The published manuscript is available at EurekaSelect via https://www.eurekaselect.com/openurl/content.php?genre=article&doi=10.2174/1568026620666191226092431

Current Topics in Medicinal Chemistry, 2020, *20*, 305-317

RESEARCH ARTICLE

MCDCalc: Markov Chain Molecular Descriptors Calculator for Medicinal Chemistry

Paula Carracedo-Reboredo[1,2,3], Ramiro Corona[3], Mikel Martinez-Nunes[3], Carlos Fernandez-Lozano[1,2], Georgia Tsiliki[4], Haralambos Sarimveis[5,6], Eider Aranzamendi[3], Sonia Arrasate[3], Nuria Sotomayor[2], Esther Lete[3], Cristian Robert Munteanu[1,2] and Humbert González-Díaz[7,8,*]

[1]*Department of Computer Science and Information Technologies, Faculty of Computer Science, University of A Coruña, CITIC, Campus Elviña s/n, 15071, A Coruña, Spain;* [2]*Group of Artificial Neural Networks and Adaptative Systems, Medical Imaging, and Diagnostic Radiology (RNASA-IMEDIR), Institute of Biomedical Research of Coruña (INIBIC), Hospital Complex of University of A Coruña (CHUAC), Sergas, University of Coruña (UDC), Xubias de arriba 84, 15006, A Coruña, Spain;* [3]*Department of Organic Chemistry II, University of the Basque Country UPV/EHU, 48940, Leioa, Bilbao, Spain;* [4]*Institute for the Management of Information Systems, ATHENA Research and Innovation Centre, 15125, Athens, Greece;* [5]*School of Chemical Engineering, National Technical University of Athens, Zografou, Campus, 15780, Athens, Greece;* [6]*Pharma-Informatics Unit, ATHENA Research and Innovation Centre, 15125, Athens, Greece;* [7]*Basque Center for Biophysics, University of the Basque Country UPV/EHU, 48940, Leioa, Bilbao, Spain;* [8]*IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain*

Abstract: ***Aims*:** Cheminformatics models are able to predict different outputs (activity, property, chemical reactivity) in single molecules or complex molecular systems (catalyzed organic synthesis, metabolic reactions, nanoparticles, *etc.*).

***Background*:** Cheminformatics models are able to predict different outputs (activity, property, chemical reactivity) in single molecules or complex molecular systems (catalyzed organic synthesis, metabolic reactions, nanoparticles, *etc.*).

***Objective*:** Cheminformatics prediction of complex catalytic enantioselective reactions is a major goal in organic synthesis research and chemical industry. Markov Chain Molecular Descriptors (MCDs) have been largely used to solve Cheminformatics problems. There are different types of Markov chain descriptors such as Markov-Shannon entropies (Shk), Markov Means (Mk), Markov Moments (πk), *etc.* However, there are other possible MCDs that have not been used before. In addition, the calculation of MCDs is done very often using specific software not always available for general users and there is not an R library public available for the calculation of MCDs. This fact, limits the availability of MCMD-based Cheminformatics procedures.

***Methods*:** We studied the enantiomeric excess ee(%)[Rcat] for 324 α-amidoalkylation reactions. These reactions have a complex mechanism depending on various factors. The model includes MCDs of the substrate, solvent, chiral catalyst, product along with values of time of reaction, temperature, load of catalyst, etc. We tested several Machine Learning regression algorithms. The Random Forest regression model has R2 > 0.90 in training and test. Secondly, the biological activity of 5644 compounds against colorectal cancer was studied.

***Result*:** We developed very interesting model able to predict with Specificity and Sensitivity 70-82% the cases of preclinical assays in both training and validation series.

***Conclusion*:** The work shows the potential of the new tool for computational studies in organic and medicinal chemistry.

ARTICLE HISTORY

Received: July 25, 2019
Revised: September 17, 2019
Accepted: September 17, 2019

DOI:
10.2174/1568026620666191226092431

Keywords: Molecular descriptors, Markov chains, Singular values, Online tool, R-script, Chiral catalyst, Enantioselectivity, α-Amidoalkylation reactions, Biological activity, Colorectal cancer.

1. INTRODUCTION

Cheminformatics models are able to predict different outputs (activity, property, chemical reactivity) in complex molecular systems (metabolic reactions) [1], nanoparticles [2], *etc*. Specifically, the prediction of chemical reactivity of complex reactions in organic synthesis is a goal of major importance for both basic research and chemical industry. Cheminformatics methods may be very useful in the prediction of chemical outcomes in stereoselective reactions [3]. Sigman *et al*. reported some of the pioneer works for the prediction of enantiomeric ratios of the products [4-6]. More

*Address correspondence to this author at the Department of Organic Chemistry II, University of the Basque Country UPV/EHU, P.O.Box 644, 48080, and IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain; E-mail: humberto.gonzalezdiaz@ehu.es

1873-4294/20 $65.00+.00

recently, Cheminformatics methodologies have been applied to predict the enantioselectivity of different types of reactions. Some of these reactions are alkylation [7, 8] allylation [9], propargylation [10, 11] intramolecular carbolithiation [12], dehydrogenative Heck-type C-C and C-N coupling reactions [13-16], Heck-Heck cascade reactions [17], asymmetric copper-catalyzed cyclopropanation of alkenes [18] and Henry reaction [19]. On the other hand, colorectal cancer (CRC) is the third most commonly occurring cancer in men and the second in women, having a mortality of approximately 56 % of the patients [20] Although a number of compounds for anti-CRC activity have been synthetized and tested, the possibility of coming across an effective drug is still too low. Moreover, this process led to a notable economical and time loss [21].

Markov Chain Molecular Descriptors (MCDs) have been largely used to solve Cheminformatics problems. There are different types of Markov chain descriptors such as Markov-Markov Means (M_k), Shannon entropies (Sh_k), Markov Moments (π_k), *etc*. [22]. However, there are other possible indices that have not been used before. For instance, singular values of matrices have been used before in Cheminformatics [23]. Nevertheless, until the best of our knowledge, there are no reports of the uses of Singular Values (SV_k) of Markov matrices as molecular descriptors. In addition, the calculation of MCDs is done very often using specific software not ever available for general users and there is not an R library public available for the calculation of MCDs. This fact limits the availability of general Cheminformatics procedures for organic synthesis researchers using MCDs. In this sense, the development of new tools which are publicly available for the calculation of molecular descriptors in general and specifically of MCDs is a promising area of research. The free accessibility to these tools may promote the development of Cheminformatics models for areas of research less explored before with this type of technique.

In this work, we developed the first library in R for the calculation of MCDs. We also report here the first public web server for the calculation of MCDs online. This online tool includes the calculation of a new class of MCDs called Markov Singular indices. We report two case studies in Cheminformatics and other areas of interest can be combined with promising results. In the first case study, we illustrate the use of Markov matrix singular probability values as molecular descriptors, for the first time. With these descriptors, we modeled the enantioselective organic reactions. This constitutes a practical example of the use of the MCDs, the R library, and the online tool in Organic Chemistry and Catalysis. In the second case study, we illustrated the use of Markov matrix mean values as molecular descriptors in the study of compounds active against Colon Rectal Cancer (CRC). The works open a new paradigm on the applications of online tools to the study of either chemical reactivity or biological activity using MCDs. In (Fig. **1**), we depict the general but simplified workflow of the present paper.

2. MATERIALS AND METHODS

2.1. RMarkov.mol Package

We propose an implementation in R of the algorithm for calculation of MCDs in the RMarkov.mol package. RMarkov.mol can calculate two drug topological indices (TIs) families: Markov Mean Properties (MMPs) using RDMarkov Means function and Markov Singular Values of Transition Probabilities (MMSVs) using RDMarkov Singulars function. Both types of TIs are using molecular graph topology with 4 atom physical properties to encode molecu-

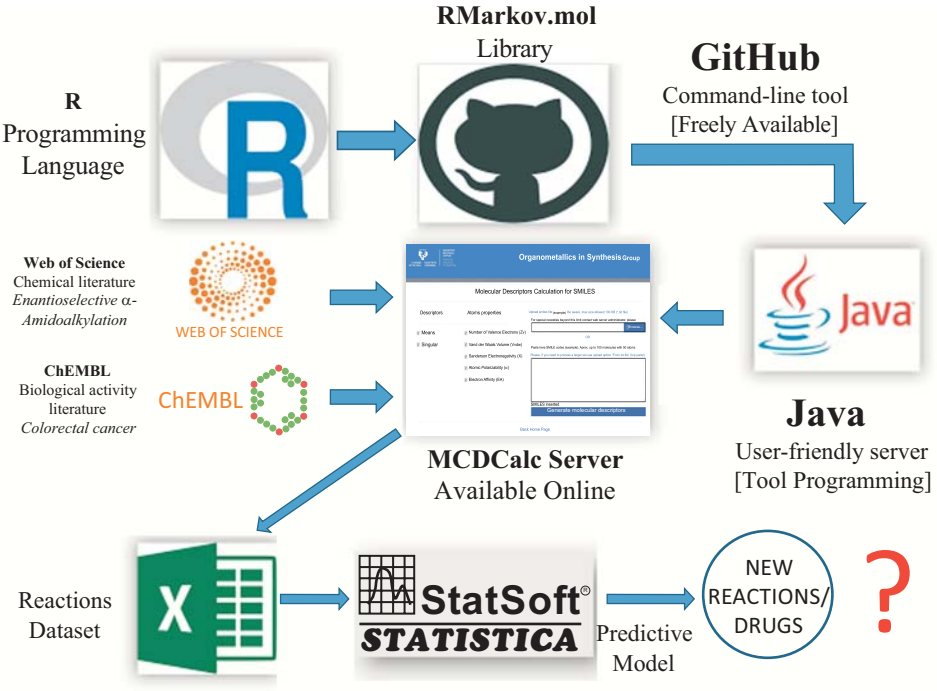


Fig. (1). General workflow of the present paper.

