ACS OMEGA

Article

Cite This: ACS Omega 2019, 4, 3122−3132

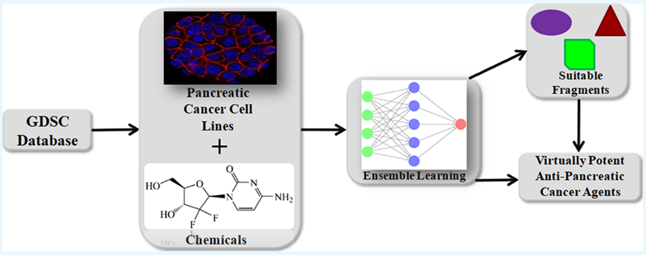
http://pubs.acs.org/journal/acsodf

Multicellular Target QSAR Model for Simultaneous Prediction and Design of Anti-Pancreatic Cancer Agents

Alejandro Speck-Planche*

Research Program on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM), 08003 Barcelona, Spain

Supporting Information


GDSC Database
Pancreatic Cancer Cell Lines
+
HO
O
N
NH2
HO
F
F
Chemicals
Ensemble Learning
Suitable Fragments
Virtually Potent Anti-Pancreatic Cancer Agents

ABSTRACT: Pancreatic cancers are widely recognized as a group of neoplasms with one of the poorest prognoses in oncology research. Despite the advances achieved in drug design and development, there is no effective cure for pancreatic cancers, and the current chemotherapeutic regimens increase the survival rate by only a few months. As an integral part of all modern drug discovery campaigns, computer-aided approaches can represent a promising alternative change to accelerate the early discovery of potent anti-pancreatic cancer agents. To date, however, most of the efforts made so far have focused on small series of structurally related chemicals, where the anti-pancreatic cancer activity has been measured against only one cancer cell line. In addition, no rational insight has been provided in the sense of unveiling the physicochemical aspects and the structural features that the molecules should possess to increase the anti-pancreatic cancer activity. This work reports the first multicellular target QSAR model based on ensemble learning (mct-QSAR-EL) that allows the simultaneous prediction and design of molecules with activity against different pancreatic cancer cell lines, which exhibit different degrees of sensitivity to chemical treatment. The mct-QSAR-EL model displayed sensitivities and specificities higher than 80% in both training and test sets. The physicochemical and structural interpretations of the molecular descriptors in the model permitted the selection of several fragments with potentially positive contributions to the increase of the anti-pancreatic cancer activity. These fragments were then assembled to design new molecules. The designed molecules were predicted as multicell line inhibitors by the mct-QSAR-EL model, and these results converged with the predictions performed by recently reported models. The designed molecules complied with Lipinski's rule of five and its variants.

1. INTRODUCTION

Pancreatic cancers remain as one of the most aggressive groups of neoplasms across the world. At the same time, pancreatic cancers are associated with a very poor prognosis; the early diagnosis is difficult, and most patients have already progressed to not resectable and incurable statuses at diagnosis.[1] The lethality involving pancreatic cancers is evidenced by the very high mortality rate which is almost equal to the incidence rate; the overall 5-year survival rate is approximately 6% (2−9%).[2] According to a recent report, pancreatic cancers are responsible for causing 331 000 deaths per year, ranking as the 7th leading cause of cancer death, and 11th among the most common cancers.[3] In addition, the reduced chemotherapeutic options to treat pancreatic cancers have become ineffective because of the emergence of resistance.[4−6] Altogether, the design of anti-pancreatic cancer agents constitutes one of the most challenging tasks for the scientific community working in oncology.

Nowadays, computational approaches constitute an integral part of the drug development campaigns,[7] accelerating the early discovery of effective therapeutics, while serving as great allies of powerful experimental methods such as high-throughput screening.[8] Nevertheless, in the context of pancreatic cancer research, the few recent computational models reported in the literature[9−12] have at least one of the following limitations. First, a series of structurally related molecules are used to develop the model. This aspect restricts the applications of the model because of the small chemical space that is analyzed. Second, the anticancer activity is predicted against only one pancreatic cancer cell line. This factor is detrimental because different pancreatic cancer cell lines can exhibit varying degrees of sensitivity/resistance to a same chemical. Last, the models published to date cannot provide sufficiently clear information regarding structural requirements that a molecule should posses to increase the cytotoxicity and versatility against the different pancreatic cancer cell lines.

In the last 5 years, several research groups have emphasized the need to use multitarget models for quantitative structure−activity relationships (mt-QSAR) as an attempt to solve the aforementioned drawbacks. In this context, some computational models that involve the application of the mt-QSAR

Received: December 31, 2018
Accepted: January 30, 2019
Published: February 12, 2019

DOI: 10.1021/acsomega.8b03693

Table 1. Molecular Descriptors Present in the mct-QSAR-EL Model

molecular descriptor	concept
$D[TssAq_1(PSA)]c_t$	deviation of the total stochastic atom-based quadratic index of order 1 weighted by the polar surface area and depending on the pancreatic cancer cell lines against which the chemicals were assayed
$D[LnsAq_7(HYD)A]c_t$	deviation of the local nonstochastic atom-based quadratic index of order 7 weighted by the hydrophobicity, being based on the atoms acting as hydrogen bond acceptors and their chemical environments, and depending on the pancreatic cancer cell lines against which the chemicals were assayed
$D[LnsAq_4(HYD)C]c_t$	deviation of the local nonstochastic atom-based quadratic index of order 4 weighted by the hydrophobicity, being based on the aliphatic carbon atoms and their chemical environments, and depending on the pancreatic cancer cell lines against which the chemicals were assayed
$D[LnsAq_0(HYD)D]c_t$	deviation of the local nonstochastic atom-based quadratic index of order 0 weighted by the hydrophobicity, being based on the atoms acting as hydrogen bond donors and their chemical environments, and depending on the pancreatic cancer cell lines against which the chemicals were assayed
$D[LnsAq_7(HYD)D]c_t$	deviation of the local non-stochastic atom-based quadratic index of order 7 weighted by the hydrophobicity, being based on the atoms acting as hydrogen bond donors and their chemical environments, and depending on the pancreatic cancer cell lines against which the chemicals were assayed
$D[LssAq_2(E)G]c_t$	deviation of the local stochastic atom-based quadratic index of order 2 weighted by the electronegativity, being based on halogen atoms and their chemical environments, and depending on the pancreatic cancer cell lines against which the chemicals were assayed
$D[LssAq_3(E)M]c_t$	deviation of the local stochastic atom-based quadratic index of order 3 weighted by the electronegativity, being based on carbon atoms of terminal methyl groups and their chemical environments, and depending on the pancreatic cancer cell lines against which the chemicals were assayed

paradigm have been developed to simultaneously predict the activity against multiple biomolecular[13−21] and/or non-biomolecular[22−28] targets. Indeed, the mt-QSAR models belong to a selected group of advanced *in silico* tools that combine perturbation theory ideas and machine learning (PTML) modeling; PTML models use mathematical operators that characterize the deviations (perturbations) of a query chemical with respect to the average (expected) values for all of the drugs/chemicals tested against the same biological entity (microorganisms, cell lines, laboratory animals, etc.), by measuring the same effect (activity, toxicity, pharmacokinetic properties, among others), or by considering the same assay protocol. Advanced PTML models (also known as multi-scale models) have been successfully employed in many diverse areas such as organic chemistry,[29,30] material science, and nanotechnology,[31−35] immunology and immunotoxicity,[24,27,36] antimicrobial research,[37−41] neuroscience,[13,22,42,43] and cancer research.[23,44−49]

Bearing in mind all these ideas, this work reports the first multicellular target QSAR model based on ensemble learning (mct-QSAR-EL) that allows the simultaneous prediction and design of molecules with activity against different pancreatic cancer cell lines that exhibit different varying degrees of sensitivity to chemical treatment. Here, a fragment-based interpretation of the molecular descriptors in the mct-QSAR model is provided with the purpose of extracting relevant information regarding the physicochemical properties and structural features that are needed for the enhancement of the anti-pancreatic cancer activity. It is shown that by assembling different molecular fragments with favorable contribution to the activity under study, different molecules can be designed as potentially effective and versatile anti-pancreatic cancer agents.

