.; Li, L. N.; Zhang, X. Y., Molecular design of anticancer drug leads based on three-dimensional quantitative structure-activity relationship. *Journal of chemical information and modeling* **2011**, 51, 1999-2006.
16. Benfenati, E.; Toropov, A. A.; Toropova, A. P.; Manganaro, A.; Gonella Diaza, R., coral Software: QSAR for Anticancer Agents. *Chemical biology & drug design* **2011**, 77, 471-6.

17. Gonzalez-Diaz, H.; Bonet, I.; Teran, C.; De Clercq, E.; Bello, R.; Garcia, M. M.; Santana, L.; Uriarte, E., ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur. J. Med. Chem.* **2007**, 42, 580-5.
18. Gonzalez-Diaz, H.; Vina, D.; Santana, L.; de Clercq, E.; Uriarte, E., Stochastic entropy QSAR for the in silico discovery of anticancer compounds: prediction, synthesis, and in vitro assay of new purine carbanucleosides. *Bioorg. Med. Chem.* **2006**, 14, 1095-107.
19. Gonzales-Diaz, H.; Gia, O.; Uriarte, E.; Hernadez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A., Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design I: discovery of anticancer compounds. *J Mol Model* **2003**, 9, 395-407.
20. Jung, M.; Kim, H.; Kim, M., Chemical genomics strategy for the discovery of new anticancer agents. *Curr. Med. Chem.* **2003**, 10, 757-62.
21. Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N., Mining the NCI anticancer drug discovery databases: genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 189-99.
22. Han, L.; Cui, J.; Lin, H.; Ji, Z.; Cao, Z.; Li, Y.; Chen, Y., Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. *Proteomics* **2006**, 6, 4023-37.
23. Chou, K. C.; Shen, H. B., Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS ONE* **2010**, 5, e11335.
24. Chou, K. C.; Shen, H. B., Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nature protocols* **2008**, 3, 153-62.
25. Cai, Y. D.; Chou, K. C., Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. *Journal of proteome research* **2005**, 4, 967-71.

26. Shen, H. B.; Chou, K. C., QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. *Journal of proteome research* **2009**, 8, 1577-84.
27. Chou, K. C.; Shen, H. B., Large-scale predictions of gram-negative bacterial protein subcellular locations. *Journal of proteome research* **2006**, 5, 3420-8.
28. Rodriguez-Soca, Y.; Munteanu, C. R.; Dorado, J.; Pazos, A.; Prado-Prado, F. J.; Gonzalez-Diaz, H., Trypano-PPI: a web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions. *Journal of proteome research* **2010**, 9, 1182-90.
29. Munteanu, C. R.; Vazquez, J. M.; Dorado, J.; Sierra, A. P.; Sanchez-Gonzalez, A.; Prado-Prado, F. J.; Gonzalez-Diaz, H., Complex network spectral moments for ATCUN motif DNA cleavage: first predictive study on proteins of human pathogen parasites. *Journal of proteome research* **2009**, 8, 5219-28.
30. Gonzalez-Diaz, H.; Saiz-Urra, L.; Molina, R.; Santana, L.; Uriarte, E., A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions. *Journal of proteome research* **2007**, 6, 904-8.
31. Gonzalez-Diaz, H.; Prado-Prado, F.; Garcia-Mera, X.; Alonso, N.; Abeijon, P.; Caamano, O.; Yanez, M.; Munteanu, C. R.; Pazos, A.; Dea-Ayuela, M. A.; Gomez-Munoz, M. T.; Garijo, M. M.; Sansano, J.; Ubeira, F. M., MIND-BEST: Web server for drugs and target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretical-experimental study of G3PDH protein from Trichomonas gallinae. *Journal of proteome research* **2011**, 10, 1698-718.
32. Concu, R.; Dea-Ayuela, M. A.; Perez-Montoto, L. G.; Bolas-Fernandez, F.; Prado-Prado, F. J.; Podda, G.; Uriarte, E.; Ubeira, F. M.; Gonzalez-Diaz, H., Prediction of enzyme classes from 3D structure: a general model and examples of experimental-theoretic scoring of peptide mass fingerprints of Leishmania proteins. *Journal of proteome research* **2009**, 8, 4372-82.

33. Aguero-Chapin, G.; Varona-Santos, J.; de la Riva, G. A.; Antunes, A.; Gonzalez-Vlla, T.; Uriarte, E.; Gonzalez-Diaz, H., Alignment-free prediction of polygalacturonases with pseudofolding topological indices: experimental isolation from Coffea arabica and prediction of a new sequence. *Journal of proteome research* **2009**, 8, 2122-8.
34. Martinez, S. G.; Tenorio-Borroto, E.; Barbabosa Pliego, A.; Diaz-Albiter, H.; Vazquez-Chagoyan, J. C.; Gonzalez-Diaz, H., PTML Model for Proteome Mining of B-cell Epitopes and Theoretic-Experimental Study of Bm86 Protein Sequences from Colima Mexico. *Journal of proteome research* **2017**.
35. Fernández, M.; Caballero, F.; Fernández, L.; Abreu, J. I.; Acosta, G., Classification of conformational stability of protein mutants from 3D pseudo-folding graph representation of protein sequences using support vector machines. *Proteins* **2008**, 70, 167-175.
36. Fernández, L.; Caballero, J.; Abreu, J. I.; Fernández, M., Amino Acid Sequence Autocorrelation Vectors and Bayesian-Regularized Genetic Neural Networks for Modeling Protein Conformational Stability: Gene V Protein Mutants. *Proteins* **2007**, 67, 834–852.
37. Caballero, J.; Fernandez, L.; Abreu, J. I.; Fernandez, M., Amino Acid Sequence Autocorrelation vectors and ensembles of Bayesian-Regularized Genetic Neural Networks for prediction of conformational stability of human lysozyme mutants. *Journal of chemical information and modeling* **2006**, 46, 1255-68.
38. Perez-Riverol, Y.; Audain, E.; Millan, A.; Ramos, Y.; Sanchez, A.; Vizcaino, J. A.; Wang, R.; Muller, M.; Machado, Y. J.; Betancourt, L. H.; Gonzalez, L. J.; Padron, G.; Besada, V., Isoelectric point optimization using peptide descriptors and support vector machines. *Journal of proteomics* **2012**, 75, 2269-74.
39. Greener, J. G.; Sternberg, M. J., AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC bioinformatics* **2015**, 16, 335.
40. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.;

Magarinos, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R., The ChEMBL database in 2017. *Nucleic Acids Res* **2017**, 45, D945-D954.

41. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **2012**, 40, D1100-7.
42. Gonzalez-Diaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M., General Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry. *Current Topics in Medicinal Chemistry* **2013**, 13, 1713-1741.
43. Blazquez-Barbadillo, C.; Aranzamendi, E.; Coya, E.; Lete, E.; Sotomayor, N.; Gonzalez-Diaz, H., Perturbation theory model of reactivity and enantioselectivity of palladium-catalyzed Heck-Heck cascade reactions. *Rsc Advances* **2016**, 6, 38602-38610.
44. Casanola-Martin, G. M.; Le-Thi-Thu, H.; Perez-Gimenez, F.; Marrero-Ponce, Y.; Merino-Sanjuan, M.; Abad, C.; Gonzalez-Diaz, H., Multi-output Model with Box-Jenkins Operators of Quadratic Indices for Prediction of Malaria and Cancer Inhibitors Targeting Ubiquitin-Proteasome Pathway (UPP) Proteins. *Current Protein & Peptide Science* **2016**, 17, 220-227.
45. Romero-Duran, F. J.; Alonso, N.; Yanez, M.; Caamano, O.; Garcia-Mera, X.; Gonzalez-Diaz, H., Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives. *Neuropharmacology* **2016**, 103, 270-278.
46. Kleandrova, V. V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. 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 Science & Technology* **2014**, 48, 14686-14694.
47. Luan, F.; Kleandrova, V. V.; Gonzalez-Diaz, 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* **2014**, 6, 10623-30.

48. Alonso, N.; Caamano, O.; Romero-Duran, F. J.; Luan, F.; Cordeiro, M. N. D. S.; Yanez, M.; Gonzalez-Diaz, H.; Garcia-Mera, X., Model for High-Throughput Screening of Multitarget Drugs in Chemical Neurosciences: Synthesis, Assay, and Theoretic Study of Rasagiline Carbamates. *Acs Chemical Neuroscience* **2013**, 4, 1393-1403.
49. Speck-Planche, A.; Cordeiro, M., Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol. Divers.* **2017**, 21, 511-523.
50. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N., Unified multi-target approach for the rational in silico design of anti-bladder cancer agents. *Anticancer Agents Med Chem* **2013**, 13, 791-800.
51. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N., Chemoinformatics in multi-target drug discovery for anti-cancer therapy: in silico design of potent and versatile anti-brain tumor agents. *Anticancer Agents Med Chem* **2012**, 12, 678-85.
52. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N., Rational drug design for anti-cancer chemotherapy: multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. *Bioorg. Med. Chem.* **2012**, 20, 4848-55.
53. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N., Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents. *Eur. J. Pharm. Sci.* **2012**, 47, 273-9.
54. Cordeiro, M. N.; Speck-Planche, A., Computer-aided drug design, synthesis and evaluation of new anti-cancer drugs. *Curr Top Med Chem* **2012**, 12, 2703-4.
55. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N., Multi-target drug discovery in anti-cancer therapy: fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorg. Med. Chem.* **2011**, 19, 6239-44.
56. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N., Multi-target drug discovery in anti-cancer therapy: fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorg. Med. Chem.* **2011**, 19, 6239-44.

57. Collier, G.; Ortiz, V., Emerging computational approaches for the study of protein allostery. *Archives of biochemistry and biophysics* **2013**, 538, 6-15.
58. 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 Systems with Applications* **2017**, 72, 306-316.
59. Gonzalez-Diaz, H.; Herrera-Ibata, D. M.; Duardo-Sanchez, 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. *Journal of chemical information and modeling* **2014**, 54, 744-55.
60. Casanola-Martin, G. M.; Le-Thi-Thu, H.; Perez-Gimenez, F.; Marrero-Ponce, Y.; Merino-Sanjuan, M.; Abad, C.; Gonzalez-Diaz, H., Multi-output model with Box-Jenkins operators of linear indices to predict multi-target inhibitors of ubiquitin-proteasome pathway. *Mol. Divers.* **2015**, 19, 347-56.
61. Hill, T.; Lewicki, P., *STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining*. StatSoft: Tulsa, 2006 Vol. 1, p 813.
62. Young, R. C.; Mitchell, R. C.; Brown, T. H.; Ganellin, C. R.; Griffiths, R.; Jones, M.; Rana, K. K.; Saunders, D.; Smith, I. R.; Sore, N. E.; et al., Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists. *Journal of medicinal chemistry* **1988**, 31, 656-71.
63. Friden, M.; Winiwarter, S.; Jerndal, G.; Bengtsson, O.; Wan, H.; Bredberg, U.; Hammarlund-Udenaes, M.; Antonsson, M., Structure-brain exposure relationships in rat and human using a novel data set of unbound drug concentrations in brain interstitial and cerebrospinal fluids. *Journal of medicinal chemistry* **2009**, 52, 6233-43.
64. Hitchcock, S. A.; Pennington, L. D., Structure-brain exposure relationships. *Journal of medicinal chemistry* **2006**, 49, 7559-83.

65. Garcia, I.; Fall, Y.; Garcia-Mera, X.; Prado-Prado, F., Theoretical study of GSK-3alpha: neural networks QSAR studies for the design of new inhibitors using 2D descriptors. *Mol. Divers.* **2011**, 15, 947-55.
66. Marrero-Ponce, Y.; Siverio-Mota, D.; Galvez-Llompart, M.; Recio, M. C.; Giner, R. M.; Garcia-Domenech, R.; Torrens, F.; Aran, V. J.; Cordero-Maldonado, M. L.; Esguera, C. V.; de Witte, P. A.; Crawford, A. D., 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, 5736-53.
67. Kuligowski, J.; Perez-Guaita, D.; Escobar, J.; de la Guardia, M.; Vento, M.; Ferrer, A.; Quintas, G., Evaluation of the effect of chance correlations on variable selection using Partial Least Squares-Discriminant Analysis. *Talanta* **2013**, 116, 835-40.
68. Bhagwanth, S.; Mishra, S.; Daya, R.; Mah, J.; Mishra, R. K.; Johnson, R. L., Transformation of Pro-Leu-Gly-NH2 peptidomimetic positive allosteric modulators of the dopamine D2 receptor into negative modulators. *ACS chemical neuroscience* **2012**, 3, 274-84.
69. Celis, M. E.; Taleisnik, S.; Walter, R., Regulation of formation and proposed structure of the factor inhibiting the release of melanocyte-stimulating hormone. *Proceedings of the National Academy of Sciences of the United States of America* **1971**, 68, 1428-33.
70. Katzenschlager, R.; Jackson, M. J.; Rose, S.; Stockwell, K.; Tayarani-Binazir, K. A.; Zubair, M.; Smith, L. A.; Jenner, P.; Lees, A. J., Antiparkinsonian activity of L-propyl-L-leucyl-glycinamide or melanocyte-inhibiting factor in MPTP-treated common marmosets. *Mov. Disord.* **2007**, 22, 715-9.
71. Palomo, C.; Aizpurua, J. M.; Benito, A.; Miranda, J. I.; Fratila, R. M.; Matute, C.; Domercq, M.; Gago, F.; Martin-Santamaria, S.; Linden, A., Development of a new family of conformationally restricted peptides as potent nucleators of beta-turns. Design, synthesis, structure, and biological evaluation of a beta-lactam peptide analogue of melanostatin. *J. Am. Chem. Soc.* **2003**, 125, 16243-60.

72. Sampaio-Dias, I. E.; Silva-Reis, S. C.; Garcia-Mera, X.; Brea, J.; Loza, M. I.; Alves, C. S.; Algarra, M.; Rodriguez-Borges, J. E., Synthesis, Pharmacological, and Biological Evaluation of MIF-1 Picolinoyl Peptidomimetics as Positive Allosteric Modulators of D2R. *ACS chemical neuroscience* **2019**, 10, 3690-3702.
73. Ferreira da Costa, J.; Caamaño, O.; Fernández, F.; García-Mera, X.; Sampaio-Dias, I. E.; Brea, J. M.; Cadavid, M. I., Synthesis and allosteric modulation of the dopamine receptor by peptide analogs of l-prolyl-l-leucyl-glycinamide (PLG) modified in the l-proline or l-proline and l-leucine scaffolds. *European Journal of Medicinal Chemistry* **2013**, 69, 146-158.
74. Sampaio-Dias, I. E. S., C. A. D.; García-Mera, X.; da Costa, J. F.; Caamaño, O.; Rodríguez-Borges, J. E., *Org. Biomol. Chem.* **2016**, 14, 11065-11069.
75. Sampaio-Dias, I. E.; Sousa, C. A. D.; Silva-Reis, S. C.; Ribeiro, S.; Garcia-Mera, X.; Rodriguez-Borges, J. E., Highly efficient one-pot assembly of peptides by double chemoselective coupling. *Organic & Biomolecular Chemistry* **2017**, 15, 7533-7542.
76. Liang, X.; Haynes, B. S.; Montoya, A., Acid-Catalyzed Ring Opening of Furan in Aqueous Solution. *Energy & Fuels* **2018**, 32, 4139-4148.
77. Verma, V.; Mann, A.; Costain, W.; Pontoriero, G.; Castellano, J. M.; Skoblenick, K.; Gupta, S. K.; Pristupa, Z.; Niznik, H. B.; Johnson, R. L.; Nair, V. D.; Mishra, R. K., Modulation of agonist binding to human dopamine receptor subtypes by L-prolyl-L-leucyl-glycinamide and a peptidomimetic analog. *The Journal of pharmacology and experimental therapeutics* **2005**, 315, 1228-36.
78. Christopoulos, A.; Kenakin, T., G protein-coupled receptor allosterism and complexing. *Pharmacological reviews* **2002**, 54, 323-74.
79. Bhagwanth, S.; Mishra, R. K.; Johnson, R. L., Development of peptidomimetic ligands of Pro-Leu-Gly-NH(2) as allosteric modulators of the dopamine D(2) receptor. *Beilstein Journal of Organic Chemistry* **2013**, 9, 204-214.

80. Vartak, A. P.; Skoblenick, K.; Thomas, N.; Mishra, R. K.; Johnson, R. L., Allosteric modulation of the dopamine receptor by conformationally constrained type VI beta-turn peptidomimetics of Pro-Leu-Gly-NH2. *J Med Chem* **2007**, 50, 6725-9.

81. Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V., Virtual computational chemistry laboratory--design and description. *Journal of computer-aided molecular design* **2005**, 19, 453-63.

82. Gonzalez-Diaz, H.; Perez-Montoto, L. G.; Ubeira, F. M., Model for vaccine design by prediction of B-epitopes of IEDB given perturbations in peptide sequence, in vivo process, experimental techniques, and source or host organisms. *Journal of immunology research* **2014**, 2014, 768515.

83. Gonzalez-Diaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M., 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, 1713-41.

84. Gottlieb, H. E.; Kotlyar, V.; Nudelman, A., NMR Chemical Shifts of Common Laboratory Solvents as Trace Impurities. *The Journal of Organic Chemistry* **1997**, 62, 7512-7515.

Chapter 03.
Anticancer Compounds

3. CHAPTER 03. ANTICANCER COMPOUNDS

Paper 3. Bediaga H, Arrasate S, González-Díaz H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Comb Sci*. 2018 Nov 12; 20 (11): 621-632. doi: 10.1021/acscombsci.8b00090.

3.1. CHAPTER 03. ABSTRACT.

Determining thetarget proteins of new anti-cancer compounds is a very important task in Medicinal Chemistry. In this sense, chemists need to carry out assays with multiple experimental conditions (c_j). These assays cover >70 different biological activity parameters (c_0), >300 different drug targets (c_1), >230 cell lines (c_2), and 5organismsof assay (c_3) and/or organisms of the target (c_4), *etc*.These assays in the raw data set downloaded from ChEMBL correspond to 2499 of ovarian cancer, 45833leukemia, 6227breast, 3499colon, 3159 lung, 2750 prostate, 601 melanoma,492stomach, 134liver cancer, *etc*. This is a very complex dataset to be rationalized by researchers in order to extract useful relationships and predict new compounds. We developed a new IFPTML model for this ChEMBLdataset of preclinical assays of anti-cancer compounds.This is a simple but very powerful linear model with only three variables, AUROC = 0.872, Specificity = Sp(%) = 90.2, Sensitivity = Sn(%) = 70.6, and overall Accuracy = Ac(%) = 87.7 in training series. The model also have Sp(%) = 90.1, Sn(%) = 71.4, and Ac(%) = 87.8 in external validation series. The model use PT operators based on multi-condition moving averages to capture all the complexity of the dataset. We also compared the model with non-linear ANN models obtaining similar results.This confirms the hypothesis of a linear relationship between the PT operators and the classification as anti-cancer compound. This model is a simple but versatile tool for the prediction of the targets of anti-cancer compounds taking into consideration multiple variations in preclinical assays.

Keywords: ChEMBL; Anti-cancer compounds; Perturbation Theory; Machine Learning.

3.2. CHAPTER 03. INTRODUCTION.

The World Health Organization (WHO) pointed out that Cancer is still among the more dangerous diseases nowadays. Classic anticancer compounds use to have a high cytotoxicity and multiple cellular targets. Regrettably, they also attack normal cells causing untoward effects. Recently, selective anticancer compounds have been introduced to target specific abnormalities in a cancer cell. Nevertheless, taking into consideration the number different mechanisms and possible targets it is unlikely that a single chemical become one hundred percent effective. In this context, it is more probable that new drug leads will be more effective if combined with other chemicals in order to cover multiple action mechanisms [126].Medicinal chemists could use experimental techniques and/or computational techniques to predict new drugs against different targets [53]. Specifically, in Machine Learning (ML) [127-129] techniques we can calculate different molecular descriptors codify the chemical structure of chemical compounds [70-86].

Unfortunately, classic methods fail to account for large "Big Data (BD)" sets of preclinical assays. BD sets of assays of anticancer compounds are difficult to study due to the high complexity of the data in addition to the huge volume. For example, the ChEMBL database compiles BD sets of very heterogeneous preclinical assays [88, 89]. ChEMBL BD sets of anti-cancer compounds cover multiple biological activity parameters (Potency, IC_{50}, K_i, K_m, *etc.*), different cellular lines, organisms of the protein target, organism of assay, *etc.* [130-132]. In fact, many researchers have developed Cheminformatics models for the discovery of anticancer compounds. Nevertheless, almost all of these models are specific for homologous series of compounds, one target, and/or unique cell line. Other models uses heterogeneous series of compounds but are unable to incorporate information about the target, cell line, organism of assay, *etc.* [54-69].

On the other hand,Perturbation Theory (PT) ideas combined with ML methods (PT + ML = IFPTML models)may be used to solve this kind of problems in compound discovery[82, 90]. In fact, IFPTML models have been used in Medicinal Chemistry, Proteomics, Nanotechnology, *etc.* for modeling large data sets with BD features[8-13]. More recently, Speck-Planche and Cordeiro*et al.* have developed Cheminformatics models for anticancer compounds incorporating BD specs like

organisms, cell lines, *etc*. They have studied large data sets of compounds with BD specs form ChEMBL database. They have pre-processed these data sets to calculate PT operators. Later, they used a ML technique to process the data and obtaining a IFPTML predictive model.

However, in all cases the IFPTML models reported focused on one specific types of cancer (such as breast, brain, prostate, or bladder cancer) each model [91-98]. Nevertheless, there are no reports of general-purpose IFPTML models for anticancer compounds including data for multiple types of cancer at the same time, until the best of our knowledge. In this work, we develop the first general-purpose IFPTML model for the prediction of anticancer compounds against several types of cancers. In **Figure 1** we illustrate the general workflow used in this paper to obtain the IFPTML model.

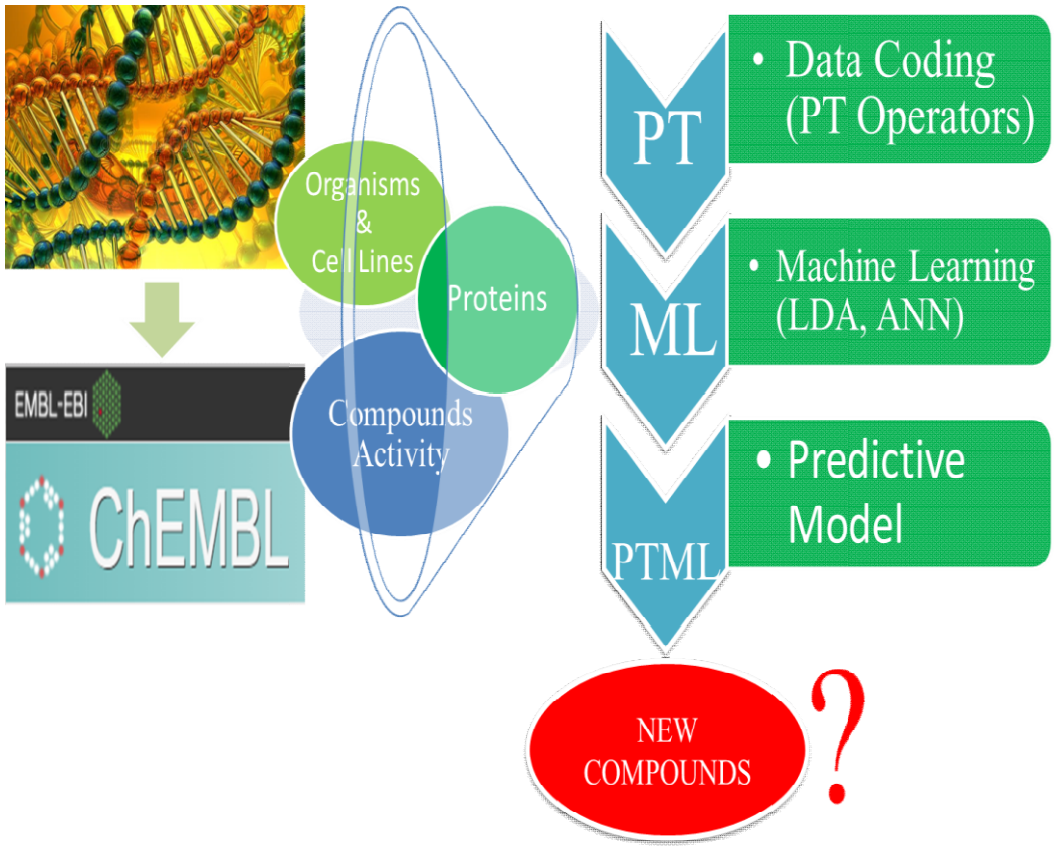


Figure 1. General workflow used in this paper to obtain the IFPTML model

3.3. IFPTML linear model with one-condition moving averages.

IFPTML model starts with the expected value of activity and adds the effect of different perturbations in the system. Consequently, the model has two types of input variables: the expected-value function $f(v_{ij})_{expt}$ and the PT operators $\Delta D_k(c_j)$. The input variable $f(v_{ij})_{expt}$ represents the expected value of biological activity for one compound in n different experimental conditions $\mathbf{c}_j = (c_0, c_1, c_2, \ldots c_j \ldots c_{max})$. Please, note that $\mathbf{c}_j$ (in **boldface**) denotes a vector of multiple conditions and c_j denotes a single condition. The other PT operators are Moving Averages (MA) calculated for one condition at time. It means that we can calculate the PT operators as follow $\Delta D_k(c_j) = D_{ki} - \langle D_k(c_j) \rangle$. They depend on the value of the molecular descriptorsD_{ki} of type k used to quantify the structure of the i^{th} drug. In this work, the specific molecular descriptors used are D_1 = ALOGP (n-Octanol/Water Partition Coefficient) and D_2 = TPSA (Polar Surface Area). These values where taken directly from ChEMBL data set. The PT operators measure the deviation of $D_{ki}$from the expected value of $\langle D_k(c_j) \rangle$ (average value) of this descriptor for different sets of drugs $c_j^{100\text{-}102}$. The equation of the best model found using this kind of PT operators is the following:

$$\begin{aligned} f(v_{ij})_{calc} = & -5.94193 + 14.73111 \cdot f(v_{ij})_{expt} \\ & -0.07986 \cdot \Delta D_1(c_0) \\ & -0.08724 \cdot \Delta D_1(c_1) \\ & -0.00628 \cdot \Delta D_1(c_2) \\ & +0.03577 \cdot \Delta D_1(c_4) \\ & +0.00651 \cdot \Delta D_1(c_5) \end{aligned} \qquad (1)$$

$$n = 87701 \qquad \chi^2 = 42891.07 \qquad p < 0.05$$

In **Table 1** we included a more detailed explanation about all the input variables analyzed to seek the PTML-LDA linear model. The output of the model $f(v_{ij})_{calc}$ is a scoring function of the

value v_{ij} of biological activity of the i^{th} drug in the conditions of assay $\mathbf{c}_j$. In the particular case of an LDA model $f(v_{ij})_{calc}$is not in the range 0-1 and it is not a probability. However, for a given value of $f(v_{ij})_{calc}$the LDA algorithmcan calculate the respective values of posterior probabilities $p(f(v_{ij})= 1)_{pred}$. The LDA algorithm uses the Mahalanobis's distance metric to calculate these probabilities [48].After calculating $p(f(v_{ij})= 1)_{pred}$ we can easily calculate the Boolean function $f(v_{ij})_{pred}= 1$ when $p(f(v_{ij})= 1)_{pred}> 0.5$ or $f(v_{ij})_{pred}= 0$ otherwise. The values of $f(v_{ij})_{pred}= 1$ or 0 have to be compared with the respective observed values $f(v_{ij})_{obs}= 1$ or 0 to calculate the Sn, Sp, and Ac of the model. It is straight forward to realize that when $f(v_{ij})_{pred} = f(v_{ij})_{obs}$ the case is correctly classified [48]. The present model presented high values of Specificity Sp = 89.9, Sensitivity Sn = 70.8, and overall Accuracy Ac = 87.4 in training series. The model presented very similar values of Sn, Sp, and Ac in external validation series, see **Table 1**. These values are in the range considered as useful for classification models with application in Medicinal Chemistry [107]. The data points (compound-assay pair) used in validation series have not been used to train the model.

