biology

MDPI

Article

Prediction of Antimalarial Drug-Decorated Nanoparticle Delivery Systems with Random Forest Models

Diana V. Urista [1], Diego B. Carrué [2], Iago Otero [2], Sonia Arrasate [1], Viviana F. Quevedo-Tumailli [2,3], Marcos Gestal [2,4], Humbert González-Díaz [1,5,6] and Cristian R. Munteanu [2,4,*]

1 Department of Organic Chemistry II, University of Basque Country (UPV/EHU), Sarriena w/n, 48940 Leioa, Spain; diana_urista_marquez@hotmail.com (D.V.U.); sonia.arrasate@ehu.es (S.A.); humberto.gonzalezdiaz@ehu.es (H.G.-D.)
2 RNASA-IMEDIR, Computer Science Faculty, CITIC, University of A Coruna, Campus Elviña s/n, 15071 A Coruña, Spain; d.bcarrue@gmail.com (D.B.C.); i.otero.coto@gmail.com (I.O.); viviana.quevedo@udc.es (V.F.Q.-T.); marcos.gestal@udc.es (M.G.)
3 Universidad Estatal Amazónica UEA, Km. 2 1/2 vía Puyo a Tena (paso lateral), Puyo 160150, Pastaza, Ecuador
4 Biomedical Research Institute of A Coruña (INIBIC), Hospital Teresa Herrera, Xubias de Arriba 84, 15006 A Coruña, Spain
5 IKERBASQUE, Basque Foundation for Science, Alameda Urquijo 36, 48011 Bilbao, Spain
6 Basque Centre for Biophysics CSIC-UPVEHU, University of Basque Country UPV/EHU, Barrio Sarriena, 48940 Leioa, Spain
* Correspondence: c.munteanu@udc.es

Received: 24 June 2020; Accepted: 27 July 2020; Published: 30 July 2020

Abstract: Drug-decorated nanoparticles (DDNPs) have important medical applications. The current work combined Perturbation Theory with Machine Learning and Information Fusion (PTMLIF). Thus, PTMLIF models were proposed to predict the probability of nanoparticle–compound/drug complexes having antimalarial activity (against Plasmodium). The aim is to save experimental resources and time by using a virtual screening for DDNPs. The raw data was obtained by the fusion of experimental data for nanoparticles with compound chemical assays from the ChEMBL database. The inputs for the eight Machine Learning classifiers were transformed features of drugs/compounds and nanoparticles as perturbations of molecular descriptors in specific experimental conditions (experiment-centered features). The resulting dataset contains 107 input features and 249,992 examples. The best classification model was provided by Random Forest, with 27 selected features of drugs/compounds and nanoparticles in all experimental conditions considered. The high performance of the model was demonstrated by the mean Area Under the Receiver Operating Characteristics (AUC) in a test subset with a value of 0.9921 ± 0.000244 (10-fold cross-validation). The results demonstrated the power of information fusion of the experimental-centered features of drugs/compounds and nanoparticles for the prediction of nanoparticle–compound antimalarial activity. The scripts and dataset for this project are available in the open GitHub repository.

Keywords: decorated nanoparticles; drug delivery; antimalarial compounds; big data; Perturbation Theory; Machine Learning; ChEMBL database

1. Introduction

Drug-decorated nanoparticles (DDNPs) are among the most interesting nanomaterials, with a broad range of medical applications. Many of them are used in drug delivery systems for different

types of chemical compounds. These systems have numerous advantages, since there are countless combinations of drugs and nanoparticles that can be effective in treating different conditions. At the same time, they have some weaknesses. For example, the synthesis of nanoparticles can sometimes be expensive, or it can involve a lot of time that can increase with the number of samples. For this reason, in order to improve the possibility of forming effective pairs, there is a need for computational models.

Recently, some researches have been focusing on finding DDNPs that show antimalarial properties. For instance, silver and gold nanoparticles, that were synthesized from leaf and bark extracts of Myrtaceae, exhibited an effective antiplasmodial activity [1], and exopolysaccharide coated ZnO nanoparticles (EPS-ZnO NPs) presented functional effects against malaria vectors [2]. Therefore, this study aims to design a useful computational model that allows a good prediction of the antimalarial activity of varied drug–nanoparticle pairs.

Moreover, a brand new method for data fusion in nanotechnology, bio-molecular sciences, chemistry and big data analysis has been proposed in different works: it integrates Perturbation Theory (PT) and Machine Learning (ML) [3–13], using distinct PT operators to analyze changes in the varied non-structural and structural conditions of a test at once (PTML). A few of these PT operators represent the generalization of a classic cheminformatics approach introduced by Corwin Hansch [14]. He noticed the significant potential of using predictive methodologies to resolve multivariate questions in medicinal chemistry. Hansch's classic approach allows one to search for models with multiple physicochemical conditions so as to foretell the biological activity of compounds, and these models possibly include quadratic and/or linear terms. In this process, which is a Linear Free Energy Relationship (LFER) model, most of the terms are physicochemical parameters linked with the free energy of drug ionization, binding, transport, etc. In addition, because we are fusing the information (IF) of drugs and nanoparticles, the model becomes a PTMLIF (PTML + IF).

As an illustration, the logarithmic term (logP) of the octanol/water partition coefficient (P) is presented as an estimate of the free energy of drug transport and molecular lipophilicity [15]. We can approximate logP values via chemical fragment methods (such as CLogP), or via atomic methods (such as ALogP or XLogP) [16,17]. The logarithmic terms of acidity constants (pKa) are connected to the free energy of drug ionization. Additionally, to account for more molecular properties, we can use different parameters like Polar Surface Area (PSA). Generally, for a given molecule m_i, we can utilize as input for the model several types of molecular properties, taking into account measures of molecular polarizability, lipophilicity, electronegativity, etc. [17,18]. We can define these models as:

$$f(\varepsilon_i) = \sum_{k=1}^{kmax} a_k \cdot D_k(m_i) + \sum_{k=1}^{kmax} b_k \cdot D_k(m_i)^2 + e_0 \quad (1)$$

where ε_i is the biological activity of the molecule m_i, $f(\varepsilon_i)$ is a function of the variable ε_i, $D_k(m_i)$ are the molecular descriptors of m_i, and a_k and b_k are the coefficients. This classical model works to account for changes in the chemical structure of the drug/compound using the molecular descriptors, but it does not take into account the result regarding the drug activity of perturbations in multiple experimental conditions (c_j). These include assay conditions or changes in drug chemical structure, such as c_0 = the biological parameter used (CC_{50}: the ratio of the 50% cytotoxic concentration, IC_{50}: inhibition concentration, etc.), c_1 = organism, c_2 = cell name, c_3 = assay organism, etc. An example is the large datasets found in the public database ChEMBL [19–25]. We used PTML methods to analyze a large set of over 50,000 preclinical assays of drugs. These assays incorporate drugs targeting Plasmodium. The PTML model proceeds from the classic LFER approach for drug activity. We combined the use of eight Machine Learning methods with feature selection in order to obtain a more accurate classifier for our task.