2. RESULTS AND DISCUSSION

2.1. mct-QSAR-EL Model. The mct-QSAR-EL model was found by analyzing different ensembles of artificial neural networks (ANNs) with different architectures such as linear neural networks (LNN), radial basis function (RBF), and multilayer perceptron (MLP). The objective was to find the best ensemble, that is, the most appropriate combination of ANNs that gives the best consensus predictions. The best mct-QSAR-EL model developed in this study contains the profile Output 7:[5]:1; this is an ensemble model formed by seven molecular descriptors used as inputs, five ANNs based on the RBF architecture, and it has an output variable, which is the predicted categorical value of the inhibitory activity [Pred_APCA$_i$(c_t)] against the pancreatic cancer cell lines. The symbols and definitions of the molecular descriptors present in the mct-QSAR-EL model are represented in Table 1, while all chemical and biological data are available in the Supporting Information 1 file.

The accuracy of the mct-QSAR-EL model is 85.42% in the training set, which means that 3714 out of 4348 cases/drugs were correctly classified, while the same metric for the test set is 81.99% (1188 out of 1449 chemicals). In addition, other statistical indices confirm the very good performance exhibited by the mct-QSAR-EL model. In this sense, the sensitivity [Sn(%)] and specificity [Sp(%)] values define the percentages of correctly classified active and inactive molecules, respectively; they have values higher than 80% in both training and test sets (Table 2). In addition, the classification results for each drug in the dataset used in this work are reported in the Supporting Information 2. Also, in the same table, the Matthews correlation coefficient (MCC) is higher than 0.6. If we consider that MCC can take values from −1 (the poorest performance) to 1 (perfect prediction/classification), with 0 indicating a random performance, it can be concluded that the MCC > 0.6 reflects a strong correlation between the observed and predicted values of the categorical variable of anti-pancreatic cancer activity APCA$_i$(c_t).

Table 2. Measures of Internal and External Performance of the mct-QSAR-EL Model

symbols[a]	training set	test set
N_{active}	1562	520
CC_{active}	1322	425
Sn(%)	84.64%	81.73%
$N_{inactive}$	2786	929
$CC_{inactive}$	2392	763
Sp(%)	85.86%	82.13%
MCC	0.692	0.623

[a] N_{active}—total number of active molecules; $N_{inactive}$—total number of inactive molecules; CC_{active}—molecules correctly classified as active; $CC_{inactive}$—molecules correctly classified as inactive; Sn(%)—sensitivity (percentage of molecules correctly classified as active); Sp(%)—specificity (percentage of molecules correctly classified as inactive); MCC—Matthews' correlation coefficient.

It should be noted that measures such as Sn(%) and Sp(%) only offer information regarding the global performance of the

Table 3. Different mt-QSAR/PTML Models Focused on the Prediction of Anticancer Activity

cancer type	cancers	BJ operator[a]	ML[b]	NV[c]	cases[d]	Sn(%)[e]	Sp(%)[f]	refs
multiple cancers	>10	MCMA	LDA	>10	116 934	>70 (>70)	≈90 (≈90)	23
multiple cancers	>10	MCMA	LDA	3	116 934	>70 (>70)	>90 (>90)	23
multiple cancers	>10	MCMA	ANN (LNN)	4	116 934	>80 (>80)	>80 (>80)	23
bladder cancer	1	SCMA	LDA	>10	664	>90 (>90)	>90 (>90)	44
bladder cancer	1	SCMA	ANN (RBF)	10	664	>95 (>95)	>95 (>95)	44
brain tumors	1	SCMA	LDA	>10	1236	≈90 (>85)	>90 (>85)	45
breast cancer	1	SCMA	LDA	>10	2272	>85 (≈90)	>90 (>90)	46
colorectal cancer	1	SCMA	LDA	>10	1651	>90 (>90)	>90 (>90)	47
colorectal cancer	1	SCMA	ANN (RBF)	>10	1651	>95 (>90)	>95 (>90)	47
prostate cancer	1	SCMA	LDA	>10	1668	>85 (≈90)	>90 (>90)	48
sarcoma	1	SCMA	LDA	>10	3017	>90 (>90)	>90 (>90)	49
pancreatic cancer	1	SCMA	E-ANN−RBF	5	5797	≈85 (>80)	>85 (>80)	this work

[a] Box−Jenkins operator. MCMA—multicondition moving average; MCMA considers the multiple elements of the experimental protocol at the same time (e.g., target, the measure of biological effect, and/or details such as the accuracy of the assay). SCMA—single-condition moving average; SCMA considers only one aspect of the experimental condition (e.g., the target in the case of the works mentioned here). [b] ML method used to develop the PTML models. LDA—linear discriminant analysis. ANN—artificial neural networks. LNN—linear neural networks. RBF—radial basis function. E-ANN−RBF—ensemble of artificial neural networks based on the RBF architecture. [c] Number of variables (molecular descriptors) in each PTML model. [d] Number of compounds. [e] Sensitivity. [f] Specificity. [e,f] The values in parentheses correspond to the test set.

model. We are dealing here with a model able to simultaneously classify/predict dissimilar drugs/molecules against different pancreatic cancer cell lines. Therefore, here, the local measures of these statistical indices were calculated, namely, $[\mathrm{Sn}(\%)]c_t$ and $[\mathrm{Sp}(\%)]c_t$, respectively; they depend on the drugs against which each pancreatic cancer cell lines was experimentally tested. The results gathered from the Supporting Information 3 indicate that $[\mathrm{Sn}(\%)]c_t$ is in the range 69.79−96.15% for the training set, and 63.64−100% for the test set. At the same time, $[\mathrm{Sp}(\%)]c_t$ exhibits values in the intervals 77.97−95 and 68.75−100% for the training and test sets, respectively. These local statistical indices demonstrate that the mct-QSAR-EL model continues to perform fairly well.

2.2. Comparison with Other mt-QSAR/PTML Models Focused on Anticancer Chemicals. There have been several mt-QSAR/PTML models reported in the scientific literature, which predict the inhibitory activity of chemicals against different cancer cell lines.[23,44−49] All of the models reported to date have advantages and limitations. For instance, by using the Box−Jenkins operators known as multicondition moving averages (MCMA), the first three models reported in Table 3 integrate and predict huge amounts of chemical and biological data focused on multiple cancers. This is a very favorable aspect because these models can be used as computational tools to perform virtual screening of large and heterogeneous databases. Another advantage of these three models is that they consider predictions not only against cellular targets but against cancer-related proteins. This permits to gather information regarding the potential mechanisms of action through which a chemical may act. These two advantages constitute drawbacks of the other models represented in Table 3, which employ single condition moving averages (SCMA).