lar information and Markov chains theory to include atomic intra-molecular interactions. The algorithm is derived from previous python private software, MInD-Prot but the atom weights are different: number of valence electrons, van der Waals atomic radius, covalent radius, atomic mass, van der Waals volume, Sanderson electronegativity, atomic polarizability, ionization potential, and electron affinity. The open source R package is based on ChemmineR, base, expm, and MASS packages and it is available at https://www.github.com/muntisa/RMarkov.mol. The combination of RMarkov.mol with RRegrs package generates a powerful and fast R tool for designing QSAR regression models. The current study is presenting the Web interface to the R version of the tool using only the Markov Singular Values of Transition Probabilities (MMSVs). The main algorithm of RMarkov.mol has the following steps:

- Read the SMILES Formulas Inputs
- Extraction of the atom connectivity matrix.
- The atomic properties are added as weights.
- The transition probabilities matrix is calculated (Markov chains).
- Vector - matrices products are calculated using k value as matrix powers (k = distance between interacting atoms).

Calculation of molecular descriptors for each atom property and atom type, averaging the values for all ks. The included atom properties are the following: number of valence electrons (Zv), van der Waals atomic radius (Rvdw), covalent radius (Rcov), atomic mass (m), van der Waals volume (Vvdw), Sanderson electronegativity (SAe), atomic polarizability (aPolar), ionization potential (IP), and electron affinity (EA). All TIs are calculating for six types of atoms: All (all atoms), Csat (saturated C), Cuns (unsaturated C), Hal (halogen), Het (heteroatoms) and HetNoX (heteroatoms but not halogens). The user can modify the atom properties file by adding or removing any column. The descriptors are averaged for all k values (distance between atoms that are interacting). The molecular graph is defined for each molecule as the set of nodes (atoms) and edges (chemical bonds). In the case of MMSVs, additional calculations of the singular values of the transition probabilities are used. RDMarkovMeans is calculating 54 TIs (9 atom properties x 6 atom types). RDMarkovSingulars generates a different number of TIs depending on the flag fAllKs: if fAllKs=1, it calculates 540 TIs for each atom property, atom type, k value, Min and Max values + the averages for all ks; if fAllKs=0, only 108 averaged TIs are calculated. There are 288 descriptors for all k values: (4 properties*6 atom types* 6 powers) * 2 for Min and Max. Only 48 descriptors are represented by the averaged values: (4 properties*6 atom types) * 2 for Min and Max. Specifically, the Drug Markov Singular Values of Transition Probabilities are calculated using the following steps:

- Read the inputs: SMILES formulas and atom properties.
- Get connectivity matrix (CM), nodes = atoms, edges = chemical bonds.
- Get weights vector (w) for each atom property.
- Calculate weighted matrix (W) using CM and w.
- Calculate transition probability (P) based on W.
- Calculate k powers of P; the results are Pk matrices.
- Calculate Markov Singular Values for all power, each type of atom property and atom type.
- Calculate the average values over all K values (total = 336).

2.2. RMarkov.mol Library

The main functions of RMarkov.mol permits two calls, each for one single family of MCDs: DMarkovMeans and RDMarkovSingulars. The details about the parameters of the functions are presented in the R package documentation. All these parameters have specific default values such as input file name as "SMILES.txt", output file name as "RDMarkov Singulars Results.csv", power k = 3 (distance between interaction atoms) and a flag for full or averaged descriptors (only for MMSVs). The user can modify these parameters. The following examples present custom calls of the RDMarkovMeans() and RDMarkovSingulars() functions using different input and output files, and k values. The output variables DMMs and MMSVs contain data frames with all correspondent MCDs. These molecular descriptors can be coupled with a regression R package to seek new Cheminformatics models using the same language for all the process. The code of this library is as follows:

```
> library (RMarkov.mol)
> #Run RDMarkovMeans & RDMarkovSingulars with default values
> #SFile ="XXXXX", sResultFile ="XXXX", kPower = "3")
> DMMs <- RDMarkovMeans()
> MMSVs <- RDMarkovSingulars()
>
> # Run RDMarkovMeans and RDMarkovSingulars with
> # mySMILES.txt as input, myResults.csv as results, k = 4
> DMMs <- RDMarkovMeans () SFile ="mySMILES. txt " ,
> sResultFile ="myResults.csv " ,
> kPower="4")
> MMSVs <- RDMarkovSingulars (SFile ="mySMILES.txt",
> sResultFile ="myResults.csv " ,
> kPower="4")
```

3. CHEMICAL REACTIVITY (CASE STUDY)

3.1. Data for Chemical Reactivity Study

A large benchmark dataset of α-amidoalkylation reactions was used in this work. This dataset included the Brønsted acid catalyzed intermolecular α-amidoalkylation reactions developed by our group [24, 25] and literature data [26-31] for related reactions with different types of substrates

(cyclic and bicyclic hydroxylactams), nucleophiles (enamides, indoles, *etc*.) and chiral catalysts (phosphoric acids, phosphoramides, *etc*.), under different experimental conditions.

3.2. Molecular Descriptors for Chemical Reactivity Study

The descriptors used for the current models contain information about reaction factors: catalyst configuration, additive (TMSCl), temperature, time, solvent dipole, catalyst load, and drying agent. Molecular structural factors: molecule with role q-th in the reaction, substrate, product, solvent, catalyst, and nucleophile. In Table **1**, we summarized the input variables used in the linear model. The linear models have the following form.

$$f(v_{ij})_{calc} = a_0 + \sum_{k=1}^{k_{max}} b_k \cdot V_k + \sum_{k=1}^{k_{max}} c_k \cdot D_k(m_j) \quad (1)$$

Table 1. Output *vs.* Input variables used in the model.

Variable Type	Variable	Details
Output	*ee*(%)[R_{cat}]	Enantiomeric excess using *R* catalyst
Input	f_0 = (R =1/S = -1)cat	Catalyst configuration
Reaction	f_1 = TMSCl(eq)	TMSCl additive
Operation	f_2 = T(K)	Temperature
Variables	f_3 = t(h)	Reaction time
	f_4 = Ds	Solvent dipole
	f_5 = Load(%)	Catalyst load
	f_6 = Da	Drying agent
Input	SV(w, g, Sub)	Sub = Substrate (q = 0)
Chemical	SV(w, g, Prod)	Prod = Product (q = 1)
Structure	SV(w, g, Solv)	Solv = Solvent (q = 2)
Variables	SV(w, g, Cat)	Cat = Catalyst (q = 3)
	SV(w, g, Nuc)	Nuc = Nucleophile (q = 4)

SV(w, g, m_q) = Max Singular Values (SV_{max}) for molecule (m_q) with organic chemical group g and role q-th (substrate, product, solvent, *etc*.) in the reaction.

4. BIOLOGICAL ACTIVITY (CASE STUDY)

4.1. Data for Biological Activity Study

Firstly, 5644 preclinical assays of CRC active compounds were obtained 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 values of ε_{ij} depend on the structure of the drug and also on a series of boundary conditions that delimit the characteristics of the assay $\mathbf{c}_j = (c_0, c_1, c_2, \ldots c_n)$. The first c_j is c_0 = the biological activity v_{ij} (Inhibition, GI_{50}, IC_{50}, *etc*.). 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 doing so, we discretized the values as follows: $f(v_{ij})_{obs} = 1$ when v_{ij} > cutoff and desirability of the biological activity parameter $d(c_0) = 1$. 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 a 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.

