ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Perturbation-Theory Machine Learning (PTML) Multilabel Model of the ChEMBL Dataset of Preclinical Assays for Antisarcoma Compounds

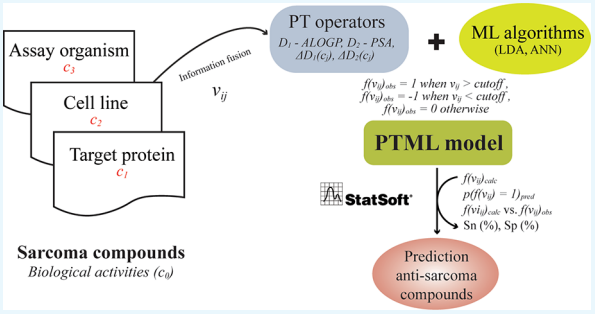
Alejandro Cabrera-Andrade,*,& Andrés López-Cortés,& Cristian R. Munteanu, Alejandro Pazos, Yunierkis Pérez-Castillo, Eduardo Tejera, Sonia Arrasate, and Humbert González-Díaz*

Cite This: *ACS Omega* 2020, 5, 27211−27220

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Sarcomas are a group of malignant neoplasms of connective tissue with a different etiology than carcinomas. The efforts to discover new drugs with antisarcoma activity have generated large datasets of multiple preclinical assays with different experimental conditions. For instance, the ChEMBL database contains outcomes of 37,919 different antisarcoma assays with 34,955 different chemical compounds. Furthermore, the experimental conditions reported in this dataset include 157 types of biological activity parameters, 36 drug targets, 43 cell lines, and 17 assay organisms. Considering this information, we propose combining perturbation theory (PT) principles with machine learning (ML) to develop a PTML model to predict antisarcoma compounds. PTML models use one function of reference that measures the probability of a drug being active under certain conditions (protein, cell line, organism, *etc.*). In this paper, we used a linear discriminant analysis and neural network to train and compare PT and non-PT models. All the explored models have an accuracy of 89.19−95.25% for training and 89.22−95.46% in validation sets. PTML-based strategies have similar accuracy but generate simplest models. Therefore, they may become a versatile tool for predicting antisarcoma compounds.



INTRODUCTION

Sarcomas are a group of malignant neoplasms of connective tissue. Although their prevalence is much lower than carcinomas, the number of cases is increasing according to the World Health Organization.[1] At the molecular level, their behavior differs from carcinomas, presenting a more varied and complex etiology. This high etiological complexity possibly stems from their mesenchymal origin, which makes it difficult to propose new therapeutic targets for the respective treatment.[2−6] Representative anticancer compounds tend to have high cytotoxicity and low cellular specificity.[7] This leads to a decreased efficiency within the treatment and a low remission rate of the disease. However, a description of new molecular markers and the constant performance of drug preclinical assays have generated large amounts of data.[8−12] This data, if adequately rationalized, may lead in turn to the design of more selective drugs, which takes into account specific drivers based on pathogenic signaling pathways. For instance, the Chemical Database of the European Molecular Biology Laboratory (ChEMBL)[13,14] contains experimental outcomes for >37,900 different preclinical assays of antisarcoma drug candidates. These assays cover a large and structurally heterogeneous series of >34,900 different chemical compounds. Furthermore, the preclinical assays have been carried out on very different experimental conditions. These experimental conditions include up to 155 different types of biological activity parameters, 36 protein targets, 43 cell lines, and 17 assay organisms. Overall, this forms a large and complex dataset susceptible to analysis so as to extract useful knowledge for drug discovery.

In this context, we can use computational techniques to explore this experimental dataset due to the evident difficulties to analyze it manually. Specifically, cheminformatics methodologies have succeeded in the discovery of new drug candidates effective in the wet-lab.[15,16] However, many models developed thus far are applied only to carcinomas and/or are focused on homologous series of compounds with one target or a single cell line.[17−26] In recent years, several studies have focused on applying these methodologies to the study of new types of antisarcoma drugs, mainly on cell lines.[27−30] However, almost all the models reported have a narrow domain of application because they focus on only one set of conditions, for instance,

Received: July 13, 2020
Accepted: October 6, 2020
Published: October 15, 2020

https://dx.doi.org/10.1021/acsomega.0c03356
ACS Omega 2020, 5, 27211−27220

Table 1. PTML Model Results

series	statistical parameter[a]	predicted statistics (%)	observed set	predicted set $f(\nu_{ij})_{pred} = 0$	predicted set $f(\nu_{ij})_{pred} = 1$
training	Sp	95.63	$f(\nu_{ij})_{obs} = 0$	25,647	1172
	Sn	79.64	$f(\nu_{ij})_{obs} = 1$	330	1291
	Ac	94.72	total	25,977	2463
validation	Sp	95.79	$f(\nu_{ij})_{obs} = 0$	8559	376
	Sn	81.62	$f(\nu_{ij})_{obs} = 1$	100	444
	Ac	94.98	total	8659	820

[a]Sn, sensitivity (%); Sp, specificity (%); Ac, accuracy (%).

Table 2. Variables Used to Fit the PTML Model

condition[a] ($\mathbf{c}_j$)	condition name	symbol	operator formula	operator information
c_0	activity type	$f(\nu_{ij})_{obs}$	=IF(AND(ν_{ij} > cutoff(c_0), $d(c_0)$ = 1), 1, IF (AND (ν_{ij} < cutoff(c_0), $d(c_0)$ = −1), 1, 0))	observed classification of the outcome ν_{ij} in the assay with conditions $\mathbf{c}_j$
c_0	activity type	$f(\nu_{ij})_{ref}$	$n(f(\nu_{ij})_{obs} = 1)/n_j$	function of reference if the observed value of probability $p(f(\nu_{ij}) = 1)_{expt}$ for the activity ν_{ij} of type c_0
$\mathbf{c}_j = [c_1,c_2,c_3]$	all conditions ($\mathbf{c}_j$)	$\Delta D_1(\mathbf{c}_j)$	$ALOGP_i - \langle ALOGP(\mathbf{c}_j)\rangle$	deviation of the molecular descriptors of hydrophobicity/lipophilicity D_1 (ALOGP) and polar surface area D_2 (PSA) from each expected value ($\langle D_1(\mathbf{c}_j)\rangle$) or ($\langle D_2(\mathbf{c}_j)\rangle$) for the conditions $\mathbf{c}_j$ (c_1 = protein target; c_2 = cell line; c_3 = assay organism)
$\mathbf{c}_j = [c_1,c_2,c_3]$	all conditions ($\mathbf{c}_j$)	$\Delta D_2(\mathbf{c}_j)$	$PSA_i - \langle PSA(\mathbf{c}_j)\rangle$	

[a]MMA operators with a subset of multiple conditions included in eq 1.

one specific property, target protein, or cell line. Thus, models where multiple conditions of assays are considered at the same time are attractive. Perturbation theory (PT) ideas with machine learning (ML) methods (PT + ML = PTML models) are particularly useful for fitting complex datasets with big data features in drug discovery, proteomics, nanotechnology, *etc.*[31−41]