Nevertheless, the main limitation of the first three models is that they do not offer a rationale that leads to the design of new molecules with potential anticancer activity. Such a limitation is the strong point of the remaining models that are focused on modeling anticancer against a defined type of cancer. In this sense, these other mt-QSAR/PTML models (including the present work) offer a fragment-based view of the structural features that are required to increase the anticancer activity. Indeed, the main contribution of the mt-QSAR/PTML model developed in this work is that it demystifies the belief regarding the inability of the nonlinear models to be interpreted in a physicochemical and structural context. The present mt-QSAR/PTML model is also exploited as a knowledge generator, allowing the guided design of new molecules with potential anticancer activity. In any case, all of the mt-QSAR/PTML models depicted in Table 3 exhibit very good performances. This suggests that they could be used as parts of a unified computational framework devoted to screen and filter the chemical space for the future discovery of anticancer agents.

2.3. Applicability Domain. As most of the computational models are employed to predict large and heterogeneous chemical libraries, it is important to estimate the reliability of the predictions. In this context, applicability domain is a well-known concept devoted to define the regions in which any model will predict accurately and reliably. A series of applicability domain approaches have been reported in the literature,[50] and to date, there is no consensus regarding the superiority of one approach over the others. Most of the applicability domain approaches are based on defining interpolation regions. Nevertheless, the consensus predictions performed by ensemble learning models such as those based on random forest have also been reported to provide reliable extrapolation regions.[50] Bearing in mind these ideas, the mct-QSAR-EL model developed here can in principle give reliable predictions in the extrapolation regions because of its ensemble learning nature.

On the other hand, the applicability domain focused on the interpolation region was also defined by the descriptor space approach. The results regarding the applicability domain based on the descriptors' space appear in the Supporting Information 4. To define such an applicability domain, the following steps were followed. First, in the subset formed by the correctly classified chemicals in the training set, the maximum and minimum values were determined. Second, the information of the maximum and minimum values of each molecular descriptor was used to calculate local scores. If a defined descriptor value of a new chemical was between the maximum and minimum values, the local score was annotated with a value equal to one; otherwise, the local score took the value of zero. As the mct-QSAR-EL model was built by considering seven molecular descriptors, the total score (calculated as the sum of the local scores) must have a value of seven in order. Only the compounds with total score equal to seven were considered to be inside the applicability domain of the model. Nine chemicals were found to be outside the applicability domain (Supporting Information 4).

2.4. Fragment-Based Interpretation of the Molecular Descriptors in the mct-QSAR-EL Model. Nowadays, QSAR models, as well as other predictive chemoinformatic tools, are used with the sole purpose of performing virtual screening of large dataset of chemicals. As a result, the ability of a predictive model to provide insightful physicochemical and structural information is frequently neglected and underestimated.[51] In this sense, here, physicochemical and structural interpretations of the molecular descriptors in the mct-QSAR-EL model will be given. Such interpretations will be elucidated by relaying on the sensitivity values of the molecular descriptors, which reflect their relative influences in the model (Figure 1). The greater the SV of a molecular descriptor, the more significant will be that descriptor.

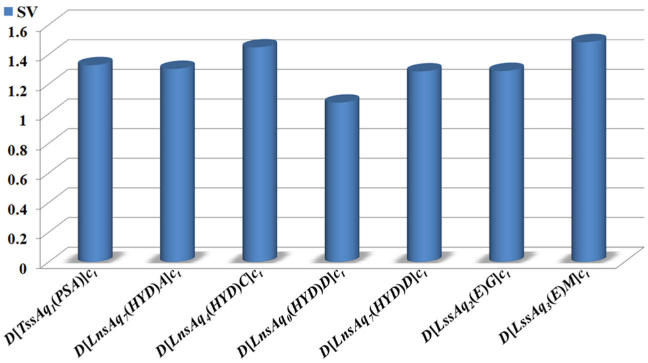


Figure 1. Significance of each molecular descriptor according to its sensitivity value (SV).

A reasonable interpretation of the molecular descriptors in a model may be provided if the model had been developed by using linear data analysis methods such as multiple linear regression or linear discriminant analysis. However, the mct-QSAR-EL model created in this study is nonlinear, which in principle may hinder the clear interpretation of the molecular descriptors. This inconvenience was solved by Speck-Planche and co-workers, who proposed an approach that allowed the extraction of physicochemical and structural information from the molecular descriptors of a nonlinear model.[52] This approach focuses on the calculation of the class-based mean values (Table 4). In this sense, for each molecular descriptor, two mean values are calculated: one for the molecules assigned and correctly predicted as active and the other for the molecules annotated and correctly classified as inactive. This procedure is applied only to those molecules present in the training set.

Notice that the mct-QSAR-EL model was able to discriminate the active molecules from the inactive ones with good accuracy. This has been possible because active and inactive molecules exhibit physicochemical properties and structural characteristics, which are specific for each class. Such a difference will be reflected in some way in the mean values of each molecular descriptor for each class. By comparing the mean values of each molecular descriptor for the two classes, one can determine a tendency that suggests how the physicochemical properties and the structural features should vary to increase the anti-pancreatic cancer activity. It should be pointed out that such a tendency is relative because of the nonlinear nature of the mct-QSAR-EL model. Nevertheless, this assumption is rigorous enough to ensure an adequate interpretation of the molecular descriptors.[52]

Table 4. Tendency of Variability of Each Molecular Descriptor in the mct-QSAR-EL Model According to the Class-Based Mean Values

class-based means	active	inactive	tendency[a]
$D[TssAq_1(PSA)]c_t$	−0.004	0.024	decrease
$D[LnsAq_7(HYD)A]c_t$	0.004	−0.030	increase
$D[LnsAq_4(HYD)C]c_t$	0.011	−0.080	increase
$D[LnsAq_0(HYD)D]c_t$	0.001	−0.006	increase
$D[LnsAq_7(HYD)D]c_t$	0.001	−0.017	increase
$D[LssAq_2(E)G]c_t$	−0.001	0.025	decrease
$D[LssAq_3(E)M]c_t$	0.016	−0.091	increase

[a] Tendency—referred to the potential variation (increase or diminution) of a molecular descriptor in order to increase the antimalarial activity.