4.2. PTML Linear Model

Perturbation-Theory Machine Learning (PTML) algorithm is useful to seek predictive models for complex datasets with multiple Big Data features [1, 32]. 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 $\mathbf{c}_j = (c_0, c_1, c_2, \ldots c_n)$ using as input a value of reference $f(v_{ij})_{ref}$ and the PT operators. PT operators similar to Box-Jenkins Moving Average (MA) measure the deviation of the compounds from the group of reference [33, 34]. The MA operators $\Delta D_k(c_j)_g$ are dependent on the conditions of assay $\mathbf{c}_j$, the type of the property studied k (electronegativity, polarizability, *etc*.), and the group of atoms considered g (All, Heteroatoms, *etc*.). It is possible to develop linear PTML models in order to predict the biological activity and/or classify compounds as active or non-active in terms of biological activity [35-40]. Using Linear Discriminant Analysis (LDA) [41] we can develop PTML-LDA linear classification models. PTML-LDA linear models have the following form.

$$f(v_{ij})_{calc} = a_0 + a_1 \cdot f(v_{ij})_{ref} + \sum_{k=1,j=0}^{k_{max} \cdot j_{max}} a_k \cdot \Delta D_k(w, g, \mathbf{c}_j) \quad (2)$$

5. RESULTS AND DISCUSSION

5.1. RMarkov.mol Library

In this work, we developed the first library in R for the calculation of MCDs. Several molecule descriptors could not be directly related to physical-chemical properties and the explanation of the Cheminformatics models become difficult. Therefore, the new RMarkov.mol tool implements two new classes of molecule descriptors that are based on physical-chemical atom properties. RMarkov.mol calls can be integrated into complex desktop and Web tools in Cheminformatics and combined with RRegrs (10 methods regression tool) can develop simple model building scripts. The RMarkov.mol library is available online as an open repository at https://www.github.com/muntisa/RMarkov.mol.

5.2. MCDCalc Desktop Software and Online Tool, Availability and System Requirements

In this work, we have also developed the first public tool for the calculation of MCDs online. The name of this tool is Markov Chemical Descriptors Calculator (MCDCalc). The MCDCalc web server is available online at the following link: http://oms.ehu.eus/CPTMLTool/mcdCalc. This tool is an online implementation of the R-script mentioned before using Java and JavaScript for the interactive behavior, Apache maven for project management, Spring for dependency injections and Thymeleaf as a server-side Java template engine. Operating system(s): Web service—platform inde-

pendent. Command-line tool/Library—Windows, Linux, MacOS. Programming language: R-script Java and JavaScript, Apache maven for project management, Spring for dependency injections, and Thymeleaf as a server-side Java template engine. Other requirements: Java 1.8. Any restrictions to use by non-academics: Password (available upon request to authors) is required for running or accessing the results using the web service. Permission of the authors is required for use in commercial applications. The online tools are freely available online (upon password request) at: http://oms.ehu.eus:8080/CPTMLTool/mcdCalc. The command-line tool/Library is available online at: https://www.github.com/muntisa/RMarkov.mol. The data used to train both the reactivity and biological activity models is available at Figshare project https://figshare.com/ account/home#/projects/32039. The reactivity and biological activity files have been uploaded with doi: https://doi.org/10.6084/m9.figshare.6260549.v3 and https://doi.org/10.6084/m9.figshare.7993118.v2, respectively. This data include values of T(K), t(h), Load(%)), SMILE codes (chemical structure) of substrate, product, catalyst, nucleophile, and observed *vs*. predicted values for reactivity model. The data also includes drugs, targets, cell lines, molecular descriptors, *etc*. for biological activity model.

In Fig. (**2A**), we show the user-friendly graphical interface of this webserver. In addition, we have developed a software desktop application for use offline, (Fig. **2B**). The desktop version is available upon request to the corresponding authors. We recommend the use of this version in case you have no connection to the internet or the webserver fails due to different reasons. For the use of the desktop version, the user should have the Java virtual machine installed. The executable is the file MMD.jar.

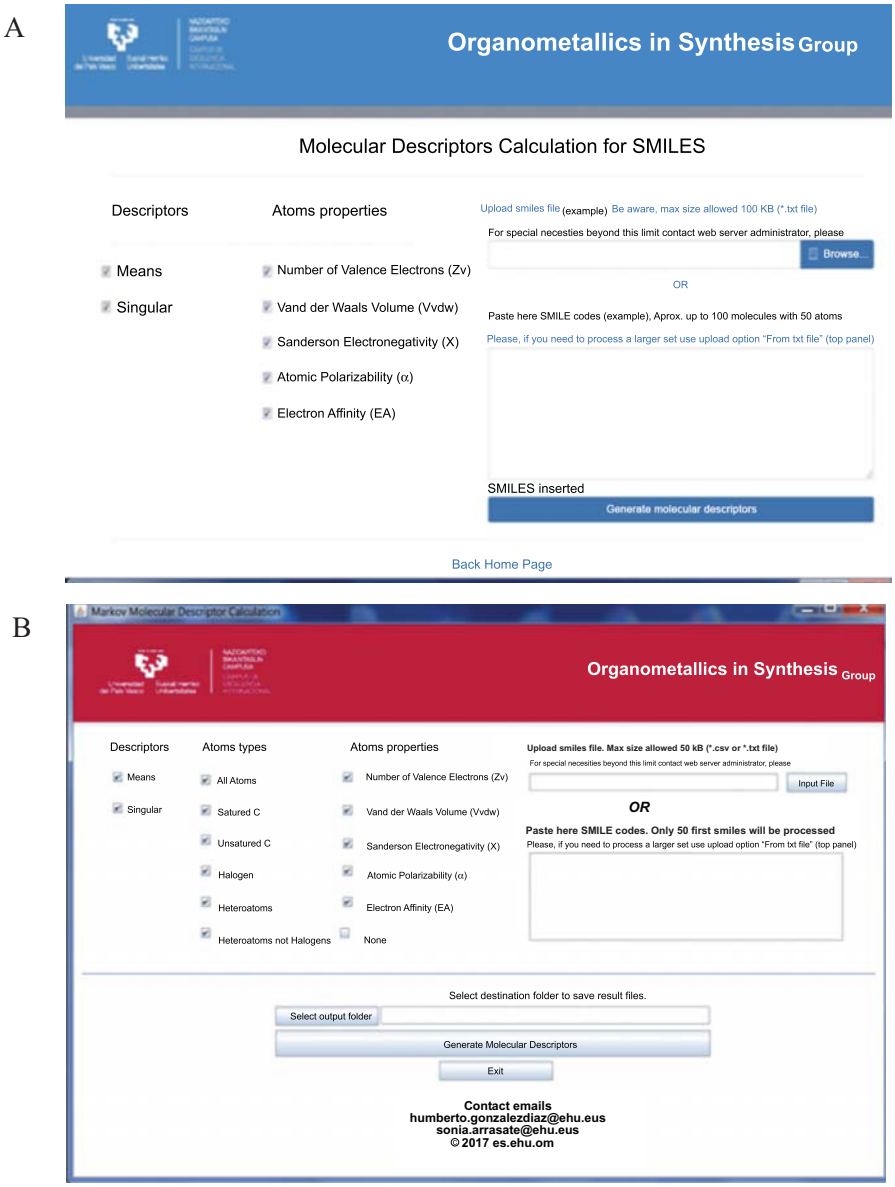
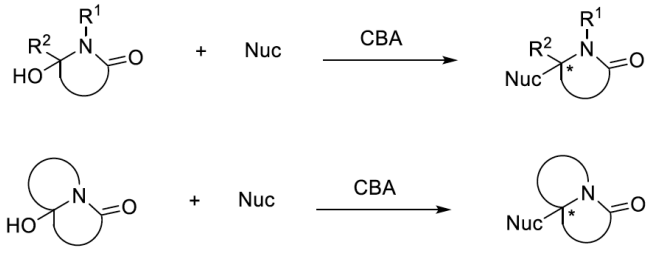


Fig. (2). (**A**) MCDCalc online tool user interface and (**B**) Executable software. *(A higher resolution / colour version of this figure is available in the electronic copy of the article).*

6. ML MODELS OF CHEMICAL REACTIVITY (CASES STUDY 1)

6.1. The α-amidoalkylation Reactions

The α -amidoalkylation reaction [42-45] is one of the most attractive methods for the stereoselective C−C bond formation and it has been widely utilized in the synthesis of a variety of complex organic molecules, including natural products and pharmaceuticals [46-48]. This method possesses several distinct advantages. The reaction is reported to have a wide nucleophile and substrate scope. In addition, the reaction is highly diastereoselective [49-51] when cyclic and bicyclic *N*-acyliminium ion intermediates are involved, which can be generated *in situ* from the corresponding hydroxylactams, using both protic and Lewis acids. This strategy is applicable to the construction of tertiary and quaternary stereocenters in an asymmetric fashion [52]. These enantioselective variants [53-55] have been developed using mainly chiral Brønsted acids (CBAs) as BINOL derived phosphoric acids and phosphoramides [56-59] as well as ureas and thioureas [60-63] as catalysts. In addition, the procedure works well with aromatics and heteroaromatics (Friedel-Crafts reactions) [64-67] enamides, silylenol ethers, *etc*. [68]. Computational chemistry has helped to understand the mechanism of these α -amidoalkylation reactions. In Scheme **1**, we depict the general idea behind the catalytic enantioselective intermolecular α -amidoalkylation reactions studied here.