2. Materials and Methods

Figure 1 presents the methodology used to build the PTMLIF classifier for the antimalarial activity of DDNPs. The methodology flow contains the following steps: (1) Get 23 properties of nanoparticle and anti-malaria drugs/compounds from the literature and public databases as initial molecular descriptors; (2) Fuse information about experimental conditions and the properties of anti-malaria drugs/compounds and nanoparticles, using an experimental-centered transformation of the original features (Box–Jenkins Moving Average operators); (3) Integrate drug/compound and nanoparticle data into the study dataset; (4) Build the baseline PTMLIF models using default parameters of the ML methods; (5) Improve the performance of the best classifier by using only the most important features (feature selection).

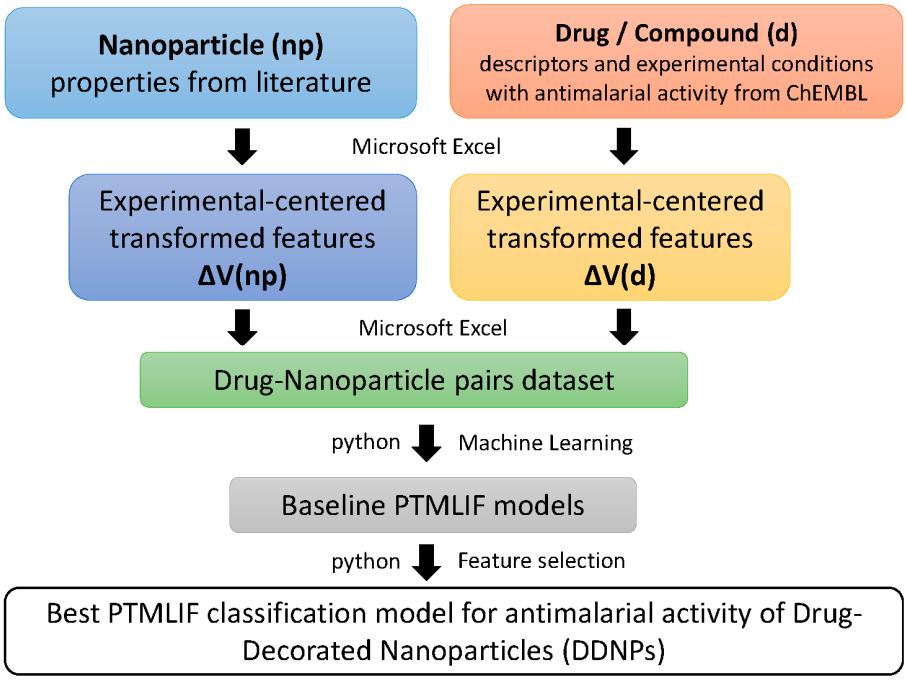

Figure 1. Workflow for development of the Perturbation Theory-Machine Learning (PTML) model.

The initial molecular descriptors were Mw, PSA and ALOGP (3 descriptors for the ChEMBL compounds/small molecules), and NMUnp, Lnp, Vnpu, Enpu, Pnpu, Uccoat, Uicoat, Hycoat, AMRcoat, TPSA(NO)coat, TPSA(Tot)coat, ALOGPcoat, ALOGP2coat, SAtotcoat, SAacccoat, SAdoncoat, Vxcoat, VvdwMGcoat, VvdwZAZcoat and PDIcoat (20 descriptors for nanoparticles). The following abbreviations were used: Mw = molecular weight; PSA = polar surface area; ALOGP = logarithmic term of the octanol/water partition coefficient; np = nanoparticle; npu = nanoparticle elemental unit (Al, SiO2, etc.); NMU = number of monomeric units in the np; V = average of atomic Van der Waal Volume for all atoms in the npu (<V(cm^3/mol)>); E = electronegativity; P(A3) = atomic polarizability; L = np large (experimental data); UC = uncoated nanoparticles; NMU = number of monomer units; HMT = Hexamethylenetetramine; TMAOH = Tetramethylammonium Hydroxide; DMEM = Dulbecco's modified Eagle's medium; coat = np coating; Uc = unsaturation count; Ui = unsaturation index; Hy = hydrophilic factor; AMR = Ghose–Crippen molar refractivity; TPSA(NO) = topological polar surface area using N,O polar contributions; TPSA(Tot) = topological polar surface area using N,O,S,P polar contributions; ALOGP2 = squared Ghose–Crippen octanol/water partition coefficient (logP^2); SAtot = total surface area from P_VSA-like descriptors; SAacc = surface area of acceptor atoms from P_VSA-like descriptors; SAdon = surface area of donor atoms from P_VSA-like descriptors; Vx = McGowan volume; VvdwMG = van der Waals volume from McGowan

volume; VvdwZAZ = van der Waals volume from Zhao–Abraham–Zissimos equation; PDI = packing density index.

2.1. ChEMBL Data Pre-Processing

We got the results of several preclinical assays from ChEMBL. The experimental measure, $\varepsilon_{ij}(d)$, used to quantify the biological activity of the i-th molecule (m_i) over the j-th objective, represents the outcome of every assay. The values of $\varepsilon_{ij}(d)$ rely on the structure of the compound, and also on certain limit conditions that mark off the properties of the assay $\mathbf{c}_j(d) = (c_0(d), c_1(d), c_2(d), \ldots, c_n(d))$. The first $c_j(d)$ is $c_0(d)$ = the biological activity; we used drugs with CC_{50}, EC_{50} and IC_{50}. Other conditions are $c_1(d)$ = organism, $c_2(d)$ = cell name, c_3 = assay organism, $c_4(d)$ = assay strain, etc. (see Table 1). We used classification techniques because the values $\varepsilon_{ij}(d)$ are not exact numbers in some cases. Furthermore, we discretized the values in this way: for IC_{50} and EC_{50} $f(v_{ij}(d))obs = 1$ when v_{ij} < cutoff and desirability of the biological activity parameter $D(c_0(d)) = -1$ (see Table 2); for CC_{50}, $f(v_{ij}(d))obs = 1$ when v_{ij} > cutoff and desirability $D(c_0(d)) = 1$, if not $f(v_{ij}(d))obs = 0$. The desirability $D(c_0(d)) = 1$ or -1 denotes that the parameter measured decreases or increases directly with a biological effect, which can be desired or not. Finally, we calculated the deviations of each condition for all drugs/compounds.