Table 1.Results of the model and input variables analyzed

Obs. Sets[a]	Stat. Param.[b]	Pred. Stat.	Predicted sets		
			n_j	$f(v_{ij})_{pred} = 0$	$f(v_{ij})_{pred} = 1$
Training series					
$f(v_{ij})_{obs}= 0$	Sp	89.9	76553	68795	7758
$f(v_{ij})_{obs} = 1$	Sn	70.8	11148	3258	7890
Total	Ac	87.4	87701		
Validation series					
$f(v_{ij})_{obs}= 0$	Sp	89.9	25598	23016	2582
$f(v_{ij})_{obs} = 1$	Sn	71.8	3635	1024	2611
Total	Ac	87.7	29233		

We used the forward-stepwise strategy [48] of variable selection to detect the more important perturbations on different conditions c_j related to the anticancer property, like organism of assay, cell line, *etc*.See details about these variables on **Table 2**.Interestingly, in this model the operators selected are only of type $\Delta D_1(c_j) = \Delta ALOGP(c_j)$. So they measure only perturbations on the value of ALOGP compared to other sub-sets c_j of drugs.The parameter ALOGP is widely used in Medicinal Chemistry because it is related to the lipophilicity of drugs and consequently to their capacity to pass through biological membranes or interact with protein hydrophobic pockets [133-136]. Specifically, the forward-stepwise strategy selected as important deviations on the kind of activity (c_0) measured for the same target (c_1), in the same cell line (c_2), and assayed on the same organism (c_4).

Table 2. PT operators used as input

c_j	Condition	Symbol	Operator Formula	Operator Information
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 a given type of activity (v_{ij})
c_0	Activity type	$\Delta D_1(c_0)$	$ALOGP_i$ - <$ALOGP(c_0)$>	$\Delta ALOGP(c_j)$ accounts for variability on chemical structure and conditions of assay c_j in terms of deviation (Δ) of the Hydrophobicity of the compound ($ALOGP_i$) from the expected value (<$ALOGP(c_j)$>) for a givenassay conditions
c_1	Target	$\Delta D_1(c_1)$	.	
c_2	Cell Line	$\Delta D_1(c_2)$	.	
c_3	Organism of Assay	$\Delta D_1(c_3)$	.	
			$ALOGP_i$ - <$ALOGP(c_j)$>	
c_4	Organism of Target	$\Delta D_1(c_4)$	.	
			.	
c_5	Target type	$\Delta D_1(c_5)$	$ALOGP_i$ - <$ALOGP(c_5)$>	

We can use this model to score the activity of a new compound in different conditions of assay. Firstly, we have to substitute the expected probability of activity $p(f(v_{ij})_{obs}=1)_{expt}$ on the equation, see **Table 3**.Note that these values change for different activities, like Potency(nM), Inhibition(%), IC_{50}(nM), GI_{50}(nM), Activity(%), TGI(nM), *etc*. Consequently, the model can predict different kinds of activity parameters for a single compound. Next, we have to substitute in the model the values of ALOGP for a new compound (taken from ChEMBL and/or calculated with software).

Table 3. One-condition parameters for selected cases

Condition c_0[a]	Input parameters used to specify c_0[b]						
Activity	$<D_1(c_0)>$	$<D_2(c_0)>$	$n_j(c_0)$	$n_j(f(v_{ij})=1)_{obs}$	$p(f(v_{ij})=1)_{ref}$	cutoff	$d(c_0)$
Potency(nM)	3.64	78.88	41934	2027	0.05	100	-1
Inhibition(%)	4.31	64.99	18698	408	0.02	142	1
IC_{50}(nM)	3.70	99.06	16052	2696	0.17	100	-1
GI_{50}(nM)	3.79	87.78	13263	1469	0.11	100	-1
Activity(%)	4.80	91.65	5112	3009	0.59	67.4	1
TGI(nM)	3.97	102.25	3931	8	0.00	100	-1
LC_{50}(nM)	4.31	89.93	3805	13	0.00	100	-1
EC_{50}(nM)	3.50	91.22	1510	246	0.16	100	-1
K_i(nM)	3.00	93.85	1363	44	0.03	100	-1

In order to calculate these expected values of probability we have to evaluate the formula $p(f(v_{ij})_{obs}=1)_{expt}= n(f(v_{ij})=1)_{obs}/n_j$. This is the ration between the number of compounds $n(f(v_{ij})=1)_{obs}$ with a desired level of activity for the condition c_j and the number of compounds n_j assayed for the same condition c_j. A compound have a desired level of activity $f(v_{ij})_{obs}=1$ when the value of activity $v_{ij}>$ cutoff for properties with desirability $d(c_0) = 1$. A compound also have a desired level of activity $f(v_{ij})_{obs}=1$ when the value of activity $v_{ij}<$ cutoff for properties with desirability $d(c_0) = -1$. Otherwise, the compound is not considered to have a desired level of activity and then $f(v_{ij})_{obs}= 0$.

It is straightforward to realize that that the property desirability $d(c_0) = 1$ for properties of the compound that we want to maximize and $d(c_0) = -1$ otherwise. The cutoff = 100 for properties with unitsinnM. If not, cutoff = $<v_{ij}>$ expected value (average) of the value of activity v_{ij}.In order to predict a new compound, we also have to substitute in the model the expected values of the molecular descriptors $<D_1(c_j)>$ for different conditions. In **Table 4**, we depict selected values of the averages $<D_1(c_j)>$. In the table, you can note that these values change for different conditions, so the model give a different result for one compound if you change this condition.

Table 4. One-condition averages and number of cases for selected conditions of assay

Condition c_1[a]	Parameters used to specify c_1		
Target	$<D_1(c_1)>$	$<D_2(c_1)>$	$n_j(c_1)$
Cellular tumor antigen p53	3.88	81.58	26104
Breast cancer type 1 susceptibility protein	3.24	74.78	15868
ATP-binding sub-family G member 2	4.09	84.03	1572
PDZ-binding kinase	3.05	92.33	891
Condition c_2[a]	Parameters used to specify c_2		
Cell Line	$<D_1(c_2)>$	$<D_2(c_2)>$	$n_j(c_2)$
MCF7	3.92	101.75	2893
LNCaP	4.68	89.60	252
MDA-MB-435	3.59	91.66	546
Condition c_3[a]	Parameters used to specify c_3		
Assay organism	$<D_1(c_3)>$	$<D_2(c_3)>$	$n_j(c_3)$
H. sapiens	3.86	83.70	111111
M. musculus	2.67	91.35	1640
Condition c_4[a]	Parameters used to specify c_4		
Organism	$<D_1(c_4)>$	$<D_2(c_4)>$	$n_j(c_4)$
H. sapiens	3.84	83.11	99590
M. musculus	2.36	90.81	1245
Condition c_5[a]	Parameters used to specify c_5		

Target type	$<D_1(c_5)>$	$<D_2(c_5)>$	$n_j(c_5)$
Cell-Line	3.99	85.83	51407
Single Protein	3.63	80.49	49001
Unchecked	4.05	88.42	14460
Organism	2.58	90.34	923

[a]*The full name of the species isHomosapiens,Mesocricetusauratus, Musmusculus, Rattusnorvegicus.*

3.4. IFPTML linear model with multi-condition moving averages.

We call your attention that the PT operators used in the previous model are based on single-condition averages (the simplest case). There is another possibility of calculating PT operators based on multi-condition averages[106]. It means, the average calculation runs over all cases with the same set of conditions $\mathbf{c}_j$. Remember that, in this context, $\mathbf{c}_j$ (with **c** in boldface) refers to a vector of multiple conditions $\mathbf{c}_j = (c_0, c_1, c_2, c_3, c_4)$. However, we can calculate the averages using different combinations of conditions. The more complex case is the use of all conditions together c_0-c_4. The equation of this model is the following.

$$\begin{aligned} f(v_{ij})_{calc} = & -5.9391530229475200 \\ & +14.80382301029070000 \cdot f(v_{ij})_{expt} \\ & -0.1086628540078630 \cdot \Delta D_1(c_0, c_1, c_2, c_3, c_4) \\ & +0.00686894639635945 \cdot \Delta D_2(c_0, c_1, c_2, c_3, c_4) \end{aligned} \quad (2)$$

$$n = 87701 \qquad \chi^2 = 43222.56 \qquad p < 0.05$$

In training series, the model presented high values of Specificity = Sp(%) = 90.2, Sensitivity = Sn(%) = 70.6, and overall Accuracy = Ac(%) = 87.7. The model was stable in external validation series with values of Sp(%) = 90.1, Sn(%) = 71.4, and Ac(%) = 87.8. LDA models are Bayesian methods in the sense that they calculate the posterior probabilities $p(f(v_{ij})= 1)_{pred}$ taking into consideration the prior probability $p(f(v_{ij})= 1)_{prior}$[48]. This is the probability with which

all compounds presented the desired classification $f(v_{ij})_{pred}= 1$. In this paper the *prior* probability was set as $p(f(v_{ij})= 1)_{prior} = 0.8$ for the best model found. Consequently, we also obtained values of AUROC = 0.87; which are notably higher than AUROC = 0.5 (value for a random classifier), **Figure 2**. This indicates that our model is not a random classifier. In fact, the value of Chi-square is $\chi^2 = 43222.56$ with a p-level < 0.05 indicating that the classifier performs a statistically significant separation of both classes ($f(v_{ij})_{obs} = 0$ *vs*. $f(v_{ij})_{obs} = 0$).In **Table 5**, we summarize the results for this new model.

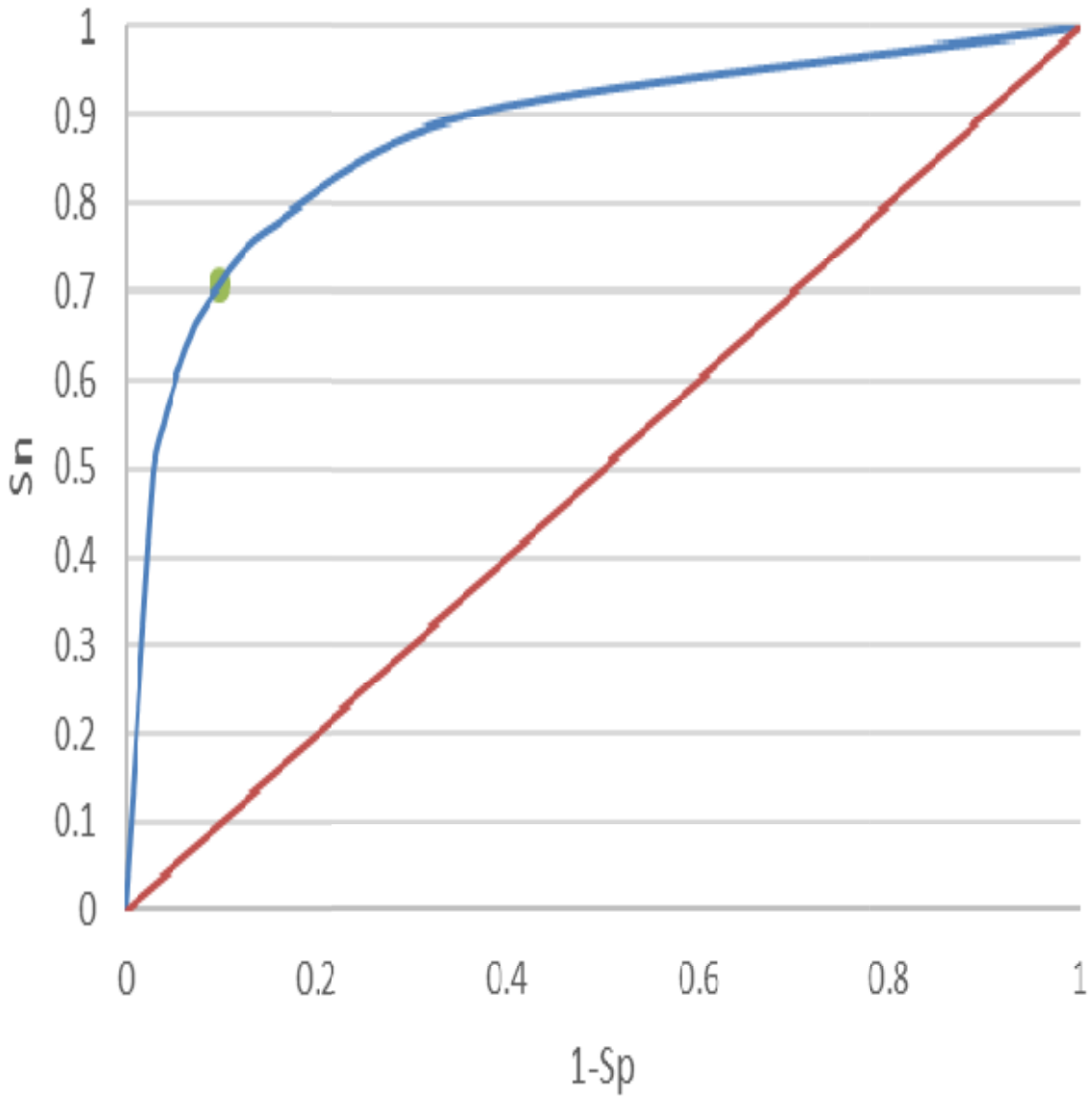


Figure 2.ROC analysis of the PTML-LDA model

Table 5. PTML-LDA model results

Data	Priors	Observed	Statistical	Predicted	$f(v_{ij})_{pred}$

Set	p(1)	Sets [a]	Parameter[b]	Statistics	n_j	0	1
t	0.80	$f(v_{ij})_{obs}= 0$	Sp(%)	90.2	76553	69083	7470
		$f(v_{ij})_{obs} = 1$	Sn(%)	70.6	11148	3274	7874
		total	Ac(%)	87.7	87701		
v		$f(v_{ij})_{obs}= 0$	Sp(%)	90.1	25598	23071	2527
		$f(v_{ij})_{obs} = 1$	Sn(%)	71.4	3635	1039	2596
		total	Ac(%)	87.8	29233		
t	0.50	$f(v_{ij})_{obs}= 0$	Sp(%)	93.7	76553	71698	4855
		$f(v_{ij})_{obs} = 1$	Sn(%)	63.7	11148	4049	7099
		total	Ac(%)	89.8			
v		$f(v_{ij})_{obs}= 0$	Sp(%)	93.4	25598	23921	1677
		$f(v_{ij})_{obs} = 1$	Sn(%)	64.3	3635	1297	2338
		total	Ac(%)	89.8			
t	0.12	$f(v_{ij})_{obs}= 0$	Sp(%)	95.8	76553	73358	3195
		$f(v_{ij})_{obs} = 1$	Sn(%)	56.1	11148	4897	6251
		total	Ac(%)	90.8			
v		$f(v_{ij})_{obs}= 0$	Sp(%)	95.6	25598	24484	1114
		$f(v_{ij})_{obs} = 1$	Sn(%)	56.4	3635	1585	2050
		total	Ac(%)	90.8			

[a]The observed classification classes are two: drugs with a desired level of biological effect observed $f(v_{ij})_{obs}= 1$or $f(v_{ij})_{obs}= 0$ otherwise. [b]Sn(%) = Sensitivity, Sp(%) = Specificity, and Ac(%) = Accuracy.

One can argue that this kind of multi-condition variable is somehow less informative because the information about all conditions of the same assay is merged together in one input variable. However, the model is very simple and powerful precisely because you can include a lot of information about assay conditions in only one variable. As in the previous case, we can also use this model to score the activity of a new compound. Now we need to use here PT operators sensible to changes on multiple conditions at the same time.They are the expected probability

$p_j(f(v_{ij}) = 1/\mathbf{c}_j)_{expt}$and the average values $<D_1(\mathbf{c}_j)>$for multiple conditions at the same time, see **Table 6**.

Table 6. Selected values of multi-condition averages

Multi-condition assays[a]				Multi-condition input parameters [b]			
Activity	Target	Cell	Assay Org.	Averages		Count &Probs	
(c_0)	(c_1)	(c_2)	(c_3)	$<D_1(\mathbf{c}_j)>$	$<D_2(\mathbf{c}_j)>$	$n_j(\mathbf{c}_j)$	$f(v_{ij})=1/\mathbf{c}_j)_{expt}$
Potency (nM)	Cellular tumor antigen p53	-	*Hs*	3.88	81.58	26104	0.08
Potency (nM)	Breast cancer type 1 susceptibility protein	-	*Hs*	3.25	74.42	15830	0.00
IC_{50}(nM)	PI3-kinase p110-alpha subunit	-	*Hs*	3.31	97.73	827	0.46
Inhibition (%)	Tubulin alpha chain	MCF7	*Hs*	2.99	83.29	18	0.00
K_i(nM)	PDZ-binding kinase	-	-	3.27	91.68	586	0.00

[a]*Hs = Homo sapiens*, *Mm = Musmusculus, Ss = Susscrofa*.[b] Please, note that multi-contidion parameters depend on a vector $\mathbf{c}_j$ (denoted in **boldface**) of conditions and not on a single condition c_j.

3.5. IFPTML linear model simulation of multi-condition space.

The IFPTML model can be used also to carry out simulations of the effect over anticancer activity of perturbations in the chemical structure and multiple conditions of assay. Using the model we can predict the values of anti-cancer activity scores $f(vij)_{calc}$ and/or probabilities of activity $p(f(v_{ij})= 1)_{pred}$. We predicted the values of $p(f(v_{ij})= 1)_{pred}$ for more than 115000 data points.Next, we can extrapolate these values for a large 3D chemical space using distance-weighted least squares[48]. In **Figure 3** we depict the results of this extrapolation in a 3D contour plot. The figure plots the values of $p(f(v_{ij})= 1)_{pred}$(out of plane axis) *vs*. the values of the MMAs $\Delta ALOGP(c_0, c_1, c_2, c_3, c_4)$ and $\Delta PSA(c_0, c_1, c_2, c_3, c_4)$.

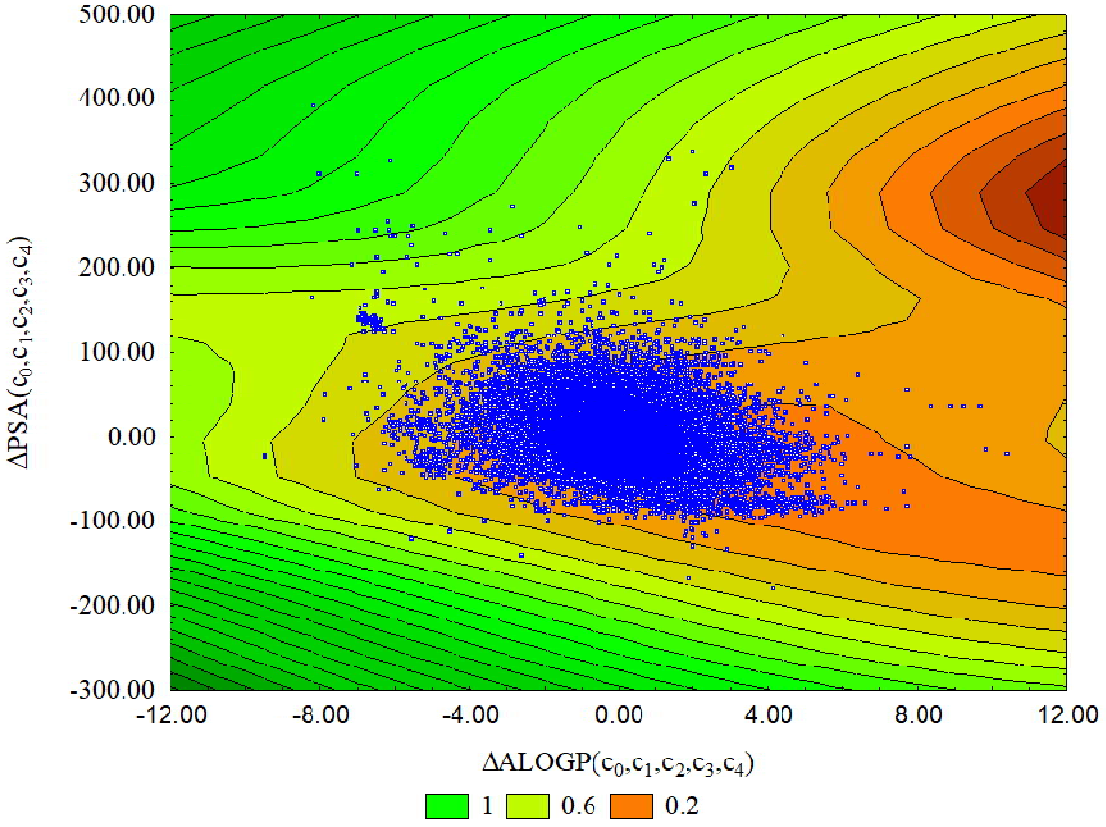


Figure 3. Heatmap for $p(f(v_{ij} = 1)_{pred}$ using distance-weighted least squares

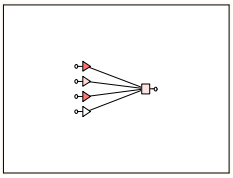
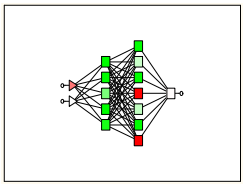
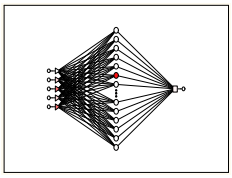
This kind of simulations can help us to discover general tendencies, if any, in the biological activity. For instance, we can see that in this 2D chemical space the values of desired biological activity appear in the area with lower $\Delta D_1(c_j)$ = ΔALOGP(c_0, c_1, c_2, c_3, c_4) and $\Delta D_2(c_j)$ = ΔPSA(c_0, c_1, c_2, c_3, c_4). It means that those chemical structure changes decreasing ALOGP (hydrophobicity) bellow the expected value <ALOGP($\mathbf{c}_j$)> are expected to increase the desired anticancer activity. Besides, chemical structure changes increasing TPSA (surface polarity) over the expected <PSA($\mathbf{c}_j$)> may increase also the desired anticancer activity. These values are coherent with the values (sign) equal to -0.109 and +0.007 (approx.) for the coefficients of ΔALOGP(c_0, c_1, c_2, c_3, c_4) and ΔPSA(c_0, c_1, c_2, c_3, c_4) in the equation of the model. Please, note that the values of <ALOGP($\mathbf{c}_j$)> and <PSA($\mathbf{c}_j$)> change for different sets of assay conditions (see **Table 6**). For instance, a compound should have a ALOGP < 3.88 and a value of TPSA > 81.58 to reach a desired level of anticancer activity in the assay of Potency(nM) for Cellular tumor antigen p53 in Humans. However, a compound should have ALOGP < 3.31 and a value of TPSA > 97.73 in the assay of IC_{50}(nM) for PI3-kinase p110-alpha subunit in Humans. In any case, according to the model the expected value of probability for the assay may modulate the final value of activity but do not change the signal of the score. Note that the coefficient of $f(v_{ij})_{expt}$ in the model is positive +14.80 (approx.) and the values of the variable are ever positive (0in the range 0-1) because $f(v_{ij})_{expt}$ = $p(f(v_{ij}) = 1/\mathbf{c}_j)_{expt}$ is a probability.

3.6. IFPTML-ANN non-linear models.

We also tested different non-linear ANN algorithms to seek PTML-ANN models potentially alternative to PTML-LDA models.The PTML-ANN models found have more balanced values of Spand Sn ≈ 80% in training and validation series, **Table 7**. All the PTML-ANN models found presented values of AUROC ≈ 0.90 for training and external validation series**.**This confirm again (like in PTML-LDA model) that the relationship among the PT operators and Anti-cancer activity is not random (AUROC = 0.5) [48]. In fact, these values are only slightly higher than the value of AUROC = 0.87 of the PTML-LDA model. The first Linear Neural Network (LNN)

models found (PTML-LNN) have four variables one more than the PTML-LDA model.The second PTML-LNN model has five variables but this does not increased significantly the values of Sp,Sn, and AUROC. We need to specify that the variables have been selected automatically by the variable selection algorithm of the ANN module of the software.The non-linear models Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) have similar performance than the linear models (LDA and LNN) in terms of Sp, Sn, and AUROC. In particular, the PTML-RBF model is notably complex having >1000 neurons in the hidden layer but even though the Sp, Sn, and AUROC values are similar to the linear models. This seems to confirm the hypothesis of the linearity of the relationship among the PT operators studied and anti-cancer activity. In closing, the ANN models have more balanced values of Sp and Sn but similar performance to the linear model.In**Figure 4**, we can see the curves for all ANN models (almost overlapped in many cases).

Table 7. PTML-ANN models results

Profile	AUROC	Set	0	1	(%)	Parm.	(%)	0	1
LNN 4:4-1:1 	0.90	0	61546	15007	80.40	Sp	80.36	20571	5027
	0.90	1	2113	9035	81.05	Sn	80.908	694	2941
MLP 2:2-5-7-1:1 	0.90	0	61337	15216	80.12	Sp	80.02	20484	5114
	0.90	1	2168	8980	80.55	Sn	80.440	711	2924
RBF 5:5-1336-1:1 	0.91	0	62622	13931	81.80	Sp	81.75	20926	4672
	0.91	1	1999	9149	82.07	Sn	81.981	655	2980

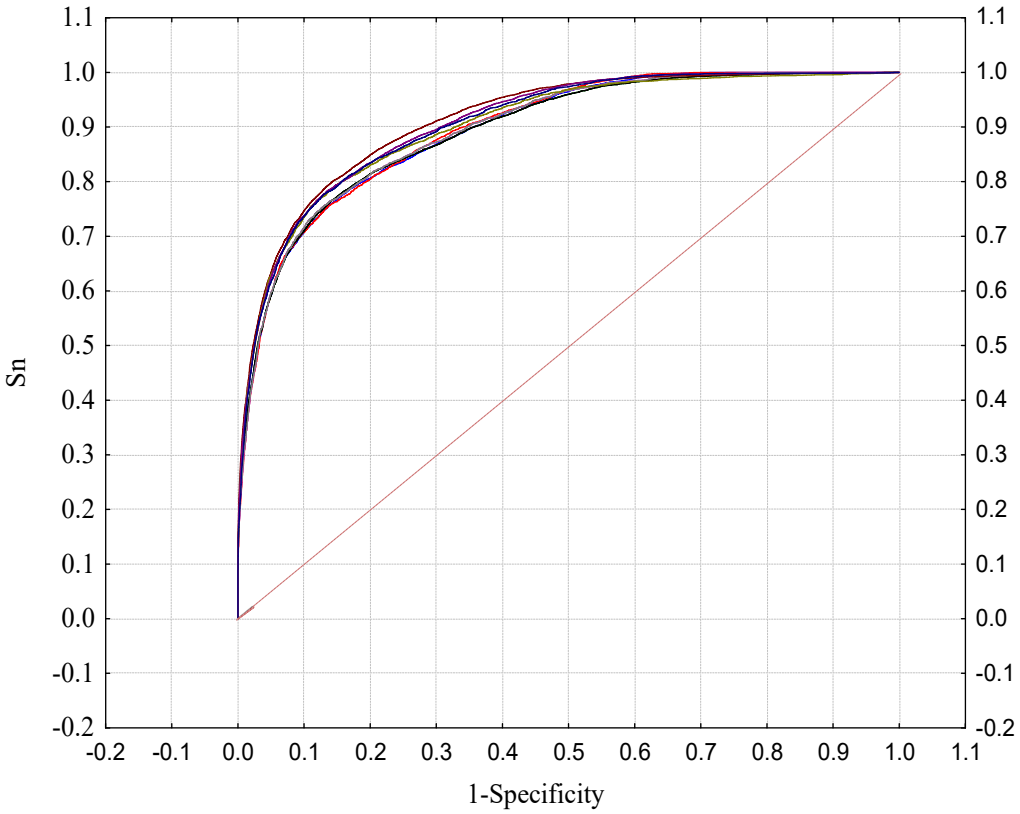


Figure 4. ROC analysis of the PTML-ANN models

The first input variable of all the PTML-ANN modelswasalsof$(v_{ij})_{expt}$ expected value of probability of a desired anticancer activity $f(v_{ij})=1$. The two structural variable have been also D_1 = ALOGP, D_2 = TPSA selected by the algorithm in some ANN models. Last, the algorithm also selected the two PT operators $\Delta D_1(\mathbf{c}_j)$ = ΔALOGP($\mathbf{c}_j$) and $\Delta D_2(\mathbf{c}_j)$ = ΔPSA($\mathbf{c}_j$) . These variables are similar to the variables in the PTML-LDA model. The result indicates that three variables seem to be enough to quantify the more significant information related to the anti-cancer activity. Remember that our PT operators are MMA; multi-condition. In all cases the sensitivity ratio was ≥ 1, indicating that these variables are significant for the model, see **Table 8**[48].