PTML models begin with one function of reference that measures the probability of a drug to be active under certain conditions (protein, cell line, organism, *etc.*). Next, PTML models use PT operators (PTOs) to account for the perturbations (deviations) of the input variables of this drug with respect to a population of drugs assayed under the same conditions. ML algorithms are used to establish the relationship between the inputs and the output variable. In cancer research, Speck-Planche *et al.* and other researchers have developed PTML-like models for different types of cancers (with an emphasis on carcinomas) such as bladder, prostate, brain, and breast cancers.[42−50] In addition, Bediaga *et al.* developed a PTML algorithm for predicting anticancer compounds using data for multiple types of carcinomas at the same time.[51] Speck-Planche *et al.* also recently developed the first PTML-like model for the prediction of antisarcoma compounds using a spectral moment approach.[52]

In any case, there are no reports of other PTML-like models for antisarcoma compounds. In this study, we carried out a comprehensive compilation, curation, and preprocessing of the ChEMBL dataset for preclinical assays of antisarcoma compounds. After that, we developed the first PTML model able to fit this complex dataset with >37,900 assays and >34,900 compounds. To the best of our knowledge, the study outperforms all previous efforts in terms of simplicity of the model and number of cases, compounds, and cell lines considered.

RESULTS AND DISCUSSION

PTML Antisarcoma Compound Model. The statistical parameters for the PTML model showed a high specificity (Sp) and sensitivity (Sn) for the training series (95.63 and 79.64, respectively). In addition, similar values were obtained for Sp (95.79) and Sn (81.62) in the validation sets. Furthermore, the p-level obtained from the chi-square (χ^2 = 16848.08) was <0.05, indicating that the model is able to perform a statistically significant separation of both classes. It is also interesting to observe the high overall accuracy (Ac) obtained in both sets: over 94% (Table 1). These results suggest that the generated model performs a statistically significant classification of antisarcoma compounds; hence, it can be considered useful for classification models with application in medicinal chemistry. The full list of biological activities (c_0) in the ChEMBL dataset of antisarcoma preclinical experimental assays is shown in Table S1.

The resulting PTML−linear discriminant analysis (LDA) model showed the following formula

$$f(\nu_{ij})_{calc} = -11.8545 + 34.8028 \cdot f(\nu_{ij})_{ref} + 0.37 \cdot D_1 - 0.0128 \cdot D_2 - 0.3616 \cdot [D_1 - \langle D_1(\mathbf{c}_j)\rangle] + 0.0191 \cdot [D_2 - \langle D_2(\mathbf{c}_j)\rangle]$$

$$n = 34955,\ \chi^2 = 16848.08,\ p < 0.001 \quad (1)$$

The PTML-LDA model was initiated by using as an input the values the function of reference $f(\nu_{ij})_{ref}$ for each compound and by adding the effect of perturbations within the system. These perturbation effects refer to the PTOs $\Delta D_k(\mathbf{c}_j)$. In eq 1, "i" and "j" are the assay and condition, respectively. Additional coefficients and terms are described in Table 2.

The parameters ALOGP and PSA are widely used in medicinal chemistry because they are related to the lipophilicity of drugs and, consequently, to their capacity to pass through biological membranes or interact with protein

https://dx.doi.org/10.1021/acsomega.0c03356
ACS Omega 2020, 5, 27211−27220

Table 3. Comparison to Other PTML Models of Anticancer Compounds

cancer type[a]	PT[b]	ML[c]	NV[d]	cases[e]	Sn(%)[f]	Sp(%)[f]	ref
sarcoma							
MSS	MMA	LDA	3	37,919	~80	>90	this work
MSS	MA	LDA	>10	3017	>90	>90	52
carcinoma							
bladder	MA	LDA	>10	664	>90	>90	44
bladder		ANN (RBF)	10	664	>95	>95	44
brain	MA	LDA	>10	1236	~90	>90	45
breast	MA	LDA	>10	2272	>85	>90	47
colorectal	MA	LDA	>10	1651	>90	>90	46
colorectal	MA	ANN (RBF)	>10	1651	>90	>90	46
prostate	MA	LDA	>10	1668	>85	>90	49
MCS	MMA	LDA	>10	116,934	>70	~90	51
MCS	MMA	LDA	3	116,934	>70	>90	51
MCS	MMA	ANN	4	116,934	>80	>80	51

[a]MSS, multiple sarcoma subtypes; MCS, multiple carcinoma subtypes. [b]PT operators used in PTML models: MMA, multicondition moving average; MA, moving average. [c]ML method used for the PTML models: LDA, linear discriminant analysis; ANN, artificial neural networks; RBF, radial basis function; LNN, linear neural networks; E-ANN (RBF), ensemble of artificial neural networks based on the RBF architecture. [d]NV, number of input variables. [e]Number of preclinical assays. [f]Approximate values for training series.

Table 4. Different Scores Calculated for the Selected Biological Activities (c_0)

activity parameter for $v_{ij}(c_0)$ (unit)	$n_j(c_0)$[a]	$\langle v_{ij}(c_0)\rangle$[b]	$d_j(c_0)$[c]	cutoff (c_0)	$n(f(v_{ij})_{obs} = 1)$[d]	$p(f(v_{ij})_{obs} = 1/c_0)$[e]
potency (nM)	31,581	19669.199	−1	100	149	0.005
IC_{50} (nM)	1808	228362.82	−1	100	177	0.098
inhibition (%)	690	39.186507	1	50	225	0.326
CC_{50} (nM)	450	134445.04	−1	100	4	0.009
activity (%)	404	52.416163	1	50	208	0.515
EC_{50} (nM)	379	63578.521	−1	100	44	0.116
TGI (%)	202	43.915842	1	50	102	0.505
T/C	173	26.556832	1	50	28	0.162
IC_{50} (μg mL^{-1})	167	64.429402	−1	60	118	0.707
T/C (%)	144	156.92153	1	50	123	0.854
GI_{50} (nM)	113	66515.131	−1	100	13	0.115
EC_{50} (μg mL^{-1})	90	60.733562	−1	60	57	0.633

[a]$n_j(c_0)$, total compounds with experimental values. [b]$\langle v_{ij}(c_0)\rangle$, average calculated of each c_0 biological activity. [c]$d_j(c_0)$, desirability value (1, −1) assigned to each c_0. [d]$n(f(v_{ij})_{obs} = 1)$, total number of biologically active compounds observed within each c_0 according to the experimental values $v_{ij}(c_0)$ reported for the parameters j. [e]$p(f(v_{ij})_{obs} = 1/c_0)$, probability of a desired biological activity within the conditions c_0.

hydrophobic pockets.[53−56] The PTML algorithm has been previously applied to the study of multiple preclinical assays of anticancer drugs. As shown in Table 3, most applications have been directed toward the most prevalent carcinomas among the global population. For instance, Speck-Planche *et al.* reported PTML-like models for bladder,[44] colorectal,[46] breast,[47] prostate[49] cancers and for multiple carcinoma subtypes.[51] In addition, PTML-like models have been tested in antibrain tumor agents.[45] Interestingly, Bediaga *et al.* demonstrated the application of a PTML on several types of carcinomas simultaneously and obtained similar Sn and Sp values as we did (>90%).[51] All these PTML-like models are able to account for changes in target proteins, cellular lines, organisms, *etc.* However, they are specific models for carcinomas, not for sarcomas.