Table 1. CHEMBL assay conditions (selected examples).

c_0 = Parameter	**n_j** [a]	**c_0 = Parameter**	**n_j**
IC_{50} nM	30,981	IC_{50} ug·mL^{-1}	4914
EC_{50} nM	10,337	CC_{50}	11,629
c_1 = Organism	**n_j**	**c_1 = Organism**	**n_j**
Plasmodium falciparum D6	564	Plasmodium falciparum K1	6066
Plasmodium falciparum	35,463	Plasmodium yoelii yoelii	36
Plasmodium berghei	471	Plasmodium cynomolgi	15
c_2 = Cell name	**n_j**	**c_2 = Cell name**	**n_j**
Erythrocyte	677	Huh-7	118
FM3A	14	L6	85
HeLa	19	MRC5	23
Hepatocyte	28	Oocyte	5
HepG2	123	Vero	156
c_3 = Assay organism	**n_j**	**c_3 = Assay organism**	**n_j**
Plasmodium falciparum	31,587	Plasmodium berghei ANKA	6
Plasmodium falciparum D10	1147	Plasmodium falciparum 3D7	2606
Plasmodium falciparum FcB1/Columbia	330	Plasmodium falciparum NF54	938
Plasmodium berghei	461	Plasmodium falciparum FCR-3/Gambia	31
c_4 = Assay Strain	**n_j**	**c_4 = Assay Strain**	**n_j**
W2mef	39	W2	8591
NF54	929	TM91C235	474
W2/Indochina	31	VS1	25
W2-Mef	16	TM90C2B	68
c_5 = Curated by	**n_j**	**c_5 = Curated by**	**n_j**
Autocuration	38,150	Expert	2794
Intermediate	5288		
c_6 = Assay Type	**n_j**	**c_6 = Assay Type**	**n_j**
F	46,179	B	53

[a] n_j indicates the number of samples for each of the conditions.

Table 2. Compound activity parameters (c_0).

c_0 = Activity (Units)	Statistical Parameters [a]						
	<LogP>	**<PSA>**	n_0	n_1	p_1	**d**	**Cutoff**
IC_{50} nM	4.128024	72.6311	30,981	8954	0.289	−1	100.0
EC_{50} nM	4.2390887	67.1602	10,337	1437	0.139	−1	100.0
IC_{50} ug.mL^{-1}	4.0724379	75.0632	4914	4889	0.994	−1	325.0
CC_{50} nM	4.0650589	67.7534	11,629	11,608	0.998	1	100.0

[a] Parameters <LogP> and <PSA> = Average value of LogP and PSA for all drugs m_i with value reported in ChEMBL dataset. These parameters are needed for the moving average calculation. Other parameters: n_0 = number of compounds that shown each different activity, n_1 = number of compounds considered as positive, $p_1 = n_1/n_0$ probability of a compound being considered positive, d = −1, 0, 1 is the desirability of the parameter, cutoff = limit for the compound being treated as active or not.

2.2. Nanoparticle Data Pre-Processing

From the literature, we collected the outcomes of many nanoparticles, and the measure ε_{ij} expresses the result of each of them. The values of ε_{ij}(np) depend on different properties of the nanoparticle, and also on some boundary conditions that delimit the characteristics of the assay c_j (np) = (c_0(np), c_1(np), c_2(np), ... , c_n(np)) (see Table 3). Again, the first c_j(d) = the biological activity, and we only used nanoparticles with CC_{50}, EC_{50} and IC_{50}, so that they could match with the biological activities of the drugs/compounds. Other conditions are c_1(np) = cell name, c_2(np) = nanoparticle shape, c_3(np) = nanoparticle medium and c_5 = surface coating. Additionally, we discretized the values in the same way that we did with drugs/compounds (see Table 4). In the end, we determined the deviations of every c_j for all nanoparticles.

Table 3. Decorated Nanoparticles assay conditions (selected examples).

c_0 = Parameter	n_j [a]	c_0 = Parameter	n_j
IC_{50} nM	29	CC_{50}	113
EC_{50} nM	30		
c_1 = Cell line	**n_j**	**c_1 = Cell line**	**n_j**
A549 (H)	23	BRL 3A (R)	4
Lycopersicon esculentum	16	3T3 (M)	9
HepG2 (H)	15	CaCo-2 (H)	6
c_2 = Shape	**n_j**	**c_2 = Shape**	**n_j**
Spherical	61	Elliptical	21
Irregular	3	Pseudo-spherical	8
Slice-shaped	3	Polyhedral	3
Needle	2	Pyramidal	10
Rod	9		
c_3 = Assay Medium	**n_j**	**c_3 = Assay Medium**	**n_j**
Dry	118	RPMI	3
H_2O	44	1% Triton X-100/H_2O	3
DMEM	3	H_2O/TMAOH	1

Table 3. *Cont.*

c_4 = Surface coating	n_j	c_4 = Surface coating	n_j
UC	125	11-mercaptoundecanoic acid	3
PEG-Si(OMe)$_3$	8	PVP	4
PVA	1	Propylamonium fragment	4
Sodium citrate	17	Undecylazide fragment	2

[a] n_j = the number of samples for each of the conditions; IC_{50} = the half maximal inhibitory concentration; EC_{50} = the concentration of a drug that gives half-maximal response; CC_{50} = the ratio of the 50% cytotoxic concentration; A549 (H) = Lung carcinoma cells; HepG2 (H) = human liver cancer cells; BRL 3A (R) = Buffalo Rat Liver cells; 3T3 (M) = Fibroblast cells; CaCo-2 (H) = human colon carcinoma cells; DMEM = Dulbecco's modified eagle medium; RPMI = Roswell Park Memorial Institute medium; TMAOH = Tetramethylammonium hydroxide; UC = uncoated; PEG-Si(OMe)3 = trimethoxysilyl poly(ethylene glycol); PVP = Polyvinylpyrrolidone; PVA = Polyvinyl alcohol.

Table 4. Nanoparticle activity parameters (c_0).

c_0 = Activity (Units)	Statistical Parameters [a]						
	<LogP>	<PSA>	n_0	n_1	p_1	d	Cutoff
EC_{50} uM	1.66	18.02	30	27	0.9	−1	25,422
IC_{50} uM	3.24	38.79	29	21	0.7241	−1	18,714
CC_{50} uM	1.63	24.97	113	21	0.1858	1	3099

[a] Parameters <LogP> and <PSA> = Average value of LogP and PSA for all nanoparticles m_i. These parameters are needed for the moving average calculation. Other parameters: n_0 = number of decorated nanoparticles that shown each different activity, n_1 = number of nanoparticles considered as positive, $p_1 = n_1/n_0$ probability of a nanoparticle being considered positive, d = −1, 0, 1 is the desirability of the parameter, cutoff = limit for the DNPs being treated as active or not.