Table 8. Sensitivity analysis of all input variables in the PTML-ANN models

PTML-ANN Models	Data Set	Input variables sensitivity ratio[a]				
		$f(v_{ij})_{expt}$	D_1	D_2	$\Delta D_1(c_0,c_1,c_2,c_3,c_4)$	$\Delta D_2(c_0,c_1,c_2,c_3,c_4)$
LNN 4:4-1:1	train	1.28		1.00	1.00	1.00
	val	1.27		1.00	1.00	1.00
MLP 2:2-5-7-1:1	train	1.12		1.01		
	val	1.13		1.01		
RBF 5:5-1336-1:1	train	1.35	1.04	1.04	1.03	1.04
	val	1.32	1.03	1.03	1.03	1.03

[a]$f(v_{ij})_{expt}$ = $p(f(v_{ij}=1)/c_0)_{expt}$expected value of probability of anticancer activity for condition c_0(biologicalactivity parameter). D_1 = ALOGP, D_2 = PSA, $\Delta D_1(c_0,c_1,c_2,c_3,c_4)$ = ALOGP - <ALOGP(c_0,c_1,c_2,c_3,c_4)>, $\Delta D_2(c_0,c_1,c_2,c_3,c_4)$ = TPSA - <PSA(c_0,c_1,c_2,c_3,c_4)>

3.7. IFPTML *vs*. other models for anticancer compounds.

As we mentioned before Speck-Planche and Cordeiro*etal.*have reported various IFPTML models for the discovery of anticancer compounds. In **Table 9** we show a comparison among the present model and all these models. Many of these models are able to include at the same time also including information about different target proteins, cellular lines, organisms, *etc*. We can note that the Sp and Sn of the previous models is higher-to-similar to the present model. However, all these models are specific for only one type of cancer [91-98]. We should note that the model reported in this paper fits a very complex and larger data set of n > 80000 cases and >10 different types of cancer.

Table 9. Comparison to other IFPTML models of anti-cancer compounds

CancerType	Cancers	PT[a]	ML[b]	NV[b]	Cases[c]	Sn(%)[d]	Sp(%)[d]	Ref
Bladder	1	MA	LDA	n.a.	n.a.	>90	>90	[92]
Brain	1	MA	LDA	n.a.	n.a.	>90	>90	[93]
Colorectal	1	MA	LDA	>10	1237	>90	>90	[94]
Breast	1	MA	LDA	>10	2272	>85	>95	[95]
Prostate	1	MA	LDA	>10	1250	>85	>95	[98]
Breast	1	MA	LDA	>10	24285	>90	>90	[137]
Multiple Cancers	>10	MA	LDA	>10	87701	>70	90	This work
		MMA	LDA	3		>70	>90	
			ANN	4		>80	>80	

[a] PT operators used, MA = Moving Average, MMA = Multi-condition Moving Average.[b]ML method used and NV = Number of input variables, n.a. = not available to authors of this work. [c]Total number of cases or in series. [d]Approximate values for training and validation series.

In fact, the data set used includes 45833for assays of compounds for the treatment of leukemia. It also includes the results of 2499assays of compounds against ovarian cancer, 6227assays for breast cancer, 3499assays for colon cancer, 3159 assays for lung cancer, 2750assays for prostate cancer, 601melanoma, 492assays for stomach cancer, and 134assays for liver cancer, *etc*. (see file SI03.xls). In addition, in all the previous cases the number of input variables is notably higher. This could be the use here of MMA operators instead of simple MA operators. Remember that MMA operators can encode multiple experimental conditions at the same time instead only one condition for each MA. The present seems to be the first IFPTML model able to predict anticancer compounds against different types of cancer.

3.8. CHAPTER 03. CONCLUSIONS.

In this research, we showed that IFPTML techniques are useful to model complex ChEMBLdatasets of anti-cancer compounds with BD characteristics (huge volume, variability, low veracity, complexity, *etc*.).The PTML-LDA models presented here are the first able to predict anticancer activity of compounds for different types of cancers.For the present dataset, PTML-LDA models with MMA operators including multiple assay conditions at the same time are more efficient resulting in simpler models. On the other hand, models with MA operators for one condition at time are also useful but the model need a higher number of variables than MMA models. In addition, non-linear IFPTML models based on ANN algorithms presented more balanced values of Sp and Sn than LDA models.

3.9. CHAPTER 03. MATERIALS AND METHODS.

ChEMBL Data pre-processing. We obtained the outcomes of many preclinical assays from ChEMBL. The result of each assay is expressed by one experimental parameter ε_{ij}used to quantify the biological activity of the i^{th} molecule (m_i) over the j-th target. The valuesofε_{ij} depends on the structure of the drug and also on a series of boundary conditions that delimit the characteristics of the assay$\mathbf{c}_j$= (c_0, c_1, c_2, …c_n). The first $c_j$is c_0 = the biological activity ε_{ij}(IC_{50}, EC_{50}, *etc*.)*per se*. Other conditions are c_1 = target protein, c_2 = organism of assay, *etc*. The values ε_{ij}compiled are not exact numbers in many cases. That is why we used classification techniques instead of regression methods. In so doing, we discretized the values as follow: $f(v_{ij})_{obs}$ = 1 when v_{ij}>cutoffand desirability of the biological activity parameter $d(c_0)$ = 1 (see **Table 1**). The value is also $f(v_{ij})_{obs}$= 1 when v_{ij}< cutoff and desirability $d(c_0)$ = -1, $f(v_{ij})_{obs}$ = 0 otherwise. The value $f(v_{ij})_{obs}$ = 1 points to and strong effect of the compound over the target. The desirability $d(c_0)$ = 1 or -1 indicates that the parameter measured increases or decreases directly with a desired or not desired biological effect.

PTML linear model. PTML modeling technique is useful to seek predictive models for complex datasets with multiple BD features [123, 124]. We can predict scoring function values $f(v_{ij})_{calc}$ for the i-th compound in the j-th preclinical assay with multiple conditions of assay $\mathbf{c}_j$=

(c_0, c_1, c_2, …c_n). PT operators similar to Box-Jenkins MA operators are used as input[82, 90]. IFPTML linear models have the following form.

$$f(\mathrm{v_{ij}})_{calc} = \mathrm{a_0} + \mathrm{a_1} \cdot f(\mathrm{v_{ij}})_{expt} + \sum_{k=1,j=0}^{k_{max},j_{max}} \mathrm{a}_{kj} \cdot \Delta \mathrm{D_k}(\mathrm{c_j}) \quad (3)$$

3.10. CHAPTER 03. AUTHORS CONTRIBUTIONS.

Conceived and designed the experiments: IESD, HGD. Performed the experiments: IESD, HGD. Analysed the data: B.H. (Thesis Author), IESD, XGM, JERB, HGD. Wrote the paper: B.H. (Thesis Author), IESD, XGM, HGD. All authors have given approval to the final version of the paper manuscript. The authors declare no competing financial interest.

3.11. CHAPTER 04. REFERENCES.

1. Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P., Comparability of mixed IC50 data–a statistical analysis. *PloS one* **2013**, 8, e61007.

2. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P., Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. *Environmental health perspectives* **2003**, 111, 1361.

3. 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, 1203-1213.

4. Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P., ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res* **2015**, 43, W612-20.

5. Pundir, S.; Martin, M. J.; O'Donovan, C.; UniProt, C., UniProt Tools. *Curr Protoc Bioinformatics* **2016**, 53, 1 29 1-15.

6. Coordinators, N. R. J. N. a. r., Database resources of the national center for biotechnology information. **2017**, 45, D12.

7. Ferreira da Costa, J.; Silva, D.; Caamano, O.; Brea, J. M.; Loza, M. I.; Munteanu, C. R.; Pazos, A.; Garcia-Mera, X.; Gonzalez-Diaz, 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 chemical neuroscience* **2018**, 9, 2572-2587.

8. Blazquez-Barbadillo, C.; Aranzamendi, E.; Coya, E.; Lete, E.; Sotomayor, N.; Gonzalez-Diaz, H., Perturbation theory model of reactivity and enantioselectivity of palladium-catalyzed Heck-Heck cascade reactions. *Rsc Advances* **2016**, 6, 38602-38610.

9. Casanola-Martin, G. M.; Le-Thi-Thu, H.; Perez-Gimenez, F.; Marrero-Ponce, Y.; Merino-Sanjuan, M.; Abad, C.; Gonzalez-Diaz, H., Multi-output Model with Box-Jenkins Operators of Quadratic Indices for Prediction of Malaria and Cancer Inhibitors Targeting Ubiquitin-Proteasome Pathway (UPP) Proteins. *Current Protein & Peptide Science* **2016**, 17, 220-227.

10. Romero-Duran, F. J.; Alonso, N.; Yanez, M.; Caamano, O.; Garcia-Mera, X.; Gonzalez-Diaz, H., Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives. *Neuropharmacology* **2016**, 103, 270-278.

11. Kleandrova, V. V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. 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 Science & Technology* **2014**, 48, 14686-14694.

12. Luan, F.; Kleandrova, V. V.; Gonzalez-Diaz, 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* **2014**, 6, 10623-30.

13. Alonso, N.; Caamano, O.; Romero-Duran, F. J.; Luan, F.; Cordeiro, M. N. D. S.; Yanez, M.; Gonzalez-Diaz, H.; Garcia-Mera, X., Model for High-Throughput Screening of Multitarget Drugs in Chemical Neurosciences: Synthesis, Assay, and Theoretic Study of Rasagiline Carbamates. *Acs Chemical Neuroscience* **2013**, 4, 1393-1403.

14. Ambure, P.; Halder, A. K.; Gonzalez Diaz, H.; Cordeiro, M., QSAR-Co: An Open Source Software for Developing Robust Multitasking or Multitarget Classification-Based QSAR Models. *Journal of chemical information and modeling* **2019**, 59, 2538-2544.

15. Bernabe Ortega-Tenezaca, V. Q.-T., Humbert González-Díaz, FRAMA 1.0: Framework for Moving Average Operators Calculation in Data Analysis. *In MOL2NET, International Conference Series on Multidisciplinary Sciences* **2017**, 3.

16. Bediaga, H.; Arrasate, S.; Gonzalez-Diaz, H., PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS combinatorial science* **2018**, 20, 621-632.

17. Nocedo-Mena, D.; Cornelio, C.; Camacho-Corona, M. D. R.; Garza-Gonzalez, E.; Waksman de Torres, N.; Arrasate, S.; Sotomayor, N.; Lete, E.; Gonzalez-Diaz, H., Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks. *Journal of chemical information and modeling* **2019**, 59, 1109-1120.

18. Vasquez-Dominguez, E.; Armijos-Jaramillo, V. D.; Tejera, E.; Gonzalez-Diaz, H., Multioutput Perturbation-Theory Machine Learning (PTML) Model of ChEMBL Data for Antiretroviral Compounds. *Molecular pharmaceutics* **2019**.

19. Levy, V.; Grant, R. M., Antiretroviral Therapy for Hepatitis B Virus-HIV-Coinfected Patients: Promises and Pitfalls. *Clinical Infectious Diseases* **2006**, 43, 904-910.

20. Benhamou, Y., Antiretroviral therapy and HIV/hepatitis B virus coinfection. *Clinical infectious diseases* **2004**, 38, S98-S103.

21. Yang, R.; Gui, X.; Xiong, Y.; Gao, S.-c.; Yan, Y., Impact of hepatitis B virus infection on HIV response to antiretroviral therapy in a Chinese antiretroviral therapy center. *International Journal of Infectious Diseases* **2014**, 28, 29-34.

22. Quevedo-Tumailli, V. F.; Ortega-Tenezaca, B.; Gonzalez-Diaz, H., Chromosome Gene Orientation Inversion Networks (GOINs) of Plasmodium Proteome. *Journal of proteome research* **2018**, 17, 1258-1268.

23. Martinez-Arzate, S. G.; Tenorio-Borroto, E.; Barbabosa Pliego, A.; Diaz-Albiter, H. M.; Vazquez-Chagoyan, J. C.; Gonzalez-Diaz, H., PTML Model for Proteome Mining of B-Cell Epitopes and Theoretical-Experimental Study of Bm86 Protein Sequences from Colima, Mexico. *Journal of proteome research* **2017**, 16, 4093-4103.

24. Concu, R.; MN, D. S. C.; Munteanu, C. R.; Gonzalez-Diaz, H., PTML Model of Enzyme Subclasses for Mining the Proteome of Biofuel Producing Microorganisms. *Journal of proteome research* **2019**, 18, 2735-2746.

25. Blay, V.; Yokoi, T.; Gonzalez-Diaz, H., Perturbation Theory-Machine Learning Study of Zeolite Materials Desilication. *Journal of chemical information and modeling* **2018**, 58, 2414-2419.

26. Organization for Economic Co-operation and Development (OECD). Guidance document on the validation of (quantitative) structure-activity relationship ((Q)SAR) models. OECD Series on Testing and Assessment. *69. OECD Document ENV/JM/MONO* **2007**, 55-65.

27. Speck-Planche, A.; Cordeiro, M. N., Simultaneous modeling of antimycobacterial activities and ADMET profiles: a chemoinformatic approach to medicinal chemistry. *Curr Top Med Chem* **2013**, 13, 1656-65.

28. Speck-Planche, A.; Cordeiro, M. N., Chemoinformatics for medicinal chemistry: in silico model to enable the discovery of potent and safer anti-cocci agents. *Future medicinal chemistry* **2014**, 6, 2013-28.

29. Speck-Planche, A.; Cordeiro, M. N. D. S., De novo computational design of compounds virtually displaying potent antibacterial activity and desirable in vitro ADMET profiles. *Med. Chem. Res.* **2017**, 26, 2345-2356.

30. Speck-Planche, A.; Kleandrova, V. V.; Ruso, J. M.; Cordeiro, M. N., First Multitarget Chemo-Bioinformatic Model To Enable the Discovery of Antibacterial Peptides against Multiple Gram-Positive Pathogens. *Journal of chemical information and modeling* **2016**, 56, 588-98.

31. Speck-Planche, A.; Cordeiro, M., Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol. Divers.* **2017**.

32. Kennard, R. W.; Stone, L. A., Computer aided design of experiments. *Technometrics* **1969**, 11, 137-148.

33. Venkatasubramanian, V.; Sundaram, A., Genetic algorithms: introduction and applications. *Encyclopedia of Computational Chemistry* **2002**, 2.

34. Rogers, D.; Hopfinger, A. J., Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *Journal of Chemical Information and Computer Sciences* **1994**, 34, 854-866.

35. Hemmateenejad, B.; Akhond, M.; Miri, R.; Shamsipur, M., Genetic algorithm applied to the selection of factors in principal component-artificial neural networks: application to

QSAR study of calcium channel antagonist activity of 1, 4-dihydropyridines (nifedipine analogous). *Journal of Chemical Information and Computer Sciences* **2003**, 43, 1328-1334.

36. Hasegawa, K.; Miyashita, Y.; Funatsu, K., GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *Journal of Chemical Information and Computer Sciences* **1997**, 37, 306-310.

37. Ambure, P.; Roy, K., Understanding the structural requirements of cyclic sulfone hydroxyethylamines as hBACE1 inhibitors against Aβ plaques in Alzheimer's disease: a predictive QSAR approach. *RSC Advances* **2016**, 6, 28171-28186.

38. Gramatica, P.; Chirico, N.; Papa, E.; Cassani, S.; Kovarich, S., QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. *Journal of Computational Chemistry* **2013**, 34, 2121-2132.

39. Gao, H., Application of BCUT metrics and genetic algorithm in binary QSAR analysis. *Journal of Chemical Information and Computer Sciences* **2001**, 41, 402-407.

40. Sutherland, J. J.; O'brien, L. A.; Weaver, D. F., Spline-fitting with a genetic algorithm: A method for developing classification structure− activity relationships. *Journal of Chemical Information and Computer Sciences* **2003**, 43, 1906-1915.

41. Snedecor, G.; Cochran, W., Statistical Methods Oxford and IBH publishing co. *New Delhi* **1967**, 593.

42. Breiman, L., Random forests. *Machine learning* **2001**, 45, 5-32.

43. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H., The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* **2009**, 11, 10-18.

44. Wilks, S. S., Certain generalizations in the analysis of variance. *Biometrika* **1932**, 471-494.

45. Fawcett, T., An introduction to ROC analysis. *Pattern recognition letters* **2006**, 27, 861-874.

46. Fisher, R. A., *The design of experiments*. Oliver And Boyd; Edinburgh; London: 1937.

47. Roy, K.; Kar, S.; Ambure, P., On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems* **2015**, 145, 22-29.

48. Hill, T.; Lewicki, P., *STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining*. StatSoft: Tulsa, 2006 Vol. 1, p 813.

49. Nussinov, R.; Tsai, C. J., Allostery in disease and in drug discovery. *Cell* **2013**, 153, 293-305.

50. Csermely, P.; Nussinov, R.; Szilagyi, A., From allosteric drugs to allo-network drugs: state of the art and trends of design, synthesis and computational methods. *Curr Top Med Chem* **2013**, 13, 2-4.

51. Szilagyi, A.; Nussinov, R.; Csermely, P., Allo-network drugs: extension of the allosteric drug concept to protein- protein interaction and signaling networks. *Curr Top Med Chem* **2013**, 13, 64-77.

52. Csermely, P.; Korcsmaros, T.; Kiss, H. J.; London, G.; Nussinov, R., Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.* **2013**, 138, 333-408.

53. Taylor, P. J., *Comprehensive Medicinal Chemistry*. 1990 ed.; Pergamon Press: Oxford, 1990; Vol. 4, p 241-294.

54. Wang, J.; Yun, D.; Yao, J.; Fu, W.; Huang, F.; Chen, L.; Wei, T.; Yu, C.; Xu, H.; Zhou, X.; Huang, Y.; Wu, J.; Qiu, P.; Li, W., Design, synthesis and QSAR study of novel isatin analogues inspired Michael acceptor as potential anticancer compounds. *Eur. J. Med. Chem.* **2018**, 144, 493-503.

55. Pogorzelska, A.; Slawinski, J.; Zolnowska, B.; Szafranski, K.; Kawiak, A.; Chojnacki, J.; Ulenberg, S.; Zielinska, J.; Baczek, T., Novel 2-(2-alkylthiobenzenesulfonyl)-3-(phenylprop-2-ynylideneamino)guanidine derivatives as potent anticancer agents - Synthesis, molecular structure, QSAR studies and metabolic stability. *Eur. J. Med. Chem.* **2017**, 138, 357-370.

56. Slawinski, J.; Szafranski, K.; Pogorzelska, A.; Zolnowska, B.; Kawiak, A.; Macur, K.; Belka, M.; Baczek, T., Novel 2-benzylthio-5-(1,3,4-oxadiazol-2-yl)benzenesulfonamides with anticancer activity: Synthesis, QSAR study, and metabolic stability. *Eur. J. Med. Chem.* **2017**, 132, 236-248.

57. Murahari, M.; Kharkar, P. S.; Lonikar, N.; Mayur, Y. C., Design, synthesis, biological evaluation, molecular docking and QSAR studies of 2,4-dimethylacridones as anticancer agents. *Eur. J. Med. Chem.* **2017**, 130, 154-170.

58. Ruddarraju, R. R.; Murugulla, A. C.; Kotla, R.; Chandra Babu Tirumalasetty, M.; Wudayagiri, R.; Donthabakthuni, S.; Maroju, R.; Baburao, K.; Parasa, L. S., Design, synthesis, anticancer, antimicrobial activities and molecular docking studies of theophylline containing acetylenes and theophylline containing 1,2,3-triazoles with variant nucleoside derivatives. *Eur. J. Med. Chem.* **2016**, 123, 379-396.

59. Singh, H.; Kumar, R.; Singh, S.; Chaudhary, K.; Gautam, A.; Raghava, G. P., Prediction of anticancer molecules using hybrid model developed on molecules screened against NCI-60 cancer cell lines. *BMC Cancer* **2016**, 16, 77.

60. Pingaew, R.; Prachayasittikul, V.; Worachartcheewan, A.; Nantasenamat, C.; Prachayasittikul, S.; Ruchirawat, S.; Prachayasittikul, V., Novel 1,4-naphthoquinone-

based sulfonamides: Synthesis, QSAR, anticancer and antimalarial studies. *Eur. J. Med. Chem.* **2015**, 103, 446-59.

61. Anwer, Z.; Gupta, S. P., A QSAR study on some series of anticancer tyrosine kinase inhibitors. *Medicinal chemistry (Shariqah, United Arab Emirates)* **2013**, 9, 203-12.

62. Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J., SMILES-based QSAR approaches for carcinogenicity and anticancer activity: comparison of correlation weights for identical SMILES attributes. *Anticancer Agents Med Chem* **2011**, 11, 974-82.

63. Huang, X. Y.; Shan, Z. J.; Zhai, H. L.; Li, L. N.; Zhang, X. Y., Molecular design of anticancer drug leads based on three-dimensional quantitative structure-activity relationship. *Journal of chemical information and modeling* **2011**, 51, 1999-2006.

64. Benfenati, E.; Toropov, A. A.; Toropova, A. P.; Manganaro, A.; Gonella Diaza, R., coral Software: QSAR for Anticancer Agents. *Chemical biology & drug design* **2011**, 77, 471-6.

65. Gonzalez-Diaz, H.; Bonet, I.; Teran, C.; De Clercq, E.; Bello, R.; Garcia, M. M.; Santana, L.; Uriarte, E., ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur. J. Med. Chem.* **2007**, 42, 580-5.

66. Gonzalez-Diaz, H.; Vina, D.; Santana, L.; de Clercq, E.; Uriarte, E., Stochastic entropy QSAR for the in silico discovery of anticancer compounds: prediction, synthesis, and in vitro assay of new purine carbanucleosides. *Bioorg. Med. Chem.* **2006**, 14, 1095-107.

67. Gonzales-Diaz, H.; Gia, O.; Uriarte, E.; Hernadez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A., Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design I: discovery of anticancer compounds. *J Mol Model* **2003**, 9, 395-407.

68. Jung, M.; Kim, H.; Kim, M., Chemical genomics strategy for the discovery of new anticancer agents. *Curr. Med. Chem.* **2003**, 10, 757-62.

69. Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N., Mining the NCI anticancer drug discovery databases: genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 189-99.

70. Han, L.; Cui, J.; Lin, H.; Ji, Z.; Cao, Z.; Li, Y.; Chen, Y., Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. *Proteomics* **2006**, 6, 4023-37.

71. Chou, K. C.; Shen, H. B., Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS ONE* **2010**, 5, e11335.

72. Chou, K. C.; Shen, H. B., Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nature protocols* **2008**, 3, 153-62.

73. Cai, Y. D.; Chou, K. C., Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. *Journal of proteome research* **2005**, 4, 967-71.

74. Shen, H. B.; Chou, K. C., QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. *Journal of proteome research* **2009**, 8, 1577-84.

75. Chou, K. C.; Shen, H. B., Large-scale predictions of gram-negative bacterial protein subcellular locations. *Journal of proteome research* **2006**, 5, 3420-8.

76. Rodriguez-Soca, Y.; Munteanu, C. R.; Dorado, J.; Pazos, A.; Prado-Prado, F. J.; Gonzalez-Diaz, H., Trypano-PPI: a web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions. *Journal of proteome research* **2010**, 9, 1182-90.

77. Munteanu, C. R.; Vazquez, J. M.; Dorado, J.; Sierra, A. P.; Sanchez-Gonzalez, A.; Prado-Prado, F. J.; Gonzalez-Diaz, H., Complex network spectral moments for ATCUN motif DNA cleavage: first predictive study on proteins of human pathogen parasites. *Journal of proteome research* **2009**, 8, 5219-28.

78. Gonzalez-Diaz, H.; Saiz-Urra, L.; Molina, R.; Santana, L.; Uriarte, E., A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions. *Journal of proteome research* **2007**, 6, 904-8.

79. Gonzalez-Diaz, H.; Prado-Prado, F.; Garcia-Mera, X.; Alonso, N.; Abeijon, P.; Caamano, O.; Yanez, M.; Munteanu, C. R.; Pazos, A.; Dea-Ayuela, M. A.; Gomez-Munoz, M. T.; Garijo, M. M.; Sansano, J.; Ubeira, F. M., MIND-BEST: Web server for drugs and target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretical-experimental study of G3PDH protein from Trichomonas gallinae. *Journal of proteome research* **2011**, 10, 1698-718.

80. Concu, R.; Dea-Ayuela, M. A.; Perez-Montoto, L. G.; Bolas-Fernandez, F.; Prado-Prado, F. J.; Podda, G.; Uriarte, E.; Ubeira, F. M.; Gonzalez-Diaz, H., Prediction of enzyme classes from 3D structure: a general model and examples of experimental-theoretic scoring of peptide mass fingerprints of Leishmania proteins. *Journal of proteome research* **2009**, 8, 4372-82.

81. Aguero-Chapin, G.; Varona-Santos, J.; de la Riva, G. A.; Antunes, A.; Gonzalez-Vlla, T.; Uriarte, E.; Gonzalez-Diaz, H., Alignment-free prediction of polygalacturonases with pseudofolding topological indices: experimental isolation from Coffea arabica and prediction of a new sequence. *Journal of proteome research* **2009**, 8, 2122-8.

82. Martinez, S. G.; Tenorio-Borroto, E.; Barbabosa Pliego, A.; Diaz-Albiter, H.; Vazquez-Chagoyan, J. C.; Gonzalez-Diaz, H., PTML Model for Proteome Mining of B-cell Epitopes and Theoretic-Experimental Study of Bm86 Protein Sequences from Colima Mexico. *Journal of proteome research* **2017**.

83. Fernández, M.; Caballero, F.; Fernández, L.; Abreu, J. I.; Acosta, G., Classification of conformational stability of protein mutants from 3D pseudo-folding graph representation of protein sequences using support vector machines. *Proteins* **2008**, 70, 167-175.

84. Fernández, L.; Caballero, J.; Abreu, J. I.; Fernández, M., Amino Acid Sequence Autocorrelation Vectors and Bayesian-Regularized Genetic Neural Networks for Modeling Protein Conformational Stability: Gene V Protein Mutants. *Proteins* **2007**, 67, 834–852.

85. Caballero, J.; Fernandez, L.; Abreu, J. I.; Fernandez, M., Amino Acid Sequence Autocorrelation vectors and ensembles of Bayesian-Regularized Genetic Neural Networks for prediction of conformational stability of human lysozyme mutants. *Journal of chemical information and modeling* **2006**, 46, 1255-68.

86. Perez-Riverol, Y.; Audain, E.; Millan, A.; Ramos, Y.; Sanchez, A.; Vizcaino, J. A.; Wang, R.; Muller, M.; Machado, Y. J.; Betancourt, L. H.; Gonzalez, L. J.; Padron, G.; Besada, V., Isoelectric point optimization using peptide descriptors and support vector machines. *Journal of proteomics* **2012**, 75, 2269-74.

87. Greener, J. G.; Sternberg, M. J., AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC bioinformatics* **2015**, 16, 335.

88. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magarinos, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R., The ChEMBL database in 2017. *Nucleic Acids Res* **2017**, 45, D945-D954.

89. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **2012**, 40, D1100-7.

90. Gonzalez-Diaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M., General Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry. *Current Topics in Medicinal Chemistry* **2013**, 13, 1713-1741.

91. Speck-Planche, A.; Cordeiro, M., Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol. Divers.* **2017**, 21, 511-523.

92. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N., Unified multi-target approach for the rational in silico design of anti-bladder cancer agents. *Anticancer Agents Med Chem* **2013**, 13, 791-800.

93. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N., Chemoinformatics in multi-target drug discovery for anti-cancer therapy: in silico design of potent and versatile anti-brain tumor agents. *Anticancer Agents Med Chem* **2012**, 12, 678-85.

94. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N., Rational drug design for anti-cancer chemotherapy: multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. *Bioorg. Med. Chem.* **2012**, 20, 4848-55.

95. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N., Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents. *Eur. J. Pharm. Sci.* **2012**, 47, 273-9.

96. Cordeiro, M. N.; Speck-Planche, A., Computer-aided drug design, synthesis and evaluation of new anti-cancer drugs. *Curr Top Med Chem* **2012**, 12, 2703-4.

97. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N., Multi-target drug discovery in anti-cancer therapy: fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorg. Med. Chem.* **2011**, 19, 6239-44.

98. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N., Multi-target drug discovery in anti-cancer therapy: fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorg. Med. Chem.* **2011**, 19, 6239-44.

99. Collier, G.; Ortiz, V., Emerging computational approaches for the study of protein allostery. *Archives of biochemistry and biophysics* **2013**, 538, 6-15.

100. 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 Systems with Applications* **2017**, 72, 306-316.

101. Gonzalez-Diaz, H.; Herrera-Ibata, D. M.; Duardo-Sanchez, 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. *Journal of chemical information and modeling* **2014**, 54, 744-55.

102. Casanola-Martin, G. M.; Le-Thi-Thu, H.; Perez-Gimenez, F.; Marrero-Ponce, Y.; Merino-Sanjuan, M.; Abad, C.; Gonzalez-Diaz, H., Multi-output model with Box-Jenkins operators of linear indices to predict multi-target inhibitors of ubiquitin-proteasome pathway. *Mol. Divers.* **2015**, 19, 347-56.

103. Young, R. C.; Mitchell, R. C.; Brown, T. H.; Ganellin, C. R.; Griffiths, R.; Jones, M.; Rana, K. K.; Saunders, D.; Smith, I. R.; Sore, N. E.; et al., Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists. *Journal of medicinal chemistry* **1988**, 31, 656-71.

104. Friden, M.; Winiwarter, S.; Jerndal, G.; Bengtsson, O.; Wan, H.; Bredberg, U.; Hammarlund-Udenaes, M.; Antonsson, M., Structure-brain exposure relationships in rat

and human using a novel data set of unbound drug concentrations in brain interstitial and cerebrospinal fluids. *Journal of medicinal chemistry* **2009**, 52, 6233-43.

105. Hitchcock, S. A.; Pennington, L. D., Structure-brain exposure relationships. *Journal of medicinal chemistry* **2006**, 49, 7559-83.

106. Garcia, I.; Fall, Y.; Garcia-Mera, X.; Prado-Prado, F., Theoretical study of GSK-3alpha: neural networks QSAR studies for the design of new inhibitors using 2D descriptors. *Mol. Divers.* **2011**, 15, 947-55.

107. Marrero-Ponce, Y.; Siverio-Mota, D.; Galvez-Llompart, M.; Recio, M. C.; Giner, R. M.; Garcia-Domenech, R.; Torrens, F.; Aran, V. J.; Cordero-Maldonado, M. L.; Esguera, C. V.; de Witte, P. A.; Crawford, A. D., 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, 5736-53.

108. Kuligowski, J.; Perez-Guaita, D.; Escobar, J.; de la Guardia, M.; Vento, M.; Ferrer, A.; Quintas, G., Evaluation of the effect of chance correlations on variable selection using Partial Least Squares-Discriminant Analysis. *Talanta* **2013**, 116, 835-40.

109. Bhagwanth, S.; Mishra, S.; Daya, R.; Mah, J.; Mishra, R. K.; Johnson, R. L., Transformation of Pro-Leu-Gly-NH2 peptidomimetic positive allosteric modulators of the dopamine D2 receptor into negative modulators. *ACS chemical neuroscience* **2012**, 3, 274-84.

110. Celis, M. E.; Taleisnik, S.; Walter, R., Regulation of formation and proposed structure of the factor inhibiting the release of melanocyte-stimulating hormone. *Proceedings of the National Academy of Sciences of the United States of America* **1971**, 68, 1428-33.

111. Katzenschlager, R.; Jackson, M. J.; Rose, S.; Stockwell, K.; Tayarani-Binazir, K. A.; Zubair, M.; Smith, L. A.; Jenner, P.; Lees, A. J., Antiparkinsonian activity of L-propyl-

L-leucyl-glycinamide or melanocyte-inhibiting factor in MPTP-treated common marmosets. *Mov. Disord.* **2007**, 22, 715-9.

112. Palomo, C.; Aizpurua, J. M.; Benito, A.; Miranda, J. I.; Fratila, R. M.; Matute, C.; Domercq, M.; Gago, F.; Martin-Santamaria, S.; Linden, A., Development of a new family of conformationally restricted peptides as potent nucleators of beta-turns. Design, synthesis, structure, and biological evaluation of a beta-lactam peptide analogue of melanostatin. *J. Am. Chem. Soc.* **2003**, 125, 16243-60.

113. Sampaio-Dias, I. E.; Silva-Reis, S. C.; Garcia-Mera, X.; Brea, J.; Loza, M. I.; Alves, C. S.; Algarra, M.; Rodriguez-Borges, J. E., Synthesis, Pharmacological, and Biological Evaluation of MIF-1 Picolinoyl Peptidomimetics as Positive Allosteric Modulators of D2R. *ACS chemical neuroscience* **2019**, 10, 3690-3702.

114. Ferreira da Costa, J.; Caamaño, O.; Fernández, F.; García-Mera, X.; Sampaio-Dias, I. E.; Brea, J. M.; Cadavid, M. I., Synthesis and allosteric modulation of the dopamine receptor by peptide analogs of l-prolyl-l-leucyl-glycinamide (PLG) modified in the l-proline or l-proline and l-leucine scaffolds. *European Journal of Medicinal Chemistry* **2013**, 69, 146-158.

115. Sampaio-Dias, I. E. S., C. A. D.; García-Mera, X.; da Costa, J. F.; Caamaño, O.; Rodríguez-Borges, J. E., *Org. Biomol. Chem.* **2016**, 14, 11065-11069.

116. Sampaio-Dias, I. E.; Sousa, C. A. D.; Silva-Reis, S. C.; Ribeiro, S.; Garcia-Mera, X.; Rodriguez-Borges, J. E., Highly efficient one-pot assembly of peptides by double chemoselective coupling. *Organic & Biomolecular Chemistry* **2017**, 15, 7533-7542.

117. Liang, X.; Haynes, B. S.; Montoya, A., Acid-Catalyzed Ring Opening of Furan in Aqueous Solution. *Energy & Fuels* **2018**, 32, 4139-4148.

118. Verma, V.; Mann, A.; Costain, W.; Pontoriero, G.; Castellano, J. M.; Skoblenick, K.; Gupta, S. K.; Pristupa, Z.; Niznik, H. B.; Johnson, R. L.; Nair, V. D.; Mishra, R. K.,

Modulation of agonist binding to human dopamine receptor subtypes by L-prolyl-L-leucyl-glycinamide and a peptidomimetic analog. *The Journal of pharmacology and experimental therapeutics* **2005**, 315, 1228-36.

119. Christopoulos, A.; Kenakin, T., G protein-coupled receptor allosterism and complexing. *Pharmacological reviews* **2002**, 54, 323-74.

120. Bhagwanth, S.; Mishra, R. K.; Johnson, R. L., Development of peptidomimetic ligands of Pro-Leu-Gly-NH(2) as allosteric modulators of the dopamine D(2) receptor. *Beilstein Journal of Organic Chemistry* **2013**, 9, 204-214.

121. Vartak, A. P.; Skoblenick, K.; Thomas, N.; Mishra, R. K.; Johnson, R. L., Allosteric modulation of the dopamine receptor by conformationally constrained type VI beta-turn peptidomimetics of Pro-Leu-Gly-NH2. *J Med Chem* **2007**, 50, 6725-9.

122. Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V., Virtual computational chemistry laboratory--design and description. *Journal of computer-aided molecular design* **2005**, 19, 453-63.

123. Gonzalez-Diaz, H.; Perez-Montoto, L. G.; Ubeira, F. M., Model for vaccine design by prediction of B-epitopes of IEDB given perturbations in peptide sequence, in vivo process, experimental techniques, and source or host organisms. *Journal of immunology research* **2014**, 2014, 768515.

124. Gonzalez-Diaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M., 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, 1713-41.

125. Gottlieb, H. E.; Kotlyar, V.; Nudelman, A., NMR Chemical Shifts of Common Laboratory Solvents as Trace Impurities. *The Journal of Organic Chemistry* **1997**, 62, 7512-7515.

126. Graham, L. P., *An Introduction to Medicinal Chemistry*. Oxford Univ Pr: 2017.

127. Vilar, S.; Santana, L.; Uriarte, E., Probabilistic neural network model for the in silico evaluation of anti-HIV activity and mechanism of action. *J. Med. Chem.* **2006**, 49, 1118-1124.

128. Santana, L.; Uriarte, E.; Gonzalez-Diaz, H.; Zagotto, G.; Soto-Otero, R.; Mendez-Alvarez, E., A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins. *J. Med. Chem.* **2006**, 49, 1149-56.

129. Santana, L.; Gonzalez-Diaz, H.; Quezada, E.; Uriarte, E.; Yanez, M.; Vina, D.; Orallo, F., Quantitative structure-activity relationship and complex network approach to monoamine oxidase A and B inhibitors. *J. Med. Chem.* **2008**, 51, 6740-51.

130. Svensson, F.; Bender, A.; Bailey, D., Fragment-Based Drug Discovery of Phosphodiesterase Inhibitors. *J. Med. Chem.* **2017**.

131. Stumpfe, D.; Dimova, D.; Bajorath, J., Computational Method for the Systematic Identification of Analog Series and Key Compounds Representing Series and Their Biological Activity Profiles. *J. Med. Chem.* **2016**, 59, 7667-76.

132. Jansen, C.; Kooistra, A. J.; Kanev, G. K.; Leurs, R.; de Esch, I. J.; de Graaf, C., PDEStrIAn: A Phosphodiesterase Structure and Ligand Interaction Annotated Database As a Tool for Structure-Based Drug Design. *J. Med. Chem.* **2016**, 59, 7029-65.

133. Chittchang, M.; Gleeson, M. P.; Ploypradith, P.; Ruchirawat, S., Assessing the drug-likeness of lamellarins, a marine-derived natural product class with diverse oncological activities. *Eur. J. Med. Chem.* **2010**, 45, 2165-72.

134. Hansch, C.; Verma, R. P., A QSAR study for the cytotoxic activities of taxoids against macrophage (MPhi)-like cells. *Eur. J. Med. Chem.* **2009**, 44, 274-9.

135. Roy, K.; Pratim Roy, P., Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. *Eur. J. Med. Chem.* **2009**, 44, 2913-22.

136. Sarkar, A.; Anderson, K. C.; Kellogg, G. E., Computational analysis of structure-based interactions and ligand properties can predict efflux effects on antibiotics. *Eur. J. Med. Chem.* **2012**, 52, 98-110.

137. Speck-Planche, A.; Cordeiro, M. N., Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol. Divers.* **2017**.

Chapter 04.
LAGA Software

4. CHAPTER 04. LAGA SOFTWARE

Paper 1. Ortega-Tenezaca, B., Quevedo-Tumailli, V., Bediaga, H., Collados, J., Arrasate, S., Madariaga, G., Munteanu, C., Cordeiro, M., & González-Díaz, H. (2020). IFPTML Multi-Label Algorithms: Models, Software, and Applications. *Current Topics in Medicinal Chemistry*, *20*(25), 2326–2337. doi: https://doi.org/10.2174/1568026620666200916122616

4.1. CHAPTER 04. ABSTRACT

IFPTML models have been demonstrated in the previous chapter to be of interest for multi-output anti-cancer compounds prediction in pre-clinical assays. From the previous chapter we concluded that user-firendly software are necessary if we want that Pharmaceutical industry and medicinal chemistry experimentalists apply our IFPTML drugs discovery models. In this chapter we report the implementation of the IFPTML anti-cancer drugs model (see details in previous chapter) in the LAGA software.

Keywords

Drug Discovery, Cheminformatics, Multi-target models, Machine Learning, Phyton.

4.2. CHAPTER 04. INTRODUCTION.

Form the previous chapter and the section above we can infer that we may need to Database languages and tools to process the information more efficiently.[1,2] In fact, database language like Structured Query Language (SQL) developed by IBM in 1970, are ideal for processing the information contained in the datasets used in IFPTML modeling.[3] Databases allow large amounts of information to be stored, consulted and modified, which implies a need to develop mechanisms to manage said information. In addition, the reliability of the stored data must be guaranteed, as well as its security. To do this, information began to be stored in operating system files. These files contained irremovable information, so new files had to be created every time the stored data had to be modified. With this information management method, several problems arose.[1,4]

- Data redundancy and inconsistency due to files being created at different times by different developers.
- Difficult access to the data, since sometimes new needs arose that required new ways of accessing the data, which were not foreseen.
- Data isolation, caused by having several different files to store a group of data.
- Dependence on the correct functioning of the computer system, since if there was a drop in the electrical network while the data was being modified, the information could be lost or remain in an inconsistent state.
- Anomalies when accessing several users at the same time. If more than one user accessed the database and made a change, information could remain in a contradictory situation.
- Security problems, since not all users must be authorized to access all the information stored in the database.

Today, the SQL language contains a system of data definition and manipulation statements, ensures the integrity of the databases, a control system for the so-called transactions and an access authorization system.[3] This system allows more dynamic and secure access and management of information than its predecessor systems. The data stored in a SQL database can

be numeric or character strings; and, within the numeric variables, they can be integers or real numbers. In addition, each variable can have certain characteristics or restrictions that set the amount of memory that it can occupy, being able to restrict its length, in the case of character strings, or its number of digits and decimals, in the case of real numbers. Based on this, there are the following domain types:[3]

- char(n): Fixed-length character string n.
- varchar(n): Variable-length character string with a maximum of n characters.
- int: Integer number.
- numeric(p,d): Fixed point real number, with p digits (including sign and comma), of which d are from the decimal part.
- real (, double precision): Floating point number. If desired, it can be specified to be double precision.
- float(n): Floating point number with precision of n digits.

In addition to basic domains like the ones above, new data types can also be defined. With this method, it is possible to assign constraints to variables, in order to define unique characteristics that make their definition unambiguous. In this way, SQL databases become an almost indispensable tool for Object Oriented Programming, since each object has characteristics that can be defined by variables created by a user.SQL databases are structured tables, where the relationship between the data set is defined. Each database can be made up of one or more tables, allowing grouping relationships of data sets under a global relationship. For example, assuming a database where the telephone numbers of all the residents of a neighborhood and the telephone companies to which they belong are stored, the database would contain the following type of tables, see **Table 1**.

Table 1. Fragment of the fictitious database that contains the telephone number of the residents o

Name	1st last name	2nd last name	Address	Phone	Company
name 1	*Surname 1.1*	*Surname 1.2*	*Address 1*	*Number 1*	*company 1*
name 2	*Surname 2.1*	*Surname 2.2*	*address 2*	*number 2*	*company 2*
name 3	*Surname 3.1*	*Surname 3.2*	*address 3*	*Number 3*	*company 3*

Each row in the table is called a record. Therefore, each record will be the one that collects the relationship between the values of each variable; and the table will be the one that collects the relationship between records. It can happen that variables from more than one different record have the same value, which causes that ambiguity that SQL is intended to avoid. Following the example of the previous table, it can happen that two neighbors have the same name. To avoid this fact, there is the primary cave, which can be any type of variable, with the particularity that each value is unique, assigning one to each record (see **Table 2**).

Table 2. Fragment of the fictitious database.

primary key	Name	1st last name	2nd last name	Address	Phone	Company
key 1	*name 1*	*Surname 1.1*	*Surname 1.2*	*Address 1*	*Number 1*	*company 1*
key 2	*name 2*	*Surname 2.1*	*Surname 2.2*	*address 2*	*number 2*	*company 2*
key 3	*name 3*	*Surname 3.1*	*Surname 3.2*	*address 3*	*Number 3*	*company 3*

As already mentioned above, you can also record more information about one of the variables in a table by creating another table within the same database. Following the same example, in the database of the telephone data of the neighbors, another table can be stored where each record is the information of each company, thus obtaining a global relationship within the same database (see **Table 3**).

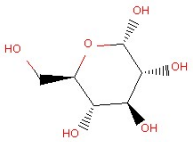
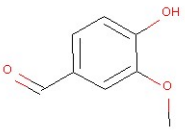
Table 3. Fragment of the table of companies of the fictitious database

primary key	Company	country of registration	Number of clients	Annual benefits (€)
key 1	*name 1*	*country 1*	*Clients 1*	*benefits 1*
key 2	*name 2*	*country 2*	*clients 2*	*benefits 2*
key 3	*name 3*	*country 3*	*clients 3*	*benefits 3*

4.3. LAGA software development.

As already mentioned, the multi-condition IFPTML model requires a program for its implementation. To do this, a tool has been developed in Python language that allows analyzing large groups of activities, for which reason this project has been named Large Activity Group Analyzer (LAGA).4 First, the user must decide if they want to enter a single molecule against several different test ones, or a list of molecules against a single condition Figure 1. For this, it has been necessary to achieve a simple format that allows the introduction of large molecules in a simple way, and that, in addition, can be saved in small files. This has been achieved using the Simplified Molecular Input Line Entry Specification (SMILES) code format, which encodes molecules in a single line of text (see **Table 4**).[5,6]

Table 4. SMILES code and structure of glucose, vanillin, and nicotine molecules.

Name	Molecular formula	*SMILES*	Structure
Glucose	$C_6H_{12}O_6$	OC[C@@H](O1)[C@@H](O)[C@@H](O)[C@@H](O)[C@@H](O)1	
Vanillin	$C_8H_8O_3$	O=Cc1ccc(O)c(OC)c1	

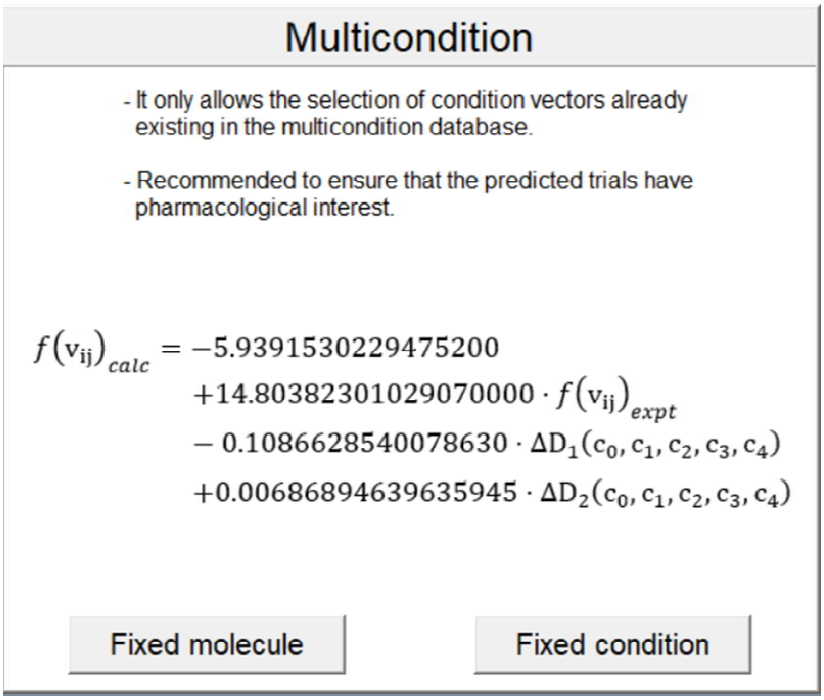


Figure 1.Model execution mode selection window.

Secondly, it is necessary to calculate the log Py -values PSAfor the input molecules. For this, the chemoinformatics library *mordred 1.2.0* has been usedwhich allows introducing *SMILES codes* to convert them into objects, interpreted by the library as molecules, and calculate their molecular descriptors, https://github.com/mordred-descriptor/mordred.[7]Third, LAGA allows the user to enter one or more condition vectors $\boldsymbol{c_j}$. As the implemented model does not allow any combination of conditions, a database has been developed in SQLite where the $\boldsymbol{c_j}$corresponding valus of$f(v_{ij})_{ref}$, $\langle \log \mathrm{P} \rangle_j$and $\langle \mathrm{PSA} \rangle_j$, which will be consulted later by the model, have been registered.[3]Once the variables (*SMILES codes* and vectors $\boldsymbol{c_j}$) have been introduced, the model is executed and the value is calculated $f(v_{ij})_{calc}$. After obtaining this value, the probability that the input molecules will be classified in the group of favorable assays $p = [f(v_{ij})_{pred} = 1]$ is calculated and presented on the screen, together with the assay information (input molecule and conditions) and the calculated values of the descriptors. Finally, the user can view the results in

more detail, to know the values $f(v_{ij})_{ref}$, $\langle \log P \rangle_j$ and $\langle PSA \rangle_j$ corresponding to the tests carried out, and save them in the form of a text file.In the**Figure 2**LAGA execution process has been represented by means of a flowchart. So that everything works correctly and the user can easily use the program, a GUI has been created with the *Tkinter*library.[8]

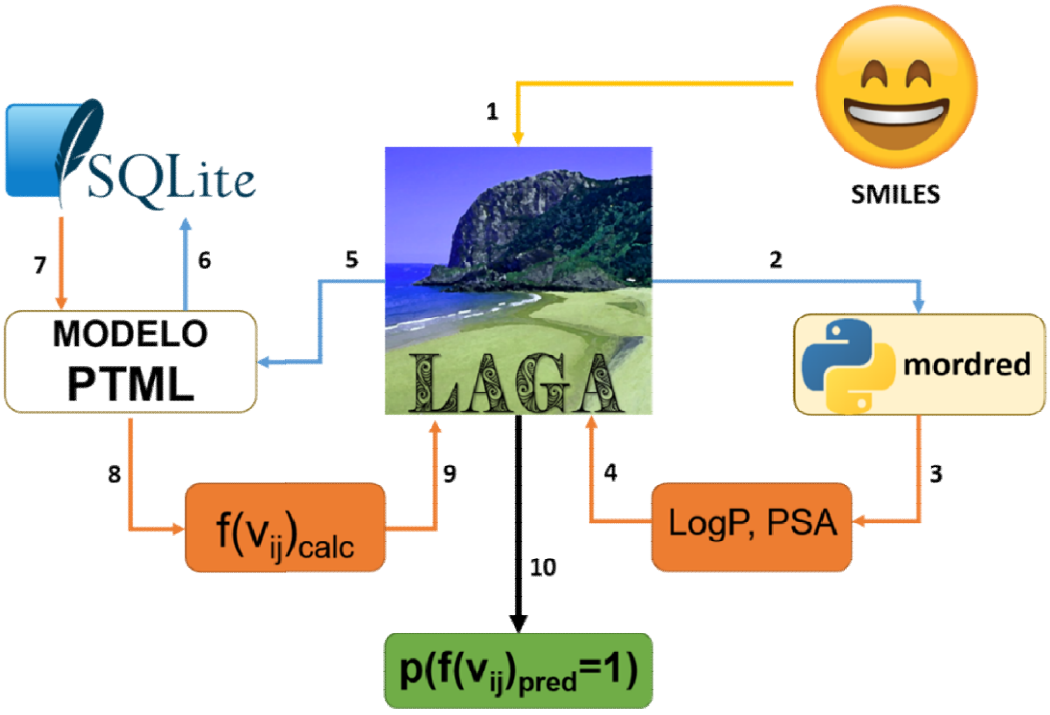


Figure 2. LAGA software flow diagram corresponding to the execution.

4.4. SMILES codess reading.

To introduce the molecules as SMILES codes, they can either be written directly in the text box or uploaded in a text file with a .txt extension (**Figure 3**). Once the molecules are entered in text format, the MolFromSmiles function is called to convert them into objects that Python interprets as molecules. This function returns an array containing the coordinates of the groups of atoms that make up the molecule. To do this, it needs to know the properties of the elements that make it up, so it uses its own database, RDKit database, and a PostgreSQL database

manager.[9]Once the text is converted to molecules, the functions of the mordred library are called for descriptor calculation.

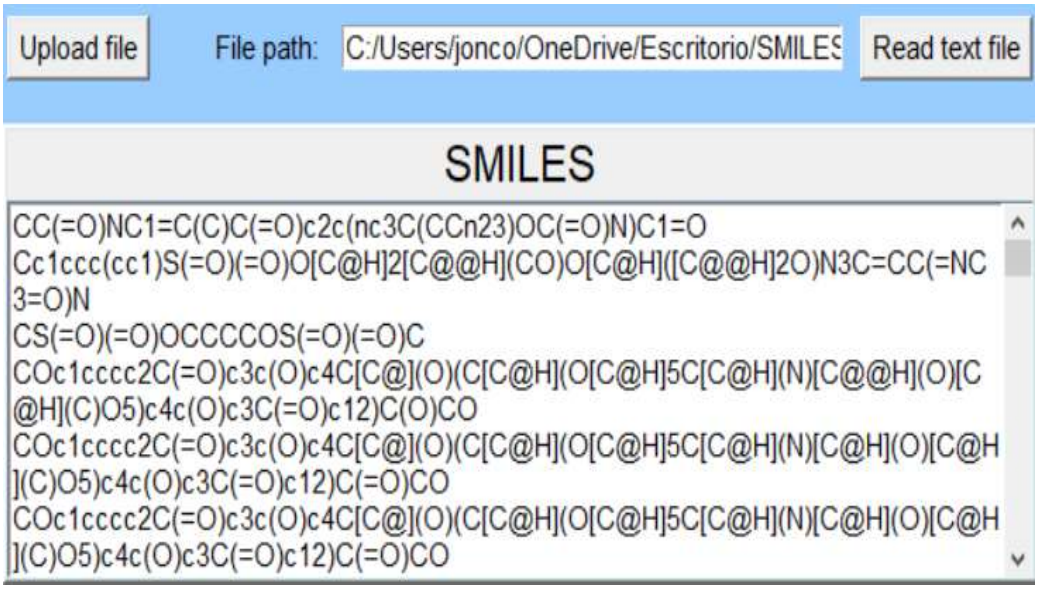


Figure 2: SMILES introduction window.

4.5. Calculationof molecular descriptors with mordredlibrary.

In order to calculate the logp and tpsa to be fed into the model, the molecular descriptor calculator from the mordred library is used. It is a library designed to calculate up to 1800 two-dimensional and three-dimensional molecular descriptors, in order to design software for the implementation of qsar models. The mordred library is made up of 42 modules, for the calculation of 42 types of molecular descriptors. Among the most important are the acidity constants calculator (acidbase), the benzenes and aromatic bonds counter (aromatic), the counter of different types of atoms (atomcount), the polarizability calculator (polarizability), the rings counter (ringcount), the calculator using the slogp method and the calculator of the tpsa.7-11

4.6. CHEMBL *vs*. Mordredlibrary molecular descriptors.

ChEMBL database for anticancer compounds. These values were calculated using computational methods, so for a correct implementation of the ifptml model, it would be

necessary to use the same algorithm. Firstly, the calculation methods for the partition coefficient are different, since chembl has used the ALOGP method and *Mordred*software uses the SLOGP method.[10]Next, **Table 5** compares the characteristics of both methods.

Table 5. Comparison between the ALogP and SLogP methods.

Method	Characteristics
ALogP	Each atom has a contribution to logPand that the final value is the sum of each contribution. A database is necessary where the contribution of each type of atom is collected. Suitable for smaller molecules and with simple aromaticity.
SLogP	Modification of the AlogP method. Consider the contribution value of each type of atom, that of its neighbors and correction factors. A database is necessary where the contribution of each type of atom is collected. Suitable for larger molecules and with more complex aromaticity.

Due to the differences between the two methods shown in **Table 8**, all the values of the partition coefficient of the molecules in the database have been calculated using the Mordred function SLogP. Then a comparison has been made between the values of ALogP and SLogP.
Second, the algorithm implemented by ChEMBL for the calculation of TPSA values is unknown. This was also a problem, as not all the values in the database matched those calculated by mordred. For this reason, a comparison has been made between the ChEMBL TPSA values and those calculated with the mordred TPSA function. Once the difference between the values of the database and the calculated ones has been demonstrated, a reparameterization of the IFPTML model has been carried out. In order to find a relationship between the methods used to calculate the ALogP values of the original database with the algorithm implemented by mordred to calculate the SLogP values, the SLogP values for the 47530 molecules have been calculated.