Bearing in mind all of the ideas discussed in this subsection, although physicochemical and structural information will be gathered from the different molecular descriptors, they will also be interpreted from a fragment-based point of view. This means that when interpreting each molecular descriptors, different fragments/functional groups will be mentioned as examples of substructural alerts that can positively contribute to the increment of the anti-pancreatic cancer activity. This idea is supported by the fact that it is well established that any topological (graph-based) index of a molecule can be represented as a linear combination of the frequency with which diverse fragments (connected and disconnected) appear in the molecule.[53] The use of fragment-based approaches has been suggested to guide medicinal chemists in the fast identification of 2D pharmacophores.[43]

The molecular descriptors employed in this work consider the bond order, and the number represented as a subscript in each of them indicates the topological distance, that is, the number of bonds (without counting the bond multiplicity) that exist between any two atoms. There are four molecular descriptors that contain information regarding the role of the hydrophobicity (HYD) in the biological activity under study. These are $D[LnsAq_7(HYD)A]c_t$, $D[LnsAq_4(HYD)C]c_t$, $D[LnsAq_0(HYD)D]c_t$, and $D[LnsAq_7(HYD)D]c_t$. These descriptors are weighted by the "joint hydrophobic contribution", a term that indicates the multiplication of the HYD values of any two atoms according to the HYD scale proposed by Ghose−Crippen approach.[54−57] In this sense, in a molecule, different atom types are generated by considering a defined atom and its neighbors; hydrogen and halogens atoms are classified by the hybridization and oxidation states of the carbon atom to which they are attached. On the other hand, carbon atoms are classified by their hybridization state and depending on whether their chemical environment includes other carbons or heteroatoms; for the case of the hydrogen atoms, the heteroatoms attached to a carbon atom in α position are also considered.[54−57]

From one side, $D[LnsAq_7(HYD)A]c_t$ and $D[LnsAq_7(HYD)D]c_t$ are the fourth and sixth most influential descriptors in the

DOI: 10.1021/acsomega.8b03693
ACS Omega 2019, 4, 3122−3132

mct-QSAR model, respectively. They both characterize the increment of the joint hydrophobic contributions of any two atoms placed at the topological distance equal to 7. However, in $\boldsymbol{D[LnsAq_7(HYD)A]c_t}$, at least one of the two atoms should act as a hydrogen bond acceptor (HBA) and in $\boldsymbol{D[LnsAq_7(HYD)D]c_t}$, at least one of the two atoms should be a hydrogen bond donor (HBD). Examples of HBAs are all nitrogen and oxygen atoms, excluding nitrogens with a formal positive charge and/or higher oxidation states, as well as pyrrolic nitrogens. Fragments such as 4-(pyridin-4-yl)pyridine, 2-(4-(dimethylamino)phenyl)acetonitrile, and 2-(4-methoxybenzyl)-3-methyl-1*H*-pyrrole can increase $\boldsymbol{D[LnsAq_7(HYD)A]c_t}$. On the other hand, HBDs are $-OH$, $-NH_2$, $-NHR$ (with R being an alkyl group), and fragments such as (2-methylnaphthalen-7-yl)methanamine, (4-isopropylphenyl)methanamine, and (1,2,3,4-tetrahydronaphthalen-7-yl)methanol can increase the value of $\boldsymbol{D[LnsAq_7(HYD)D]c_t}$. It should be noted that there can be fragments, which can increase both $\boldsymbol{D[LnsAq_7(HYD)A]c_t}$ and $\boldsymbol{D[LnsAq_7(HYD)D]c_t}$ (Figure 2).

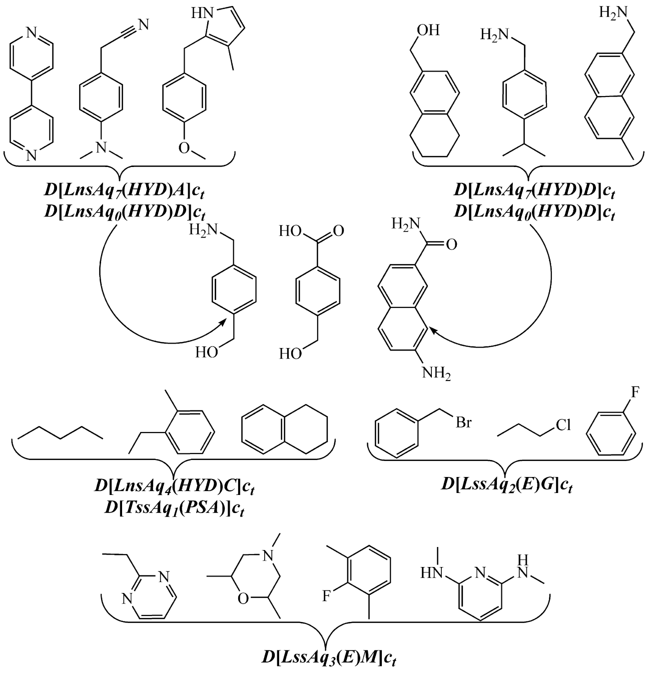


Figure 2. Fragments directly selected from the physicochemical and structural interpretations of the molecular descriptors in the mct-QSAR-EL model. Some fragments are favorable for more than one molecular descriptor.

Another HYD-based descriptor is $\boldsymbol{D[LnsAq_4(HYD)C]c_t}$, which indicates the augmentation of the joint hydrophobic contribution of any two atoms placed at a topological distance equal to 4, with one of these atoms being an aliphatic carbon. In this sense, $\boldsymbol{D[LnsAq_4(HYD)C]c_t}$ is the second most important descriptor, and fragments such as aliphatic chains containing five carbon atoms, dialkyl substituted benzenes, or fused systems containing aliphatic and aromatic rings can increase $\boldsymbol{D[LnsAq_4(HYD)C]c_t}$. Finally, the other descriptor focused on hydrophobic factors is $\boldsymbol{D[LnsAq_0(HYD)D]c_t}$ (the lowest importance), which emphasizes the increase of the number of atoms acting as HBDs.

In the model developed here, there are two descriptors that depend on electronic factors. One of them is $\boldsymbol{D[LssAq_2(E)G]c_t}$, which involves the decrease of the electronegative in those regions where any two atoms are placed at the topological distance equal to 2. Having the fifth greatest significance in the model, $\boldsymbol{D[LssAq_2(E)G]c_t}$ can be decreased by fragments containing as few halogens as possible or no halogens at all. Some fragments such as of mono-halo-substituted benzenes and alkyl or benzyl halides (Figure 2) can decrease the value of $\boldsymbol{D[LssAq_2(E)G]c_t}$. The electronic aspects of the molecules are also characterized by $\boldsymbol{D[LssAq_3(E)M]c_t}$. This is the most important descriptor in the model, and it indicates the increase of the electronegativity in those regions where any two atoms are placed at a topological distance equal to 3, with one of these atoms belonging to a methyl group. The fragments known to increase $\boldsymbol{D[LssAq_3(E)M]c_t}$ are 2-ethylpyrimidine, 2,4,6-trimethylmorpholine, 2-fluoro-1,3-dimethylbenzene, and N^2,N^6-dimethylpyridine-2,6-diamine.

Last, we have the descriptor $\boldsymbol{D[TssAq_1(PSA)]c_t}$ (the third most influential), which contains information regarding the diminution of the global polar surface area (PSA) by considering each atom and its adjacent neighbors. Any aliphatic or aromatic without heteroatoms able to form hydrogen bonds can decrease the value of $\boldsymbol{D[TssAq_1(PSA)]c_t}$.

2.5. Designing Versatile Inhibitors against Multiple Pancreatic Cancer Cell Lines.

Despite the successful applications, the computational models based on ML methods as tools for virtual screening and drug design[58−61] have been heavily criticized because they are treated as black boxes. They also require specialized personnel with high programming/codification skills to develop the adequate algorithms for prediction. On the other hand, ML models are not capable of providing an interpretation of the physicochemical and/or structural features of the chemicals in the databases, which they intend to analyze and predict. In addition, there are other factors such as the uncertainty associated with the experimental data and the yet unresolved issue that the molecular descriptors reported to date can properly characterize a reduced fraction of the information contained in the molecules. As a result, any model devoted to performing predictions of external databases will perform reliable predictions to some extent.