It is worth noting that to the best of our knowledge, Speck-Planche *et al.*[52] seem to be the only researchers to have reported a previous PTML-like model for sarcomas thus far. In their study, the prediction model in external validation resulted in Ac (90.78) and Sp (90.65) values that were lower than what was obtained in our model (Ac = 94.98 and Sp = 95.79). However, our PTML algorithm showed a lower sensitivity in external validation data (81.62%) than the model obtained by Speck-Planche *et al.* (91.74%). Even when our model had a much lower number of variables and used a stricter cut-off definition for activity class (i.e., IC_{50} = 0.1 μM instead 1 μM), these aspects alone cannot explain the sensitivity reduction.

The generated PTML-LDA model (eq 1) has important characteristics that allow it to be used within research focused on drug discovery. One of the main advantages of our model is the considerable reduction of input variables for the construction of the algorithm through the inclusion of PTOs. This reduction allowed us to work on datasets with a large amount of information, to define cut-off values, and to calculate the probability of belonging to a class, whether this was a prediction for active compounds (1) or inactive compounds (0). In this way, the Sn or Sp values of the model can be adjusted according to the delimited cut-offs. An ideal prediction model has a reasonable trade-off between Sn and Sp. This means that a high sensitivity is achieved by accepting a relatively low Sp and, conversely, a high Sp is reached by compromising Sn. Sp is synonymous with a true-negative rate, which is related to the false-positive rate,[30] so a high specificity in a prediction model for drug discovery implies that it is unlikely to get a positive result in a drug that does not have a desired biological activity. Thus, a positive

https://dx.doi.org/10.1021/acsomega.0c03356
ACS Omega 2020, 5, 27211−27220

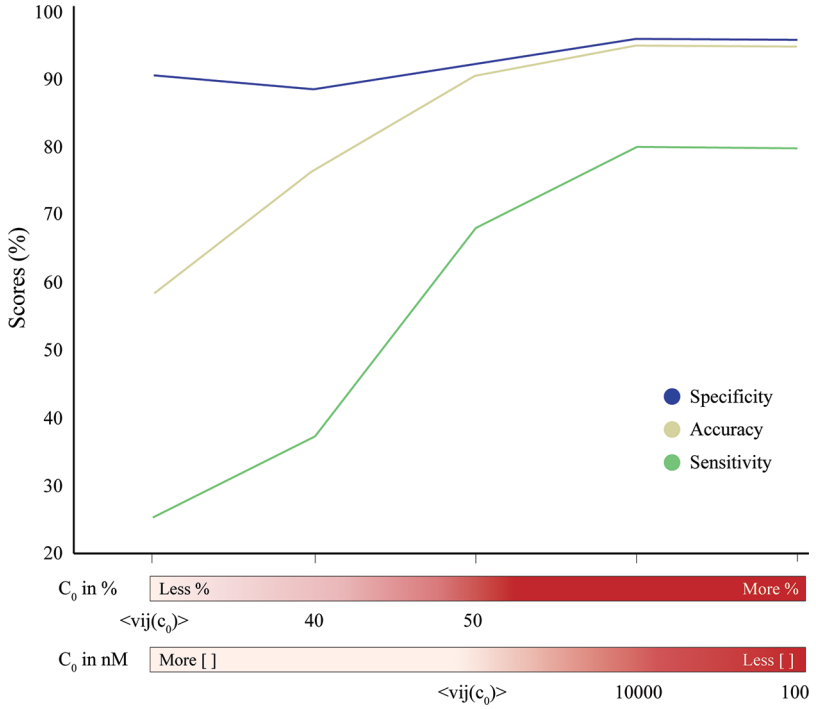


Figure 1. Variation of the specificity, sensitivity, and accuracy values according to the cut-offs implemented. The variation of these scores based on the biological activities c_0 is included in the *x*-axis. Biological activities c_0 expressed in % (e.g., inhibition, activity, tumor growth inhibition, *etc.*) and those expressed in nM (e.g., potency, IC_{50}, CI_{50}, *etc.*) are described. The final model is obtained by applying cut-off values of 50 for c_0 expressed in % and 100 for c_0 expressed in nM.

outcome in a specific model is quite informative in a drug discovery scenario.

On the other hand, a main attribute is the possible combination of several experimental conditions for the prediction of new compounds. In this sense, Speck-Planche *et al.*[52] used around 3000 interactions derived from 14 cell lines and only considered IC_{50} assays for their model. However, we modeled 37,919 interactions cases comprising 36 protein targets, 43 cell lines, and 17 assay organisms. We also included several different assay types (Table 4). The modeling task we have is more complex not only because of the increment in the chemical diversity but also the wide type of heterogeneity in the interactions (i.e., target types and organisms). The two models cannot be compared in this scenario and our reduction in the ability to detect the true-positive cases (Sn) could be a consequence of this data complexity and also the modeling strategy.

PTML Cut-Off Scanning Study. As mentioned above, the cut-off implemented in the model is a rigorous value that, at the experimental level, is important if one desires to increase effectiveness in the process of discovering antisarcoma drugs. A restricted value promotes high certainty in the prediction of active compounds for achieving a desired biological action under multiple test conditions.[57−59] Furthermore, a strict cut-off can decrease the rate of predicted false positives; therefore, if the assay is to be implemented, then it needs a higher sensitivity or higher specificity. This value can be modeled depending on the experimental conditions one wishes to apply. This cut-off value also influences the accuracy within our model. As observed in Figure 1, when using the average $\langle v_{ij}(c_0)\rangle$ calculated for each c_0, the Ac is not a desirable score. These low statistical values are mainly influenced by the low Sn in the prediction. By increasing the rigor, the model improves its prediction values for the active compounds (1). When looking at these results, our prediction algorithm not only takes into account several experimental conditions but also restricts the prediction of compounds to those that have true biological activity.

PTML vs ML Model Comparison. Most multitasking or multilabel ML methods are useful for predicting multiple categorical outputs for the same set of input continuous variables.[60,61] However, our problem was a little different: we had to develop an ML model with only two possible outputs, $f(v_{ij})_{pred}$ = 1 or 0, for the same set of input variables. That meant that our model was not multitasking for a single case with a set of input variables containing multiple continuous variables plus multiple categorical input variables. However, we had multiple combinations of input categorical variables or levels for the same set of input continuous variables. Hence, our model was multilabel in the input categorical variables for the same set of input continuous variables. To illustrate this fact, we developed here a comparison of our PTML-LDA model vs classic ML using multiple labeling categorical variables. As seen in Figure 2A, the performance of our PTML-LDA model compared to a classic ML-LDA demonstrates similar values based on Sp, Sn, and Ac. Similarly, when developing neural networks (NN), the results of PTML-NN (Figure 2B) and ML-NN (Figure 2C) are quite similar. One of the advantages of our PTML model is the inclusion of PTOs, which greatly reduces the number of variables to generate the algorithm. Thus, although the statistics of all the models generated are quite similar, the PTML methodology allows for the reduction of variables from 164 variables in classic ML methods to only 5 in the PTML model. All the PTML and non-PTML model results are described in Table S2.