2.3. Combine Data PRE-Processing

Once both databases were done, we combined them by doing pairs with the same experimental conditions, for example, a CC50 with a CC50 nanoparticle. In addition, we used the same method to discretize each formed pair. Thus, a dataset of 107 input features and 249,992 examples will be used to build ML classification models. The positive (1) and negative control cases (0) were assigned as follows: if desirability function d(c0) = −1, then $c_{ij} = 1$ when $\varepsilon_{ij} < 100$ nM or $\varepsilon_{ij} <$ average $<\varepsilon_{ij}>$ for properties not measured in nM. In addition, if $d(c_0) = 1, 0$, then $c_{ij} = 1$ when $\varepsilon_{ij} >$ average value $<\varepsilon_{ij}>$. An extra input feature (prob = probability) was created as the probability of c0 for compound–nanoparticle pairs (count of the number of compound–nanoparticle pairs for each c0 activity type/total number of pairs). The name of the final features in the dataset has the format [d_/np_][original descriptor name]([experimental condition]). For example:

- d_DPSA(c2) = difference (D) between original values of PSA descriptor and the mean of PSA values in experimental condition c2 (for drugs/compounds, d_);
- np_DLnp(c4) = difference between Lnp value and the mean of Lnp values in experimental condition c4 (for nanoparticles, np_).

2.4. Machine Learning Methods

The study is done using eight Machine Learning scikit-learn classifiers to find the best classifier able to predict the probability of a nanoparticle–compound pair highly express antimalarial activity:

1. KNeighborsClassifier = KNN—k-nearest neighbors: one of the most well-known non-parametric classifiers in the ML field. It assigns an unclassified sample to the same class as the nearest of the k samples in the training set [26];

2. SVC(linear) = SVM linear—support vector classifier with linear kernels: the input data is non-linearly mapped to a higher dimensionality space, where a linear decision surface can be established [27];
3. SVC = SVM—support vector classifier with non-linear RBF kernels: the real problems tend not to have a linear solution, and SVM can handle this limitation by using nonlinear kernel functions such as Gaussian radial basis (RBF) [28];
4. LogisticRegression = LR—Logistic regression [29] is a linear model which can estimate the probability of a binary response using different factors;
5. LinearDiscriminantAnalysis = LDA—linear discriminant analysis [30]: a statistical supervised method that is based on the projection of data to a lower dimension to maximize the scatter between classes versus the scatter within each class. This projection facilitates the separation of the data;
6. DecisionTreeClassifier = DT—Decision Tree uses a series of decision rules inferred from the features as a tree of rules. Thus, the paths from root to leaf represent classification rules [31];
7. RandomForestClassifier = RF—Random forest [32] is an ensemble method that aggregates several decision trees (parallel trees). Each tree is generated using a bootstrap sample drawn randomly from the original dataset using a classification or regression tree (CART) method and the Decrease Gini Impurity (DGI) as the splitting criterion [33]. RF is characterized by low bias and low correlation between individual trees, and high variance;
8. XGBClassifier = XGB—XGBoost a tree-based ensemble method wherein weak classifiers are added to correct errors (sequential trees [34]). This classifier demonstrate excellent performances through the Kaggle competition projects [35].

2.5. ML Workflow

The features were standardized by removing the mean and scaling to unit variance, using the standard scaler in *scikit-learn*. A stratified 10-fold cross-validation was performed, preserving the percentages of samples for each class. As the dataset samples were not balanced, class weights were computed for each class using $N/(k*n_i)$, where N is the total number of samples, k the number of classes and n_i the number of samples belonging to the class i. This results in weights of 0.63778 for class 1 and 2.31448 for class 2. The model's performance was measured using Area Under the Receiver Operating Characteristics (AUC).

Given the results obtained in the baseline, the workflow has continued only with the best model. From this point on, a feature selection was done using the mean impurity decrease, which is already implemented in sklearn. This metric is calculated using the weighted gini impurity decreases for all nodes, averaged over all trees [33]. Thus, a feature selection was done using ExtraTreesClassifier [36] with n_estimators = 100, class weights and 10-fold CV (see Feature-Selection.ipynb [37]). We chose this tree-based method to select the most important features because extra trees (sometimes named extreme random trees) offer a higher performance in the presence of noisy features [38]. Our custom feature selection algorithm keeps at least one feature for each experimental condition for drugs/compounds and nanoparticles, and the probability feature (if the automatic selector eliminates them).

The simplest PTML linear models will be the first classifiers to test for complex datasets with multiple BD characteristics [39,40]. We can approximate function values $f(v_{ij}(d)$ and $v_{ij}(np))_{calc}$ for the i-th drug–nanoparticle pair in the j-th preclinical assay with multiple conditions of assay c_j. As input, we used PT operators that can also be Box–Jenkins Moving Average (MA) operators [41,42]. PTML linear models have the following generic form:

$$f(v_{ij}(d), v_{ij}(np))_{calc} = a_0 + a_1 \cdot f(v_{ij}(d), v_{ij}(np))_{expt} + \sum_{k=1,j=0}^{k_{max},j_{max}} a_{kj} \cdot \Delta D_k\left(d_j, c_j\right) + \sum_{k=1,j=0}^{k_{max},j_{max}} b_{kj} \cdot \Delta D_k(np_j, c_j) \quad (2)$$

Additional results have been provided in order to explain the predictions with the best model using Shapley values [43] (SHAP_test.ipynb). All the scripts for baseline AUCs, feature selection and the final model are available as an open repository at https://github.com/d-bcarrue/NanoDrugsMalaria [37].

3. Results and Discussion

In the present work, we created a PTML model to predict the activity of organic compounds assembled of some nanoparticles used against malaria disease. In doing so, we expanded the idea behind Hansch's analysis and searched models with applications to nanomedicine. As a proof-of-concept test, we investigated a huge number dataset of drugs downloaded from ChEMBL, and another dataset of nanoparticles. Those datasets contain (see materials and methods) the outcomes of many experimental pharmacological assays.

The model supposes that the changes in drug–nanoparticle binding occur thanks to perturbations in the input boundary conditions of both nanoparticles and drugs. We focused only on a nanoparticle–drug/compound binding pseudo-constant ($v_{ij(d)}$,$v_{ij(np)}$), defined by us, to quantify the probability of a nanoparticle–drug/compound pair highly expressed against malarial activity. This PTML model begins with a reference value, $f\left(v_{ij_d}, v_{ij_{np}}\right)_{expt}$, and then adds the effects of perturbations in the structure of the compound or conditions of the assay, and the properties of the nanoparticle and its coating. Other input terms used here are the perturbation terms ΔLogP and ΔPSA, which are similar to the Moving Average (MA) functions utilized in the Box–Jenkins models in time series (Box and Jenkins, 1970). Example of MAs are the deviations of PSA and logP of compounds/drugs and nanoparticles from the expected values of these parameters for assays under the same conditions c_j. For example, DLogP = LogP(m_i) − LogP(c_j), where LogP(c_j) is the average of LogP(m_i) for all molecules, m_i, in the same assay with a set of conditions c_j.