Then, a dispersion of points of ALogP has been plotted against SLogP, with its respective linear fit. The cutoff point with the ordinate axis has been set at ALogP=0, since both algorithms should calculate a null partition coefficient for molecules that do not contain atoms with non-zero contribution.As can be seen in **Figure 4**, the slope of the line is 0.9622. Since the slope in the case that both algorithms produce the same results is 1, this indicates that they give different results, as expected. On the other hand, the value of R^2 = 0.7781 indicates that the correlation between both groups is lower, so it is not convenient to calculate the values of SLogP and convert them to ALogP in order to implement the model.

In order to find a relationship between the methods used to calculate the ALogP values of the original database with the algorithm implemented by mordred to calculate the SLogP values, the SLogP values for the 47530 molecules have been calculated. Then, a dispersion of points of ALogP has been plotted against SLogP, with its respective linear fit. The cutoff point with the ordinate axis has been set at ALogP=0 since both algorithms should calculate a null partition coefficient for molecules that do not contain atoms with non-zero contribution. As can be seen in **Figure 4**, the slope of the line is 0.9622. Since the slope in the case that both algorithms produce the same results is 1, this indicates that they give different results, as expected. On the other hand, the value of R^2 = 0.7781 indicates that the correlation between both groups is lower, so it is not convenient to calculate the values of SLogP and convert them to ALogP in order to implement the model.

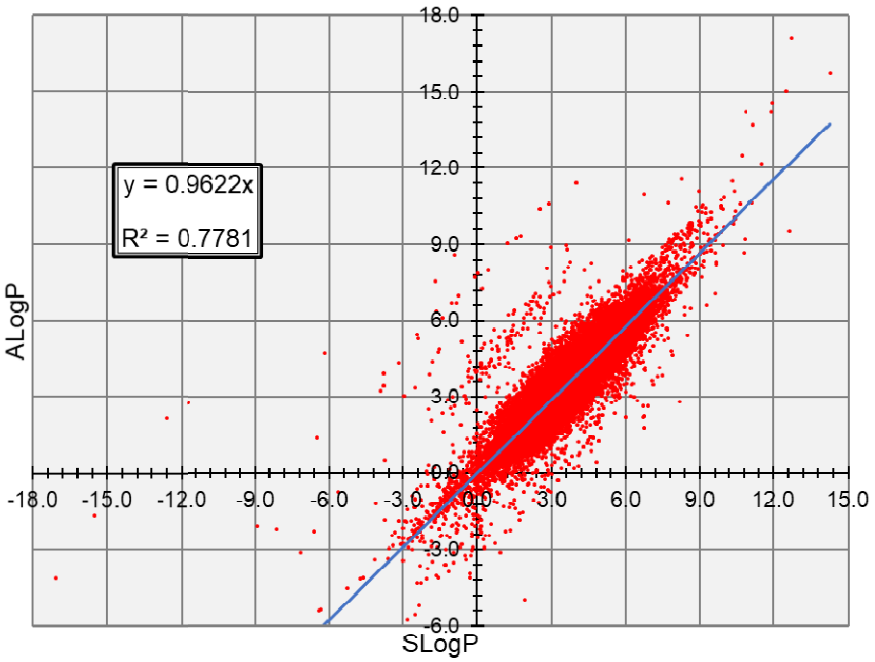


Figure 4. Representation of ALogP vs. SLogP

The **Figure 5** represents 1 out of 1000 data. In it, it can be seen how, for higher values of the distribution coefficient, the slope moves away from unity. This is because the SLogP method also considers contributions from neighboring atoms which adds to the contribution of an individual atom. A low correlation is also observed in this graph. The results obtained from both representations suggest a reparameterization of the IFPTML model.

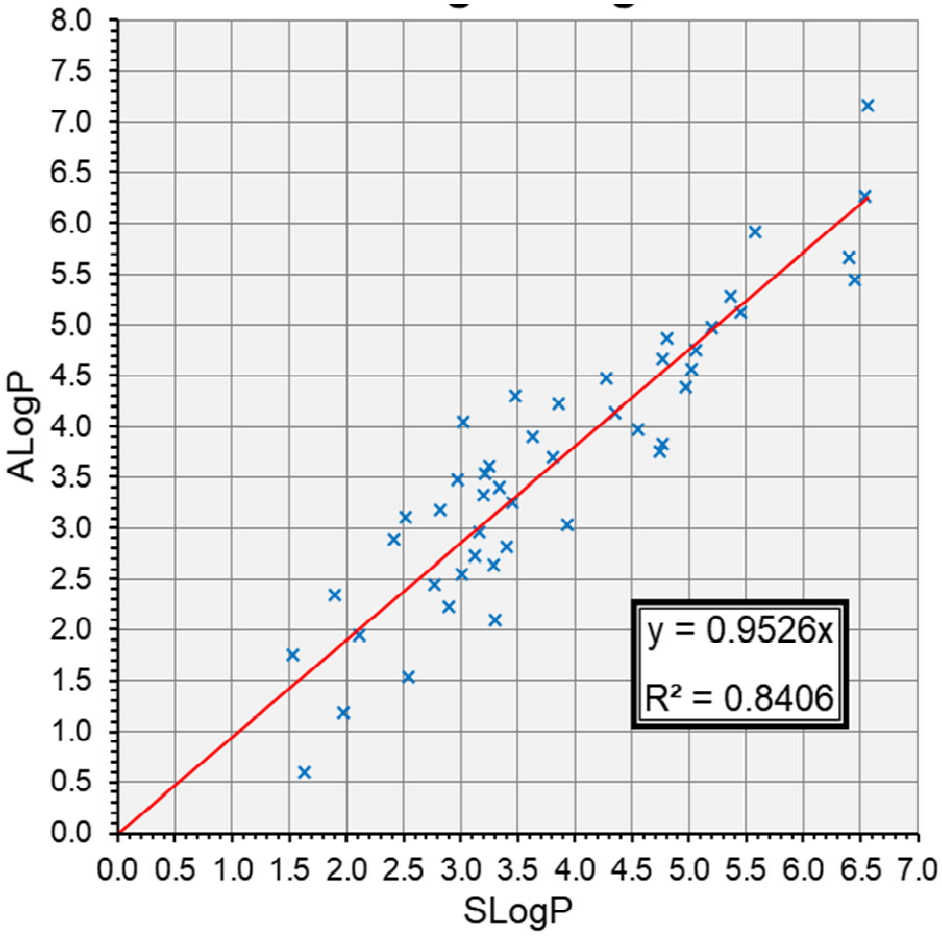


Figure 5. Plot of ALogP vs. SLogP using 1 in 1000 data.

Next, in order to find a relationship between the methods used to calculate the TPSA values from the original database with the algorithm implemented by Mordred for the same calculation, the TPSA values for the 47,530 molecules have been calculated. Next, a dispersion of TPSA points has been plotted against TPSA, with their respective linear fit. The cut-off point with the ordinate axis has been set at TPSA =0, since the minimum TPSA of a molecule is 0, regardless of the calculation method.

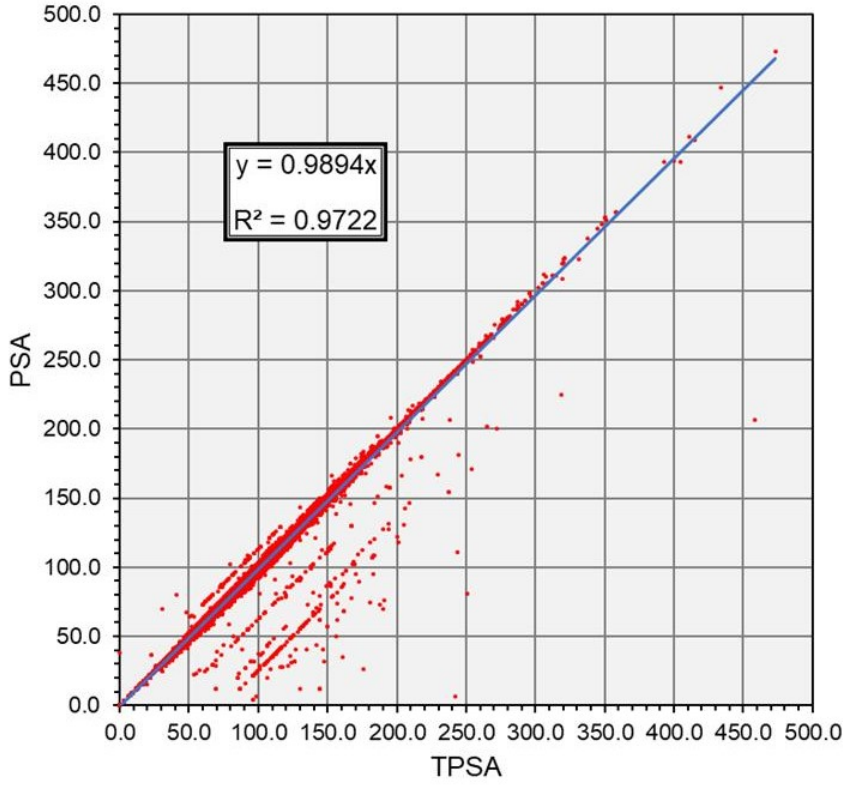


Figure 6. Representation of TPSA vs. TPSA

First, three families of points can be observed. The most numerous tends to intersect the ordinate axis at PSA=0; the other two, on the other hand, are much smaller than the previous one and tend to cross the ordinate axis in negative values. This implies that the algorithm used to calculate the TPSA values from the ChEMBL database does not consider a minimal polar surface in some groups of molecules. Second, we observe that the slope of the line is 0.9894. In this case it is closer to unity than in the case of the partition coefficient, which indicates that the TPSA calculation methods are similar. On the other hand, the value of R 2 = 0.9722 indicates that the correlation between both groups is high. The **Figure 7** represents 1 out of 1000 data. In it, a close linear relationship is observed, since the two least numerous families of points have been neglected. The value of R 2 = 0.9986 indicates that the correlation between the TPSA and TPSA values is strong.

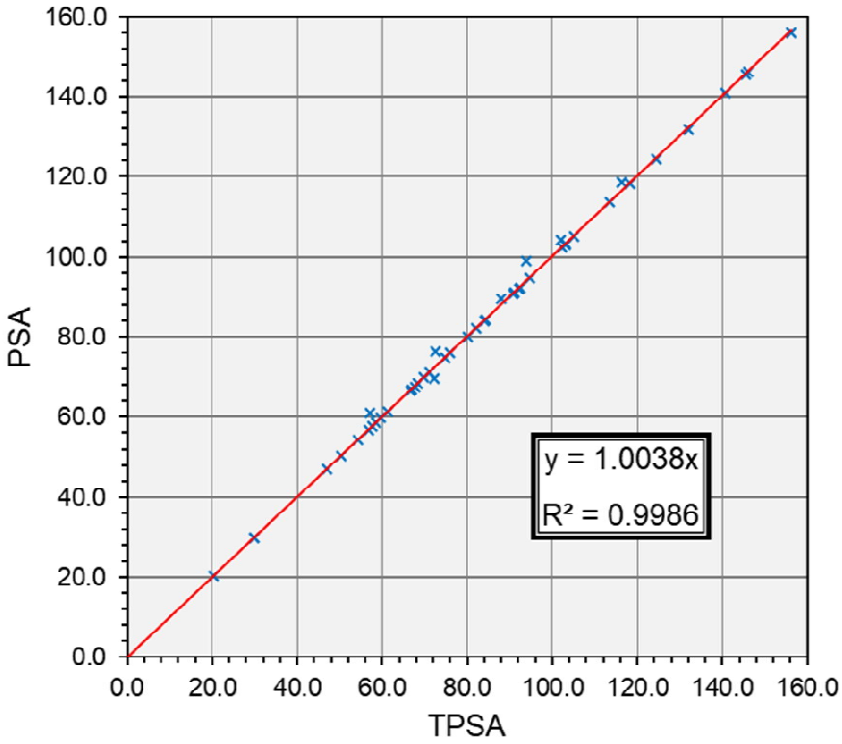


Figure 7. TPSA vs. TPSA plot using 1 in 1000 data

4.7. IFPTML model reparameterization.

Despite the strong correlation and linear dependence indicated by both graphs, the appearance of three families of points suggests a reparameterization of the model. Mordred descriptor calculator, and SQLite database has been created where the same data from the ChEMBL database have been registered with the new descriptor values. The database has been named PTMLmulticonditon dataset db, where the data has been divided into two tables: training, which contains the trials that have been used in the parameterization; and testing, which contains the data to be used in the validation. Afterwards, a new parameterization of the model with the data from the training table has been carried out with Statistica. The expressions that have been obtained for $f_1(v_{ij})_{calc}$ and $f_0(v_{ij})_{calc}$, with which the probability $p(f(v_{ij})_{pred}=1)$ will be calculated later, are the following classification functions:

$$f_1(v_{ij})_{calc} = -6.1560731987274 + 17.706853 \cdot f(v_{ij})_{ref} - 0.111993 \cdot \Delta D_1(c_j) + 0.000455 \cdot \Delta D_2(c_j) \quad (13)$$

$$f_0(v_{ij})_{calc} = -0.251481 + 2.975200 \cdot f(v_{ij})_{ref} - 0.001753 \cdot \Delta D_1(c_j) - 0.000043 \cdot \Delta D_2(c_j) \quad (14)$$

Subtracting equation (14) from (13) , the following discriminant function is obtained:

$$f(v_{ij})_{calc} = -5.904592 + 14.731652 \cdot f(v_{ij})_{ref} - 0.110240 \cdot \Delta D_1(c_j) + 0.000497 \cdot \Delta D_2(c_j) \quad (15)$$

4.8. IFPTML model posterior probabilities calculation.

In order to obtain a method for calculating the probability that the trial is classified as favourable, different references have been consulted. The Statistica manual [19] does not specify any function for calculating probability, so an alternative function has been proposed [20]. First of all, one must be aware of the geometric meaning of the problem at hand. In this classification problem, the set of trials used for the parameterization of the IFPTML model form a dispersion of points in a one-dimensional space, where the dimension along which they are distributed are their values $f(v_{ij})_{calc}$. Therefore, the value $f(v_{ij})_{calc}$ for all trials in the parameterization database has been calculated. Afterwards, the averages and $\langle f_1(v_{ij})_{calc} \rangle$ and $\langle f_0(v_{ij})_{calc} \rangle$for the group $f(v_{ij})_{obs}$=0and the group f(vij)obs=1respectively, have been calculated. In this way, the positions of the centroids of both groups in space have been established. To find a probability function, one must take into account the following characteristics that it must have. The value of the probability function $p(f(v_{ij})_{pred}=1)$must be maximum, and therefore tend to 1, when $f_1(v_{ij})_{calc} \rightarrow f_1(v_{ij})_{max}$. The value of the probability function $p(f(v_{ij})_{pred} = 1)$ must be maximun, and therefore

tend to 1, when $f_1(v_{ij})_{calc} \rightarrow f_1(v_{ij})_{max}$ and $f_0(v_{ij})_{calc} \rightarrow f_0(v_{ij})_{min}$. The value of the probability function $p(f(v_{ij})_{pred} = 1)$ must be minimun, and therefore tend to 0, when $f_1(v_{ij})_{calc} \rightarrow f_1(v_{ij})_{min}$ and $f_0(v_{ij})_{calc} \rightarrow f_0(v_{ij})_{max}$. The prior probability π_1 for $p(f(v_{ij})_{pred} = 1)$ should force the trials predicted by the model to be favourable. Therefore, considering the above, the following probability function is proposed, where π_0=1-π_1:

$$p\left(f(v_{ij})_{pred} = 1\right) = \frac{\pi_1 e^{f_1(v_{ij})_{calc}}}{\pi_1 e^{f_1(v_{ij})_{calc}} + \pi_0 e^{f_0(v_{ij})_{calc}}} \qquad (16)$$

Operating, the following expression is obtained for the probability as a function of the function of the equation (15). Note that if both priors were equal, the probability would be the form of the sigmoid function (see **Figure 8**).

$$p\left(f(v_{ij})_{pred} = 1\right) = \frac{\pi_1}{\pi_1 + \pi_0 e^{-f(v_{ij})_{calc}}} \qquad (17)$$

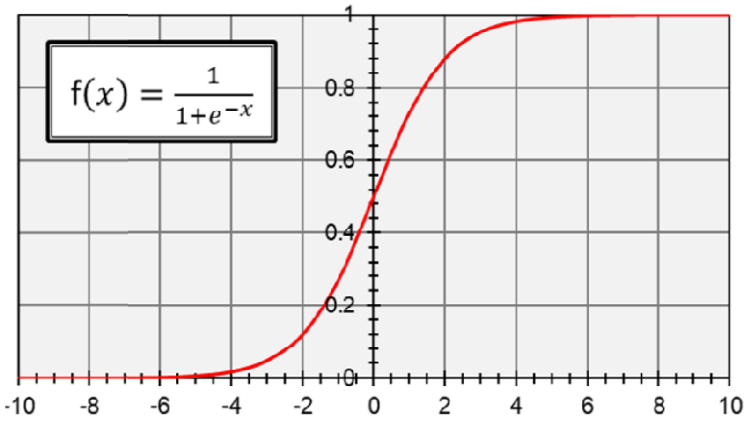


Figure 8. Sigmoid function.

To verify the correct functioning of LAGA and the validity of the function $p(f(v_{ij})_{pred} = 1)$, a prediction has been made for the group of tests in the testing table of the PTML multiconditon dataset database, which contains 29233 cases. Firstly, a sensitivity Sn = 62%, a specificity Sp = 94% and an accuracy Ac = 90% have been observed. Finally, these are the equations that LAGA uses in the implementation of the IFPTML and calculation of the probability that a trial is

classified in the group of favorable trials. This new reparameterized model presents a specificity Sp=89.9%, sensitivity Sn=70.6%and precision Ac=87.4%. On validation, it returns values of Sp=89.9%, Sn=71.6%and Ac=87.7%.

In it, it can be seen how the probability function adjusts to a sigmoid displaced from the origin, remaining centered on the point $f(v_{ij})_{calc}$ = -1.5. It must be considered that negative values of the discriminant function indicate that the input trial is closer to the group of unfavorable cases, and that positive values indicate that it is closer to the favorable cases. Thus, the fact that the probability function is centered on a negative value implies that, through this choice of the probability function, the classification of trials as favorable cases is favored.

From **Figure 9** it can also be seen that for the input trials for which $f(v_{ij})_{calc}$ = 0 the probability of being classified as favorable is 0.8, precisely the *prior probability* that had been established. This implies that for the trials that the discriminant function places at the same distance from the centroid of the set of favorable cases and from the centroid of the set of unfavorable cases, they are classified as favorable results. Since the *Statistica program* gives a probability value for the *training data set,* the probability for that data set has been calculated with equation (17). Later, this probability has been represented against that indicated by *Statistica*software. This second probability function is unknown, since this program uses a private method, so with this representation it is intended to observe if there is any relationship with the proposed probability function.

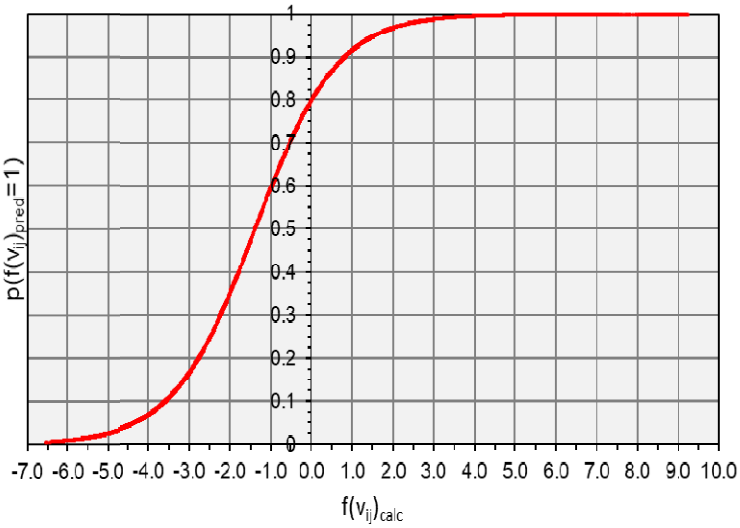


Figure9. Probability function.

The **Figure 10**shows that there is a non-linear relationship between both probabilities. No points are observed outside the curve, so both functions depend on the same value $f(v_{ij})_{calc}$. Therefore, the chosen probability function is correct to the first approximation.

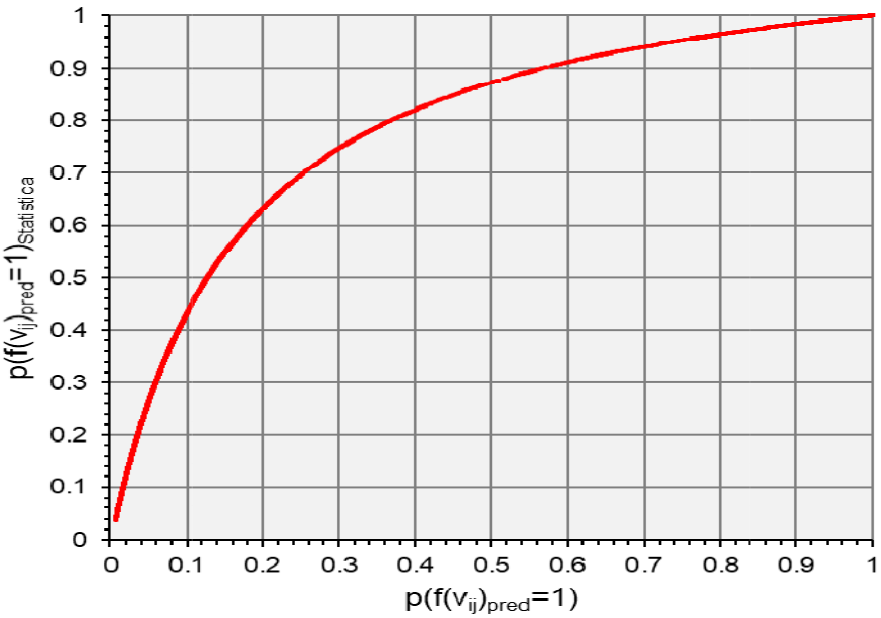


Figure 10. Probability given by *Statistica*versus the proposed probability.

4.9. IFPTML model implementation

For the implementation of the model, an SQLite database (multicondition.db) has been developed where all the combinations of conditions have been registered. This database is constantly consulted by LAGA to show the user the conditions that they can choose (See **Figure 11**). First, the user must choose the activity or activities that he wishes to classify and, based on this choice, multicondition.db is consulted and the combinations that do not exist for the chosen activities disappear from the lists of conditions.

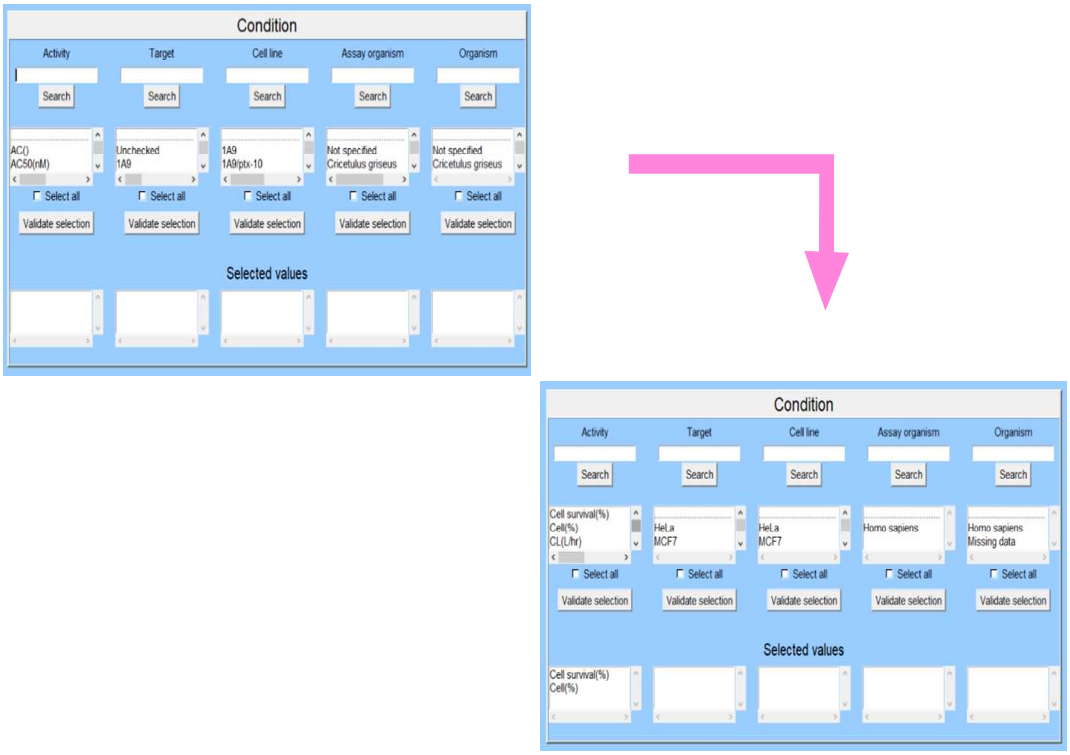


Figure 11. Selection of conditions.

Once the SMILES codes have been entered and the conditions selected, the model is executed. To do this, a module (IFPTML) has been built in which the functions corresponding to equations (13), (14), and (15) are stored. This module communicates with the SQLite multicondition.db database to query the values of the disturbances corresponding to the selected conditions vector.

Finally, the values of the discriminant functions and the probability are calculated (the probability function is discussed in the next section, equation (17)). So that the user is aware of the results, a summary of the results obtained is returned in the results window of the interface, specifying the test conditions, the probability that the test belongs to the favorable cases and the classification of each test. as favorable or unfavorable. In addition, a text file is created where the results are saved in detail in a table generated by tabulations. If the user wants to see the results in more detail, they can view the file that has been created automatically. In this way, you can save the results in a table format to view them with any spreadsheet program.

4.10. CHAPTER 04. CONCLUSIONS.

The LAGA program was intended to develop a tool to implement the multi-condition PTML-LDA model, which predicts the success of preclinical trials through a discriminant function. In order for this model to be accessible to users, a graphical interface has been developed in which to introduce all the input variables and in which, at the end, the results are presented in the form of probability of success. It must be considered that LAGA favors the prediction of positive results, since a high *prior probability has been chosen*. In this way, possible favorable assays, but close to the set of unfavorable assays, will be classified in the first group, so good candidates for *in vitro assays will not be lost*.On the other hand, it is intended to include in LAGA a new PTML-LFER model based on the Linear Discriminant Analysis (LDA) algorithm, which allows combinations of conditions outside the database. With the inclusion of this model, LAGA will be a complete tool for predicting the success of preclinical trials.

4.11. CHAPTER 04. MATERIALS AND METHODS.

The following materials have been used for the development of this software. The IFPTML model for anticancer compounds developed by Bediaga H., Arrasate S. and H. González-Díaz in previous papers and re-parametrized in this chapter. The ChEMBL database was used in the same verion reported in the previous works [4]. In order to code the program, we used the

following Python Libraries. The *SQLite 3* library was used for the creation and management of the database used by the program.[3] The *rdkit*library was used for reading input molecules.[9] The *Mordred 1.2.0* library was used for for the calculation of molecular descriptors.[7] Last the *Tkinter*library was used for building the Graphical User Interface *(*GUI).[8]

4.12. CHAPTER 04. AUTHORS CONTRIBUTIONS

Conceived and designed the experiments: HGD. Performed the experiments: J.C, B.H. (Thesis Author). Analysed the data: J.C, B.H. (Thesis Author). Wrote the paper: J.C, B.H. (Thesis Author), A.S., HGD. All authors have given approval to the final version of the paper manuscript. The authors declare no competing financial interest.