It should be pointed out that each model will always attempt to predict new molecules according to two aspects: the predetermined values of the molecular descriptors used to create it and the hierarchy of the molecular descriptors in the model. The first aspect is in some way controlled by the different applicability domain approaches. However, the second aspect is as crucial as the first one, and it is usually neglected. Notice that the molecular descriptors in the training set will have different degrees of hierarchy/importance in a model, while embodying different physicochemical and structural information that may need to be changed with the aim of improving a biological effect under study. In the end, it should be expected that only if a query chemical follows such a degree of the hierarchy in the model, this should be predicted with a higher accuracy, and the prediction should have a strong agreement with the experimental result. Because of the importance of understanding the second aspect mentioned here, recent approaches focused on the fragment-based topological design have emerged as a promising alternative to optimize chemical space.[18,51,62−65] Such approaches provide the computational models with a better way to "control" the chemical space to be predicted; it is suggested that to achieve a higher certainty in the predictions, the chemicals must be designed according to the physicochemical and structural interpretations of the models.[18,51,62−65]

Here, six molecules were designed according to the following steps (Figure 3). First, from each molecular descriptor present in the model, a fragment assumed to be favorable for that descriptor was selected. Second, the new molecules were designed by connecting and/or fusing the different favorable fragments. Third, if needed, certain atoms or functional groups were added to the structure of the molecule. As the last and most crucial step, all of the procedures involving the three previous steps were carried out by considering the physicochemical and structural interpretations of the molecular descriptors in the mct-QSAR-EL model.

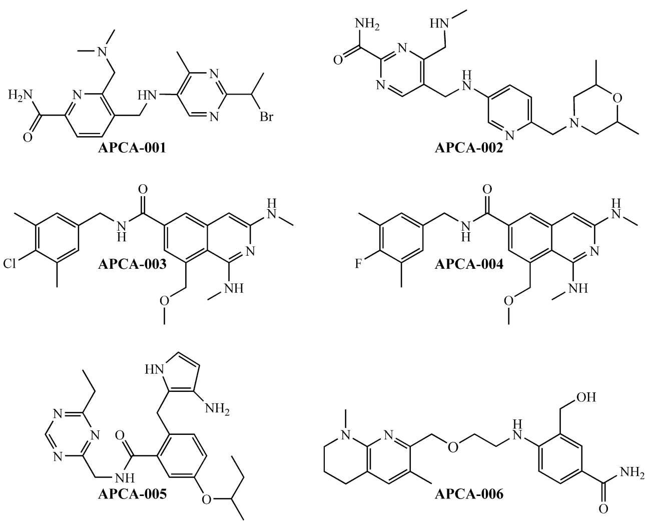


Figure 3. Molecules designed by the mct-QSAR-EL model using a fragment-based topological approach.

The six designed molecules fall inside the applicability domain of the model (Supporting Information 5). The molecules APCA-001, APCA-002, and APCA-006 were predicted as anti-pancreatic cancer chemicals (IC_{50} < 6.45 μM) against the 31 pancreatic cancer cell lines used in this study. At the same time, the molecule APCA-004 was predicted as cytotoxic against 28 out of 31 pancreatic cancer cell lines. Interestingly, the molecule APCA-003, being an analogue of APCA-004, was predicted as active against only nine cell lines. This result demonstrates that the molecular descriptors used to build the mct-QSAR-EL model are very sensitive to small changes in the physicochemical properties and structural aspects of the molecules. Of course, this prediction does not exclude APCA-003 as a multicellular target inhibitor, but it indicates that its IC_{50} value against the other 22 pancreatic cell lines may be higher than 6.45 μM. Finally, although the same rules were followed when designing the six molecules, APCA-005 was predicted as inactive against all pancreatic cancer cell lines. Again, this result leads to two potential conclusions. From one side, APCA-005 may be active against many cell lines but with IC_{50} > 6.45 μM, so it may undergo further design. On the other hand, APCA-005 may be discarded because the design did not yield the desired result. The predictions performed by the mct-QSAQR-EL model against the designed molecules are available in the Supporting Information 5.

Despite their relative simple structures, the six designed molecules may represent new families of chemicals for future biological testing. In this sense, three out of six molecules were predicted as inhibitors against all of the pancreatic cell lines reported here: APCA-001, APCA-002, and APCA-006. To gain more knowledge regarding the novelty of these molecules, a search was conducted in widely recognized chemical databases such as ChEMBL,[66] ZINC,[67] and ChemSpider.[68] A similarity cutoff was set to be ≥0.8. The search did not produce any result of similar molecules, except for the case of the molecule APCA-006 for which one match was found (ChemSpider ID: 10169697). However, the molecule found is a system of four fused rings, with no HBDs, while APCA-006 has several rotatable bonds, and three atom/functional groups were able to act as HBDs. Regardless of the criterion used, from the point of view of a medicinal chemist, there is not chemical similarity between APCA-006 and its apparent match ChemSpider ID: 10169697.

2.6. Druglikeness. Assessing the druglikeness is an essential step in the early discovery of potentially active and safer chemicals. In this sense, the Lipinski's rule of five[69] is a well-established method to evaluate if a molecule has chemical properties and physical properties that would make it an orally active drug in humans. This rule states that to achieve an adequate oral bioavailability, a molecule should have the following requirements: no more than 5 HBD, 10 or fewer HBA, a molecular weight (MW) lower than 500 Da, and an octanol−water partition coefficient (Clog *P*) not greater than 5. At the same time, a second variant of the aforementioned rule, known as the Ghose filter,[70] suggests Clog *P* in the range −0.4 to +5.6, molar refractivity in the interval 40−130 cm^3/mol, MW in the range 180−480 Da, and the number of atoms (nAT) from 20 to 70. On the other hand, Veber and co-workers have criticized the cutoff value of atomic weight (AW), indicating that the molecules with PSA lower than ≤140 $Å^2$ and the number of rotatable bonds (RBN) to be ≤10 are likely to have good oral bioavailability.[71] All of the

Table 5. Molecular Properties Calculated for the Molecules Designed by the mct-QSAR-EL Model

ID[a,b]	HBD	HBA	MW (Da)	Clog *P*	MR (cm^3/mol)	nAT	PSA ($Å^2$)	RBN
APCA-001	3	7	407.32	1.543	103.55	48	97.03	7
APCA-002	4	9	399.5	−0.151	112.77	58	118.29	8
APCA-003	3	6	426.95	5.687	126.41	57	75.28	7
APCA-004	3	7	410.49	5.117	121.82	57	75.28	7
APCA-005	4	7	408.51	1.084	119.25	58	118.81	9
APCA-006	4	7	384.48	1.618	112.82	56	100.71	8

[a] Hydrogen bond donors (HBDs) and hydrogen bond acceptors (HBAs) were calculated manually. In addition, fluorine atoms were considered as HBAs, while the pyrrolic nitrogens were not. [b] MW—molecular weight; Clog *P*—logarithm of the octanol/water coefficient; MR—molar refractivity; nAT—number of atoms (hydrogen atoms included); PSA—polar surface area; RBN—number of rotatable bonds.

properties mentioned in these three rules were estimated for the six designed molecules (Table 5); HBD and HBA were calculated manually, Clog P was calculated by the program ChemDraw Ultra v8.0.,[72] and the other properties were determined by the software Marvin Sketch v15.11.16.0 from ChemAxon.[73] Except APCA-003 and its analog APCA-004, which are very hydrophobic, the other molecules obeyed the Lipinski's rule of five and its variants.