https://dx.doi.org/10.1021/acsomega.0c03356
ACS Omega 2020, 5, 27211−27220

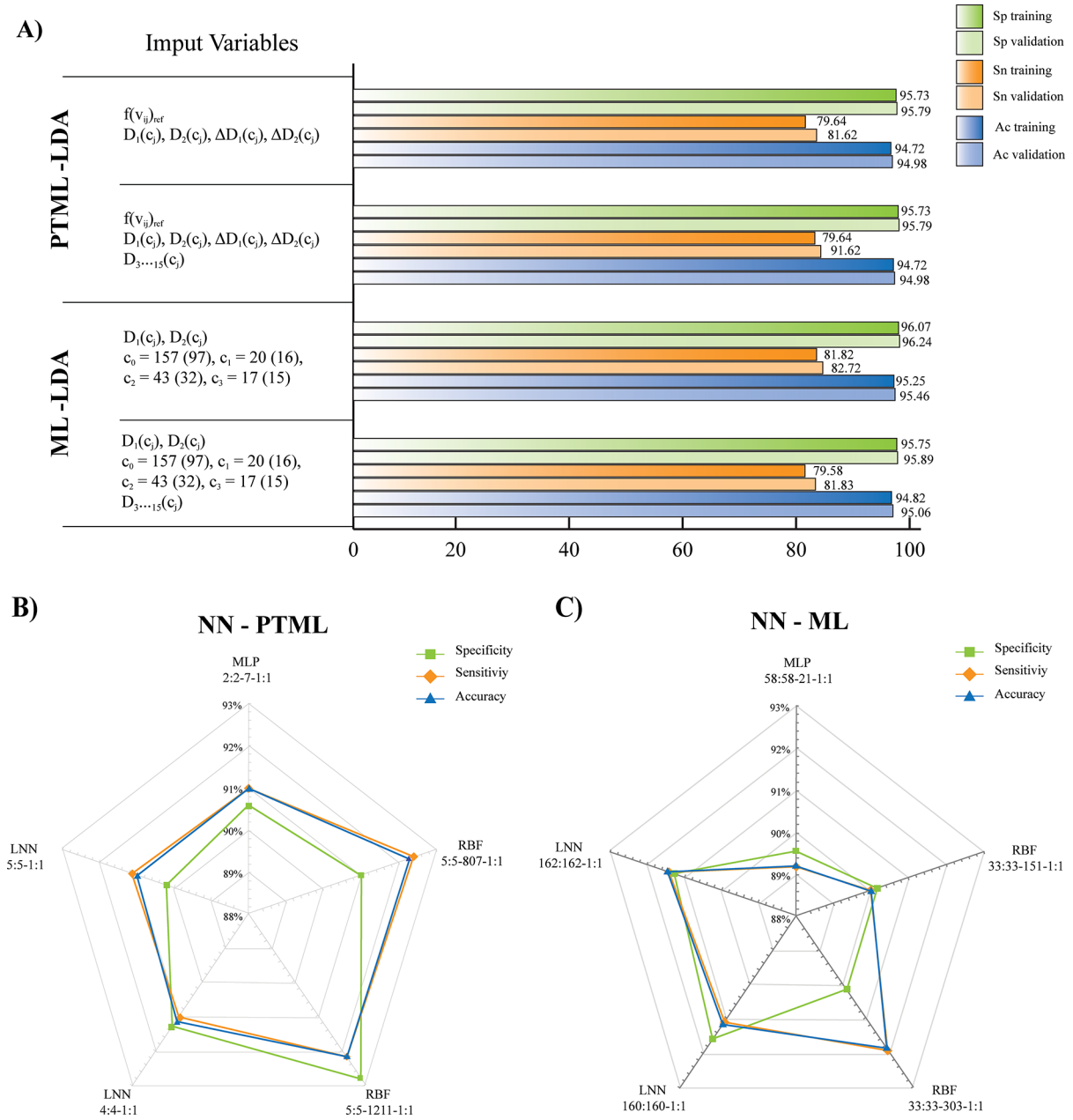


Figure 2. PTML vs ML models. Comparison of sensitivity, specificity, and accuracy of all the generated models. (A) Prediction values of PTML-LDA and ML-LDA models using different types of input variables: $f(v_{ij})_{pred}$ is the function of reference; $D_1(c_j)$ and $D_2(c_j)$ are the ALOGP and PSA descriptors, respectively; $\Delta D_1(c_j)$ and $\Delta D_1(c_j)$ are the deviations of the molecular descriptors of ALOGP and PSA, respectively; D_3, ..., $D_{15}(c_j)$ are the 12 BCUT molecular descriptors calculated from ChemAxon. Unlike the PTML model, the ML model is calculated with conditions c_1, c_2, and c_3 as a separated set of categorical variables. (B) Prediction values between the neural network-PTML (NN-PTML) and (C) NN-ML models. The NN obtained were multilayer perceptron (MLP), linear neural network (LNN), and radial basis function network (RBF).

PTML vs ML Model with Other Descriptors. Previous studies have considered a wide variety and quantity of molecular descriptors in PTML models. For example, for sarcoma modeling, Speck-Planche *et al.*[52] used 423 descriptors followed by a feature selection strategy. Similarly, 289 descriptors were used in a PTML model on breast cancer.[47] We used this approach as a strategy to compare the performance of PTML model vs classic ML techniques including new molecular descriptors (Figure 2A). In this ML study, we included 12 BCUT molecular descriptors (D_k, with $k > 2$) as an input, which were not used in the previous model, and 162 categorical (dummy) variables (C_k). These C_k have been used to label the multiple conditions of the assays c_j (organisms, proteins, cell lines, *etc.*). One must remember that D_1 = ALOGP and D_2 = PSA. The new molecular descriptors were D_3, D_4, ..., D_{14}. The expansion of the variables together with the ML strategies yielded good results but did not outperform what was obtained for the PTML-LDA anti-sarcoma model (as seen in Figure 2A and Table S2) and the number of variables increased to 174 input variables in total. This suggests that by adding different molecular descriptors and probably feature selection strategies, acceptable models for drug discovery can be built. However, our PTML-LDA model based on D_1 and D_2 is a simpler yet effective model.