4.13. CHAPTER 04. REFERENCES.

1. H. Bediaga, S. Arrasate, and H. González-Díaz, ACS Comb.Sci. 2018,20, 621.
2. H. Gonzalez-Diaz, S. Arrasate, A. Gomez-Sanjuan, N. Sotomayor, E. Lete, L. Besada-Porto, and J. Ruso, CurrentTopics in Medicinal Chemistry13 ,**1713** (2013).
3. SQLite3 ;www.sqlitetutorial.net/sqlite-python (07/06/2019).
4. Ortega-Tenezaca, B., Quevedo-Tumailli, V., Bediaga, H., Collados, J., Arrasate, S., Madariaga, G., Munteanu, C. R., Cordeiro, M., & González-Díaz, H..*CurrentTopics in Medicinal Chemistry*, **2020**, *20*(25), 2326–2337.
5. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. **1988**, 28, 31-36.
6. Weininger, D.; Weininger, A.; Weininger, J.L. SMILES. 2. Algorithm for generation of unique SMILES notation J. Chem. Inf. Comput. Sci. **1989**, 29, 97-101.
7. Moriwaki H, Tian Y-S, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. Journal of Cheminformatics **2018**, 10, 4. doi: 10.1186/s13321-018-0258-y
8. JE Grayson, *Python and Tkinter Programming* (Manning, Greenwich, CT, 2000)
9. Open source chemoinformatics*rdkit*; http://www.rdkit.org (10/17/2018).

10. SA Wildman and GM Crippen.Prediction of Physicochemical Parameters by Atomic Contributions. J. Chem. Info. and Comput. Sci. **1999**, 39, 868-873.
11. P. Ertl, B. Rohde, and P. Selzer, Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. Journal of Medicinal Chemistry 2000, **43**, 3714-3717.

Chapter 05.

IKERDATA S.L. Company

5. CHAPTER 5. IKERDATA S.L. COMPANY

5.1. CHAPTER 05. ABSTRACT.

IKERDATA S.L., VAT No.: B16731879, PIC No.: 885528036, ZITEK, UPVEHU, Rectorate building is an Inter-University Startup Supported by Biscay Provincial Council, and Researchers of University of Basque Country (UPV/EHU), IKERBASQUE Foundation, Greater Bilbao, Basque Country, and UDC University of Coruña, Spain. The company offers the following services: Artificial Intelligence (AI), Scientific Consulting, Computing, Software Development, Consortium Partnering, Regulatory Affairs Counseling, Publication, and Training Services.

Keywords: Startup, Technology Transfer, Artificial Intelligence, Sciencitific Computing

5.2. IKERDATA S.L. company scope

IKERDATA S.L. offers the following services: Artificial Intelligence (AI), Scientific Consulting, Computing, Software Development, Consortium Partnering, Regulatory Affairs Counseling, Publication, and Training Services. The company has the following Area of Expertise: Software Development and Computing (AI, ML, Cheminfo., Bioinfo., Complex Networks) Commisioned Research. Consulting on Regulatory Affairs (Patents, Copyright, GDPR Data Protection). We Can Also Arrange/Sub-Contracting Experimental Scientific Services (Organic Synthesis, HPLC, NMR, IR Spectroscopy, CryoEM Imaging, Pharmacological Assays...) with Partner University Labs to Corroborate Computing Outcomes. This company was created based on the know how accumaltaed by the founder partners. Know How: We Develop Our Own Scientific Computing Algorithms and Software Tools with Original Artificial Intelligence, Statistics, Complex Networks Analysis, Bioinformatics and Cheminformatics Algorithms.

5.3. IKERDATA S.L. company organization chart.

The company has nine founder patners including 4 researchers of The University of Basque Country (UPV/EHU), 3 researchers of University of Coruña (UDC) and 2 additional partners focused in administration and/or communication. The organization chart includes two administrators and a cosultive Supporting Advisory Board (SAB). The two administrators are the Chief Excecutive Officer (CEO) with researche expertise and one administrator who assist the CEO in company administration, accounting, etc.

5.4. IKERDATA S.L. partners & scientific advisory board.

The founder partners (see **Figure 5**) are thos original partners that participated in the creation of the company. Among them, the members of the Sientific Advisory Board (SAB) are those partners who are uniersity researchers or professors offering scientific consulting services to the company. SAB members are not allowed to have decision-making positions in the company.

They have only the 10% of the company participation (per partner) according to public servant legislation in Spain.

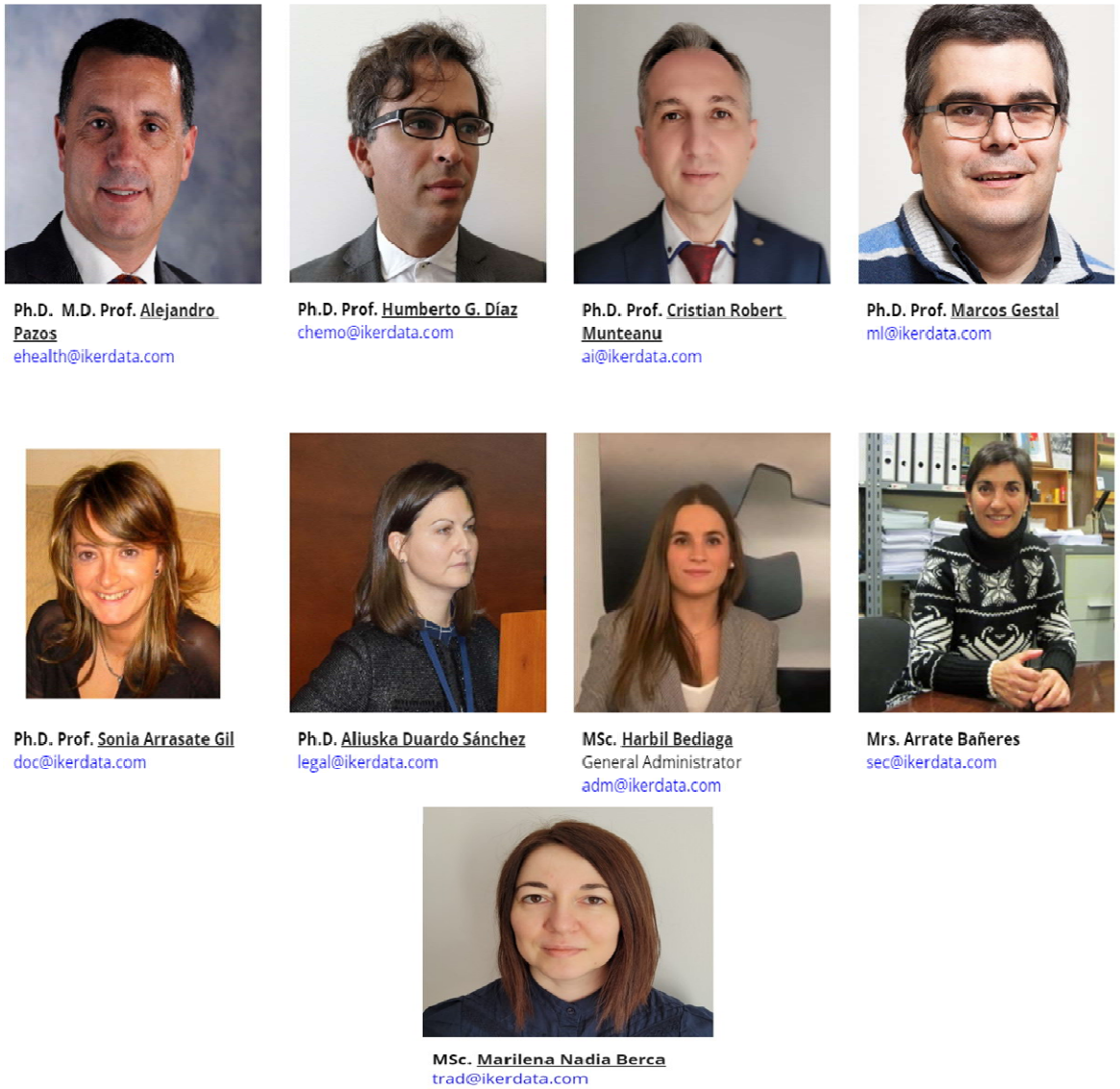


Figure 1. IKERDATA S.L. funder partners

5.5. IKERDATA S.L. company services.

We are focused on Analysis Services, Consulting, Training, Association of Research Consortiums, Knowledge Diffusion, etc. We specialize in the application of advanced computing techniques and procedures, including Artificial Intelligence (AI), statistics, complex networks, bioinformatics, chemoinformatics, pharmacoinformatics,... for decision-making and optimization of resources in the field of health, the Pharmaceutical Industry (early discovery and repositioning of drugs or vaccine design); in the validation and discovery of Biomarkers, Nanobiotechnology, Biotechnology, the fuel industry and biomedical engineering. At IKERDATA, S.L. we develop, validate and commercialize predictive models whose functions have been specifically designed to meet the demand of each company or project and we design the necessary tools to facilitate their use. In any case, ensuring the adequate transfer of technology to the recipient of our products.

5.6. IKERDATA S.L. products and their applications.

This can facilitate the creation by the company of the following types of products: We Can Offer Scientific Computing Services or Implement, Validate, and Transfer User-Friendly Software Packages Offering Solutions to Practical Problems Tailored to Client Necessities. These products may have different applications: Decision Making and Resources Optimization in Drug Discovery, Vaccine Development, Biomarkers Validation, Pharmaceutical Formulation, Catalyst Design, Synthetic Routes Optimization, Nanosystems Design, Materials Discovery, Synthesis Optimization, Biofuel Blending, Food, Nutraceuticals, and Cosmetics Production, Toxicity Prediction, and Eco-Toxicity, Risk Assesment, *etc*.

5.7. IKERDATA S.L. Companystandars

In their practice the company is begning to work and/or trying to adhere/implement the following regulations, industry standards, principles, ethical policies, *etc*. European General Data Protection Regulation (GDPR), (2016), FAIR Principles for software, FAIR principles for data, Sustainability Development Goals (SDG), White Paper in AI of European Union, USA FDA Modernization act. (2023), REACH (1907/2006), OECD QSAR (2004), Cosmetics (1223/2009), Pesticides (1107/2009),*etc*.

5.8. IKERDATA S.L. Employment generation (until 2023, july).

The company has generated until now a total of 7 part-time contracts. These contracts are of two types: (1) contracts generated with funds obtained from projects developed for clients and (2) young researchers training contracts funded with public funds. In total 3 of the contracts generated until now are of type (1) and 4 of the contracts are of type (2). The 4 types (2) contracts have been funded with the NEXTGENERATIONEU INVESTIGO-IKERDATA Programme/Project sponsored by the the European Commision with NextGenerationEU funds from Recovery Plan for Europe by means of Basque Government (EuskoJaurlaritza), and Lanbide (Basque Employment Service). We have used the grant to hire 4 young female researchers (2 of them in 2022-Oct and other two to begin in 2023-Jan-01). Each one of them will be in charge of different research projects with collaborating institutions and also management activities. The young researcher contracted has enrolled in PhD programs of the UPV/EHU and UDC. Project Principal Investigator (PI): Prof. H. González-Díaz (UPV/EHU, Company Co-Founder), Project Co-PI: Prof. S. Arrasate (UPV/EHU, Company Co-Founder), UDC Project supervisor: Prof. Alejandro Pazos (UDC, Company Co-Founder), Project Manager: IKERDATA S.L. CEO H. Bediaga, Project Legal Affairs PI: IKERDATA S.L. Legal Affairs Director PhD. A. Duardo-Sánchez, Project Budget: **264000 EUR**.

5.9. IKERDATA S.L. Homepage and social media.

The company has a landing homepage published online, https://www.ikerdata.com/en. The homepage describes the About information, partners, collaboration and contracting modalities, etc. The about company information is the following. IKERDATA, S.L. is a "start-up" company founded by professors from the UPV / EHU, the IKERBASQUE Foundation and the University of A Coruña (UDC), with the support of ZITEK, the program to support entrepreneurship on the Bizkaia Campus, to put in value its results of Research and development in transfer actions that help to improve the quality of care of public services and the competitiveness of the productive sector in our environment or area of influence. The company has also a Linkedin page used to announce company activities and news, make products and services advertisement, contact potential users, promote the company trademark, etc., https://www.linkedin.com/company/ikerdata/.

III. Thesis Conclusions

II. PART 03. THESIS CONCLUSIONS

6.1. THESIS OVERALL CONCLUSIONS

At the beginning of thisthesis, we have set a series of goal related to the development (traing/validation) of new models based on the IFPTML algorithm for the discovery of new drugs (anticancer, allosteric modulators). In addition, we intended to implement one of the drug designsIFPTMLmodels developed in a beta version of new user-friendly software for end users. Finally, we proposed to create a new startup company focused on the development of new products (software), services (data analysis, scientific computing, consulting) related to the application of IFPTML in different health and industry problems. Therefore, the conclusions of this thesis are in consonance with the goals set:

- We were able to develop (train/validate) new linear and non-linear multi-output IFPTML models for the prediction of potential anti-cancer drugs.

- We were able to develop (train/validate) new linear and non-linear multi-output IFPTML models for the study of allosteric modulating drug trials.

- We were able to develop anew software, called LAGA.PTML, in which we implemented IFPTML algorithms for the prediction of anti-cancer compounds with auserfliendly interface to allow application by the end users from pharmaceutical industry and/or Cheminformatics and Medicinal Chemistry researchers.

- We were able to launch a new company called IKERDATA S.L. as a case of technological transference of IFPTML modeling algorithms to society.

6.2. FUTURE DEVELOPMENTS

- Develop (train/validate) new linear and non-linear multi-output IFPTML models for the prediction of anti-cancer drugs taking into account information from Clinical Trials and other pharmaceutical formulation components and additivies different from the active principal ingredient.
- Develop (program) a beta version of the LAGA.PTML software, in which IFPTML algorithms for the prediction of final pharmaceutical forms of anti-cancer compounds are implemented for the first time considering information from Clinical Trials and other pharmaceutical formulation components and additivies different from the active principal ingredient.
- Develop (train/validate) new linear and non-linear multi-output IFPTML models for the study of allosteric modulating drugs taking into consideration the concencation of the compound, time of assay, type of allosteric modelation, different classes of molecular descriptors, target protein sequence, etc.
- Consolidate the inter-university company IKERDATA S.L. based on IFPTML data analysis techniques and others carry out the transference from university to company of the new software developed according to the regulations stablished by the Offices of Transference of both universities implicated.

6.3. ABBREVIATIONS LIST

ANOVA =**An**alysis **o**f **Va**riance

AUROC = **R**eceiver **O**perating **C**haracteristic **C**urve

ANN = **A**rtificial **N**eural **N**etworks

BD = **B**ig **D**ata

BLR = **B**inary **L**ogistic **R**egression

CHEMBL = Chemicaldatabase of **E**uropean **M**olecular **B**iology **L**aboratory

DEPT = **D**istortionless **E**nhancement by **P**olarization **T**ransfer

DMF =*N*,*N*-**dim**ethyl**f**ormamide

DOS = **D**iversity-**O**riented **S**ynthesis

ESI-TOF =**E**lectrospray**I**onization **– T**ime **o**f **F**light

FRAMA = **F**ramework **M**oving **A**verage

HRMS = High-**R**esolution **M**ass **S**pectrometry

IF = Information Fusion

IFPTML = **I**nformation **F**usion and **P**erturbation-**T**heory **M**achine **L**earning

LAGA = **L**arge **A**ctivity **G**roup **A**nalyzer

LDA = **L**inear **D**iscriminant **A**nalysis

LNN = **L**inear **N**eural **N**etwork

LOGP = **Log**arithm of the n-Octanol/Water **P**artition coefficient

LOGR =**Lo**gistic **R**egression

MA =**M**oving **A**verage

MIC = **M**inimun **I**nhibitory **C**oncentratiom

MLP = **M**ulti-**L**ayer **P**erceptron

MMA = **M**ulti-condtion **M**oving **A**verage

ML =**M**achine **L**earning

mGluR = **M**etabotropic **Glu**tamate **R**eceptor

MRNs = **M**etabolic **R**eaction **N**etworks

NCBI = **N**ational **C**enter for **B**iotechnology **I**nformation

NMR =**N**uclear **M**agnetic **R**esonance

NPs =**N**ano-**P**artícles

PAMs =**P**ositive **A**llosteric **M**odulators

PINs =**P**rotein **I**nteraction **N**etworks

PT = **P**erturbation **T**heory

PTOs = **P**erturbation **T**heory **O**perators

QSAR = **Q**uantitative**S**tructure-**A**ctivity **R**elationship

RBF = **R**adial **B**asis **F**unction

SMILES = **S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem

SQL = Structured **Q**uery **L**anguage

TPSA = Topological **P**olar **S**urface **A**rea

WHO = **W**orld **H**ealth **O**rganization

IV.Annexes

Annex I.

Related Papers

ANNEX 1.JOURNAL PAPERS RELATED TO THIS THESIS (FIRST PAGE).

A1.1. Publication1.

Ortega-Tenezaca, B., Quevedo-Tumailli, V., Bediaga, H., Collados, J., Arrasate, S., Madariaga, G., Munteanu, C. R., Cordeiro, M., & González-Díaz, H. (2020). IFPTML Multi-Label Algorithms: Models, Software, and Applications. *Current Topics in Medicinal Chemistry*, *20*(25), 2326–2337.

REVIEW ARTICLE

PTML Multi-Label Algorithms: Models, Software, and Applications

Bernabe Ortega-Tenezaca[1,2,3], Viviana Quevedo-Tumailli[1,2,3], Harbil Bediaga[4], Jon Collados[4,5], Sonia Arrasate[4], Gotzon Madariaga[5], Cristian R Munteanu[1,6,7], M. Natália D.S. Cordeiro[3,*] and Humbert González-Díaz[4,8,9,*]

[1]RNASA-IMEDIR, Computer Science Faculty, University of A Coruna, 15071 A Coruña, Spain; [2]Universidad Estatal Amazónica UEA, Puyo, Pastaza, Ecuador; [3]LAQV@REQUIMTE, Department of Chemistry and Biochemistry, University of Porto, 4169 007 Porto, Portugal; [4]Department of Organic and Inorganic Chemistry, University of Basque Country UPV/EHU, 48940 Leioa, Spain; [5]Department of Condensed Matter Physics, University of Basque Country UPV/EHU, 48940 Leioa, Spain; [6]Biomedical Research Institute of A Coruña (INIBIC), University Hospital Complex of A Coruña (CHUAC), 15006 A Coruña, Spain; [7]Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC), Campus de Elviña s/n, 15071 A Coruña, Spain; [8]Basque Center for Biophysics CSIC- UPV/EHU, University of Basque Country UPV/EHU, 48940 Leioa, Spain; [9]IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Biscay, Spain

ARTICLE HISTORY

Received: June 18, 2020
Revised: July 19, 2020
Accepted: July 20, 2020

DOI:
10.2174/1568026620666200916122616

Abstract: By combining Machine Learning (ML) methods with Perturbation Theory (PT), it is possible to develop predictive models for a variety of response targets. Such combination often known as Perturbation Theory Machine Learning (PTML) modeling comprises a set of techniques that can handle various physical, and chemical properties of different organisms, complex biological or material systems under multiple input conditions. In so doing, these techniques effectively integrate a manifold of diverse chemical and biological data into a *single* computational framework that can then be applied for screening lead chemicals as well as to find clues for improving the targeted response(s). PTML models have thus been extremely helpful in drug or material design efforts and found to be predictive and applicable across a broad space of systems. After a brief outline of the applied methodology, this work reviews the different uses of PTML in Medicinal Chemistry, as well as in other applications. Finally, we cover the development of software available nowadays for setting up PTML models from large datasets.

Keywords: Drug Discovery, Cheminformatics, Multi-target models, Large data sets, PTML, Perturbation theory, Machine learning.

1. INTRODUCTION

Quantitative Structure-Activity Relationships (QSAR) modelling is a widely employed computational approach that aims at predicting endpoint response(s) (*e.g.* activity, property, or toxicity) of chemicals based on their encoding features (*descriptors*), and it is playing an increasingly key role in drug or material design.

Any response value of a chemical compound may vary considerably when determined using different experimental protocols or when applying the same experimental protocol but in different conditions, such as laboratory, environmental, time, and even if different biological measures are employed like *IC*50, *EC*50, *Ki*, *etc.* [1, 2]. Determining the response value of new chemical compounds is a particularly important task in Medicinal Chemistry but simultaneously, highly demanding both in terms of time and resources. Currently, studies are conducted on Cheminformatics models to predict physicochemical properties of small organic molecules, proteins, proteomes, and complex systems. It is useful for reducing time and research resources in laboratories. Different authors have applied the combination of PT, and ML to obtain PTML models on biological systems [3].

In such case, one should highlight the ChEMBL database, which is nowadays a well-recognized resource in the field of drug discovery and medicinal chemistry research. In fact, this database curates and stores standardized bioactivity, molecules, targets, and drug data retrieved from multiple sources, as well as from the primary medicinal chemistry literature [4]. It includes, in addition, multiple conditions of assays, such as different experimental parameters, biological assays, target proteins, cell lines, assay organisms, etc. Other databases that exist and comprise

*Address correspondence to these authors at the LAQV@REQUIMTE, Department of Chemistry and Biochemistry, University of Porto, 4169-007 Porto, Portugal; Tel: +351 927599268; E-mail: ncordeir@fc.up.pt and Department of Organic and Inorganic Chemistry, and Basque Center for Biophysics CSIC-UPVEHU, University of Basque Country UPV/EHU, 48940 Leioa, Spain; IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Biscay, Spain; Tel: +34 94 601 3547;
E-mail: humberto.gonzalezdiaz@ehu.eus

1568-0266/20 $65.00+.00

A1.2. Publication2.

Sampaio-Dias IE, Rodríguez-Borges JE, Yáñez-Pérez V, Arrasate S, Llorente J, Brea JM, Bediaga H, Viña D, Loza MI, Caamaño O, García-Mera X, González-Díaz H. Synthesis, Pharmacological, and Biological Evaluation of 2-Furoyl-Based MIF-1 Peptidomimetics and the Development of a General-Purpose Model for Allosteric Modulators (ALLOPTML). *ACS Chem Neurosci*. 2021 Jan 6;12(1):203-215. doi: 10.1021/acschemneuro.0c00687.

ACS Chemical Neuroscience

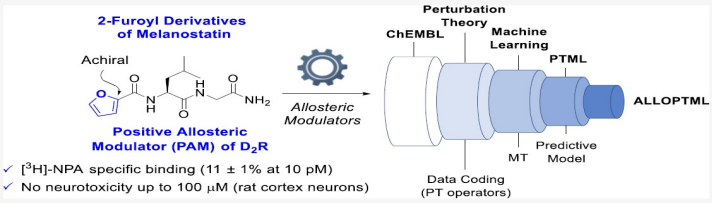
pubs.acs.org/chemneuro Research Article

Synthesis, Pharmacological, and Biological Evaluation of 2-Furoyl-Based MIF-1 Peptidomimetics and the Development of a General-Purpose Model for Allosteric Modulators (ALLOPTML)

Ivo E. Sampaio-Dias,* José E. Rodríguez-Borges, Víctor Yáñez-Pérez, Sonia Arrasate, Javier Llorente, José M. Brea, Harbil Bediaga, Dolores Viña, María Isabel Loza, Olga Caamaño, Xerardo García-Mera,* and Humberto González-Díaz*

Cite This: *ACS Chem. Neurosci.* 2021, 12, 203−215 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: This work describes the synthesis and pharmacological evaluation of 2-furoyl-based Melanostatin (MIF-1) peptidomimetics as dopamine D_2 modulating agents. Eight novel peptidomimetics were tested for their ability to enhance the maximal effect of tritiated *N*-propylapomorphine ([^{3}H]-NPA) at D_2 receptors (D_2R). In this series, 2-furoyl-L-leucylglycinamide (**6a**) produced a statistically significant increase in the maximal [^{3}H]-NPA response at 10 pM (11 ± 1%), comparable to the effect of MIF-1 (18 ± 9%) at the same concentration. This result supports previous evidence that the replacement of proline residue by heteroaromatic scaffolds are tolerated at the allosteric binding site of MIF-1. Biological assays performed for peptidomimetic **6a** using cortex neurons from 19-day-old Wistar-Kyoto rat embryos suggest that **6a** displays no neurotoxicity up to 100 μM. Overall, the pharmacological and toxicological profile and the structural simplicity of **6a** makes this peptidomimetic a potential lead compound for further development and optimization, paving the way for the development of novel modulating agents of D_2R suitable for the treatment of CNS-related diseases. Additionally, the pharmacological and biological data herein reported, along with >20 000 outcomes of preclinical assays, was used to seek a general model to predict the allosteric modulatory potential of molecular candidates for a myriad of target receptors, organisms, cell lines, and biological activity parameters based on perturbation theory (PT) ideas and machine learning (ML) techniques, abbreviated as ALLOPTML. By doing so, ALLOPTML shows high specificity Sp = 89.2/89.4%, sensitivity Sn = 71.3/72.2%, and accuracy Ac = 86.1%/86.4% in training/validation series, respectively. To the best of our knowledge, ALLOPTML is the first general-purpose chemoinformatic tool using a PTML-based model for the multioutput and multicondition prediction of allosteric compounds, which is expected to save both time and resources during the early drug discovery of allosteric modulators.

KEYWORDS: *Allosteric modulators, artificial neural networks, big data, ChEMBL, machine learning, Melanostatin, multitarget models, perturbation theory*

1. INTRODUCTION

Dopamine receptors belong to a complex monoaminergic family of G protein-coupled receptors (GPCRs) represented by five distinct receptors (D_{1-5} receptors),[1] which are grouped into D_1-like (related to excitatory neurotransmission and composed by D_1 and D_5 isoforms) and D_2-like receptors (associated with inhibitory neurotransmission, comprising D_2, D_3, and D_4

Received: October 22, 2020
Accepted: December 7, 2020
Published: December 21, 2020

https://dx.doi.org/10.1021/acschemneuro.0c00687
ACS Chem. Neurosci. 2021, 12, 203−215

A1.3. Publication 3.

Bediaga H, Arrasate S, González-Díaz H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Comb Sci*. 2018 Nov 12; 20 (11): 621-632. doi: 10.1021/acscombsci.8b00090.

ACS Combinatorial Science

Research Article

Cite This: *ACS Comb. Sci.* 2018, 20, 621−632

pubs.acs.org/acscombsci

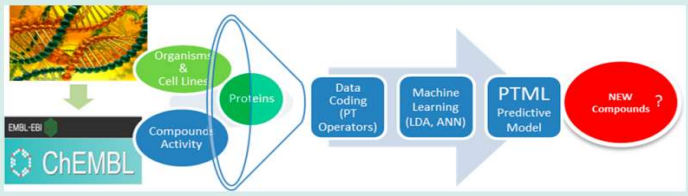
PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer

Harbil Bediaga,[†] Sonia Arrasate,*[,†] and Humbert González-Díaz*[,†,‡]

[†]Department of Organic Chemistry II, University of Basque Country UPV/EHU, 48940, Leioa, Spain

[‡]IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain

Supporting Information



ABSTRACT: Determining the target proteins of new anticancer compounds is a very important task in Medicinal Chemistry. In this sense, chemists carry out preclinical assays with a high number of combinations of experimental conditions ($\mathbf{c}_j$). In fact, ChEMBL database contains outcomes of 65 534 different anticancer activity preclinical assays for 35 565 different chemical compounds (1.84 assays per compound). These assays cover different combinations of $\mathbf{c}_j$ formed from >70 different biological activity parameters (c_0), >300 different drug targets (c_1), >230 cell lines (c_2), and 5 organisms of assay (c_3) or organisms of the target (c_4). It include a total of 45 833 assays in leukemia, 6227 assays in breast cancer, 2499 assays in ovarian cancer, 3499 in colon cancer, 3159 in lung cancer, 2750 in prostate cancer, 601 in melanoma, etc. This is a very complex data set with multiple Big Data features. This data is hard to be rationalized by researchers to extract useful relationships and predict new compounds. In this context, we propose to combine perturbation theory (PT) ideas and machine learning (ML) modeling to solve this combinatorial-like problem. In this work, we report a PTML (PT + ML) model for ChEMBL data set of preclinical assays of anticancer compounds. This is a simple linear model with only three variables. The model presented values of area under receiver operating curve = AUROC = 0.872, specificity = Sp(%) = 90.2, sensitivity = Sn(%) = 70.6, and overall accuracy = Ac(%) = 87.7 in training series. The model also have Sp(%) = 90.1, Sn(%) = 71.4, and Ac(%) = 87.8 in external validation series. The model use PT operators based on multicondition moving averages to capture all the complexity of the data set. We also compared the model with nonlinear artificial neural network (ANN) models obtaining similar results. This confirms the hypothesis of a linear relationship between the PT operators and the classification as anticancer compounds in different combinations of assay conditions. Last, we compared the model with other PTML models reported in the literature concluding that this is the only one PTML model able to predict activity against multiple types of cancer. This model is a simple but versatile tool for the prediction of the targets of anticancer compounds taking into consideration multiple combinations of experimental conditions in preclinical assays.