3. CONCLUSION

The urgent need to find effective treatments against pancreatic cancers has paved the way toward the application of advanced computational approaches to accelerate the early discovery of antineoplastic agents. More efforts should be focused on extracting relevant phenomenological information from the predictive *in silico* models. The mct-QSAR-EL model reported here constitutes a good example of how the anticancer activity can be predicted with high accuracy against multiple pancreatic cancer cell lines. The fragment-based physicochemical and structural interpretation of the molecular descriptors in the model was essential in the design of virtually new molecules. Four of the six molecules designed by the mct-QSAR-EL model were predicted as highly versatile anti-pancreatic cancer chemicals, and they are now available to the scientific community as potential leads for future biological assays. This work also permits to envisage new horizons in pancreatic cancer research through the development and use of focused chemical libraries, which can find wide utility in other therapeutic areas within oncology.

4. MATERIALS AND METHODS

4.1. Database and Molecular Descriptors.

All chemical and biological data were extracted from the database known as Genomics of Drug Sensitivity in Cancer (GDSC).[74−76] The dataset contains 231 FDA-approved or investigational drugs experimentally tested against 31 pancreatic cancer cell lines. In the moment of retrieving the data, not all drugs were tested against all pancreatic cancer cell lines, and therefore, the dataset reported here contains 5797 statistical cases. That happens because, as stated by the GDSC team, in many instances, a significant proportion of the cell lines were only partially responsive or resistant to a given drug within the range of experimental screening concentrations, and therefore, the confidence of these inhibition values is reduced. Notice that a statistical case (or just a case) is a drug tested against a defined pancreatic cancer cell line. Only the most confident activity values were used in this study. The anti-pancreatic activity was measure as IC_{50}, that is, the inhibitory concentration leading to reduction of 50% cell viability.

Each statistical case was annotated as active [$APCA_i(c_t) = 1$] or inactive [$APCA_i(c_t) = -1$], with $APCA_i(c_t)$ being a binary categorical variable that denoted the inhibitory activity of the ith case/drug against a specific pancreatic cancer cell line (c_t). The annotations were made by considering the cutoff value of 6.45 μM. Thus, a drug was assigned as active its inhibitory activity was $IC_{50} \leq 6.45$ μM; otherwise, the drug was annotated as inactive. It should be pointed out that in early drug discovery programs, compounds are tested for their anticancer activity using a defined assay protocol at a screening concentration of 10 μM;[77] the hit compounds found during the primary screen are subjected to an additional confirmatory screen (following the same assay protocol) but at a lower concentration. As it can be observed, the IC_{50} cutoff value chosen in this work is lower than the classically used screening concentration, and that will have a positive impact on the rigor of a model when searching for high potency inhibitors. In addition, statistically speaking, this cutoff value $IC_{50} \leq 6.45$ μM prevents the disproportioned imbalance between the number of drugs assigned as active and the number of drugs annotated as inactive.

The chemical structures of the drugs were deposited in a txt file in the form of SMILES codes. The file was manually changed to *.smi. Then, with the purpose of obtaining information regarding the 2D connectivity of the drugs, the program Standardizer v15.11.16.0 was used to convert the *.smi file to *.sdf.[78] Subsequently, the molecular descriptors known as the total and local atom-based quadratic indices from the nonstochastic and stochastic adjacency matrices[79−81] were calculated by using the program QUBILs-MAS v1.0.[82] These descriptors have been suggested to exhibit a similar or superior discriminant power[83] to that shown by combining several families of molecular descriptors calculated by popular programs such as DRAGON.[84] These topological (graph-based) indices can be calculated according to the following equations

$$\mathrm{TmfAq}_k(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{k}[a_{\mathrm{mf}}]_{ij}\cdot x_i \cdot x_j \tag{1}$$

In eq 1, $\mathrm{TmfAq}_k(x)$ represents the total quadratic index of order k, which considers each atom i and its jth neighbors at the topological distance $d = k$. Notice that in the symbol, the abbreviation "mf" is associated with the matrix form; mf = ns for nonstochastic indices, while mf = ss for the stochastic counterparts. Also, the symbol x defines the different atomic physicochemical properties: HYD, electronegativity (E), AW, PSA, and polarizability. On the other hand, ${}^{k}[a_{\mathrm{mf}}]_{ij}$ is used to describe the presence (or absence) of adjacency between any two atoms in a molecule. The local nonstochastic and stochastic quadratic indices [$\mathrm{LmfAq}_k(x)Z$] can be determined by using the same formalism

$$\mathrm{LmfAq}_k(x)Z = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{k}[a_{\mathrm{mf}}]_{ijZ}\cdot x_i \cdot x_j \tag{2}$$

Here, most of the symbols have the same meanings as in the case of eq 1. The only difference is that eq 2 is focused on defined molecular fragments based on specific atom types (Z) such as HBDs, HBAs, and others. In total, 576 quadratic indices were calculated by the program QUBILs-MAS v1.0 using a predefined configuration [algebraic form: quadratic; constrains: atom-based; matrix forms: nonstochastic and stochastic; maximum order: 7; cutoff: keep all; groups: total and local (HBAs and donors, aliphatic carbon atoms, aromatic carbons, terminal methyl groups, halogens, and heteroatoms); properties: already mentioned in the previous paragraph; aggregation: Manhattan distance].

The quadratic indices calculated in eqs 1 and 2 cannot differentiate the effect of the chemical structures of the drugs on their inhibitory activities when they are assayed against the different pancreatic cancer cell lines. In this sense, several works have demonstrated that the Box−Jenkins operators can be used to generate new molecular descriptors able to consider both the chemical structure and the set of assay conditions (e.g., target, measure of activity or toxicity, etc.) under which a chemical has been tested[29,31,36,51,62,85,86]

$$\mathrm{avgQI}(c_t) = \frac{1}{n(c_t)} \times \sum_{a=1}^{n(c_t)} \mathrm{QI}_a \tag{3}$$

In eq 3, QI_a refers to any type of quadratic index of a drug calculated in eqs 1 and 2. At the same time, $\mathrm{avgQI}(c_t)$ is the average of a quadratic index for all of the drugs/chemicals that, being experimentally tested against the same pancreatic cancer cell line, were annotated as active. Thus, $n(c_t)$ represents the number of drugs assigned as active and assayed by considering a specific pancreatic cancer cell line. Then, in a second step, the following equation is derived

$$\mathrm{DQI}_a(c_t) = \frac{\mathrm{QI}_a - \mathrm{avgQI}(c_t)}{\mathrm{QI}_{\mathrm{MX}} - \mathrm{QI}_{\mathrm{MN}}} \tag{4}$$

In eq 4, $\mathrm{DQI}_a(c_t)$ is the deviation of a quadratic index, and it measures how much a drug/chemical tested against a defined pancreatic cancer cell line, structurally deviates from the set of drugs assigned as active and tested against the same cell line. On the other hand, $\mathrm{QI}_{\mathrm{MX}}$ and $\mathrm{QI}_{\mathrm{MN}}$ are the maximum and minimum values for each quadratic index, respectively.