https://dx.doi.org/10.1021/acsomega.0c03356
ACS Omega 2020, 5, 27211−27220

Table 5. Multiple-Condition Averages for All Sarcoma Assays

assay condition (c_j)[a]			parameter		
c_1 = protein (*gene*)	c_2 = cell line	c_3 = assay organisms[b]	$n_j(c_j)$	$\langle D_1(c_j)\rangle$	$\langle D_2(c_j)\rangle$
O75874 (*IDH1*)	MD	*H. sapiens*	31,581	3.778	70.597
MD	MD	*M. musculus*	1440	2.67	103.712
MD	U2OS	*H. sapiens*	746	4.421	78.325
MD	HOS	*H. sapiens*	637	3.603	89.517
MD	MD	*H. sapiens*	375	3.846	69.876
MD	SAOS-2	*H. sapiens*	358	4.882	81.659
MD	Sarcoma-180	*M. musculus*	271	1.108	83.68
MD	MG-63	*H. sapiens*	241	2.965	111.864
MD	M5076	*M. musculus*	197	3.033	114.886
MD	HT-1080	*H. sapiens*	170	2.826	97.731
MD	143B	*H. sapiens*	131	1.283	141.735
MD	MD	*Pseudomonas aeruginosa*	130	0.277	142.432
MD	MD	MD	126	1.898	93.448
MD	rhabdomyosarcoma cell	*H. sapiens*	116	4.036	77.177
MD	CCRF S^{-180}	*M. musculus*	109	0.978	140.984
P13053 (*Vdr*)	MD	*Rattus norvegicus*	64	5.844	60.476
MD	MES-SA	*H. sapiens*	64	2.956	89.631
MD	MD	RSV	61	1.277	127.944
MD	6C3HED	*M. musculus*	60	3.09	97.831
MD	C3H/3T3	MMSV	50	0.327	139.359
P35354 (*PTGS2*)	MD	*H. sapiens*	49	3.515	69.152
MD	A204	*H. sapiens*	44	1.189	106.655
P03359 (*pol*)	MD	WMSV	44	6.786	204.629
MD	MD	*Gallus gallus*	43	0.516	106.529
P37231 (*PPARG*)	MD	*H. sapiens*	40	5.33	83.835
MD	MD	MMSV	39	0.213	166.782
Q07869 (*PPARA*)	MD	*H. sapiens*	37	5.364	81.891
Q13443 (*ADAM9*)	MD	*H. sapiens*	35	2.914	91.186
MD	MD	*R. norvegicus*	34	5.245	64.58
MD	fibroblast	MMSV	33	−1.224	150.956
MD	MD	enterovirus	33	6.348	38.332
MD	MD	human herpesvirus 1	31	6.27	57.306
MD	791T cell line	*H. sapiens*	28	−1.179	139.194
MD	C3H/3T3	*M. musculus*	28	1.745	115.047
P08253 (*MMP2*)	MD	*H. sapiens*	28	3.31	112.85
MD	MD	human enterovirus 71	28	1.967	124.221
P04637 (*TP53*), Q00987 (*MDM2*)	SJSA-1	*H. sapiens*	27	5.213	49.453
P06401 (*PGR*)	MD	*H. sapiens*	26	4.494	32.958
MD	HL-60	*H. sapiens*	25	3.81	33.754

[a]MD, missing data. [b]RSV, Rous sarcoma virus; MLV, murine leukemia virus; MMSV, Moloney murine sarcoma virus; WMSV, Woolly monkey sarcoma virus.

Multiple-Condition Averages in the PTML Antisarcoma Model. In total, we found 83 possible combinations of multiple conditions for all the included sarcoma assays. As shown in Table 5, the $n_j(c_j)$ with the highest number of entries corresponded to tests on human cell lines and on cell lines in *Mus musculus*. The multicondition moving averages (MMAs) used here, $\langle D_1(c_j)\rangle$ and $\langle D_2(c_j)\rangle$, vary significantly along all combinations. However, the anticancer compounds observed for the human osteosarcoma cell lines U2OS, HOS, SAOS-2, MG-63, and 143B and for the fibrosarcoma cell line HT-1080 were in a range of $\langle D_1(c_j)\rangle$ of 1.2−3.7. A similar range was observed in compounds tested in *M. musculus* ($\langle D_1(c_j)\rangle$ = 1−3). Interestingly, when comparing these values with the variation of $\langle D_2(c_j)\rangle$, tests on virus lines, such as Moloney murine sarcoma virus and Woolly monkey sarcoma virus, had higher means (between 140 and 205). Since the ALOGP coefficient is a measure widely used in drug discovery to assess the degree of absorption, distribution in the body, penetration across biological membranes, metabolism, and excretion, this range identified in our results is an important space for the prediction of antisarcoma drugs.[62,63] Likewise, the range of PSA evidenced in viral line assays may be a better space for this coefficient if it is desired to predict new compounds in these experimental conditions. This may be interesting when defining the validation of a certain antisarcoma compound. Thus, if a compound is significantly predicted in an experimental animal or human cell lines, then it will be possible to propose validations at the preclinical level or in clinical trials, respectively.

How to Use the PTML Model in Practice. The model is capable of scoring the activity of a single compound under different assay conditions. To predict a new compound, first, we have to substitute the expected values of function of reference $f(v_{ij})_{ref} = p(f(v_{ij} = 1)_{expt}$ in the model. As

https://dx.doi.org/10.1021/acsomega.0c03356
ACS Omega 2020, 5, 27211−27220

aforementioned, this is the probability of the compound being active for a given biological activity parameter (c_0) (see Table 2). Next, we need to substitute into the equation the values of molecular descriptors D_1 = ALOGP and D_2 = PSA of the compound (chemical structure), calculated with the same algorithm used in the ChEMBL dataset. Last, we have to substitute into the equation the average values (expected values) of the molecular descriptors $\langle D_1(\mathbf{c}_j)\rangle$ for the specific subset of conditions of the assay $\mathbf{c}_j$ we want to predict. In Table 5, we show some selected values of these averages with >25 assays reported. It can be noted that the most populated assays in *Homo sapiens* in the dataset were those *in vitro* assays that targeted the protein O75874 (*IDH1*) and that targeted the cell line U2OS. Upon inspecting Table 5, we can see that $\langle D_k(\mathbf{c}_j)\rangle$ values change for different subsets of conditions $\mathbf{c}_j$. Consequently, when we substitute the different $\langle D_k(\mathbf{c}_j)\rangle$ values into the model for the same compound, we can calculate different scores $f(v_{ij})_{calc}$ of biological activity of the same compound under multiple assay conditions. The full list of the values of $\langle D_k(\mathbf{c}_j)\rangle$ appears in Table S3.

CONCLUSIONS

In this research work, we generated a PTML-LDA model constructed with antisarcoma assays obtained from ChEMBL and a heterogeneous set of different cell lines, organisms, and targets. As far as we know, this constitutes the first time that this kind of model was tested for sarcoma comprising 34,955 chemical compounds and 37,919 assays. The PTML-LDA model was compared with classic ML approaches like the neural network and also with non-PT consideration. The rate of true positives and true negatives is similar when comparing PTML-LDA to other prediction models. PTML-LDA reduces the amount of input variables (ALOGP and PSA) needed, thus increasing the simplicity and interpretability of the model.