Scheme 1. Catalytic enantioselective intermolecular α-amidoalkylation reactions.

6.2. Chemical Reactivity RMarkov.mol ML Linear Model

However, there are no Cheminformatics models for this reaction using MCDs. The understanding of how the different parameters affect its stereochemical outcome is still difficult to rationalize. Therefore, we sought to develop computational chemistry methods for the prediction of the enantioselectivity of this type of intermolecular α-amidoalkylation reactions. We used the previous indices calculated with RMarkov.mol as input for a Multivariate Linear Regression (MLR) model. Therefore, the resulted dataset contains 156 features and 324 examples/cases. The output of the model is the parameter *ee*(%)[R_{cat}]. This parameter is equal to the enantiomeric excess of the reaction using a catalyst of configuration *R*. Consequently, in the cases of reactions reported in the literature with *R*-catalysts *ee*(%)[R_{cat}] = *ee*(%) enantiomeric excess. Conversely, in the cases of reactions enantiomeric excess *ee*(%) reported for an *S*-catalyst *ee*(%)[R_{cat}] = - *ee*(%)[S_{cat}] = - *ee*(%). Therefore, all the values of enantiomeric excess predicted with this model are for reactions using an *R*-catalyst. In order to predict the value for S-catalyst, we only have to multiply the output of the equation by -1. The best linear model found have a $R^2 = 0.828$ with Fisher ratio F = 92.07 and $p < 0.05$. These are promising values because the model is statistically significant ($p < 0.05$) and explains more than 80% of the variance. In Table **2**, we summarized the values of the parameters for each variable in the model.

Table 2. Results of the linear model.

Input Variables	Param.	Std.Err	t	p
a_0	10347.6	2319.8	4.5	0.001
T(K)	-0.1	0.1	-1.4	0.165
t(h)	0.0	0.1	0.3	0.789
Load(%)	0.2	0.5	0.5	0.628
SV(SAe,HetNoX,Cat)	-7394.4	1881.6	-3.9	<0.05
SV(aPolar,Csat,Cat)	18.8	5.1	3.7	<0.05
SV(aPolar,HetNoX,Cat)	-947.1	108.7	-8.7	<0.05
SV(aPolar,Csat,Nuc)	-18.8	4.6	-4.1	<0.05
SV(Zv,Csat,Prod)	1128.5	70.4	16.0	<0.05
SV(aPolar,Csat,Prod)	-1035.2	114.4	-9.0	<0.05
SV(EA,Csat,Prod)	-285.3	66.9	-4.3	<0.05
SV(Vvdw,HetNoX,Sub)	-11202.3	2947.3	-3.8	0.001
SV(aPolar,HetNoX,Sub)	11109.8	3030.9	3.7	<0.05

This result confirms the hypothesis of a linear relationship between the new molecular descriptors SV_{max} and the *ee*(%)[R_{cat}] of the Brønsted acid-catalyzed α-amidoalkylation reactions. Notably, all the input variables encoding structural information (SV_{max} values) are statistically significant with p-values <0.05. However, the value of R^2 could be improved, in principle, and more importantly, some input variables are not statistically significant. For instance, T(K), t(h), Load(%) have p-values higher than 0.05. In (Fig. **3**), we depict the Pareto's chart of t-values for coefficients input variables in this model (Sigma-restricted parameterization).

6.3. Chemical Reactivity RMarkov.mol & RRegrs Models

We also used the SV_{max} values calculated with RMarkov.mol as input for the RRegrs package in order to find better regression models. Several different regression methods have been tested and the R2/RMSE results are presented in Table **3** (averaged over 10 data splits). We founded the best four results with the following regression algorithms Partial Least Squares (PLS), Neural Network (NN), Support Vector Machines (SVM), and Random Forest (RF). As shown in Fig. (**4**), the RRegrs package has been used to test

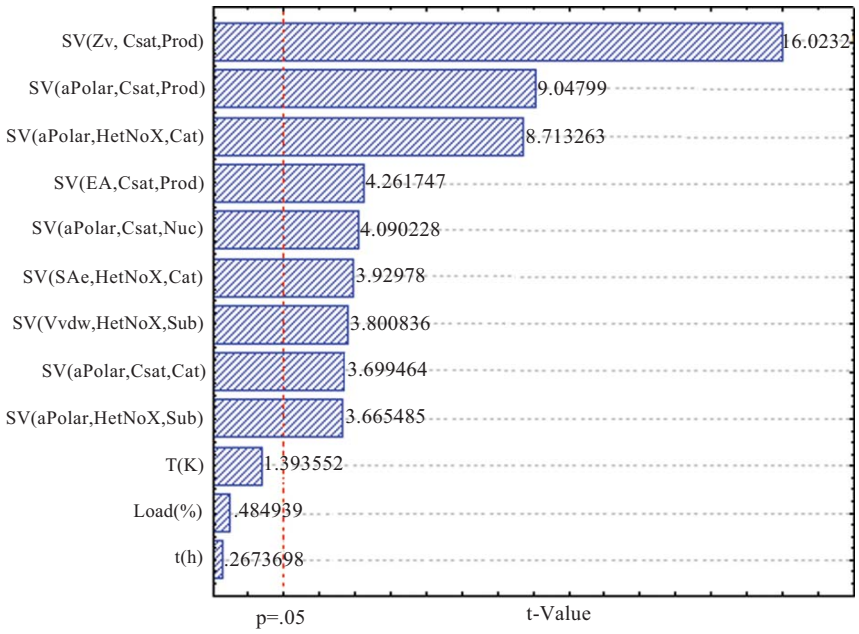


Fig. (3). Pareto chart for input variables.

Table 3. RRegrs results for enantiomeric excess using R catalyst prediction.

Method	Training		Test	
	R^2	RMSE	R^2	RMSE
RF	0.907	0.101	0.926	0.093
SVM	0.868	0.122	0.866	0.128
NN	0.849	0.132	0.829	0.143
PLS	0.801	0.155	0.775	0.167

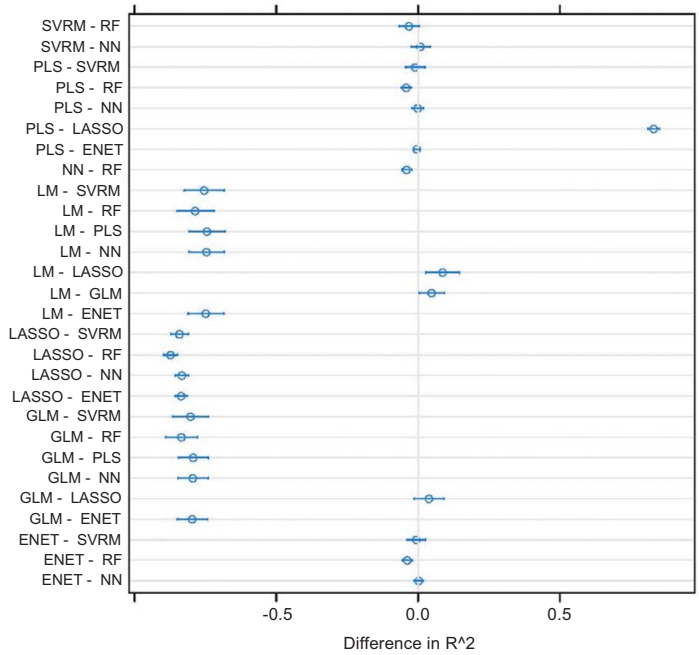


Fig. (4). Model differences for the training set (data split 10). Differences for R^2 among the package machine learning models are presented. The average performance with two-sided confidence limits is plotted as derived by the Student t-test with Bonferroni multiplicity correction.

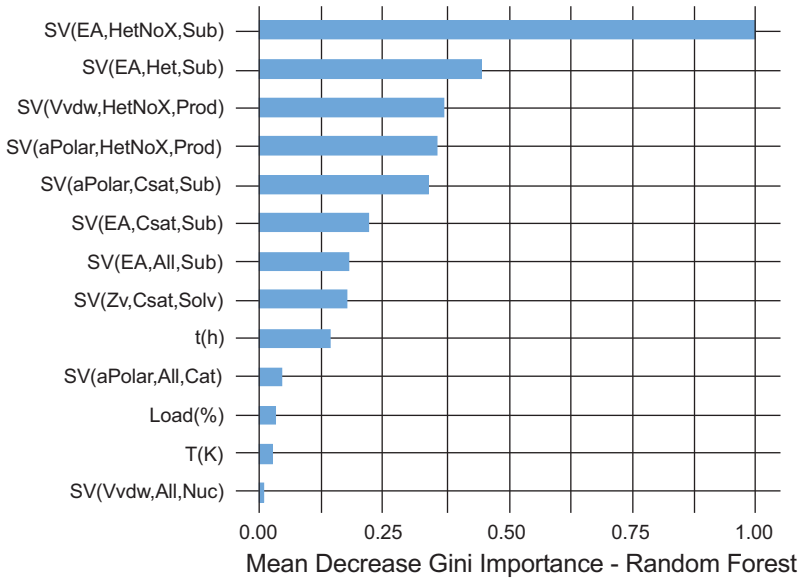


Fig. (5). Mean decrease Gini importance of the main variables selected by RF. We range this value from zero to one in order to simply understand the influence of each variable in the final model. *(A higher resolution / colour version of this figure is available in the electronic copy of the article).*

LASSO, LM, ENET and GLM. We can observe in Fig. (**4**) that Lasso Regression (LASSO), Linear Multi-regression (LM) and Generalized Linear Model (GLM) models are not fitting the dataset very well. We used as input variables SV_{max} values for all the molecules involved in the reaction: substrate, nucleophile, catalyst and product.