Using eight ML classifiers, the AUC values have been calculated (10-fold CV). The results are presented in Table 5. The best model was obtained with RF, and the AUC is 0.9844 ± 0.0007. Figure 2 represents the box-plot for the baseline AUC values of the ML methods (10-fold CV). The AUC values for the 10 splits have short ranges, especially RF. This suggests that the AUCs for all ML methods are stable within each fold. In addition, the high difference between the RF and the other methods (box-plots are far from overlapping) demonstrated that it is statistically significant.

Table 5. Area Under the Receiver Operating Characteristics (AUC) for baseline classification models.

ML Method.	Classifier	AUC Mean + sd
KNN	KNeighborsClassifier	0.8994 ± 0.0022
SVM linear	SVC(linear)	0.8949 ± 0.0019
SVM	SVC(rbf)	0.9223 ± 0.0007
LR	LogisticRegression	0.8946 ± 0.0013
LDA	LinearDiscriminantAnalysis	0.8939 ± 0.0015
DT	DecisionTreeClassifier	0.9277 ± 0.0021
RF	RandomForestClassifier	0.9844 ± 0.0007
XGB	XGBClassifier	0.9242 ± 0.0017

KNN = k-nearest neighbors; SVM linear = support vector classifier with linear kernels; SVM = support vector classifier with non-linear kernels; LR = Logistic regression; LDA = linear discriminant analysis; DT = Decision Tree; RF = Random forest; XGB = XGBoost.

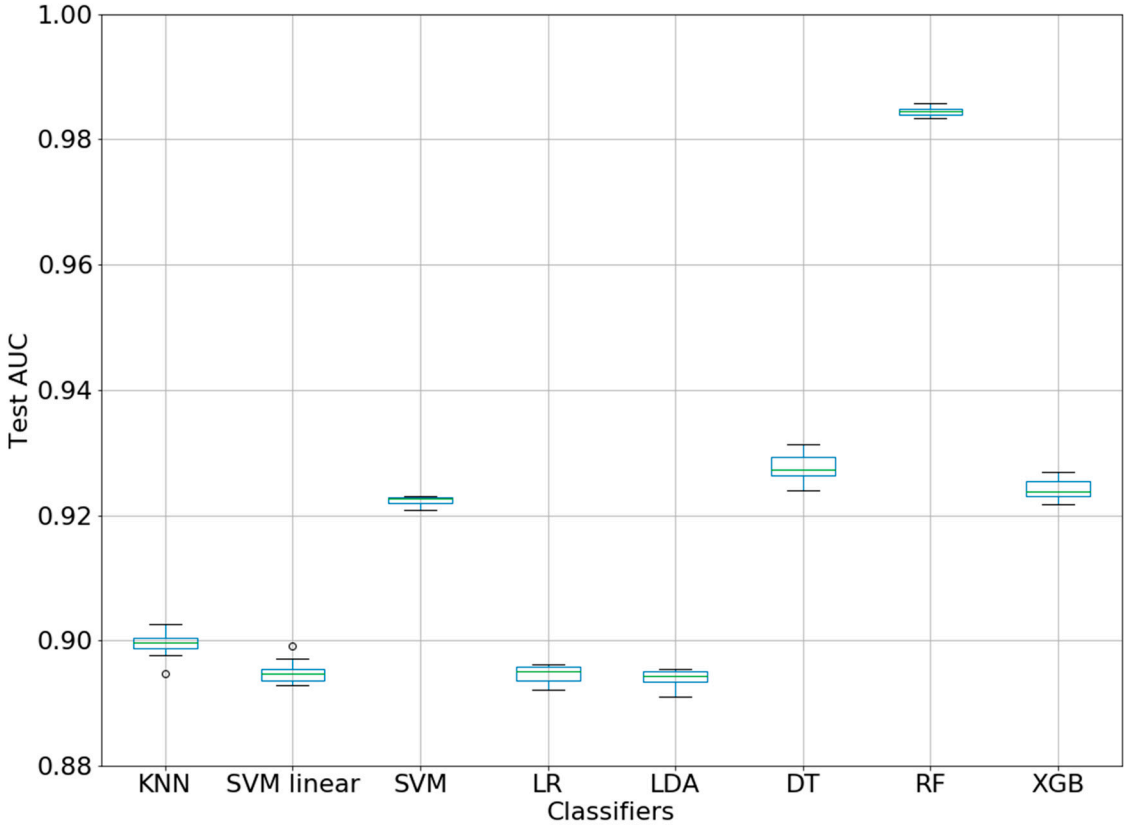

Figure 2. Box-plot for AUC values of ML classifiers (10-fold CV).

In the next step, we reduced the number of features in order to improve the AUC of the RF model. Thus, a feature selection was done using ExtraTreesClassifier. The 27 features were selected from an initial 107:

- One np–compound pair feature: prob;
- 5 np features using 5 experimental conditions (c0-c4): np_DVnpu(c0), np_DUccoat(c1), np_DVnpu(c2), np_DPnpu(c3), np_DPnpu(c4);
- 21 drug/compound features using 7 experimental conditions (c0-c6): d_DMw(c0), d_DALOGP(c0), d_DPSA(c0), d_DMw(c1), d_DALOGP(c1), d_DPSA(c1), d_DMw(c2), d_DALOGP(c2), d_DPSA(c2), d_DMw(c3), d_DALOGP(c3), d_DPSA(c3), d_DMw(c4), d_DALOGP(c4), d_DPSA(c4), d_DMw(c5), d_DALOGP(c5), d_DPSA(c5), d_DMw(c6), d_DALOGP(c6) and d_DPSA(c6).

Remarkably, this is the first model combining both Perturbation Theory and MAs in a QSBR study of relevant nanoparticle–drug/compound pairs used as an antimalarial delivery system. We determined the more relevant perturbations under different experimental conditions, c_j, related to the antimalarial property by using a RF. Casually, in this model most of the used operators are of PSA and ALOGP type. Therefore, they measure only perturbations in the value of ALOGP with respect to other subsets, c_j, of drugs and nanoparticles. ALOGP is a relevant parameter used in medicinal chemistry because it is related to lipophilicity and the capacity to cross biological membranes.