KEYWORDS: *ChEMBL, anticancer compounds, perturbation theory, machine learning, artificial neural networks, big data, multitarget models*

INTRODUCTION

The World Health Organization (WHO) pointed out that cancer is still among the more dangerous diseases nowadays. Classic anticancer compounds use to have a high cytotoxicity and multiple cellular targets. Regrettably, they also attack normal cells causing untoward effects. Recently, selective anticancer compounds have been introduced to target specific abnormalities in a cancer cell. Taking into consideration that cancer is a multifactor disease, there is an increasing interest in multitarget compounds able to target multiple intracellular pathways. Nevertheless, taking into consideration the number different mechanisms and possible targets it is unlikely that a single chemical become one hundred percent effective. In this context, it is more probable that new drug leads will be more effective if cover multiple action mechanisms. This arises multiple challenges. First, in order to validate the selectivity of

Received: July 1, 2018
Revised: September 12, 2018
Published: September 21, 2018

DOI: 10.1021/acscombsci.8b00090
ACS Comb. Sci. 2018, 20, 621−632

Annex II.

Congresses

ANNEX 2. CONGRESSES RELATED TO THIS THESIS (FIRST PAGE)

A2.1. Congress 1.

Harbil Bediaga. PTML-LDA One-Condition Model For The Design Of New Anti-Cancer Compounds. Published: 22 September 2019 ByMdpi In Mol2net'19, Conference On Molecular, Biomed., Comput. & Network Science And Engineering, 5th Ed. Congress Usedat-05: Usa-Europe Data Analysis Training Bilbao, Spain-Cambridge, Uk-Miami, Usa, 2019. Doi: 10.3390/Mol2net-05-06256, Https://Sciforum.Net/Paper/View/6256

MOL2NET, 2019, 4, ISSN: 2624-5078 1
http://sciforum.net/conference/mol2net-05/usedat-07

MDPI sciforum

USEDAT: USA-Europe Data Analysis Training School

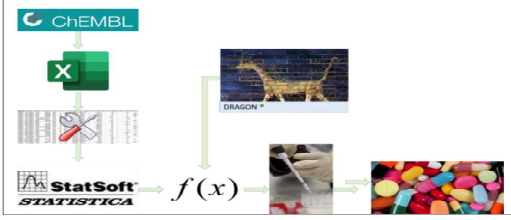
MOL2NET, International Conference Series on Multidisciplinary Sciences

PTML-LDA One-Condition model for the design of new anti-cancer compounds

Harbil Bediaga Bañeres

[1]Department of Organic Chemistry II, Faculty of Science and Technology, University of Basque Country (UPV/EHU), 48940, Leioa, Biscay, Spain.

Graphical Abstract



Abstract.

Data from preclinical assays of ChEMBL are obtained. This data is treated mathematically and a PTML-LDA model is obtained in which the conditions are analyzed individually. With this equation and the values of the descriptors of new molecules, the activity of the compound to be tested could be predicted.

Introduction

The models that analyze the Quantitative Structure-Activity Relationship (QSAR) are very useful when designing new compounds. Using the ALogP and PSA descriptors to define the structure of the compounds, we obtained a model based on the perturbation theory (PT) and machine learning (ML). The model is based on the method of classification by linear discriminant analysis (LDA).

A2.2. Congress 2.

LAGA: New software for new drug design using Perturbation Theory and Machine Learning techniques. Jon Collados, Harbil Bediaga. Published: 16 June 2020 by MDPI in MOL2NET'20, Conference on Molecular, Biomed., Comput. & Network Science and Engineering, 6th ed. congress USEDAT-06: USA-Europe Data Analysis Training Program Workshop, Bilbao, Spain-Cambridge, UK-Miami, USA, 2020, doi: 10.3390/mol2net-06-06867, https://sciforum.net/paper/view/6867

MOL2NET, **2020**, 6, ISSN: 2624-5078
http://sciforum.net/conference/mol2net-06

MDPI sciforum

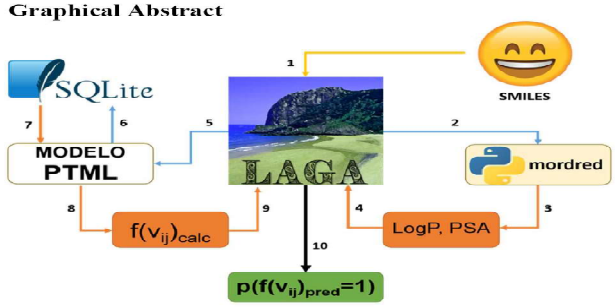
MOL2NET, International Conference Series on Multidisciplinary Sciences
Insert the title of the workshop here

LAGA: New software for new drug design using Perturbation Theory and Machine Learning techniques

Jon Collados [a], Harbil Bediaga [b]

[a] *Department of Organic Chemistry II, University of Basque Country UPV/EHU, 48940, Leioa, Spain.*
[b] *Department of Physical Chemistry, University of Basque Country UPV/EHU, 48940, Leioa, Spain.*

Graphical Abstract



Abstract.
The main objective of this project is the development of a useful computational tool for future preclinical trials. The implemented model will be a PTML-LFER model, for the prediction of the pharmacological activity of a certain molecule, or list of molecules, under multiple test conditions.

Introduction

In silico methods, which are based on computer simulations to obtain models capable of predicting whether a new compound is going to be active or not. Nowadays, more and more research groups are working on these types of methods that, with the improvement of the data calculation and processing capacity of computers, are increasingly accurate and effective.

One type of model is the Quantitative Structure-Activity Relationship (QSAR) models, which mathematically relate the molecular structure of the drug to be studied to its activity. These models are mathematical relationships, both linear and non-linear, of multiple variables. Within the QSAR models, there are the so-called Linear Free Energy Ratio (LFER) models. On the other hand, Perturbation-Theory Machine Learning (PTML) type QSAR models have also been developed that combine ideas taken from Perturbation Theory (PT) and Machine Learning (ML) methods. Finally, the PTML-LFER methods combine both approaches for the treatment of preclinical data with multiple assay specifications. In many cases, they require a long time to perform manual calculations, making it necessary to develop new specialized software that allows these models to be used quickly and easily.

A2.3. Congress3.

USEDAT-NEURODAT'21 IBRO-PERC Training Program. Rosario Isasi, Anna Lydia Svalastog, YassetPerez-Riverol, ShiraKnafo, AliuskaDuardo-Sanchez, Sonia Arrasate, David Quesada, Cristian Munteanu, Harbil Bediaga, Francisco Romero-Duran, MariaNatália D. S. Cordeiro, Bairong Shen, HumbertGonzalez-Diaz. Published: 11 January 2022 by MDPI in MOL2NET'21, Conference on Molecular, Biomed., Comput. & Network Science and Engineering, 7th ed. congress USE.DAT-07: USA-Europe Data Analysis Trends Congress, Cambridge, UK-Bilbao, Basque Country-Miami, USA, 2021. Doi: 10.3390/mol2net-07-12139, https://sciforum.net/paper/view/12139

MOL2NET, 2021, 7, ISSN: 2624-5078 1
https://mol2net-07.sciforum.net/

MDPI sciforum

USEDAT-07: USA-Europe Data Analysis Training Congress, Cambridge, United Kingdom-Bilbao, Basque Country-Miami, USA, 2021

USEDAT-NEURODAT'21 IBRO-PERC Training Program

Prof. Rosario Isasi;[1] **Prof.** Anna Lydia Svalastog (Pshyco-Soc Work);[2,3] **Dr.** Yasset Perez-Riverol;[4] **Prof.** Shira Knafo (Ph.D., M.D. Neurosci.);[5,6,7] **Dr.** Aliuska Duardo-Sanchez;[8] **Prof.** Sonia Arrasate;[9] **Prof.** David Quesada,[10] **Prof.** Cristian R Munteanu,[11] **MSc.** Harbil Bediaga Bañeres,[12] **Dr.** Francisco J Romero-Duran (PhD. M.D. Neurology);[13] **Prof.** Natalia D.S. Cordeiro;[14] **Prof.** Bairong Shen,[14] **Prof.** Alejandro Pazos (Ph.D., M.D.);[15] **Prof.** Humbert Gonzalez-Diaz (USEDAT-NEURODAT'21 Coord.)[5,7,9,*]

[1] Dept. of Human Genetics, Miller School of Medicine, Univ. of Miami, **Miami**, **United States**.
[2] Faculty of Health and Welfare, Østofold University College, **Østofold, Norway**.
[3] Dept. of Public Health and Caring Sciences, Uppsala University, **Uppsala, Sweden.**
[4] European Bioinformatics Institute (EMBL-EBI), **Cambridge**, **United Kingdom.**
[5] BIOFISIKA: Basque Center for Biophysics (CSIC, UPV/EHU), **Basque Country, Spain**.
[6] Ben-Gurion University of the Negev Faculty of Health Sciences, Beer Sheva, Southern, **Israel**.
[7] IKERBASQUE, Basque Foundation for Science, Bilbao, **Basque Country, Spain**.
[8] Chair in Law & The Human Genome Research Group, Fac. of Law, Univ. of Basque Country (UPV/EHU), Leioa, **Basque Country, Spain**.
[9] Dept. Org. and Inorganic Chemistry, Faculty of Science and Technology, Univ. of Basque Country (UPV/EHU), **Basque Country, Spain**.
[10] Dept. of Mathematics, Miami Dade College (MDC), **Miami**, **United States**.
[11] Institute for Biomedical Research (INIBIC), Univ. of Coruña (UDC), Coruña, **Spain.**
[12] IKERDATA S.L., ZITEK, UPVEHU Univ. of Basque Country, **Basque Country, Spain**.
[13] IMQ Neurology Service, Zorrotzaurre Clinic, Bilbao, **Basque Country, Spain.**
[14] Dept. of Chemistry and Biochemistry, University of Porto, **Porto, Portugal.**
[15] Institutes for Systems Genetics, West China Hospital, Sichuan University, **Chengdu, China**.

Annex III.

Resumens

(Extended)

ANNEX 3. CONGRESSES RELATED TO THIS THESIS (FIRST PAGE)

A3.1. Resumen.

La irrupción en escena de las tecnologías de detección de alto rendimiento (HTS) de fármacos ha provocado una explosión en el informe de datos de ensayos preclínicos para nuevos compuestos hit-to-lead con potencial como ingredientes farmacéuticos activos (API) en la industria farmacéutica. El análisis de todos estos datos con técnicas de Inteligencia Artificial (IA) puede conducir al desarrollo de nuevos modelos predictivos. Estos modelos pueden utilizarse a su vez para predecir compuestos más específicos y seguros con la consiguiente reducción de costes en tiempo y recursos en el desarrollo de APIs. Sin embargo, el análisis de AI de esto presenta muchos de los desafíos de los problemas de Big Data. Significa, en pocas palabras, problemas de análisis de datos con cuestiones relacionadas con Volumen, Velocidad, Veracidad, Variabilidad, Valor y Complejidad (5V + C). La primera y la segunda V se explican más o menos por sí mismas y los problemas de Variabilidad, Veracidad, Valor y Complejidad se refieren a datos con problemas de falta de datos, tendencias no consistentes, errores, informes contradictorios, interrelaciones como co-linealidad/co-linealidad, etiquetas dependientes que forman redes complejas, extrapolación de información (múltiples especies, múltiples salidas, múltiples escalas), perturbaciones en múltiples variables de entrada/salida, problemas de etiquetado múltiple, etc.

De hecho, el modelado de relaciones cuantitativas entre estructura y actividad / Quantitative Structure Activity Relationship (QSAR) es un enfoque computacional ampliamente utilizado que tiene como objetivo predecir la(s) respuesta(s) de punto final (por ejemplo, actividad, propiedad o toxicidad) de sustancias químicas en función de sus características de codificación (descriptores), y está jugando un papel cada vez más importante en el diseño de fármacos o materiales. Cualquier valor de respuesta de un compuesto químico puede variar considerablemente cuando se determina usando diferentes protocolos experimentales o cuando se aplica el mismo protocolo experimental pero en diferentes condiciones, como laboratorio, medio ambiente, tiempo e incluso si se emplean diferentes medidas biológicas como IC_{50}, EC_{50}, K_i, *etc*.

La determinación del valor de respuesta de nuevos compuestos químicos es una tarea especialmente importante en Química Médica, pero, a la vez, muy exigente tanto en tiempo como en recursos. Actualmente, se realizan estudios sobre modelos quimioinformáticos para predecir propiedades fisicoquímicas de pequeñas moléculas orgánicas, proteínas, proteomas y sistemas complejos. Para tal fin, se debe destacar la base de datos ChEMBL, que hoy en día es un recurso muy reconocido en el campo del descubrimiento de fármacos y la investigación en química médica. De hecho, esta base de datos selecciona y almacena datos estandarizados de bioactividad, moléculas, dianas y fármacos obtenidos de múltiples fuentes, así como de la literatura primaria de química médica. Incluye además múltiples condiciones de ensayos, como diferentes parámetros experimentales, ensayos biológicos, proteínas diana, líneas celulares, organismos de ensayo, etc. Otras bases de datos que existen y que contienen información diversa son el Centro Nacional de Información Biotecnológica (NCBI) y el Universal ProteinResource (UnitProt), ambos permiten fusionar su información con la proveniente de ChEMBL en un conjunto de datos para un objeto de estudio. UniProt, por ejemplo, es un recurso integral para secuencias de proteínas y datos de anotaciones que actúan sobre fármacos.Por otro lado, NCBI proporciona un amplio conjunto de recursos en línea para información y datos biológicos, incluida la base de datos de secuencias de ácidos nucleicos GenBank, y la base de datos PubMed de citas y resúmenes de revistas de ciencias de la vida publicadas.

En este contexto, nuestro grupo introdujo el Algoritmo de Fusión de Information, Teoría de Perturbación, y Aprendizaje Automático / Information Fusion, Perturbation Theory, and Machine Learning (IFPTML) para facilitar el desarrollo de modelos de salida múltiple de lectura transversal capaces de predecir múltiples resultados de compuestos químicos/fármacos en ensayos preclínicos. Tenga en cuenta que estas herramientas se pueden usar de forma independiente o combinada para resolver un problema particular de tipo combinatorio. Esto permite hacer un estudio racional de datos complejos para extraer relaciones útiles y predecir nuevas sustancias químicas.

El modelado de PT, por ejemplo, permite predecir la(s) respuesta(s) de punto final de un compuesto químico de consulta o sistema material bajo múltiples condiciones experimentales y/o

teóricas basadas en la(s) respuesta(s) de punto final de un sistema de referencia conocido.Para hacerlo, PT se combina con el enfoque de promedio móvil de Box-Jenkins, fusionando características únicas de los sistemas y simplificando las dificultades de administrar toda la información. En cuanto a las herramientas de ML, estas se han utilizado en la investigación de fármacos o materiales desde al menos los años 90, brindando soluciones rápidas y precisas a una variedad de problemas. En cuanto a la combinación de estos últimos, es decir, las herramientas de modelado predictivo IFPTML, se han aplicado ampliamente en química médica, proteómica, nanotecnología, etc. para hacer frente a grandes conjuntos de datos heterogéneos con numerosas características.

Recientemente, se han lanzado tres soluciones de software para automatizar el proceso de obtención de modelos utilizando PT, ML e IF, a saber: QSAR-Co 14, LAGA y FRAMA. QSAR-Co es un software útil para abordar algunos de los problemas críticos que generalmente se descuidan durante el desarrollo de modelos robustos de objetivos múltiples basados en clasificaciones. Nuestro grupo también ha informado sobre un software llamado SOFT.PTML que es una plataforma de propósito general para el modelado IFPTML. Sin embargo, aún quedan muchos aspectos por cubrir. Muchos problemas, como el descubrimiento de compuestos anticancerígenos y los estudios de ensayos de compuestos alostéricos, no se han analizado con algoritmos IFPTML. Además, aún no se ha informado sobre software fácil de usar específico para estos problemas. LAGA es un software desarrollado para el diseño de fármacos recurriendo tanto a la teoría de perturbaciones como a técnicas de aprendizaje automático. FRAMA ha sido desarrollado para permitir calcular descriptores y configurar condiciones de múltiples etiquetas para resolver varios problemas de diseño. Estos últimos incluyen, por ejemplo, la simplificación del análisis de cualquier conjunto de datos con una gran cantidad de características, la detección de la actividad de nuevos compuestos químicos, etc. En general, el objetivo de este artículo de revisión es explorar los avances en los últimos años sobre las técnicas utilizadas para configurar modelos predictivos IFPTML para química médica u otras aplicaciones, prestando especial atención a la disponibilidad de software fácil de usar para agilizar su uso.

Principalmente la metodología general se desarrolla en dos etapas. La primera etapa comprende *el preprocesamiento de datos.* Tras recuperar la información de interés de una base de datos que es preprocesada según criterios de valor. La segunda etapa se refiere a la aplicación de técnicas de modelado. Esto es útil para buscar modelos predictivos para conjuntos de datos complejos con múltiples características de Big Data. Finalmente, la metodología permite desarrollar modelos lineales IFPTML para predecir la actividad biológica o clasificar compuestos como activos o no nativos en términos de actividad biológica, etc.

En un trabajo reciente revisamos varios modelos con diferentes aplicaciones en Química Médica. Todos los resultados de los parámetros estadísticos como la sensibilidad, la especificidad y la precisión de los modelos IFPTML revisados en este documento de revisión son superiores al 70 %. El valor promedio más alto es 89.0% para entrenamiento y validación en el modelo IFPTML de los autores *Nocedo et al.;* con un máximo n_j igual a 56377. El valor promedio más bajo es 75.1% en el modelo IFPTML de los autores *Ferreira et al.;* con un $n_{jmínimo}$ igual a 18258 que por coincidencia son de los mismos autores en ambos casos. En las próximas secciones vamos a discutir algunos de estos modelos con más detalle.

Nocedo -Mena *et al.* reportaron por primera vez el aislamiento y caracterización de terpenos del *Cissus*planta *incisa.*Se obtienen los resultados de la base de datos ChEMBL, que contiene 160 000 resultados de pruebas preclínicas de actividad antimicrobiana para 55 931 compuestos con más de 365 parámetros de actividad biológica, 90 cepas bacterianas, 25 especies bacterianas, y el conjunto de datos de Leong y Barabási incluye 40 MRN de microorganismos Los investigadores combinaron IFPTML con la técnica IF para desarrollar el primer modelo PTMLIF. El mejor modelo lineal encontrado presentó valores de especificidad Sp(%) = 90,31 / 90,40, y sensibilidad Sn(%) = 88,14 / 88,07 en series de entrenamiento/validación como se muestra en la tabla 3. La tabla 4 muestra una comparación entre el PT-LDA obtenido y parte de la literatura, como el modelo ANN y BLR. Finalmente, se determinó experimentalmente la actividad antibacteriana de los terpenos. Los compuestos más activos fueron fitol y α- amirina , con MIC = 100 µg /ml *Enterococcus*resistente a la vancomicina*faecium y Acinetobacterbaumannii*resistente a carbapenémicos . El modelo fue útil para predecir la

actividad de estos compuestos frente a otros microorganismos con diferentes MRN para encontrar otros objetivos potenciales.

Donde, $f(v_{ij})_{calc}$es el valor de la función que puede predecir la actividad biológica del i-ésimo compuesto analizado en el j-ésimo ensayo preclínico con condiciones $c_j = (c_0 , c_1)$ contra la sésima especie de bacteria con MRN $_s$. El modelo tiene cuatro tipos de variables de entrada. El primer tipo es la función de valor esperado $f(v_{ij})_{ref}$. El segundo tipo son los valores de$Sh_k(Fármaco_i)$ que se utilizaron para cuantificar la estructura de los compuestos químicos. Y por último, dos tipos de operadores PT son el término $\Delta Sh_k(Ensayo_j)c_j$, y el otro tipo es el término $\Delta Sh_k(MRN_s)c_j$. Vásquez-Domínguez *et al*. propuso el desarrollo de un nuevo modelo predictivo que define proteínas diana de nuevos compuestos antirretrovirales. ChEMBL registra más de 140 000 ensayos preclínicos experimentales ARV (VIH, HTLV, SIV, HBV, MLV, RSV, FeLV) para 56 105 compuestos, que cubren combinaciones con 359 parámetros de actividad biológica, 55 accesiones de proteínas, 83 líneas celulares, 64 organismos de ensayo y 773 subtipos o cepas. Además, ha incluido 5.277 ensayos para el virus de la hepatitis B. El modelo IFPTML desarrollado alcanzó valores considerables en sensibilidad (%) 73,05/73,10, especificidad (%) 86,61/87,17 y precisión (%) 75,84/75,98 en series de entrenamiento/validación como se muestra en la tabla 3. Compararon modelos IFPTML alternativos con diferentes Operadores PT como covarianza, momentos exponenciales y términos.El modelo desarrollado aplicado a los ARV calcula la probabilidad de interacción de una molécula i con diferentes retrovirus bajo un conjunto de múltiples condiciones de ensayo $\mathbf{c}_j$. El VHB se ha incluido como la presencia de coinfecciones con el VIH y el VHB en los pacientes son frecuentes. El VIH prolonga la viremia del VHB, aumenta las tasas de cronicidad , también el riesgo de cirrosis y la morbilidad relacionada con el hígado. Por ello, se debe coordinar el tratamiento de ambas infecciones.Algunos estudios han encontrado medicamentos ARV efectivos y tienen una actividad significativa en el tratamiento de ciertos tipos de VHB resistentes en pacientes coinfectados con VIH/VHB. El grupo de Yang sugiere que, en caso de coinfección, la terapia ARV debe incluir agentes con actividad tanto contra el VIH como contra el VHB. Los operadores de PT de múltiples condiciones se calcularon utilizando medias móviles

combinatorias o múltiples (MMA). Donde, $f(v_{ij})_{calc}$es el valor de la función que calcula la probabilidad de interacción de una molécula *i* con diferentes retrovirus bajo un conjunto de múltiples condiciones de ensayo c_j aplicadas a tratamientos ARV. El modelo tiene tres tipos de variables de entrada. El primer tipo es la función de valor referenciado $f(v_{ij})_{ref}$. que representa el valor de actividad biológica de la molécula *m* bajo c_j subconjunto de múltiples condiciones. El segundo y los tres tipos son ΔD_k, y $\Delta D_k(c_j)$ se añaden a la ecuación efectos de perturbación en la estructura de la molécula.

Ferreira da Costa et al. diseñó un modelo destinado a predecir las interacciones fármaco-proteína (DPI) para las proteínas objetivo involucradas en las vías de la dopamina. El conjunto de datos tiene un total de 50.000 casos. El presente trabajo reporta la síntesis orgánica, caracterización química y ensayo farmacológico de una nueva serie de compuestos peptidomiméticos de compuestos peptidomiméticos*L- prolil -L- leucil - glicinamida*(PLG) . Se muestran los resultados generales de los subconjuntos de entrenamiento y validación. En la serie de entrenamiento, el modelo presentó valores altos de Especificidad = Sp (%) = 72,8, Sensibilidad = Sn (%) = 72,4 y Precisión General = Ac (%) = 72,7 como se muestra en la tabla 3. El modelo se mantuvo estable en serie de validación externa con valores de Sp (%) = 72,7, Sn (%) = 71,4 y Ac (%) = 72,6.

Quevedo-Tumailli*et al.* definió un nuevo tipo de red compleja llamada GOIN que codifica patrones de inversión de corto y largo alcance de la orientación de pares de genes en el cromosoma sobre *Plasmodiumfalciparum (Pf)* . Estas redes tienen un promedio de 383 nodos (genes) y 1314 enlaces (pares de genes con orientación inversa). Se encontraron algunas comunidades de genes que codifican proteínas relacionadas con RIFIN. El modelo IFPTML discrimina el tipo RIFIN de otras proteínas. Los parámetros de los GOINs y Centralidades se utilizaron como valores de entrada. El modelo presenta valores de sensibilidad y especificidad del 70-80% en las series de entrenamiento y validación externa, respectivamente. En conclusión, la relevancia biológica de la inversión de la orientación génica no depende directamente de la información de la secuencia genética. La entrada es una variable de centralidad llamada cercanía C_{clo}. Su centralidad mide la desviación del gen_i en el $cromosoma_k$ con respecto al valor promedio

esperado de cercanía para todos los genes en el mismo cromosoma $_k$. Esto es $f(v_{ij})_{ref}$ es igual a $C_{clo}(Gene_i, Chr_k)$ - $<C_{clo}(Chr_k)>$.

Martínez- Arzate *et al.* ha desarrollado un modelo IFPTML para descubrir nuevos epítopos de células B útiles para el diseño de vacunas y para predecir puntuaciones de epítopos inmunogénicos en diferentes condiciones experimentales. El modelo utiliza como entrada la secuencia del péptido *q* y la actividad del epítopo. La información recuperada contiene cambios estructurales en 83683 secuencias peptídicas (Seq) determinadas en ensayos experimentales informados en la base de datos IEDB, e involucran 1448 organismos (Org), 323 organismos huésped (Host), 15 tipos de procesos in vivo (Proc), 28 técnicas experimentales (Tech), más 505 aditivos adyuvantes (Adj). El modelo tiene precisión, sensibilidad y especificidad entre 71 y 80% para entrenamiento y series de validación externa.

Concu*et al.* desarrolló un modelo para predecir un conjunto de enzimas que pertenecen a la levadura Pichia . Se ha aplicado a un conjunto de datos de 19 187 enzimas que representan las 59 subclases presentes en el Protein Data Bank (PDB). Además, los autores desarrollaron modelos IFPTML basados en ANN para predecir pares enzima-enzima de secuencias de consulta de plantilla con una precisión, especificidad y sensibilidad superiores al 90 % tanto para la serie de entrenamiento como para la de validación.

QSAR-Co es un software independiente de libre acceso para llevar a cabo estudios basados en clasificaciones considerando diferentes condiciones experimentales según corresponda. Cabe señalar que **QSAR-Co** es una forma abreviada de "quimioinformática con condiciones", siendo esta última una de las características clave de este software, aunque también se pueden desarrollar modelos de quimioinformática basados en clasificación simple sin condiciones. Otra razón que motivó el desarrollo de este software fue proporcionar una plataforma distinta para derivar modelos quimioinformáticos basados en clasificación siguiendo todas las pautas recomendadas por la OCDE, es decir, modelos quimioinformáticos robustos. El software consta de dos módulos: 1) el módulo de desarrollo de modelos y 2) el módulo de pantalla/predicción. El software ' **QSAR-Co** ' versión 1.0.0 es una herramienta independiente disponible gratuitamente para descargar enla página web de QSAR-Co.Tiene dos módulos (' *desarrollo de modelos'* y '

detección/predicción') que están disponibles en el software, y ahora discutiremos todos los pasos y las funcionalidades asociadas en cada módulo. En el módulo de ' *desarrollo de modelos* ', el software proporciona todos los pasos básicos que están involucrados en el desarrollo de un modelo Cheminformatics basado en clasificación, que también incluye examinar y tratar los datos de entrada para varias condiciones experimentales, si corresponde.