4.2. Development of the mt-QSAR-EL Model. The different steps leading to the creation of the mct-QSAR-EL model are depicted in Figure 4. The dataset formed by the 5797 cases was randomly split into two series: training and test set. The training set contained 4348 cases, 1562 active, and 2786 inactive. The set represented the 75% of the total data, and it was employed to search for the best mct-QSAR-EL model. At the same time, the test set was composed of 1449 cases (remaining 25% of the total data), 520 active, and 929 inactive; the test set was used to demonstrate the predictive power of the mct-QSAR-EL. The ML method used in this study to generate the model was based on an ensemble of ANNs. The software IMMAN was employed to select the molecular descriptors with the highest discriminant power according to their differential Shannon entropies.[87] While extracting the most desirable descriptors, the correlations between them were analyzed; the cutoff interval $-0.7 < \mathrm{PCC} < 0.7$ was used to indicate lack of correlation, with PCC being the Pearson's correlation coefficient.[88]

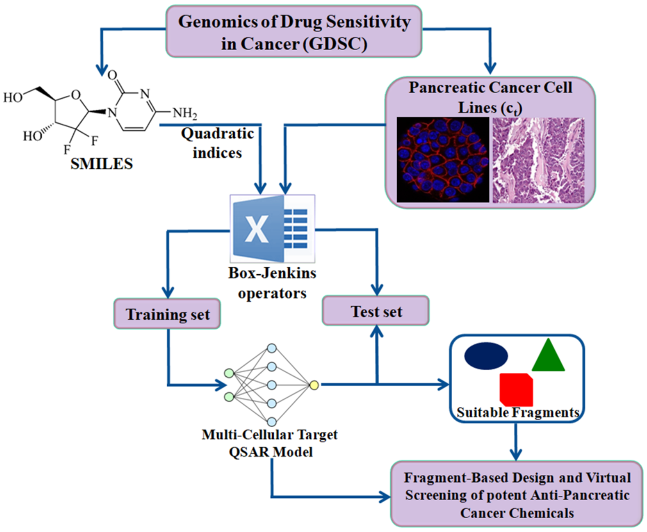

Genomics of Drug Sensitivity in Cancer (GDSC)
HO
O
N
NH2
HO
F
F
SMILES
Quadratic indices
Pancreatic Cancer Cell Lines (c_t)
Box-Jenkins operators
Training set
Test set
Suitable Fragments
Multi-Cellular Target QSAR Model
Fragment-Based Design and Virtual Screening of potent Anti-Pancreatic Cancer Chemicals

Figure 4. Steps used to develop the mct-QSAR-EL model.

Using the highest discriminant descriptors derived from the program IMMAN, the Intelligent Problem Solver of the ANNs package of the program STATISTICA v6.0[89] was used to generate the best mct-QSAR-EL. A first run was performed to choose the most appropriate ANNs architectures and the numbers of ANNs constituting the ensembles. Also, the diversity of the ANNs was checked; the ANNs forming the ensemble needed to have different numbers of neurons in their hidden layers, as well as different training and test errors, and different values of the statistical indices Sn(%), and Sp(%), Ac(%), and MCC.[90] These statistical indices were used to assess the internal quality (training set) and the predictive power (test set) of the mt-QSAR-EL model. Last, the sensitivity analysis was performed to rank the descriptors previously selected by IMMAN according to their significances in the mct-QSAR-EL model.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.8b03693.

Chemical and biological data, molecular descriptors, averages, and related measures (XLSX)

Input data and classification results (XLSX)

Local measures of the statistical quality and performance of the mct-QSAR-EL model (PDF)

Applicability domain of the mct-QSAR-EL model (XLSX)

Molecules designed by the mct-QSAR-EL model: prediction results and assessment of the applicability domain (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: alejspivanovich@yahoo.es. Phone: +34 678042076.

ORCID

Alejandro Speck-Planche: 0000-0002-9544-9016

Notes

The author declares no competing financial interest.

■ ACKNOWLEDGMENTS

A.S.-P. acknowledges the Spanish Juan de la Cierva program (grant: FJCI-2015-25572) for the financial support.

■ REFERENCES

(1) Hsieh, M. H.; Sun, L.-M.; Lin, C.-L.; Hsieh, M.-J.; Hsu, C.-Y.; Kao, C.-H. Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. *Cancer Manage. Res.* **2018**, *10*, 6317−6324.

(2) Chang, J. S.; Chen, L.-T.; Shan, Y.-S.; Chu, P.-Y.; Tsai, C.-R.; Tsai, H.-J. The incidence and survival of pancreatic cancer by histology, including rare subtypes: a nation-wide cancer registry-based study from Taiwan. *Cancer Med.* **2018**, *7*, 5775−5788.

(3) Ilic, M.; Ilic, I. Epidemiology of pancreatic cancer. *World J. Gastroenterol.* **2016**, *22*, 9694−9705.

(4) Swayden, M.; Iovanna, J.; Soubeyran, P. Pancreatic cancer chemo-resistance is driven by tumor phenotype rather than tumor genotype. *Heliyon* **2018**, *4*, No. e01055.

(5) Cintas, C.; Douché, T.; Therville, N.; Arcucci, S.; Ramos-Delgado, F.; Basset, C.; Thibault, B.; Guillermet-Guibert, J. Signal-

targeted therapies and resistance mechanisms in pancreatic cancer: Future developments reside in proteomics. *Cancers* **2018**, *10*, 174.

(6) Biancur, D. E.; Kimmelman, A. C. The plasticity of pancreatic cancer metabolism in tumor progression and therapeutic resistance. *Biochim. Biophys. Acta, Rev. Cancer* **2018**, *1870*, 67−75.

(7) 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.

(8) Jahnke, W.; Erlanson, D. A. *Fragment-Based Approaches in Drug Discovery*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2006.

(9) Deokar, H.; Deokar, M.; Wang, W.; Zhang, R.; Buolamwini, J. K. QSAR studies of new pyrido[3,4-b]indole derivatives as inhibitors of colon and pancreatic cancer cell proliferation. *Med. Chem. Res.* **2018**, *27*, 2466−2481.

(10) Satbhaiya, S.; Chourasia, O. P. Scaffold and cell line based approaches for QSAR studies on anticancer agents. *RSC Adv.* **2015**, *5*, 84810−84820.

(11) Kumar, R.; Chaudhary, K.; Singla, D.; Gautam, A.; Raghava, G. P. S. Designing of promiscuous inhibitors against pancreatic cancer cell lines. *Sci. Rep.* **2014**, *4*, 4668.

(12) 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.

(13) 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.

(14) Prado-Prado, F.; Garcia-Mera, X.; Escobar, M.; Alonso, N.; Caamano, O.; Yanez, M.; Gonzalez-Diaz, H. 3D MI-DRAGON: new model for the reconstruction of US FDA drug- target network and theoretical-experimental studies of inhibitors of rasagiline derivatives for AChE. *Curr. Top. Med. Chem.* **2012**, *12*, 1843−1865.

(15) Prado-Prado, F.; García-Mera, X.; Abeijón, P.; Alonso, N.; Caamaño, O.; Yáñez, M.; Gárate, T.; Mezo, M.; González-Warleta, M.; Muiño, L.; Ubeira, F. M.; González-Díaz, H. Using entropy of drug and protein graphs to predict FDA drug-target network: theoretic-experimental study of MAO inhibitors and hemoglobin peptides from Fasciola hepatica. *Eur. J. Med. Chem.* **2011**, *46*, 1074−1094.