Figure 3 shows the mean impurity reduction for each of the features in both the original order and the sorted. This mean impurity decrease was obtained using an Extra Trees classifier with 100 trees and weighted classes, and this model was applied in a stratified 10-fold CV. The horizontal dashed line indicates the threshold used as a selection filter. After checking that the probability feature is present, since this is strictly necessary in Perturbation Theory, we check that all experimental conditions are reflected in the selected subset. After filtering, the experimental conditions, c2 and c4, of the nanoparticles were not included. Therefore, we selected the characteristic with the highest mean

impurity decrease for both experimental conditions, and added it to the previous selection, marking them in pink. The unsorted plot presents the features on the x-axis in the order they were presented into the dataset. For a better comparison of the selected feature mean impunities (the ones above the cutoff), the ordered version of the plot was presented too.

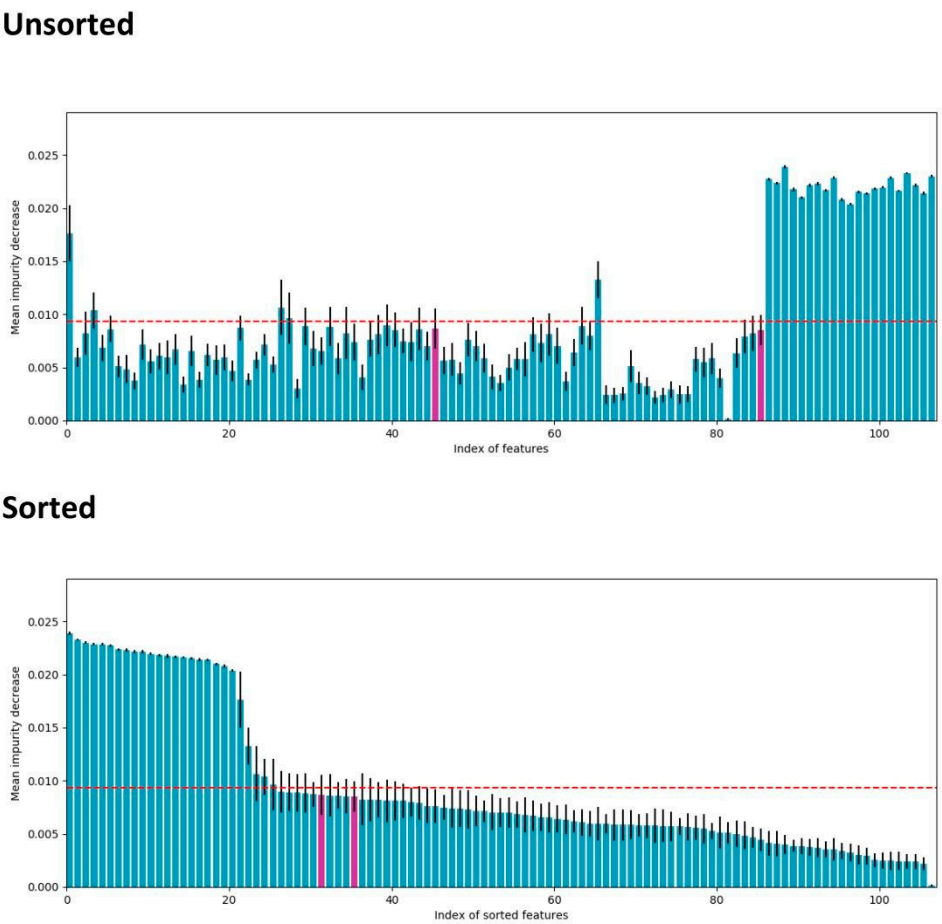

Figure 3. Feature selection using ExtraTrees: mean impurity decrease by feature, in original order and in descending order; pink bars represent features reintroduced after initial filtering.

Therefore, with only 27 selected features (from an initial 107), the mean test AUC for the RF classifier increased to 0.9921 ± 0.000244 (from 0.9844 ± 0.0007). This model shows a very good performance for a PTML model.

The feature selection showed that the current classifier prefers perturbations (MAs) of the logarithmic term of the octanol/water partition coefficient (ALOGP), polar surface area (PSA) and molecular weight (Mw) for compounds/drugs under all experimental conditions, such as activity type (c0), organism (c1), cell name (c2), assay organism (c3), assay strain (c4), type of curation (c5) and assay type (c6). In the case of the nanoparticles, the model selected the perturbations of the average of atomic Van der Waal volume for all atoms in the np (Vnpu) with activity type (c0) and with shape (c2), unsaturation count (Uccoat) in cell line (c1), atomic polarizability (Pnpu) with assay medium (c3) and surface coating (c4). Thus, we can conclude that the perturbations of the following molecular descriptors under different experimental conditions are important for anti-malaria drug-decorated nanoparticles: polarity of both components/drugs and nanoparticles, mass of compounds/drugs, volume, shape and coating unsaturation of nanoparticles.

A linear model is easily interpreted, but it is not always the most accurate model. Therefore, the complex models use different tools in order to avoid a "black box" model. Shapley values and SHAP

(Shapley Additive explanations) values are the proposed solution for the best RF model. The Shapley values represent the average of the marginal contribution across all permutations, a method for quantifying the contribution of the features to the final model. Thus, the SHAP method is able to explain the output of a machine learning model by:

- global interpretability: how much each feature contributes, either positively or negatively, to the output variable;
- local interpretability: each case/instance gets its own SHAP values in order to explain why a case has a specific prediction, and the contribution of the features to this instance.

The global interpretability is presented by the correlation of the features with the output variable or the positive/negative impact using SHAP values (Figure 4). The ordered average impact of the features on the model output for each class, and the local interpretability for different instances/cases, are included in the GitHub repository with the new script (SHAP_test.ipynb).

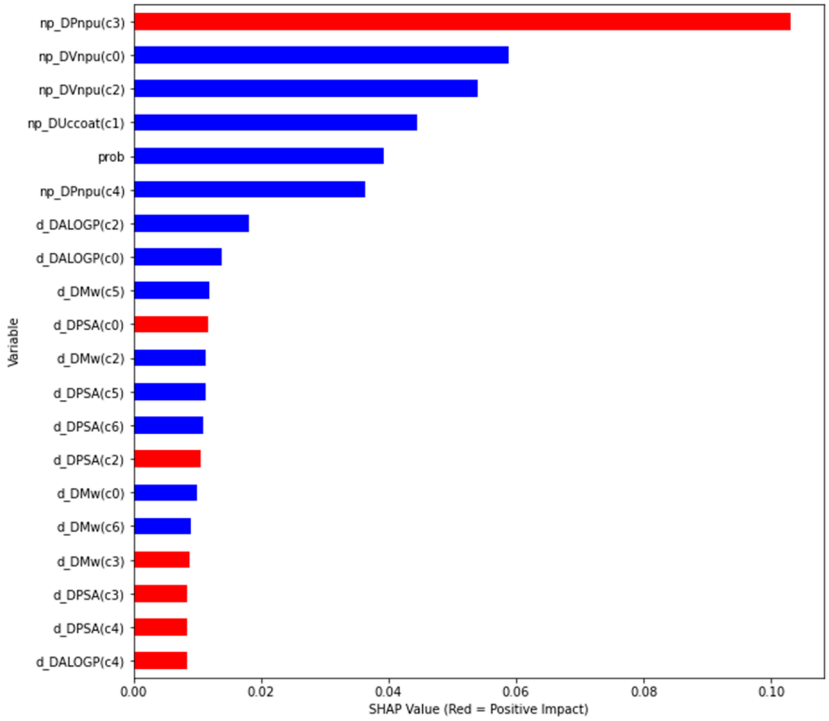

Figure 4. Feature impact on the output variable for the best model based on mean SHAP values.