No correlated features have been removed. We used as input variables SV_{max} values for all the molecules involved in the reaction: substrate, nucleophile, catalyst, and product. Ten random splits of the data were performed (75% train and 25% test) along with 100 Y-randomization runs for the validation of RF, the best model.

The PLS and NN methods have values of R^2 lower or similar ($R^2 = 0.775$ and $R^2 = 0.829$) to the value for the MLR method ($R^2 = 0.828$) described before during the test phase. The SVM-radial method ($R^2 = 0.866$) has a value of R^2 slightly higher than the MLR method. The best result in the test was achieved by the RF method with $R^2 = 0.926$. This value implies that the model explains more than 90% of variance; which is above 10% more than the MLR method. Furthermore, we found three models that outperformed MLR: RF, SVM and NN. In Fig. (**5**), we summarized the main variables of the best model (RF) using the so-called Gini importance. This index can be calculated to assess the importance of each variable on the final model.

7. PTML MODEL OF BIOLOGICAL ACTIVITY (CASE STUDY 2)

7.1. PTML Model of CRC Active Compounds

PTML model correlates the expected activity value and includes different perturbation parameters in the system. Thus, the model is constructed by two types of input variables: the observed-value function $f(v_{ij})_{obs}$ and the PT operators $\Delta D(w,g,c_j)$. After different calculations, the best model was found to be the one expressed in the following equation:

$$f(\mathbf{v_{ij}})_{calc} = -2.98050 + 6.00959 \cdot f(\mathbf{v_{ij}})_{expt} + 2.95301 \cdot \Delta D(Z, Hal, \boldsymbol{c_j}) - 0.19241 \cdot \Delta D(\mathrm{Vvdw}, All, \boldsymbol{c_j}) - 4.72848 \cdot \Delta D(EA, Hal, \boldsymbol{c_j})$$

The input variable $f(v_{ij})_{obs}$ is related to the previously observed value of biological activity for the reported compound in different combinations of experimental conditions $\mathbf{c}_j = (c_0, c_1, c_2, \ldots c_j \ldots c_{max})$. In our case, PTML-LDA algorithm gave the best results, including the most important parameters that are a measured type of activity, studied cell line and assay organism. These results were directly obtained from ChEMBL data set. The reported online free available Rmarkov.mol server gave the indices, obtaining a dataset containing 30 features and 5644 cases. The output of the model $f(v_{ij})_{obs}$ combines the value v_{ij} of biological activity of the i^{th} studied compound in different combinations of conditions of assay $c_{0,1,2}$. Among all the possibilities that include the use of LM, and in this particular case, the algorithm can calculate the probabilities by using Mahalanobis's distance metric to calculate the probability [41] for a given value of $f(v_{ij})_{calc}$. The use of forward-stepwise strategy [41] of variable selection was also performed to select the more important perturbations on different conditions c_j related to anti-colorectal cancer. After calculating $p(f(v_{ij}) = 1)_{pred}$, the Boolean function $f(v_{ij})_{pred} = 1$ can easily be calculated when $p(f(v_{ij}) = 1)_{pred} > 0.5$ or $f(v_{ij})_{pred} = 0$. The values of $f(v_{ij})_{pred} = 1$ or 0 are compared with the respective observed values $f(v_{ij})_{obs} = 1$ or 0 to calculate the Sn, Sp, and Ac of the model for the selected cutoff. Finally, when $f(v_{ij})_{pred} = f(v_{ij})_{obs}$, the case can be classified as correct [41]. The presented model gave moderated values of Specificity Sp = 70.5 and high values of Sensitivity Sn = 80.2, with an overall Accuracy Ac = 74.1 in training series. The model presented slightly higher values of Sn, Sp, and Ac in the external validation series as shown in Table **4**. These values are in accordance with clas-

Table 4. Results of the model and input variables analyzed.

Obs.	Stat.	Pred.	Predicted Sets		
Sets[a]	**Param.**[a]	**Stat.**[a]	n_j	$f(v_{ij})_{pred} = 0$	$f(v_{ij})_{pred} = 1$
Training series					
$f(v_{ij})_{obs} = 0$	Sp	70.5	2685	1894	791
$f(v_{ij})_{obs} = 1$	Sn	80.2	1548	306	2033
Total	Ac	74.1	4233		
Validation series					
$f(v_{ij})_{obs} = 0$	Sp	72.1	906	653	253
$f(v_{ij})_{obs} = 1$	Sn	81.6	505	93	412
Total	Ac	75.5	1411		

[a]Obs. Sets = Observed sets, Stat. Param. = Statistical parameter, Pred. Stat. = Predicted statistics.

Table 5. One-condition averages, cutoff, desirability $d(c_0)$, *etc.*, for selected biological parameters.

Condition c_0[a]	Input Parameters Used to Specify c_0				
Activity	$n_j(c_0)$	$n_j(f(v_{ij})=1)_{obs}$	$p(f(v_{ij})=1)_{expt}$	**cutoff**	$d(c_0)$
Inhibition(%)	2744	1582	0.577	70.00	1
GI_{50}(nM)	1305	113	0.087	50	-1
IC_{50}(nM)	664	133	0.200	50	-1
TGI(nM)	314	0	0.000	50	-1
LC_{50}(nM)	237	1	0.004	50	-1
IC_{50}(ug.mL-1)	100	98	0.980	50	-1
Activity(%)	89	34	0.382	75.00	1
EC_{50}(nM)	37	2	0.054	50	-1
ED_{50}(ug ml-1)	33	26	0.788	50	-1
Ratio	25	6	0.240	32.22	1
AC_{50}(nM)	23	0	0.000	50	-1
GI(uM)	16	16	1.000	50	-1
ID_{50}(nM)	13	11	0.846	50	-1
TCS_{50}(uM)	12	9	0.750	50	-1
MG MID(uM)	11	9	0.818	50	-1
SI	11	3	0.273	1.83	1
ID_{50}(M)	10	10	1.000	50	-1

[a]Condition c_0 = the type of activity parameter measured.

sifiying the model with application in Medicinal Chemistry [69]. It is important to mention that the data points (compound-assay pair) used in validation series have not been used to train the model.

The input parameters used to specify c_0 are resumed in Table **5**. This model is useful for the prediction of the activity of new compounds for different organisms and cell lines. Moreover, this paper reported that free available Rmarkov.mol server can calculate MCDs that could help in

Table 6. Comparison to other PTML models of anti-cancer compounds.

Cancer Type	Cancers	PT[a]	ML[b]	NV[b]	Cases[c]	Sn(%)[d]	Sp(%)[d]	Refs.
Colorectal	1	MMA	LDA	4	4233*(train)*	>70	>80	This work
Colorectal	1	MA	LDA	>10	1237*(train)*	>90	>90	[74]
Breast	1	MA	LDA	>10	24285*(total)*	>90	>90	[70, 71]
Bladder	1	MA	LDA	n.a.	n.a.	>90	>90	[72]
Brain	1	MA	LDA	n.a.	n.a.	>90	>90	[73]
Breast	1	MA	LDA	>10	2272*(total)*	>85	>95	[75]
Prostate	1	MA	LDA	>10	1250*(train)*	>85	>95	[77]
Multiple	>10	MA	LDA	>10	87701*(total)*	>70	90	This work
Cancers	-	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 in training and/or validation series. [d] Approximate values for training and validation series.

the discovery of new anticolorectal cancer drugs, decreasing the compounds that may be synthesized for an active drug.

The experimental probability was calculated according to the formula $p(f(v_{ij})_{obs} =1)_{expt} = n(f(v_{ij})=1)_{obs}/n_j$, that is the ratio between the number of compounds that are up from the selected desired level of activity for each selected condition in the total number of compounds of the condition. Whenever the v_{ij} > cutoff and the desirability $d(c_0) = 1$, the compound may be selected as $f(v_{ij})_{obs}= 1$. In the same way, when v_{ij} < cutoff and $d(c_0) = -1$ the compound is selected as favourable, whereas the other cases will be assigned as $f(v_{ij})_{obs}= -1$. For this reason, $f(v_{ij})_{obs}$ has a linear dependence with the cutoff, and may be appropriately chosen. For this case study, cutoff =50 for properties with units in nM whereas in the rest of the cases, average was used.