METHODS

ChEMBL Data Curation and Preprocessing. In total, we downloaded >370,000 outcomes for preclinical assays of antisarcoma drug candidates from the ChEMBL database. The keywords (fields) used for the search were as follows: Sarcoma (Assay) and also keywords for more relevant cell osteosarcoma lines MG-63, U2O2, HOS, SAOS-2, and 143B. After that, we carried out a data fusion of the datasets obtained into one single raw dataset. The working dataset was curated by eliminating all duplicated entries. We also eliminated all cases with missing values of biological activity (v_{ij}) and/or molecular descriptors. The molecular descriptors used were the same as those precalculated by the ChEMBL database where D_1 = log*P* and D_2 = PSA.[13,14] The final dataset obtained after curation contained 37,919 cases comprising 36 protein targets, 43 cell lines, and 17 assay organisms (Table S1). For comparison and exploration with other models, we additionally computed 12 BCUT molecular descriptors[64] with ChemAxon (http://www.chemaxon.com). The classical unweighted Burden descriptors as well as those weighted by charge and hydrogen bond properties were calculated. The lowest and the three highest eigenvalues were used for descriptor calculation.

To train the model, we split this dataset into two data subsets: training and validation series. We performed a random, stratified, and representative selection of training/validation cases. To accomplish this task, we sorted the cases by n_j (from highest to lowest) as well as by assay conditions: biological activity, protein accession, cell line, and assay organism (alphabetically from A to Z). After this, we selected every fourth case (1 out of 4) to form a training subset (75% of cases) and validation subset (25% of cases). The result of each experimental assay is the value obtained from the quantification of each biological activity and named v_{ij} ("i" and "j" represent the assay and conditions, respectively). Each biological activity depends on the conditions $\mathbf{c}_j$ (c_0, c_1, c_2, ..., c_n) used in each assay. Thus, the conditions taken into account in the data preprocessing were c_0 = biological activity, c_1 = protein accession, c_2 = cell line, and c_3 = assay organism. From v_{ij}, each experimental assay was discretized based on the desirability $d(c_0)$. This variable was defined as 1 when the result of the desired biological activity depended on an increased value of v_{ij} and −1 when the desired biological activity depended on a lower value of v_{ij}. Thus, the discretized value $f(v_{ij})_{obs}$ was calculated as follows: $f(v_{ij})_{obs}$ = 1 when v_{ij} > cut-off and $d(c_0)$ = 1. The function $f(v_{ij})_{obs}$ = 1 when v_{ij} < cut-off and $d(c_0)$ = −1; otherwise, $f(v_{ij})_{obs}$ = 0. The value $f(v_{ij})_{obs}$ = 1 refers to a strong effect of the compound over the target. Since $d(c_0)$ has a direct relationship with $f(v_{ij})_{obs}$, we applied a rational cut-off for each c_0, which will be discussed later. Briefly, the cut-off for properties related to drug concentrations and described in nM (potency, IC_{50}, CC_{50}, EC_{50}, GI_{50}, *etc.*) was set at 100. For properties described in % (inhibition, activity, TGI, among others), the cut-off was set at 50. Last, to calculate the probability of these expected values, we evaluated the relationship between the total number of the observed $n(f(v_{ij}) = 1)_{obs}$ within the level of biological activity desired for the condition $\mathbf{c}_j$ and the total number of compounds n_j that were described in that same condition. In this sense, we have that $p(f(v_{ij})_{obs} = 1)_{expt} = n(f(v_{ij}) = 1)_{obs}/c_0$.

PTML Linear Model. The multicondition moving averages (MMAs) are PTOs similar to Box−Jenkins moving average operators. However, MMAs are PTOs accounting for perturbations (changes) in multiple conditions $\mathbf{c}_j$ at the same time, while MA quantifies changes in only one condition. By using linear discriminant analysis (LDA),[65] we obtained a PTML-LDA equation as follows

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

The model generates an output score $f(v_{ij})_{calc}$ that refers to a score function for a biological activity v_{ij} under the assay conditions $\mathbf{c}_j$. The LDA algorithm includes the Mahalanobis' distance metric,[65] which makes it possible to infer predictive values through a probability calculation $p(f(v_{ij}) = 1)_{pred}$. For the variable selection, we detected specific perturbations within the conditions $\mathbf{c}_j$ that will be adjusted to anticancer properties through a forward-stepwise strategy.[65] Such conditions as c_1 = protein accession, c_2 = cell line, and c_3 = assay organism were significant, so we took them into consideration in our model. Through $p(f(v_{ij}) = 1)_{pred}$, we predicted the activity of each compound by applying the function $f(v_{ij})_{pred}$ = 1 when $p(f(v_{ij}) = 1)_{pred} > 0.5$ or $f(v_{ij})_{pred}$ = 0.

For comparison, we also used a strategy that is not based on perturbation theory. In this sense, besides the molecular descriptors, we added conditions c_1, c_2, and c_3 as a separate set of categorical variables. A total of 237 variables were needed to represent all conditions. Filtering using the variance of each

https://dx.doi.org/10.1021/acsomega.0c03356
ACS Omega 2020, 5, 27211−27220

variable leads to a total of 162 variables, including ALOGP and PSA.

The evaluation of the discriminant model was calculated from Wilks' lambda (Λ) as follows

$$\Lambda = \left[\frac{1}{1+\lambda}\right]$$

where Λ is chi-square distributed for df = $(k-1)$, k is equal to the number of parameters estimated, and $\lambda = \left[\frac{\sum(Z_j - Z)^2}{\sum(Z_{ij} - Z_i)^2}\right]$.

For ML, besides LDA, we also used neural networks (NN) with different architectures. STATISTICA software was used in both cases. The final networks obtained were multilayer perceptron (MLP), linear neural network (LNN), and radial basis function network (RBF). All these ML strategies were applied with perturbation and nonperturbation theory. The predicted 1 or 0 values were used to determine the specificity or true-negative rate (Sp), sensitivity or true-positive rate (Sn), and accuracy (Ac) when compared to the observed values. Thus, when $f(\nu_{ij})_{pre} = f(\nu_{ij})_{obs}$, the cases were determined to be correct.[65]

The metrics to evaluate the performance of all the prediction models were Ac, Sn, and Sp using the following formulae

$$\text{Ac} = \frac{\text{number of correctly classified compounds}}{\text{total number of compounds}}$$

$$\text{Sn} = \frac{\text{number of correctly classified active compounds}}{\text{total number of active compounds}}$$

$$\text{Sp} = \frac{\text{number of correctly classified inactive compounds}}{\text{total number of inactive compounds}}$$

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.0c03356.