El software permite calcular operadores de promedios móviles de Box-Jenkins para descriptores moleculares. El enfoque se discutió en detalle anteriormente. $_{Al}$ hacerlo, calcula los descriptores de promedio móvil para un descriptor molecular *Di* de compuestos individuales ' *i* '. El término derivado se denomina operador de Box-Jenkin , y estos descriptores modificados capturan la información sobre las estructuras químicas y el elemento específico de la condición experimental (c_j) bajo las cuales se analizaron las muestras. Estos descriptores modificados se calculan mediante el software QSAR-Co y se utilizan en los pasos posteriores de desarrollo del modelo Cheminformatics. Opcionalmente, se puede realizar un pretratamiento de datos para eliminar los descriptores no informativos que pueden no tener una contribución significativa en la construcción del modelo. También puede dividir el conjunto de datos en conjuntos de entrenamiento y prueba, de modo que en pasos posteriores el conjunto de entrenamiento se emplee para el desarrollo y la selección del modelo, mientras que el conjunto de prueba se emplea para la validación del modelo. Hay una opción para repetir la misma división aleatoria para reproducir el desarrollo del modelo utilizando el mismo valor inicial en la configuración. En los enfoques racionales, se proporcionan dos técnicas en el software, es decir, el algoritmo de Kennard-Stone y el método de división basado en la distancia euclidiana. El software también elimina los descriptores menos discriminatorios. QSAR-Co también proporciona ' *Algoritmo genético* ' como técnica de selección de variables para desarrollar modelos de 'Análisis discriminante lineal' (LDA). El algoritmo genético (GA) es una técnica bien conocida que se utiliza a menudo en el desarrollo de modelos de quimioinformática basados en regresión, así como para el desarrollo de modelos de quimioinformática basados en clasificación. En la actualidad, el software proporciona dos técnicas de aprendizaje automático para desarrollar modelos sólidos de quimioinformática basados en clasificación, análisis discriminante lineal

(LDA), y Random Forest es un algoritmo de aprendizaje automático supervisado que consiste en una colección o conjunto de predictores de árboles de decisión simples. En este software, hemos utilizado la biblioteca java Weka versión 3-9-3 para realizar Random Forest. Las métricas de validación como λ de Wilk proporcionan una medida de la importancia de la discriminación lograda. Se puede diseñar una matriz de confusión usando la información de la clase de respuesta real y predicha obtenida del modelo bajo evaluación, el software también brinda parámetros como Sensibilidad, Especificidad, relación de Fisher, etc., y realiza el análisis de la curva de características operativas del receptor (ROC) y prueba de aleatorización *Y*. Además, el software también realiza un análisis del dominio de aplicabilidad (AD). Por último, en el módulo 2 del software podemos realizar análisis predictivos de nuevos compuestos químicos.

Por otro lado, el software IFPTML.SOFT y su aplicación central FRAMA software versión 1.0.0 es una nueva aplicación de escritorio en entorno Windows. Fue desarrollado para el tratamiento de información química organizada en agrupamiento, y variables continuas. Utilice la información almacenada en hojas de cálculo. El tiempo de carga de la información hasta FRAMA es directamente proporcional al tamaño del archivo y la cantidad de información. La operación de unión de variables permite crear nuevas variables de agrupación uniendo dos o más variables existentes preferidas por el usuario. Se realiza mediante procesamiento por lotes en la cola de operaciones seleccionadas. El procesamiento mencionado agrega las nuevas variables al conjunto de variables disponibles en el contexto de trabajo. Una vez subido el archivo, y unidas las variables, se procede a la selección y clasificación de variables de agrupación y variables continuas. Se permite probar si existen datos anómalos como nulos o vacíos y tomar una decisión sobre su tratamiento. En la agrupación de variables, los valores anómalos pueden ser reemplazados por el valor “MD”. En variables continuas se puede sustituir por la media de los valores de la columna. También es posible en ambos tipos de variables eliminar casos.

Después de pretratar los datos, las variables seleccionadas pueden someterse a operaciones por lotes. En función de la naturaleza de las variables, se pueden realizar operaciones de agrupación de variables como Identidad, cuenta, probabilidad, Entropía de Shanon . Operaciones de transformación de variables continuas como: Identidad, exponencial, valor absoluto, potencia

numérica, logaritmo, probabilidad mínima máxima, z-score. Operaciones básicas para variables continuas como: Suma, producto, diferencia, división. Operaciones paramétricas como: Máximo, mínimo, promedio, suma, desviación estándar, multiplicaciones por una constante. Finalmente, operadores entre variables de agrupación y variables continuas como Media móvil, Suma por agrupación, Desviación estándar, Probabilidad mínima máxima, Puntuación Z. Es posible establecer los valores de entrenamiento y validación usando un patrón de caracteres. La letra T se utiliza para representar la formación y la V para la validación. El patrón predeterminado es TTTV que expresa el 75 % de los valores para entrenamiento y el 25 % para valores de validación. FRAMA 1.0.0 permite realizar la operación de análisis de regresión lineal, basada en la selección de variables de entrada independientes y la variable dependiente de salida. El procesamiento da como resultado un archivo CSV. El archivo resultante contiene los nombres de las variables, los valores de las constantes, el error estándar, el estadístico T, el valor de p (T), el estadístico F, el valor de p (F), el límite superior de confianza, el límite inferior de confianza, el índice, la intersección, Prueba T, Prueba F. La información de regresión se muestra con R^2 (r-cuadrado), r^2 ajustado, F-Test, Z-test, Chi-Squared Test. Es posible generar variables de media móvil multietiquetas y operadores que se utilizarán en Perturbación Teoría del aprendizaje automático.

Por último, a pesar de la potencial aplicación en la industria, no se ha desarrollado ninguna startup (al inicio de esta tesis) para la transferencia de la tecnología IFPTML a la industria. En consecuencia, el objetivo de esta tesis se centra en primer lugar en el desarrollo de nuevos modelos IFPTML de ensayos de compuestos anticancerígenos y compuestos alostéricos. A continuación, informamos sobre el desarrollo del software LAGA, fácil de usar, para la predicción extrapolada de compuestos anticancerígenos. Por último, se describe la planificación, creación, estructura, servicios, etc. de IKERDATA S.L una nueva startup interuniversitaria enfocada a la transferencia de tecnología IFPTML a empresas gallegas y del País Vasco en primera instancia con perspectivas de proyección hacia España, Europa y a escala global en última instancia.

A3.2. Resumo.

A irrupción en escena das tecnoloxías de detección de alto rendemento (HTS) de fármacos provocouunha explosión no informe de datos de ensaios preclínicos para novoscompostos hit-to-lead con potencial como ingredientes farmacéuticos activos (API) na industria farmacéutica. A análise de todos estes datos con técnicas de Intelixencia Artificial (IA) pode conducir aodesenvolvemento de novos modelos preditivos. Estes modelos poden utilizarse á súa vez para predicircompostosmáis específicos e seguros coa consecuente redución de custos en tempo e recursos no desenvolvemento de APIs. Con todo, a análise de AI disto presenta moitos dos desafíos dos problemas de Big Data. Significa, en poucas palabras, problemas de análises de datos con cuestións relacionadas con Volume, Velocidade, Veracidade, Variabilidade, Valor e Complexidade (5V + C). A primeira e a segunda V explícansemáisou menos por si mesmas e os problemas de Variabilidade, Veracidade, Valor e Complexidaderefírense a datos con problemas de falta de datos, tendencias non consistentes, erros, informes contraditorios, interrelacións como co-linealidad/co-linealidad, etiquetas dependentes que forman redes complexas, extrapolación de información (múltiples especies, múltiples saídas, múltiples escalas), perturbacións en múltiples variables de entrada/saída, problemas de etiquetaxe múltiple etc.

De feito, o modelado de relacións cuantitativas entre estrutura e actividade / Quantitative Structure Activity Relationship (QSAR) é un enfoque computacional amplamente utilizado que ten como obxectivo predicir a(s) resposta(s) de punto final (por exemplo, actividade, propiedadeoutoxicidade) de substancias químicas en función das súas características de codificación (descriptores), e está a xogar un papel cada vez máis importante no deseño de fármacos oumateriais. Calquera valor de respostadun composto químico pode variar considerablemente cando se determina usando diferentes protocolos experimentaisou cando se aplica o mesmo protocolo experimental pero en diferentes condicións, como laboratorio, medio ambiente, tempo e mesmo se seempregan diferentes medidas biolóxicas como IC50, EC50, Ki etc. A determinación do valor de resposta de novoscompostos químicos é unha tarefa especialmente importante en Química Médica pero, á vez, moiesixente tanto en tempo como en

recursos. Actualmente, realízanseestudos sobre modelos quimioinformáticos para predicir propiedades fisicoquímicas de pequenas moléculas orgánicas, proteínas, proteomas e sistemas complexos. Para tal fin, débese destacar a base de datos ChEMBL, que hoxe en día é un recurso moirecoñecido no campo do descubrimento de fármacos e a investigación en química médica. De feito, esta base de datos selecciona e almacena datos estandarizados de bioactividad, moléculas, dianas e fármacos obtidos de múltiples fontes, así como da literatura primaria de química médica. Inclúeademais múltiples condicións de ensaios, como diferentes parámetros experimentais, ensaiosbiolóxicos, proteínas diana, liñas celulares, organismos de ensaio etc. Outras bases de datos que existen e que conteñen información diversa son o Centro Nacional de Información Biotecnolóxica (NCBI) e o Universal ProteinResource (UnitProt), ambos permiten fusionar a súa información coa proveniente de ChEMBLnunconxunto de datos para un obxecto de estudo. UniProt, por exemplo, é un recurso integral para secuencias de proteínas e datos de anotacións que actúan sobre fármacos. Doutra banda, NCBI proporciona un amplo conxunto de recursos en liña para información e datos biolóxicos, incluída a base de datos de secuencias de acedos nucleicos GenBank, e a base de datos PubMed de citas e resumos de revistas de ciencias da vida publicadas. Neste contexto, o noso grupo introduciu o Algoritmo de Fusión de InformationFusion, Teoría de Perturbación e Aprendizaxe Automático / InformationFusion, Perturbation Theory, and Machine Learning (IFPTML) para facilitar o desenvolvemento de modelos de saída múltiple de lectura transversal capaces de predicir múltiples resultados de compostos químicos/fármacos en ensaios preclínicos. Teña en conta que estas ferramentas pódense usar de forma independente ou combinada para resolver un problema particular de tipo combinatorio. Normalmente, recórreseásseguintescombinacións: IFPTML (PT + ML) ou PTMLIF (PT + ML + IF). Isto permite facer un estudo racional de datos complexos para extraer relacións útiles e predicir novas substancias químicas.

O modelado de PT, por exemplo, permite predicir a(s) resposta(s) de punto final dun composto químico de consulta ou sistema material baixo múltiples condiciónsexperimentais e/o teóricas baseadasna(s) resposta(s) de punto final dun sistema de referencia coñecido. Para facelo, PT combínaseco-enfoque de media móbil de Box-Jenkins, fusionando características únicas dos

sistemas e simplificando as dificultades de administrar toda a información. En canto ásferramentas de ML, estas utilizáronsena investigación de fármacos oumateriais desde polo menos os anos 90, brindando solucións rápidas e precisas a unhavariedade de problemas. En canto á combinación destes últimos, é dicir, as ferramentas de modelado preditivo IFPTML, aplicáronseamplamente en química médica, proteómica, nanotecnoloxía, etc. para facer fronte a grandes conxuntos de datos heteroxéneos con numerosas características.

Recentemente, lanzáronse tres soluciones de software para automatizar o proceso de obtención de modelos utilizando PT, ML e IF, a saber: QSAR-Co, LAGA e FRAMA. QSAR-Co é un software útil para abordar algúns dos problemas críticos que xeralmente se descoidan durante o desenvolvemento de modelos robustos de obxectivos múltiples baseados en clasificacións. O noso grupo taméninformou sobre un software chamado SOFT.PTML que é unha plataforma de propósito xeral para o modelado IFPTML. Con todo, aínda quedan moitos aspectos por cubrir. Moitos problemas, como o descubrimento de compostosanticancerígenos e os estudos de ensaios de compostos alostéricos, non se analizaron con algoritmos IFPTML. Ademais, aínda non se informou sobre software fácil de usar específico para estes problemas. LAGA é un software desenvolvido para o deseño de fármacos recorrendo tanto á teoría de perturbacións como a técnicas de aprendizaxe automática. FRAMA foidesenvolvido para permitir calcular descriptores e configurar condicións de múltiples etiquetas para resolver varios problemas de deseño. Estes últimos inclúen, por exemplo, a simplificación da análise de calqueraconxunto de datos cunha gran cantidade de características, a detección da actividade de novoscompostos químicos etc. En xeral, o obxectivodeste artigo de revisión é explorar os avances nos últimos anos sobre as técnicas utilizadas para configurar modelos preditivos IFPTML para química médica ououtrasaplicacións, prestando especial atención ádispoñibilidade de software fácil de usar para axilizar o seu uso.

Principalmente a metodoloxíaxeraldesenvólvese en dúas etapas. A primeira etapa comprende o preprocesamiento de datos . Tras recuperar a información de interese dunha base de datos que é preprocesada segundo criterios de valor. A segunda etapa refírese á aplicación de técnicas de modelado. Isto é útil para buscar modelos preditivos para conxuntos de datos complexos con

múltiples características de Big Data. Finalmente, a metodoloxía permite desenvolver modelos lineais IFPTML para predicir a actividadebiolóxicaou clasificar compostos como activos ou non nativos en termos de actividadebiolóxica, etc. Nuntraballorecente revisamos varios modelos con diferentes aplicacións en Química Médica. Todos os resultados dos parámetros estatísticos como a sensibilidade, a especificidade e a precisión dos modelos IFPTML revisados neste documento de revisión son superiores ao 70 %. O valor media máis alto é 89.0% para adestramento e validación no modelo IFPTML dos autores Nocedo et ao.; cun máximo nj igual a 56377. O valor media máisbaixo é 75.1% no modelo IFPTML dos autores Ferreira et ao.; cun n j mínimo igual a 18258 que por coincidencia son dos mesmos autores en ambos os casos. Nas próximas seccións imos discutir algúns destes modelos con máis detalle.

Nocedo -Mena *et al.* reportaron por primeira vez o illamento e caracterización de terpenos do Cissus planta incisa . Obtéñense os resultados da base de datos ChEMBL, que contén 160 000 resultados de probas preclínicas de actividade antimicrobiana para 55 931 compostos con máis de 365 parámetros de actividadebiolóxica, 90 cepas bacterianas, 25 especies bacterianas, e o conxunto de datos de Leong e Barabásiinclúe 40 MRN de microorganismos Os investigadores combinaron IFPTML coa técnica IF para desenvolver o primeiro modelo PTMLIF. O mellor modelo lineal atopadopresentou valores de especificidade (%) = 90,31 / 90,40, e sensibilidade (%) = 88,14 / 88,07 en series de adestramento/validación como se mostranatáboa 3. A táboa 4 mostraunha comparación entre o PT-LDA obtido e parte da literatura, como o modelo ANN e BLR. Finalmente, determinouse experimentalmente a actividade antibacteriana dos terpenos. Os compostosmáis activos foron fitol e α- amirina, con MIC = 100 μg /ml Enterococcus resistente á vancomicina faecium e Acinetobacterbaumannii resistente a carbapenémicos. O modelo foi útil para predicir a actividadedestescompostos fronte a outros microorganismos con diferentes MRN para atoparoutrosobxectivospotenciais.

Onde, $f(v_{ij})_{calc}$ é o valor da función que pode predicir a actividadebiolóxica do i ésimo composto analizado no j ésimoensaio preclínico con condicións $cj = (c_0 , c_1)$ contra a s ésima especie de bacteria con MRN s . O modelo ten catro tipos de variables de entrada. O primeiro tipo é a función de valor esperado $f(v_{ij})$ref. O segundo tipo son os valores de Shk(Fármacoi) que

se utilizaron para cuantificar a estrutura dos compostos químicos. E por último, dous tipos de operadores PT son o termo $\Delta Sh_k(Ensayoj)_{cj}$, e o outro tipo é o termo $\Delta Sh_k(MRNs)_{cj}$.

Vásquez- Domínguez *et al.* propuxo o desenvolvementodunnovo modelo preditivo que define proteínas diana de novoscompostosantirretrovirais. ChEMBLrexistramáis de 140 000 ensaios preclínicos experimentais ARV (VIH, HTLV, SIV, HBV, MLV, RSV, FeLV) para 56 105 compostos, que cobren combinacións con 359 parámetros de actividadebiolóxica, 55 accesiones de proteínas, 83 liñas celulares, 64 organismos de ensaio e 773 subtipos ou cepas. Ademais, incluíu 5.277 ensaios para o virus da hepatite B. O modelo IFPTML desenvolvidoalcanzou valores considerables en sensibilidade (%) 73,05/73,10, especificidade (%) 86,61/87,17 e precisión (%) 75,84/75,98 en series de adestramento/validación como se mostranatáboa 3. Compararon modelos IFPTML alternativos con diferentes Operadores PT como covarianza, momentos exponenciais e termos. O modelo desenvolvido aplicado aos ARV calcula a probabilidade de interacción dunha molécula i con diferentes retrovirus baixo un conxunto de múltiples condicións de ensaio c j . O VHB incluíuse como a presenza de coinfecciones co VIH e o VHB nos pacientes son frecuentes. O VIH prolonga a viremia do VHB, aumenta as taxas de cronicidade , tamén o risco de cirrosis e a morbilidade relacionada cofígado. Por iso, débese coordinar o tratamento de ambas as infeccións. Algúnsestudosatoparon medicamentos ARV efectivos e teñenunhaactividade significativa no tratamento de certos tipos de VHB resistentes en pacientes coinfectados con VIH/VHB. O grupo de Yang suxire que, en caso de coinfección , a terapia ARV debe incluíraxentes con actividade tanto contra o VIH como contra o VHB. Os operadores de PT de múltiples condiciónscalculáronse utilizando medias móbiles combinatorias ou múltiples (MMA). Onde, $f(v_{ij})_{calc}$ é o valor da función que calcula a probabilidade de interacción dunha molécula i con diferentes retrovirus baixo un conxunto de múltiples condicións de ensaio c j aplicadas a tratamentos ARV. O modelo ten tres tipos de variables de entrada. O primeiro tipo é a función de valor referenciado $f(v_{ij})_{ref}$ que representa o valor de actividadebiolóxica da molécula m baixo c j subconjunto de múltiples condicións. O segundo e os tres tipos son ΔD_k, e $\Delta D_k(c_j)$ engádense á ecuación efectos de perturbación naestrutura da molécula.

Ferreira dá Costa *et al*. deseñou un modelo destinado a predicir as interaccións fármaco-proteína (DPI) para as proteínas obxectivo involucradas nas vías da dopamina. O conxunto de datos ten un total de 50.000 casos. O presente traballo reporta a síntese orgánica, caracterización química e ensaiofarmacolóxicodunha nova serie de compostos peptidomiméticos L- prolil -L-leucil - glicinamida (PLG). Móstranse os resultados xerais dos subconjuntos de adestramento e validación. Na serie de adestramento, o modelo presentou valores altos de Especificidade = Sp (%) = 72,8, Sensibilidade = Sn (%) = 72,4 e Precisión Xeral = Ac (%) = 72,7 como se mostranatáboa 3. O modelo mantívose estable en serie de validación externa con valores de Sp (%) = 72,7, Sn (%) = 71,4 e Ac (%) = 72,6.

Quevedo-Tumailli *et al*. definiu un novo tipo de rede complexa chamada GOIN que codifica patróns de investimento de curto e longo alcance da orientación de pares de xenes no cromosoma sobre Plasmodiumfalciparum (Pf) . Estas redes teñenunha media de 383 nodos (xenes) e 1314 ligazóns (pares de xenes con orientación inversa). Atopáronsealgunhas comunidades de xenes que codifican proteínas relacionadas con RIFIN. O modelo IFPTML discrimina o tipo RIFIN doutras proteínas. Os parámetros dos GOINs e Centralidades utilizáronse como valores de entrada. O modelo presenta valores de sensibilidade e especificidade do 70-80% nas series de adestramento e validación externa, respectivamente. En conclusión, a relevancia biolóxica do investimento da orientación génica non depende directamente da información da secuencia xenética. A entrada é unha variable de centralidade chamada proximidade C clo . A súacentralidade mide a desviación do xene i no cromosoma k con respecto ao valor media esperado de proximidade para todos os xenes no mesmo cromosoma k. Isto é $f(v_{ij})_{ref}$ igual a $C_{clo}(Gene_i,Chr_k)$ - $<C_{clo}(Chr_k)>$.

Martínez-Arzate *et al.* desenvolveu un modelo IFPTML para descubrir novos epítopos de células B útiles para o deseño de vacinas e para predicirpuntuacións de epítopos inmunogénicos en diferentes condiciónsexperimentais. O modelo utiliza como entrada a secuencia do péptido q e a actividade do epítopo. A información recuperada contén cambios estruturais en 83683 secuencias peptídicas (Seq) determinadas en ensaiosexperimentais informados na base de datos IEDB, e involucran 1448 organismos (Org), 323 organismos hóspede (Host), 15 tipos de

procesos in vivo (Proc), 28 técnicas experimentais, (Tech), máis 505 aditivos adyuvantes (Adj). O modelo ten precisión, sensibilidade e especificidade entre 71 e 80% para adestramento e series de validación externa.

Concu *et al.* desenvolveu un modelo para predicir un conxunto de encimas que pertencenao fermento Pichia . Aplicouse a un conxunto de datos de 19 187 encimas que representan as 59 subclases presentes no Protein Data Bank (PDB). Ademais, os autores desenvolveron modelos IFPTML baseados en ANN para predicir pares encima-encima de secuencias de consulta de persoalcunha precisión, especificidade e sensibilidade superiores ao 90 % tanto para a serie de adestramento como para a de validación.

QSAR-Co é un software independente de libre acceso para levar a cabo estudosbaseados en clasificacións considerando diferentes condiciónsexperimentais segundo corresponda. Cabe sinalar que QSAR-Co é unha forma abreviada de "quimioinformática con condicións", sendo esta última unha da características clave deste software, aínda que tamén se poden desenvolver modelos de quimioinformáticabaseados en clasificación simple sen condicións. Outra razón que motivou o desenvolvementodeste software foi proporcionar unha plataforma distinta para derivar modelos quimioinformáticosbaseados en clasificación seguindo todas as pautas recomendadas pola OCDE, é dicir, modelos quimioinformáticos robustos. O software consta de dous módulos: 1) o módulo de desenvolvemento de modelos e 2) o módulo de pantalla/predición. O software ' QSAR-Co ' versión 1.0.0 é unhaferramenta independente dispoñible @gratuitamente para descargar napáxina web de QSAR-Co. Ten dous módulos ('desenvolvemento de modelos' e ' detección/predición) que están dispoñibles no software, e agora discutiremos todos os pasos e as funcionalidades asociadas en cada módulo. No módulo de 'desenvolvemento de modelos', o software proporciona todos os pasos básicos que están involucrados no desenvolvementodun modelo Cheminformatics baseado en clasificación, que taméninclúe examinar e tratar os datos de entrada para varias condiciónsexperimentais, se corresponde.

O software permite calcular operadores de medias móbiles de Box-Jenkins para descriptores moleculares. O enfoque discutiuse en detalle anteriormente. Aofacelo, calcula os descriptores de media móbil para un descriptor molecular Dei de compostosindividuais 'i'. O termo derivado

denomínase operador de Box-Jenkin, e estes descriptores modificados capturan a información sobre as estruturas químicas e o elemento específico da condición experimental (cj) baixo as cales se analizaron as mostras. Estes descriptores modificados calcúlanse mediante o software QSAR-Co e utilízanse nos pasos posteriores de desenvolvemento do modelo Cheminformatics. Opcionalmente, pódese realizar un pretratamiento de datos para eliminar os descriptores non informativos que poden non ter unha contribución significativa naconstrución do modelo. Tamén pode dividir o conxunto de datos en conxuntos de adestramento e proba, de modo que en pasos posteriores o conxunto de adestramentoempréguese para o desenvolvemento e a selección do modelo, mentres que o conxunto de proba emprégase para a validación do modelo. Hai unha opción para repetir a mesma división aleatoria para reproducir o desenvolvemento do modelo utilizando o mesmo valor inicial na configuración. Nos enfoques racionais, proporciónanse dúas técnicas no software, é dicir, o algoritmo de Kennard-Stone e o método de división baseadona distancia euclidiana. O software tamén elimina os descriptores menos discriminatorios. QSAR-Co tamén proporciona ' Algoritmo xenético ' como técnica de selección de variables para desenvolver modelos de 'Análises discriminante lineal' (LDA). O algoritmo xenético (GA) é unha técnica ben coñecida que se utiliza a miúdo no desenvolvemento de modelos de quimioinformáticabaseados en regresión, así como para o desenvolvemento de modelos de quimioinformáticabaseados en clasificación. Naactualidade, o software proporciona dúas técnicas de aprendizaxe automática para desenvolver modelos sólidos de quimioinformáticabaseados en clasificación, análise discriminante lineal (LDA), e Random Forest é un algoritmo de aprendizaxe automática supervisada que consiste nunha colección ouconxunto de predictores de árbores de decisión simples. Neste software, utilizamos a biblioteca java Weka versión 3-9-3 para realizar Random Forest. As métricas de validación como λ de Wilk proporcionan unha medida da importancia da discriminación lograda. Pódese deseñar unha matriz de confusión usando a información da clase de resposta real e preditaobtida do modelo baixoavaliación, o software tamén brinda parámetros como Sensibilidade, Especificidade, relación de Fisher *etc*., e realiza a análise da curva de características operativas do receptor (ROC). Ademais, o software tamén realiza unhaanálise do dominio de aplicabilidade

(AD). Por último, no módulo 2 do software podemos realizar análisespreditivas de novoscompostos químicos.

Doutra banda, o software IFPTML.SOFT e a súa aplicación central FRAMA software versión 1.0.0 é unha nova aplicación de escritorio en contorna Windows. Foidesenvolvido para o tratamento de información química organizada en agrupamento, e variables continuas. Utilice a información almacenada en follas de cálculo. O tempo de carga da información ata FRAMA é directamente proporcional ao tamaño do arquivo e a cantidade de información. A operación de unión de variables permite crear novas variables de agrupación unindo dúas oumáis variables existentes preferidas polo usuario. Realízase mediante procesamento por lotes na cola de operacións seleccionadas. O procesamento mencionado agrega as novas variables aoconxunto de variables dispoñibles no contexto de traballo. Unha vez subido o arquivo, e unidas as variables, procédese á selección e clasificación de variables de agrupación e variables continuas. Permítese probar se existen datos anómalos como nulos oubaleiros e tomar unha decisión sobre o seutratamento. Na agrupación de variables, os valores anómalos poden ser substituídos polo valor “MD”. En variables continuas pódese substituír pola media dos valores da columna. Tamén é posible en ambos os tipos de variables eliminar casos.

Despois de pretratar os datos, as variables seleccionadas poden someterse a operacións por lotes. En función da natureza das variables, pódense realizar operacións de agrupación de variables como Identidade, conta, probabilidade, Entropía de Shanon. Operacións de transformación de variables continuas como: Identidade, exponencial, valor absoluto, potencia numérica, logaritmo, probabilidade mínima máxima, z-score. Operacións básicas para variables continuas como: Suma, produto, diferenza, división. Operacións paramétricas como: Máximo, mínimo, media, suma, desviación estándar, multiplicacións por unha constante. Finalmente, operadores entre variables de agrupación e variables continuas como Media móbil, Suma por agrupación, Desviación estándar, Probabilidade mínima máxima, Puntuación Z. É posible establecer os valores de adestramento e validación usando un patrón de caracteres. A letra T utilízase para representar a formación e a V para a validación. O patrón predeterminado é TTTV que expresa o 75 % dos valores para adestramento e o 25 % para valores de validación. FRAMA

1.0.0 permite realizar a operación de análise de regresión lineal, baseadana selección de variables de entrada independentes e a variable dependente de saída. O procesamentodá como resultado un arquivo CSV. O arquivo resultante contén os nomes das variables, os valores das constantes, o erro estándar, o estatístico T, o valor de p (T), o estatístico F, o valor de p (F), o límite superior de confianza, o límite inferior de confianza, o índice, a intersección, Proba T, Proba F. A información de regresión móstrase con R2 (r-cadrado), r2 axustado, F-Test, Z-test, Chi-Squared Test. É posible xerar variables de media móbilmultietiquetas e operadores que se utilizarán en Perturbación Teoría da aprendizaxe automática.

Por último, a pesar da potencial aplicación na industria, non se desenvolveuningunha startup (aocomezodesta tese) para a transferencia da tecnoloxía IFPTML á industria. En consecuencia, o obxectivodesta tese céntrase en primeiro lugar no desenvolvemento de novos modelos IFPTML de ensaios de compostosanticancerígenos e compostos alostéricos. A continuación, informamos sobre o desenvolvemento do software LAGA, fácil de usar, para a predición extrapolada de compostosanticancerígenos. Por último, descríbese a planificación, creación, estrutura, servizos, etc. de IKERDATA S.L unha nova startup interuniversitaria enfocada á transferencia de tecnoloxía IFPTML a empresas galegas e do País Vasco en primeira instancia con perspectivas de proxección cara a España, Europa e a escala global en última instancia. No anexo da tese inclúese un resumo ampliado de >3000 palabras en galego (Anexo 3.2.).


UNIVERSIDADE DA CORUÑA