7.2. Comparison to Other Models from the Literature

Bediaga *et al.* and Speck-Planche and Cordeiro *et al.* published before different PTML-like models for the discovery of anticancer compounds [70-77]. In Table **6**, we summarize the results obtained using these models for comparative purposes. All these PTML-like models account for perturbations (variations) on the structure of the drug and multiple assay conditions simultaneously such as target proteins, cellular lines, organisms, *etc*. We excluded classic models from the comparison because they are useful only for one specific set of conditions. Due to the difference in the datasets, this comparison focused only on the models and not on the performance of the descriptors. We can note that almost all models focus on other types of cancer. However, the model published by Speck-Planche *et al.* in 2012 is specific for CRC active compounds. It could be noted from Table **6** that our model has lower values of Sn(%) and Sp(%) but it is able to fit a training dataset above three times larger than the previous model, 4233 *vs*. 1237 preclinical assays. In this sense, the present model is expected to be able to predict a broader range of compounds and preclinical assays due to the more large and updated data set used.

CONCLUSION

MCDs have been largely used to solve Cheminformatics problems. In this work, we have developed the first library in R for the calculation of MCDs. We also report here the first public web server for the calculation of MCDs online. This online tool called MCDCalc includes the calculation of a new class of MCDs called Markov Singular values $SV_k(w,g)$ along with a classic class of MCDs called Markov mean values $D_k(w,g)$. Lastly, we have shown that SV is either useful to predict the enantiomeric excess *ee*(%)[R_{cat}] for α - amidoalkylation reactions or for the activity prediction of anti-colorectal cancer compounds. In the case study of chemical reactivity, the reactions have a complex mechanism depending on various factors. The model includes MCDs of the substrate, solvent, chiral catalyst, product along with values of time of reaction, temperature, load of catalyst, *etc.* We tested several regression algorithms. The RF regression model showed the best results. On the other hand, the case study of biological properties can lead to an alternative for the fast and rational design of colorectal cancer drug design for different organisms and cell lines.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are base of this research.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

Spanish Ministry of Economy and Competitiveness (FEDER CTQ2016-74881-P) and (CTQ2013-41229-P) and Basque Government (IT1045-16) are gratefully acknowledged for their financial support. This work is supported by "Collaborative Project in Genomic Data Integration (CICLOGEN)" PI17/01826 funded by the Carlos III Health Institute from the Spanish National Plan for Scientific and Technical Research and Innovation 2013-2016 and the European Regional Development Funds (FEDER). This project was also supported by the General Directorate of Culture, Education and University Management of Xunta de Galicia ED431D 2017/16 and "Drug Discovery Galician Network" Ref. ED431G/01 and the "Galician Network for Colorectal Cancer Research" (Ref. ED431D 2017/23) and finally by the Spanish Ministry of Economy and Competitiveness for its support with the funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503) and the European Regional Development Funds (FEDER) by the European Union.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

C.R. Munteanu acknowledges kind attention of Prof. Egon Willighagen and the financial support of the Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, Maastricht, The Netherlands, during his postdoctoral research.

REFERENCES

[1] Diez-Alarcia, R.; Yáñez-Pérez, V.; Muneta-Arrate, I.; Arrasate, S.; Lete, E.; Meana, J. J.; González-Díaz, H. Big data challenges targeting proteins in gpcr signaling pathways; combining ptml-chembl models and [35s]gtpγs binding assays. *ACS Chem. Neurosci.,* **2019**, *10*(11), 4476-4491. https://doi.org/10.1021/acschemneuro.9b00302.

[2] Santana, R.; Zuluaga, R.; Gañan, P.; Arrasate, S.; Onieva, E.; Gonzalez-Diaz, H. Designing nanoparticle release systems for drug-vitamin cancer co-therapy with multiplicative perturbation-theory machine learning (PTML) models. *Nanoscale,* **2019**, *3*(45), 21811-21823. https://doi.org/10.1039/c9nr05070a.

[3] Santiago, C. B.; Guo, J. Y.; Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.,* **2018**, *9*(9), 2398-2412. https://doi.org/10.1039/c7sc04679k.

[4] Harper, K. C.; Bess, E. N.; Sigman, M. S. Multidimensional steric parameters in the analysis of asymmetric catalytic reactions. *Nat. Chem.,* **2012**, *4*(5), 366-374. https://doi.org/10.1038/nchem.1297.

[5] Harper, K. C.; Sigman, M. S. Using physical organic parameters to correlate asymmetric catalyst performance. *J. Org. Chem.,* **2013**, *78*(7), 2813-2818. https://doi.org/10.1021/jo4002239.

[6] Bess, E. N.; Bischoff, A. J.; Sigman, M. S.; Jacobsen, E. N. Designer substrate library for quantitative, predictive modeling of reaction performance. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*(41), 14698-14703. https://doi.org/10.1073/pnas.1409522111.

[7] Huang, H.; Zong, H.; Bian, G.; Song, L. Constructing a quantitative correlation between n-substituent sizes of chiral ligands and enantioselectivities in asymmetric addition reactions of diethylzinc with benzaldehyde. *J. Org. Chem.,* **2012**, *77*(22), 10427-10434. https://doi.org/10.1021/jo3016715.

[8] Huang, H.; Zong, H.; Shen, B.; Yue, H.; Bian, G.; Song, L. QSAR analysis of the catalytic asymmetric ethylation of ketone using physical steric parameters of chiral ligand substituents. *Tetrahedron,* **2014**, *70*(6), 1289-1297. https://doi.org/10.1016/j.tet.2013.12.054.

[9] Harper, K. C.; Sigman, M. S. Predicting and optimizing asymmetric catalyst performance using the principles of experimental design and steric parameters. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*(6), 2179-2183. https://doi.org/10.1073/pnas.1013331108.

[10] Harper, K. C.; Sigman, M. S. Three-dimensional correlation of steric and electronic free energy relationships guides asymmetric propargylation. *Science,* **2011**, *333*(6051), 1875-1878. https://doi.org/10.1126/science.1206997.

[11] Harper, K. C.; Vilardi, S. C.; Sigman, M. S. Prediction of catalyst and substrate performance in the enantioselective propargylation of aliphatic ketones by a multidimensional model of steric effects. *J. Am. Chem. Soc.,* **2013**, *135*(7), 2482-2485. https://doi.org/10.1021/ja4001807.

[12] Munteanu, C. R.; Dorado, J.; Pazos-Sierra, A.; Prado-Prado, F.; Pérez-Montoto, L. G.; Vilar, S.; Ubeira, F. M.; Sanchez-Gonzaléz, A.; Cruz-Monteagudo, M.; Arrasate, S. Markov entropy centrality: chemical, biological, crime, and legislative networks. In *towards an information theory of complex networks: statistical methods and applications*; Dehmer, M., Emmert-Streib, F., Mehler, A., Eds.; Birkhäuser Boston: Boston, **2011**; pp 199-258. https://doi.org/10.1007/978-0-8176-4904-3_9.

[13] Zhang, C.; Santiago, C. B.; Crawford, J. M.; Sigman, M. S. Enantioselective dehydrogenative heck arylations of trisubstituted alkenes with indoles to construct quaternary stereocenters. *J. Am. Chem. Soc.,* **2015**, *137*(50), 15668-15671. https://doi.org/10.1021/jacs.5b11335.

[14] Zhang, C.; Santiago, C. B.; Kou, L.; Sigman, M. S. Alkenyl carbonyl derivatives in enantioselective redox relay heck reactions: accessing α,β-unsaturated systems. *J. Am. Chem. Soc.,* **2015**, *137*(23), 7290-7293. https://doi.org/10.1021/jacs.5b04289.

[15] Milo, A.; Neel, A. J.; Toste, F. D.; Sigman, M. S. A data-intensive approach to mechanistic elucidation applied to chiral anion catalysis. *Science,* **2015**, *347* (6223), 737-743. https://doi.org/10.1126/science.1261043.

[16] Park, Y.; Niemeyer, Z. L.; Yu, J. Q.; Sigman, M. S. Quantifying structural effects of amino acid ligands in pd(ii)-catalyzed enantioselective c-h functionalization reactions. *Organometallics,* **2018**, *37*(2), 203-210. https://doi.org/10.1021/acs.organomet.7b00751.

[17] Blázquez-Barbadillo, C.; Aranzamendi, E.; Coya, E.; Lete, E.; Sotomayor, N.; González-Díaz, H. Perturbation theory model of reactivity and enantioselectivity of palladium-catalyzed heck-heck cascade reactions. *RSC Adv.,* **2016**, *6*(45), 38602-38610. https://doi.org/10.1039/c6ra08751e.