ChEMBL dataset of antisarcoma preclinical experimental assays for the PTML model; results of the analyzed models for sarcoma biological activities; all the multiple-condition averages for all sarcoma assays (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Alejandro Cabrera-Andrade − *Grupo de Bio-Quimioinformática and Carrera de Enfermería, Facultad de Ciencias de la Salud, Universidad de Las Américas, Quito 170125, Ecuador; RNASA-IMEDIR, Computer Sciences Faculty, University of A Coruña, A Coruña 15071, Spain;* orcid.org/0000-0001-9702-6618; Email: raul.cabrera@udla.edu.ec

Humbert González-Díaz − *Department of Organic Chemistry II and Basque Center for Biophysics, University of Basque Country UPV/EHU, Leioa 48940, Biscay, Spain; Ikerbasque, Basque Foundation for Science, Bilbao 48011, Biscay, Spain;* orcid.org/0000-0002-9392-2797; Email: humberto.gonzalezdiaz@ehu.es

Authors

Andrés López-Cortés − *RNASA-IMEDIR, Computer Sciences Faculty, University of A Coruña, A Coruña 15071, Spain; Centro de Investigación Genética y Genómica, Facultad de Ciencias de la Salud Eugenio Espejo, Universidad UTE, Quito 170129, Ecuador*

Cristian R. Munteanu − *RNASA-IMEDIR, Computer Sciences Faculty, University of A Coruña, A Coruña 15071, Spain; Biomedical Research Institute of A Coruña (INIBIC), University Hospital Complex of A Coruña (CHUAC), A Coruña 15006, Spain; Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC), Campus de Elviña s/n, A Coruña 15071, Spain;* orcid.org/0000-0002-5628-2268

Alejandro Pazos − *RNASA-IMEDIR, Computer Sciences Faculty, University of A Coruña, A Coruña 15071, Spain; Biomedical Research Institute of A Coruña (INIBIC), University Hospital Complex of A Coruña (CHUAC), A Coruña 15006, Spain*

Yunierkis Pérez-Castillo − *Grupo de Bio-Quimioinformática and Escuela de Ciencias Físicas y Matemáticas, Universidad de Las Américas, Quito 170125, Ecuador*

Eduardo Tejera − *Grupo de Bio-Quimioinformática and Facultad de Ingeniería y Ciencias Aplicadas, Universidad de Las Américas, Quito 170125, Ecuador*

Sonia Arrasate − *Department of Organic Chemistry II and Basque Center for Biophysics, University of Basque Country UPV/EHU, Leioa 48940, Biscay, Spain*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c03356

Author Contributions

[&]A.C.-A. and A.L.-C. contributed equally to the study.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge research grants from Ministry of Economy and Competitiveness, MINECO, Spain (FEDER CTQ2016-74881-P), and Basque government (IT1045-16). The authors also acknowledge the support of Ikerbasque, Basque Foundation for Science. This work was supported by Universidad de Las Américas and 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 Consolidation and Structuring of Competitive Research Units−Competitive Reference Groups (ED431C 2018/49), funded by the Ministry of Education, University and Vocational Training of the Xunta de Galicia endowed with EU FEDER funds.

REFERENCES

(1) Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R. L.; Torre, L. A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394−424.

(2) Hui, J. Y. C. Epidemiology and Etiology of Sarcomas. *Surg. Clin. North Am.* **2016**, *96*, 901−914.

(3) Sidaway, P. Sarcoma: Genetic determinants of sarcoma risk revealed. *Nat. Rev. Clin. Oncol.* **2016**, *13*, 590.

(4) Thomas, D. M.; Ballinger, M. L. Etiologic, environmental and inherited risk factors in sarcomas. *J. Surg. Oncol.* **2015**, *111*, 490−495.

(5) HaDuong, J. H.; Martin, A. A.; Skapek, S. X.; Mascarenhas, L. Sarcomas. *Pediatr. Clin. North Am.* **2015**, *62*, 179−200.

(6) Yang, J.; Ren, Z.; Du, X.; Hao, M.; Zhou, W. The role of mesenchymal stem/progenitor cells in sarcoma: update and dispute. *Stem Cell Investig.* **2014**, *1*, 18.

(7) Double, J.; Barrass, N.; Barnard, N. D.; Navaratnam, V. Toxicity testing in the development of anticancer drugs. *Lancet. Oncol.* **2002**, *3*, 438−442.

(8) Yap, T. A.; Sandhu, S. K.; Workman, P.; de Bono, J. S. Envisioning the future of early anticancer drug development. *Nat. Rev. Cancer* **2010**, *10*, 514−523.

(9) Williams, R. J.; Walker, I.; Takle, A. K. Collaborative approaches to anticancer drug discovery and development: a Cancer Research UK perspective. *Drug Discovery Today* **2012**, *17*, 185−187.

(10) Heinemann, F.; Huber, T.; Meisel, C.; Bundschus, M.; Leser, U. Reflection of successful anticancer drug development processes in the literature. *Drug Discovery Today* **2016**, *21*, 1740−1744.

(11) Sun, J.; Wei, Q.; Zhou, Y.; Wang, J.; Liu, Q.; Xu, H. A systematic analysis of FDA-approved anticancer drugs. *BMC Syst. Biol.* **2017**, *11*, 87.

(12) Carvalho-Silva, D.; Pierleoni, A.; Pignatelli, M.; Ong, C.; Fumis, L.; Karamanis, N.; Carmona, M.; Faulconbridge, A.; Hercules, A.; McAuley, E.; Miranda, A.; Peat, G.; Spitzer, M.; Barrett, J.; Hulcoop, D. G.; Papa, E.; Koscielny, G.; Dunham, I. Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res.* **2019**, *47*, D1056−D1065.

(13) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930−D940.

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

(15) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **2018**, *23*, 1538−1546.

(16) Ali, M.; Aittokallio, T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys. Rev.* **2019**, *11*, 31−39.

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

(18) Pogorzelska, A.; Sławiński, J.; Żołnowska, B.; Szafrański, K.; Kawiak, A.; Chojnacki, J.; Ulenberg, S.; Zielińska, J.; Bączek, T. Novel 2-(2-alkylthiobenzenesulfonyl)-3-(phenylprop-2-ynylideneamino)-guanidine derivatives as potent anticancer agents - Synthesis, molecular structure, QSAR studies and metabolic stability. *Eur. J. Med. Chem.* **2017**, *138*, 357−370.

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

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

(21) Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. SMILES-based QSAR approaches for carcinogenicity and anticancer activity: comparison of correlation weights for identical SMILES attributes. *Anti-Cancer Agents Med. Chem.* **2011**, *11*, 974−982.

(22) González-Díaz, H.; Bonet, I.; Terán, C.; De Clercq, E.; Bello, R.; García, M. M.; Santana, L.; Uriarte, E. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur. J. Med. Chem.* **2007**, *42*, 580−585.

(23) González-Díaz, H.; Viña, D.; Santana, L.; de Clercq, E.; Uriarte, E. Stochastic entropy QSAR for the in silico discovery of anticancer compounds: prediction, synthesis, and in vitro assay of new purine carbanucleosides. *Bioorg. Med. Chem.* **2006**, *14*, 1095−1107.

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

(25) Jung, M.; Kim, H.; Kim, M. Chemical genomics strategy for the discovery of new anticancer agents. *Curr. Med. Chem.* **2003**, *10*, 757−762.

(26) Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI anticancer drug discovery databases: genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189−199.

(27) Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A. A.; Kim, S.; Wilson, C. J.; Lehár, J.; Kryukov, G. V.; Sonkin, D.; Reddy, A.; Liu, M.; Murray, L.; Berger, M. F.; Monahan, J. E.; Morais, P.; Meltzer, J.; Korejwa, A.; Jané-Valbuena, J.; Mapa, F. A.; Thibault, J.; Bric-Furlong, E.; Raman, P.; Shipway, A.; Engels, I. H.; Cheng, J.; Yu, G. K.; Yu, J.; Aspesi, P.; de Silva, M.; Jagtap, K.; Jones, M. D.; Wang, L.; Hatton, C.; Palescandolo, E.; Gupta, S.; Mahan, S.; Sougnez, C.; Onofrio, R. C.; Liefeld, T.; MacConaill, L.; Winckler, W.; Reich, M.; Li, N.; Mesirov, J. P.; Gabriel, S. B.; Getz, G.; Ardlie, K.; Chan, V.; Myer, V. E.; Weber, B. L.; Porter, J.; Warmuth, M.; Finan, P.; Harris, J. L.; Meyerson, M.; Golub, T. R.; Morrissey, M. P.; Sellers, W. R.; Schlegel, R.; Garraway, L. A. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **2012**, *483*, 603−607.