This figure is presenting:

- Feature importance using SHAP values: the variables are ranked in descending order;
- Impact on the prediction value using SHAP values on x-axis;
- Color shows whether that variable has a positive (in red) or a negative (in blue) impact on the output variable.

Thus, we can observe that the nanoparticle perturbation of molecular descriptors under experimental conditions has a high impact on the model prediction for anti-malaria drug/compounds carriers. These include perturbation of atomic polarizability (Pnpu), the average of atomic Van der Waal volume for all atoms (Vnpu) and unsaturation of coating (Uccoat). For the compounds/drugs, ALOGP has more impact than mass weight and PSA. Thus, we confirm that the molecular properties linked to polarity have the highest impact on the anti-malaria drug/compound–nanoparticle carriers. Atomic polarizability of nanoparticles has a more positive impact on the model output, and the volume of nanoparticles has only a negative impact: the optimal anti-malaria drug–np carriers should consider

nanoparticles with high atomic polarizability but small volume. In addition, the compounds/drugs should have higher polar surface areas (PSA) with a positive impact, and smaller weight mass with a negative impact, on the model output.

4. Conclusions

By combining Perturbation Theory ideas with Hansch's QSAR analysis and information fusion, we developed a multi-target PTMLIF model that is useful in classifying drugs based on their constant binding to many different nanoparticles and their capacity to act against plasmodium, which is the cause of malaria in humans. This model can help us to save experimental resources and time, since it allows the determination of which drug-decorated nanoparticles would be useful and which would not. In this way, we can prove only those with the highest probability of being active. The transformed features of drugs and nanoparticles have been used as input for eight Machine Learning methods. The best classification model has been obtained using Random Forest with only 27 selected features of drugs and nanoparticles in all the experimental conditions considered. The mean test AUC was 0.9921 ± 0.000244 (10-fold CV). The performance of the RF model demonstrated the power of the information fusion of the experimental-based features of drugs and nanoparticles for the prediction of probability, related to nanoparticle–drug/compound antimalarial activity. All the calculations can be reproduced using the scripts and dataset included in an open GitHub repository at https://github.com/d-bcarrue/NanoDrugsMalaria.

Author Contributions: D.V.U., S.A., V.F.Q.-T. and H.G.-D. developed the dataset. H.G.-D. proposed the combination of Perturbation Theory with Machine Learning and Information Fusion. D.B.C., I.O., M.G. and C.R.M. preprocessed the dataset, applied Machine Learning methodologies to find the best classification model and programmed all the scripts. The manuscript was written by H.G.-D., C.R.M., D.B.C., I.O. and M.G., and all authors reviewed it. All authors have read and agreed to the published version of the manuscript.

Funding: This research and the APC were funded by Consolidation and Structuring of Competitive Research Units—Competitive Reference Groups (ED431C 2018/49) funded by the Ministry of Education, University and Vocational Training of Xunta de Galicia endowed with EU FEDER funds.

Acknowledgments: This work was supported by the 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)—"A way to build Europe". 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 through the funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503) and the European Regional Development Funds (FEDER) by the European Union. Additional support was offered by the research grants from Ministry of Economy and Competitiveness, MINECO, Spain (FEDER CTQ2016-74881-P), Basque government (IT1045-16), and kind support of Ikerbasque, Basque Foundation for Science.

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

References

1. Dutta, P.P.; Bordoloi, M.; Gogoi, K.; Roy, S.; Narzary, B.; Bhattacharyya, D.R.; Mohapatra, P.K.; Mazumder, B. Antimalarial silver and gold nanoparticles: Green synthesis, characterization and In Vitro study. *Biomed. Pharmacother.* **2017**, *91*, 567–580. [CrossRef] [PubMed]
2. Abinaya, M.; Vaseeharan, B.; Divya, M.; Sharmili, A.; Govindarajan, M.; Alharbi, N.S.; Kadaikunnan, S.; Khaled, J.M.; Benelli, G. Bacterial exopolysaccharide (EPS)-coated ZnO nanoparticles showed high antibiofilm activity and larvicidal toxicity against malaria and zika virus vectors. *J. Trace Elem. Med. Biol.* **2018**, *45*, 93–103. [CrossRef] [PubMed]
3. Quevedo-Tumailli, V.F.; Ortega-Tenezaca, B.; González-Díaz, H. Chromosome gene orientation inversion networks (GOINs) of plasmodium proteome. *J. Proteome Res.* **2018**, *17*, 1258–1268. [CrossRef]