[18] Aguado-Ullate, S.; Urbano-Cuadrado, M.; Villalba, I.; Pires, E.; García, J. I.; Bo, C.; Carbó, J. J. Predicting the enantioselectivity of the copper-catalysed cyclopropanation of alkenes by using quantitative quadrant-diagram representations of the catalysts. *Chem. - A Eur. J.,* **2012**, *18*(44), 14026-14036. https://doi.org/10.1002/chem.201201135.

[19] Huang, H.; Zong, H.; Bian, G.; Yue, H.; Song, L. Correlating the effects of the n-substituent sizes of chiral 1,2-amino phosphinamide ligands on enantioselectivities in catalytic asymmetric henry reaction using physical steric parameters. *J. Org. Chem.,* **2014**, *79*(20), 9455-9464. https://doi.org/10.1021/jo500982j.

[20] Riihimaki, M.; Hemminki, A.; Sundquist, J.; Hemminki, K. Patterns of metastasis in colon and rectal cancer. *Sci. Rep.,* **2016**, *6*, 1-9. https://doi.org/10.1038/srep29765.

[21] Jhanwar, B.; Sharma, V.; Singla, R. K.; Shrivastava, B. QSAR - hansch analysis and related approaches in drug design. *Pharmacol. Online Newsl.,* **2011**, *1*, 306-344.

[22] Riera-Fernández, I.; Martín-Romalde, R.; Prado-Prado, F. J.; Escobar, M.; Munteanu, C. R.; Concu, R.; Duardo-Sanchez, A.; González-Díaz, H. From QSAR models of drugs to complex networks: state-of-art review and introduction of new markov-spectral moments indices. *Curr. Top. Med. Chem.,* **2012**, *8*, 927-960.

[23] Hull, R. D.; Fluder, E. M.; Singh, S. B.; Nachbar, R. B.; Kearsley, S. K.; Sheridan, R. P. (LaSSI) and comparison to TOPOSIM. *Society,* **2001**, 1185-1191.

[24] Aranzamendi, E.; Sotomayor, N.; Lete, E. Brønsted acid catalyzed enantioselective α -amidoalkylation in the synthesis of isoindoloisoquinolines. *J. Org. Chem.,* **2012**, *77*(6), 2986-2991. https://doi.org/10.1021/jo3000223.

[25] Aranzamendi, E.; Arrasate, S.; Sotomayor, N.; González-Díaz, H.; Lete, E. Chiral brønsted acid-catalyzed enantioselective α -amidoalkylation reactions: a joint experimental and predictive study. *ChemistryOpen,* **2016**, *5*(6), 540-549. https://doi.org/10.1002/open.201600120.

[26] Guo, Q. X.; Peng, Y. G.; Zhang, J. W.; Song, L.; Feng, Z.; Gong, L. Z. Highly enantioselective alkylation reaction of enamides by bronsted-acid catalysis. *Org. Lett.,* **2009**, *11*(20), 4620-4623. https://doi.org/10.1021/ol901892s.

[27] Xie, Y.; Zhao, Y.; Qian, B.; Yang, L.; Xia, C.; Huang, H. Enantioselective N-H functionalization of indoles with α,β- unsaturated γ-lactams catalyzed by chiral brønsted acids. *Angew. Chemie - Int. Ed.,* **2011**, *50*(25), 5682-5686. https://doi.org/10.1002/anie.201102046.

[28] Yu, X.; Lu, A.; Wang, Y.; Wu, G.; Song, H.; Zhou, Z.; Tang, C. Chiral phosphoric acid catalyzed asymmetric friedel-crafts alkylation of indole with 3-hydroxyisoindolin-1-one: enantioselective synthesis of 3-indolyl-substituted isoindolin-1-ones. *European J. Org. Chem.,* **2011**, *5,* 892-897. https://doi.org/10.1002/ejoc.201001408.

[29] Guo, C.; Song, J.; Huang, J. Z.; Chen, P. H.; Luo, S. W.; Gong, L. Z. Core-Structure-Oriented Asymmetric Organocatalytic Substitution of 3-Hydroxyoxindoles: Application in the Enantioselective Total Synthesis of (+)-Folicanthine. *Angew. Chemie - Int. Ed.,* **2012**, *51*(4), 1046-1050. https://doi.org/10.1002/anie.201107079.

[30] Yin, Q.; Wang, S. G.; You, S. L. Asymmetric synthesis of tetrahydro-β-carbolines *via* chiral phosphoric acid catalyzed transfer hydrogenation reaction. *Org. Lett.,* **2013**, *15*(11), 2688-2691. https://doi.org/10.1021/ol400995c.

[31] Courant, T.; Kumarn, S.; He, L.; Retailleau, P.; Masson, G. Chiral phosphoric acid-catalyzed enantioselective aza-friedel-crafts alkylation of indoles with γ -hydroxy-γ-lactams. *Adv. Synth. Catal.,* **2013**, *355*(5), 836-840. https://doi.org/10.1002/adsc.201201008.

[32] González-Díaz, H.; Pérez-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. *J. Immunol. Res.,* **2014**, *2014*. https://doi.org/10.1155/2014/768515.

[33] 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*(14), 1713-1741. https://doi.org/10.2174/1568026611313140011.

[34] Martínez-Arzate, S. G.; Tenorio-Borroto, E.; Barbabosa Pliego, A.; Díaz-Albiter, H. M.; Vázquez-Chagoyán, J. C.; González-Díaz, H. PTML model for proteome mining of b-cell epitopes and theoretical-experimental study of bm86 protein sequences from colima, mexico. *J. Proteome Res.,* **2017**, *16*(11), 4093-4103. https://doi.org/10.1021/acs.jproteome.7b00477.

[35] 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. *Curr. Protein Pept. Sci.,* **2016**, *17*(3), 220-227. https://doi.org/10.2174/1389203717999160226173500.

[36] Romero-Durán, F. J.; Alonso, N.; Yañez, M.; Caamaño, O.; García-Mera, X.; González-Díaz, H. Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of tvp1022 derivatives. *Neuropharmacology,* **2016**, *103*, 270-278. https://doi.org/10.1016/j.neuropharm.2015.12.019.

[37] Kleandrova, V. V.; Luan, F.; González-Díaz, 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. *Environ. Sci. Technol.,* **2014**, *48*(24), 14686-14694. https://doi.org/10.1021/es503861x.

[38] Luan, F.; Kleandrova, V. V.; González-Díaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. N. D. S. Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel qstr-perturbation approach. *Nanoscale,* **2014**, *6*(18), 10623-10630. https://doi.org/10.1039/c4nr01285b.

[39] Alonso, N.; Caamaño, O.; Romero-Duran, F. J.; Luan, F.; D. S. Cordeiro, M. N.; Yañez, M.; González-Díaz, H.; García-Mera, X. Model for high-throughput screening of multitarget drugs in chemical neurosciences: synthesis, assay, and theoretic study of rasagiline carbamates. *ACS Chem. Neurosci.,* **2013**, *4*(10), 1393-1403. https://doi.org/10.1021/cn400111n.

[40] Hill, T.; Lewicki, P.; Lewicki, P. *Statistics: methods and applications : a comprehensive reference for science, industry, and data mining*; StatSoft, **2006**.

[41] Speckamp, W. N.; Moolenaar, M. J. New developments in the chemistry of n-acyliminium ions and related intermediates. *Tetrahedron,* **2000**, *56*(24), 3817-3856. https://doi.org/10.1016/S0040-4020(00)00159-9.

[42] Yazici, A.; Pyne, S. G. Intermolecular addition reactions of a-acyliminium ions (Part I). *Synthesis (Stuttg).,* **2009**, *3*, 339-368. https://doi.org/10.1055/s-0028-1083325.

[43] Yazici, A.; Pyne, S. G. Intermolecular addition reactions of a-acyliminium ions (Part II). *Synthesis (Stuttg).,* **2009**, *4*, 513-541. https://doi.org/10.1055/s-0028-1083346.

[44] Martínez-Estibalez, U.; Gõmez-Sanjuan, A.; García-Calvo, O.; Aranzamendi, E.; Lete, E.; Sotomayor, N. Strategies based on aryllithium and n-acyliminium ion cyclizations for the stereocontrolled synthesis of alkaloids and related systems. *European J. Org. Chem.,* **2011**, *20-21*, 3610-3633. https://doi.org/10.1002/ejoc.201100123.

[45] Nielsen, T. E.; Meldal, M. Solid-phase synthesis of complex and pharmacologically interesting heterocycles. *Curr. Opin. Drug Discov. Dev.,* **2009**, *12*(6), 798-810. https://doi.org/10.1002/chin.201019251.

[46] Avendaño López, C.; de la Cuesta, E. Synthetic chemistry with n-acyliminium ions derived from piperazine-2,5-diones and related compounds. *Curr. Org. Synth.,* **2009**, *6*, 143-168. https://doi.org/10.2174/157017909788167310.