(28) Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. CORAL: classification model for predictions of anti-sarcoma activity. *Curr. Top. Med. Chem.* **2012**, *12*, 2741−2744.

(29) Vos, H. I.; Coenen, M. J. H.; Guchelaar, H.-J.; Maroeska, D.; te Loo, D. M. The role of pharmacogenetics in the treatment of osteosarcoma. *Drug Discovery Today* **2016**, *21*, 1775−1786.

(30) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discovery* **2019**, *18*, 463−477.

(31) 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*, 38602−38610.

(32) M Casañola-Martin, G.; 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 Quadratic Indices for Prediction of Malaria and Cancer Inhibitors Targeting Ubiquitin-Proteasome Pathway (UPP) Proteins. *Curr. Protein Pept. Sci.* **2016**, *17*, 220−227.

https://dx.doi.org/10.1021/acsomega.0c03356
ACS Omega 2020, 5, 27211−27220

(33) 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.
(34) 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.
(35) 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.
(36) Alonso, N.; Caamaño, O.; Romero-Duran, F. J.; Luan, F.; Cordeiro, M. N. D. S.; 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*, 1393−1403.
(37) Gonzalez-Diaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M. General Theory for Multiple Input-Output Perturbations in Complex Molecular Systems. 1. Linear QSPR Electronegativity Models in Physical, Organic, and Medicinal Chemistry. *Curr. Top. Med. Chem.* **2013**, *13*, 1713−1741.
(38) Kleandrova, V. V.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. D. S. Enabling the Discovery and Virtual Screening of Potent and Safe Antimicrobial Peptides. Simultaneous Prediction of Antibacterial Activity and Cytotoxicity. *ACS Comb. Sci.* **2016**, *18*, 490−498.
(39) Speck-Planche, A.; Cordeiro, M. N. D. S. Simultaneous virtual prediction of anti-*Escherichia coli* activities and ADMET profiles: A chemoinformatic complementary approach for high-throughput screening. *ACS Comb. Sci.* **2014**, *16*, 78−84.
(40) Speck-Planche, A.; Cordeiro, M. N. D. S. Speeding up Early Drug Discovery in Antiviral Research: A Fragment-Based in Silico Approach for the Design of Virtual Anti-Hepatitis C Leads. *ACS Comb. Sci.* **2017**, *19*, 501−512.
(41) Speck-Planche, A.; Cordeiro, M. N. D. S. Computer-aided discovery in antimicrobial research: In silico model for virtual screening of potent and safe anti-pseudomonas agents. *Comb. Chem. High Throughput Screening* **2015**, *18*, 305−314.
(42) 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. Diversity* **2017**, *21*, 525.
(43) Speck-Planche, A.; Cordeiro, M. N. D. S. Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol. Diversity* **2017**, *21*, 511−523.
(44) 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. *Anti-Cancer Agents Med. Chem.* **2013**, *13*, 791−800.
(45) Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. Chemoinformatics in multi-target drug discovery for anti-cancer therapy: in silico design of potent and versatile anti-brain tumor agents. *Anti-Cancer Agents Med. Chem.* **2012**, *12*, 678−685.
(46) 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. *Bioorg. Med. Chem.* **2012**, *20*, 4848−4855.
(47) 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*, 273−279.
(48) Cordeiro, M. N.; Speck-Planche, A. Computer-aided drug design, synthesis and evaluation of new anti-cancer drugs. *Curr. Top. Med. Chem.* **2012**, *12*, 2703−2704.
(49) 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. *Bioorg. Med. Chem.* **2011**, *19*, 6239−6244.
(50) Wei, D.-Q.; Selvaraj, G.; Kaushik, A. C. Computational Perspective on the Current State of the Methods and New Challenges in Cancer Drug Discovery. *Curr. Pharm. Des.* **2018**, *24*, 3725−3726.
(51) Bediaga, H.; Arrasate, S.; González-Díaz, H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Comb. Sci.* **2018**, *20*, 621−632.
(52) Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. Fragment-based QSAR model toward the selection of versatile anti-sarcoma leads. *Eur. J. Med. Chem.* **2011**, *46*, 5910−5916.
(53) Chittchang, M.; Gleeson, M. P.; Ploypradith, P.; Ruchirawat, S. Assessing the drug-likeness of lamellarins, a marine-derived natural product class with diverse oncological activities. *Eur. J. Med. Chem.* **2010**, *45*, 2165−2172.
(54) Hansch, C.; Verma, R. P. A QSAR study for the cytotoxic activities of taxoids against macrophage (MPhi)-like cells. *Eur. J. Med. Chem.* **2009**, *44*, 274−279.
(55) Roy, K.; Pratim Roy, P. Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. *Eur. J. Med. Chem.* **2009**, *44*, 2913−2922.
(56) Sarkar, A.; Anderson, K. C.; Kellogg, G. E. Computational analysis of structure-based interactions and ligand properties can predict efflux effects on antibiotics. *Eur. J. Med. Chem.* **2012**, *52*, 98−110.
(57) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational methods in drug discovery. *Pharmacol. Rev.* **2014**, *66*, 334−395.
(58) Leeson, P. Drug discovery: Chemical beauty contest. *Nature* **2012**, *481*, 455−456.
(59) Arnott, J. A.; Planey, S. L. The influence of lipophilicity in drug discovery and design. *Expert Opin. Drug Discovery* **2012**, *7*, 863−875.
(60) Yuan, H.; Paskov, I.; Paskov, H.; González, A. J.; Leslie, C. S. Multitask learning improves prediction of cancer drug sensitivity. *Sci. Rep.* **2016**, *6*, 31619.
(61) Nikolova, O.; Moser, R.; Kemp, C.; Gönen, M.; Margolin, A. A. Modeling gene-wise dependencies improves the identification of drug response biomarkers in cancer studies. *Bioinformatics* **2017**, *33*, 1362−1369.
(62) Waring, M. J. Lipophilicity in drug discovery. *Expert Opin. Drug Discovery* **2010**, *5*, 235−248.
(63) Giaginis, C.; Tsopelas, F.; Tsantili-Kakoulidou, A. The Impact of Lipophilicity in Drug Discovery: Rapid Measurements by Means of Reversed-Phase HPLC. *Methods Mol. Biol.* **1824**, *1824*, 217−228.
(64) Burden, F. R. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.
(65) Hill, T.; Lewicki, P. STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining. In *Discriminant Function Analysis*; 1st ed.; StatSoft, Inc.: 2006; 155−164.

https://dx.doi.org/10.1021/acsomega.0c03356
ACS Omega 2020, 5, 27211−27220