4. Ferreira da Costa, J.; Silva, D.; Caamaño, O.; Brea, J.M.; Loza, M.I.; Munteanu, C.R.; Pazos, A.; García-Mera, X.; González-Díaz, H. Perturbation theory/machine learning model of ChEMBL data for dopamine targets: Docking, synthesis, and assay of new L-prolyl-L-leucyl-glycinamide peptidomimetics. *ACS Chem. Neurosci.* **2018**, *9*, 2572–2587. [CrossRef] [PubMed]
5. 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*, 4093–4103. [CrossRef]
6. Liu, Y.; Tang, S.; Fernandez-Lozano, C.; Munteanu, C.R.; Pazos, A.; Yu, Y.-Z.; Tan, Z.; González-Díaz, H. Experimental study and random forest prediction model of microbiome cell surface hydrophobicity. *Expert Syst. Appl.* **2017**, 306–316. [CrossRef]
7. González-Durruthy, M.; Werhli, A.V.; Seus, V.; Machado, K.S.; Pazos, A.; Munteanu, C.R.; González-Díaz, H.; Monserrat, J.M. Decrypting strong and weak single-walled carbon nanotubes interactions with mitochondrial voltage-dependent anion channels using molecular docking and perturbation theory. *Sci. Rep.* **2017**, *7*, 13271. [CrossRef]
8. González-Durruthy, M.; Monserrat, J.M.; Rasulev, B.; Casañola-Martín, G.M.; Barreiro Sorrivas, J.M.; Paraíso-Medina, S.; Maojo, V.; González-Díaz, H.; Pazos, A.; Munteanu, C.R. Carbon nanotubes' effect on mitochondrial oxygen flux dynamics: Polarography experimental study and machine learning models using star graph trace invariants of raman spectra. *Nanomaterials* **2017**, *7*, 386. [CrossRef]
9. González-Durruthy, M.; Alberici, L.C.; Curti, C.; Naal, Z.; Atique-Sawazaki, D.T.; Vázquez-Naya, J.M.; González-Díaz, H.; Munteanu, C.R. Experimental-Computational study of carbon nanotube effects on mitochondrial respiration: In silico nano-QSPR machine learning models based on new raman spectra transform with Markov-Shannon entropy invariants. *J. Chem. Inf. Model.* **2017**, *57*, 1029–1044. [CrossRef]
10. Ran, T.; Liu, Y.; Li, H.; Tang, S.; He, Z.; Munteanu, C.R.; González-Díaz, H.; Tan, Z.; Zhou, C. Gastrointestinal spatiotemporal mRNA expression of ghrelin vs growth hormone receptor and new growth yield machine learning model based on perturbation theory. *Sci. Rep.* **2016**, *6*, 30174. [CrossRef]
11. 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*, 10623–10630. [CrossRef] [PubMed]
12. 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*, 14686–14694. [CrossRef] [PubMed]
13. Kleandrova, V.V.; Luan, F.; González-Díaz, H.; Ruso, J.M.; Melo, A.; Speck-Planche, A.; Cordeiro, M.N.D.S. Computational ecotoxicology: Simultaneous prediction of ecotoxic effects of nanoparticles under different experimental conditions. *Environ. Int.* **2014**, *73*, 288–294. [CrossRef] [PubMed]
14. Hansch, C. The Advent and Evolution of QSAR at Pomona College. *J. Comput. Aided Mol. Des.* **2011**, 495–507. [CrossRef] [PubMed]
15. Kubinyi, H. QSAR: Hansch Analysis and Related Approaches. In *Methods and Principles in Medicinal Chemistry*; Wiley: Hoboken, NJ, USA, 1993.
16. Cho, S.J.; Hermsmeier, M.A. Genetic algorithm guided selection: Variable selection and subset selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 927–936. [CrossRef]
17. Tetko, I.V.; Tanchuk, V.Y.; Kasheva, T.N.; Villa, A.E. Internet software for the calculation of the lipophilicity and aqueous solubility of chemical compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 246–252. [CrossRef]
18. Zhang, S.; Golbraikh, A.; Tropsha, A. Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. *J. Med. Chem.* **2006**, *49*, 2713–2724. [CrossRef]
19. 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–W620. [CrossRef]
20. Papadatos, G.; Overington, J.P. The ChEMBL database: A taster for medicinal chemists. *Future Med. Chem.* **2014**, *6*, 361–364. [CrossRef]

21. Bento, A.P.; Gaulton, A.; Hersey, A.; Bellis, L.J.; Chambers, J.; Davies, M.; Krüger, F.A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL bioactivity database: An update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090. [CrossRef]
22. Willighagen, E.L.; Waagmeester, A.; Spjuth, O.; Ansell, P.; Williams, A.J.; Tkachenko, V.; Hastings, J.; Chen, B.; Wild, D.J. The ChEMBL database as linked open data. *J. Cheminform.* **2013**, *5*, 23. [CrossRef] [PubMed]
23. Hu, Y.; Bajorath, J. Growth of ligand-target interaction data in ChEMBL is associated with increasing and activity measurement-dependent compound promiscuity. *J. Chem. Inf. Model.* **2012**, *52*, 2550–2558. [CrossRef] [PubMed]
24. Wassermann, A.M.; Bajorath, J. BindingDB and ChEMBL: Online compound databases for drug discovery. *Expert Opin. Drug Discov.* **2011**, *6*, 683–687. [CrossRef] [PubMed]
25. Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107. [CrossRef]
26. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, 21–27. [CrossRef]
27. Hao, J.; Ho, T.K. Machine learning made easy: A review of scikit-learn package in Python programming language. *J. Educ. Behav. Stat.* **2019**, 107699861983224. [CrossRef]
28. Patle, A.; Chouhan, D.S. SVM Kernel functions for classification. In *2013 International Conference on Advances in Technology and Engineering (ICATE)*; IEEE: New York, NY, USA, 2013; pp. 1–9.
29. Peduzzi, P.; Concato, J.; Kemper, E.; Holford, T.R.; Feinstein, A.R. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **1996**, *49*, 1373–1379. [CrossRef]
30. Cristianini, N. Fisher Discriminant Analysis (Linear Discriminant Analysis). In *Dictionary of Bioinformatics and Computational Biology*; Wiley: Hoboken, NJ, USA, 2004.
31. Swain, P.H.; Hauska, H. The decision tree classifier: Design and potential. *IEEE Trans. Geosci. Electron.* **1977**, *15*, 142–147. [CrossRef]
32. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
33. Calle, M.; Urrea, V. Letter to the editor: Stability of random forest importance measures. *Brief. Bioinform.* **2011**, *12*, 86–89.
34. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
35. Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
36. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, 3–42. [CrossRef]
37. D-Bcarrue. D-Bcarrue/NanoDrugsMalaria. Available online: https://github.com/d-bcarrue/NanoDrugsMalaria (accessed on 11 April 2019).
38. Moore, D.H. Classification and regression trees, by Leo Breiman, Jerome, H.; Friedman, Richard, A. Olshen, and Charles, J. Stone. Brooks/Cole Publishing, Monterey, 1984, 358 Pages, $27.95. *Cytometry* **1987**, 534–535. [CrossRef]
39. 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**, 1–15. [CrossRef] [PubMed]
40. González-Díaz, H.; Arrasate, S.; Gómez-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*, 1713–1741. [CrossRef] [PubMed]
41. Casañola-Martin, G.M.; Le-Thi-Thu, H.; Pérez-Giménez, F.; Marrero-Ponce, Y.; Merino-Sanjuán, M.; Abad, C.; González-Díaz, H. Multi-Output model with box-jenkins operators of linear indices to predict multi-target inhibitors of ubiquitin-proteasome pathway. *Mol. Divers.* **2015**, *19*, 347–356. [CrossRef] [PubMed]

42. Tenorio-Borroto, E.; Ramirez, F.R.; Speck-Planche, A.; Cordeiro, M.N.D.S.; Luan, F.; Gonzalez-Diaz, H. QSPR and Flow Cytometry Analysis (QSPR-FCA): Review and new findings on parallel study of multiple interactions of chemical compounds with immune cellular and molecular targets. *Curr. Drug Metab.* **2014**, *15*, 414–428. [CrossRef]
43. Roth, A.E. *The Shapley Value: Essays in Honor of Lloyd, S. Shapley*; Cambridge University Press: Cambridge, UK, 1988.