[47] Merad, J.; Lalli, C.; Bernadat, G.; Maury, J.; Masson, G. Enantioselective brønsted acid catalysis as a tool for the synthesis of natural products and pharmaceuticals. *Chem. - A Eur. J.,* **2018**, *24* (16), 3925-3943. https://doi.org/10.1002/chem.201703556.

[48] Osante, I.; Collado, M. I.; Lete, E.; Sotomayor, N. Cheminform abstract: stereodivergent synthesis of hetero-fused isoquinolines by acyliminium and metalation methods. *ChemInform,* **2010**, *33*(13). https://doi.org/10.1002/chin.200213151.

[49] González-Temprano, I.; Osante, I.; Lete, E.; Sotomayor, N. Enantiodivergent synthesis of pyrrolo[2,1-α]isoquinolines based on diastereoselective parham cyclization and α -amidoalkylation reactions. *J. Org. Chem.,* **2004**, *69*(11), 3875-3885. https://doi.org/10.1021/jo049672o.

[50] Abdullah, M. N.; Arrasate, S.; Lete, E.; Sotomayor, N. Stereoselective synthesis of thiaerythrinanes based on an α -amidoalkylation/rcm approach. *Tetrahedron,* **2008**, *64*(7), 1323-1332. https://doi.org/10.1016/j.tet.2007.11.053.

[51] Lee, Y. S.; Alam, M. M.; Keri, R. S. Enantioselective reactions of n-acyliminium ions using chiral organocatalysts. *Chem. - An Asian J.,* **2013**, *8*(12), 2906-2919. https://doi.org/10.1002/asia.201300814.

[52] Akiyama, T. Science of Synthesis: Asymmetric Organocatalysis In: *Bronsted Base and Acid Catalysts, and Additional Topics; List, B.*, Maruoka, K., Eds.; Georg Thieme Verlag: New York, **2012**; pp 169-217.

[53] Terada, M.; Momiyama, N. Asymmetric Organocatalysis In: *Brønsted Base and Acid Catalysts, and Additional Topic*s; Maruoka, K., Ed.; Georg Thieme VerlagL New York, **2012**; Vol. *2,* pp 219-278.

[54] Dalpozzo, R. Strategies for the asymmetric functionalization of indoles: an update. *Chem. Soc. Rev.,* **2015**, *44*(3), 742-778. https://doi.org/10.1039/c4cs00209a.

[55] Akiyama, T. Stronger Bronsted Acids. *Chem. Rev.,* **2007**, *107* (12), 5744-5758.

[56] Akiyama, T.; Mori, K. Stronger Brønsted Acids: Recent Progress. *Chem. Rev.,* **2015**, *115*(17), 9277-9306. https://doi.org/10.1021/acs.chemrev.5b00041.

[57] Parmar, D.; Sugiono, E.; Raja, S.; Rueping, M. Complete field guide to asymmetric binol-phosphate derived brønsted acid and metal catalysis: history and classification by mode of activation; brønsted acidity, hydrogen bonding, ion pairing, and metal phosphates. *Chem. Rev.,* **2014**, *114*(18), 9047-9153. https://doi.org/10.1021/cr5001496.

[58] Parmar, D.; Sugiono, E.; Raja, S.; Rueping, M. Erratum: complete field guide to asymmetric binol-phosphate derived brønsted acid and metal catalysis: history and classification by mode of activation; brønsted acidity, hydrogen bonding, ion pairing, and metal phosphates (chemical reviews (2014) 114:18. *Chem. Rev.,* **2017**, *117*(15), 10608-10620. https://doi.org/10.1021/acs.chemrev.7b00197.

[59] Takemoto, Y. Recognition and activation by ureas and thioureas: stereoselective reactions using ureas and thioureas as hydrogen-bonding donors. *Org. Biomol. Chem.,* **2005**, *3*(24), 4299-4306. https://doi.org/10.1039/b511216h.

[60] Doyle, A. G.; Jacobsen, E. N. Small-molecule h-bond donors in asymmetric catalysis. *Chem. Rev.,* **2007**, *107*(12), 5713-5743. https://doi.org/10.1021/cr068373r.

[61] Knowles, R. R.; Jacobsen, E. N. Attractive noncovalent interactions in asymmetric catalysis: links between enzymes and small molecule catalysts. *Proc. Natl. Acad. Sci.,* **2010**, *107*(48)„ 20678 LP-20685. https://doi.org/10.1073/pnas.1006402107.

[62] Jakab, G.; Schreiner, P. R. Comprehensive Enantioselective Organocatalysis; Dalpozzo, R., Ed.; Wiley-VCH: Weinheim, **2013**, Vol. *2,* pp 315-341.

[63] Terrasson, V.; De Figueiredo, R. M.; Campagne, J. M. Organocatalyzed Asymmetric Friedel-Crafts Reactions. *European J. Org. Chem.,* **2010**, *14,* 2635-2655. https://doi.org/10.1002/ejoc.200901492.

[64] Zeng, M.; You, S. L. Asymmetric friedel-crafts alkylation of indoles: the control of enantio- and regioselectivity. *Synlett,,* **2010**, *9*, 1289-1301. https://doi.org/10.1055/s-0029-1219929.

[65] de Figueiredo, R. M.; Campagne, J. M. Comprehensive Enantioselective Organocatalysis; Dalko, P. I., Ed.; Wiley-VCH: Weinheim„ **2013**, *3,* pp 1043-1066.

[66] P. Beletskaya, I.; D. Averin, A. Asymmetric friedel-crafts reactions of indole and its derivatives. *Curr. Organocatalysis,* **2015**, *3*(1), 60-83. https://doi.org/10.2174/2213337202666150505230013.

[67] Mazurkiewicz, R.; Październiok-Holewa, A.; Adamek, J.; Zielińska, K. *α-Amidoalkylating Agents: Structure, Synthesis, Reactivity and Application*; Elsevier: Amsterdam„ **2014**. https://doi.org/10.1016/B978-0-12-420160-6.00002-1.

[68] Marrero-Ponce, Y.; Siverio-Mota, D.; Gálvez-Llompart, M.; Recio, M. C.; Giner, R. M.; García-Domnech, R.; Torrens, F.; Arán, V. J.; Cordero-Maldonado, M. L.; Esguera, C. V. Discovery of novel anti-inflammatory drug-like compounds by aligning *in silico* and in vivo screening: the nitroindazolinone chemotype. *Eur. J. Med. Chem.,* **2011**, *46*(12), 5736-5753. https://doi.org/10.1016/j.ejmech.2011.07.053.

[69] Speck-Planche, A.; Cordeiro, M. N. D. S. Erratum to: Fragment-Based *in Silico* Modeling of Multi-Target Inhibitors against Breast Cancer-Related Proteins, *Mol. Divers.,* **2017**, *21*(3), 525. https://doi.org/10.1007/s11030-017-9766-3.

[70] Speck-Planche, A.; Cordeiro, M. N. D. S. Fragment-based *in silico* modeling of multi-target inhibitors against breast cancer-related proteins. *Mol. Divers.,* **2017**, *21*(3), 511-523. https://doi.org/10.1007/s11030-017-9731-1.

[71] Speck-Planche, A.; Kleandrova, V. V; Luan, F.; Cordeiro, M. N. D. S. Unified multi-target approach for the rational *in silico* design of anti-bladder cancer agents. *Anticancer. Agents Med. Chem.,* **2013**, *13*(5), 791-800. https://doi.org/10.2174/1871520611313050013.

[72] Speck-Planche, A.; V. Kleandrova, V.; Luan, F.; Natalia D. S. Cordeiro, M. 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*(6), 678-685. https://doi.org/10.2174/187152012800617722.

[73] Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. Rational drug design for anti-cancer chemotherapy: multi-target qsar models for the *in silico* discovery of anti-colorectal cancer agents. *Bioorganic Med. Chem.,* **2012**, *20*(15), 4848-4855. https://doi.org/10.1016/j.bmc.2012.05.071.

[74] Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. 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*(1), 273-279. https://doi.org/10.1016/j.ejps.2012.04.012.

[75] D. S. Cordeiro, M. N.; Speck-Planche, A. Editorial: computer-aided drug design, synthesis and evaluation of new anti-cancer drugs. *Curr. Top. Med. Chem.,* **2012**, *12*(24), 2703-2704. https://doi.org/10.2174/1568026612129990082.

[76] Speck-Planche, A. Multiple perspectives in anti-cancer drug discovery: from old targets and natural products to innovative computational approaches. *Anticancer. Agents Med. Chem.,* **2019**, *19*(2), 146-147. https://doi.org/10.2174/1871520619021904181050 54.

[77] Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. Multi-target drug discovery in anti-cancer therapy: fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorganic Med. Chem.,* **2011**, *19*(21), 6239-6244. https://doi.org/10.1016/j.bmc.2011.09.015.