(16) Marzaro, G.; Chilin, A.; Guiotto, A.; Uriarte, E.; Brun, P.; Castagliuolo, I.; Tonus, F.; González-Díaz, H. Using the TOPS-MODE approach to fit multi-target QSAR models for tyrosine kinases inhibitors. *Eur. J. Med. Chem.* **2011**, *46*, 2185−2192.

(17) García, I.; Fall, Y.; Gómez, G.; González-Díaz, H. First computational chemistry multi-target model for anti-Alzheimer, anti-parasitic, anti-fungi, and anti-bacterial activity of GSK-3 inhibitors in vitro, in vivo, and in different cellular lines. *Mol. Diversity* **2011**, *15*, 561−567.

(18) Speck-Planche, A. Combining ensemble learning with a fragment-based topological approach to generate new molecular diversity in drug discovery: In silico design of Hsp90 inhibitors. *ACS Omega* **2018**, *3*, 14704−14716.

(19) Speck-Planche, A.; Scotti, M. T. BET bromodomain inhibitors: fragment-based in silico design using multi-target QSAR models. *Mol. Diversity* **2018**, DOI: 10.1007/s11030-018-9890-8.

(20) 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.

(21) Speck-Planche, A.; Cordeiro, M. N. D. S. Multi-target QSAR approaches for modeling protein inhibitors. Simultaneous prediction of activities against biomacromolecules present in gram-negative bacteria. *Curr. Top. Med. Chem.* **2015**, *15*, 1801−1813.

(22) 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.

(23) 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.

(24) 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.

(25) Speck-Planche, A.; Cordeiro, M. N. D. S. Enabling virtual screening of potent and safer antimicrobial agents against noma: mtk-QSBER model for simultaneous prediction of antibacterial activities and ADMET properties. *Mini-Rev. Med. Chem.* **2015**, *15*, 194−202.

(26) Speck-Planche, A.; Cordeiro, M. 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.

(27) Tenorio-Borroto, E.; Ramirez, F.; Speck-Planche, A.; Cordeiro, M.; 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.

(28) Speck-Planche, A. Recent advances in fragment-based computational drug design: tackling simultaneous targets/biological effects. *Future Med. Chem.* **2018**, *10*, 2021−2024.

(29) Simón-Vidal, L.; García-Calvo, O.; Oteo, U.; Arrasate, S.; Lete, E.; Sotomayor, N.; González-Díaz, H. Perturbation-Theory and Machine Learning (PTML) Model for high-throughput screening of Parham reactions: experimental and theoretical studies. *J. Chem. Inf. Model.* **2018**, *58*, 1384−1396.

(30) Aranzamendi, E.; Arrasate, S.; Sotomayor, N.; González-Díaz, H.; Lete, E. Chiral Brønsted Acid-Catalyzed Enantioselective α-Amidoalkylation Reactions: A Joint Experimental and Predictive Study. *ChemistryOpen* **2016**, *5*, 540−549.

(31) Blay, V.; Yokoi, T.; González-Díaz, H. Perturbation theory-machine learning study of zeolite materials desilication. *J. Chem. Inf. Model.* **2018**, *58*, 2414−2419.

(32) Gonzalez-Durruthy, M.; Werhli, A. V.; Seus, V.; Machado, K. S.; Pazos, A.; Munteanu, C. R.; Gonzalez-Diaz, 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.

(33) 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.

(34) Concu, R.; Kleandrova, V. V.; Speck-Planche, A.; Cordeiro, M. N. D. S. Probing the toxicity of nanoparticles: a unified in silico machine learning model based on perturbation theory. *Nanotoxicology* **2017**, *11*, 891−906.

(35) Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; DS Cordeiro, M. N. Computational modeling in nanomedicine: prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model. *Nanomedicine* **2015**, *10*, 193−204.

(36) Tenorio-Borroto, E.; Peñuelas-Rivas, C. G.; Vásquez-Chagoyán, J. C.; Castañedo, N.; Prado-Prado, F. J.; García-Mera, X.; González-Díaz, H. Model for high-throughput screening of drug immunotoxicity - Study of the anti-microbial G1 over peritoneal macrophages using flow cytometry. *Eur. J. Med. Chem.* **2014**, *72*, 206−220.

(37) Herrera-Ibatá, D. M.; Pazos, A.; Orbegozo-Medina, R. A.; Romero-Durán, F. J.; González-Díaz, H. Mapping chemical structure-activity information of HAART-drug cocktails over complex networks

of AIDS epidemiology and socioeconomic data of U.S. counties. *BioSystems* **2015**, *132–133*, 20–34.

(38) Herrera-Ibata, D. M.; Orbegozo-Medina, R. A.; Gonzalez-Diaz, H. Multiscale mapping of AIDS in U.S. countries vs anti-HIV drugs activity with complex networks and information indices. *Curr. Bioinf.* **2015**, *10*, 639–657.

(39) González-Díaz, H.; Herrera-Ibatá, D. M.; Duardo-Sánchez, A.; Munteanu, C. R.; Orbegozo-Medina, R. A.; Pazos, A. ANN multiscale model of anti-HIV drugs activity vs AIDS prevalence in the US at county level based on information indices of molecular graphs and social networks. *J. Chem. Inf. Model.* **2014**, *54*, 744–755.

(40) Speck-Planche, A.; Cordeiro, M. N. Review of Current Chemoinformatic Tools for Modeling Important Aspects of CYPsmediated Drug Metabolism. Integrating Metabolism Data with Other Biological Profiles to Enhance Drug Discovery. *Curr. Drug Metab.* **2014**, *15*, 429–440.

(41) Speck-Planche, A.; Kleandrova, V. V.; Cordeiro, M. N. D. S. New insights toward the discovery of antibacterial agents: Multitasking QSBER model for the simultaneous prediction of antituberculosis activity and toxicological profiles of drugs. *Eur. J. Pharm. Sci.* **2013**, *48*, 812–818.

(42) Abeijon, P.; Garcia-Mera, X.; Caamano, O.; Yanez, M.; Lopez-Castro, E.; Romero-Duran, F.; Gonzalez-Diaz, H. Multi-target mining of Alzheimer disease proteome with Hansch's QSBR-perturbation theory and experimental-theoretic study of new thiophene isosters of rasagiline. *Curr. Drug Targets* **2017**, *18*, 511–521.

(43) Speck-Planche, A.; Luan, F.; Cordeiro, M. N. D. S. Role of Ligand-Based Drug Design Methodologies toward the Discovery of New Anti- Alzheimer Agents: Futures Perspectives in Fragment-Based Ligand Design. *Curr. Med. Chem.* **2012**, *19*, 1635–1645.

(44) Planche, A. S.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. Unified multi-target approach for the rational in silico design of anti-bladder cancer agents. *Anticancer Agents Med. Chem.* **2013**, *13*, 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. *Anticancer Agents Med. Chem.* **2012**, *12*, 678–685.

(46) 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.

(47) 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.

(48) 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.

(49) 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.

(50) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **2012**, *17*, 4791–4810.

(51) Speck-Planche, A.; Cordeiro, M. N. D. S. in silico approach for the design of virtual anti-hepatitis C leads. *ACS Comb. Sci.* **2017**, *19*, 501–512.

(52) Speck-Planche, A.; Kleandrova, V. QSAR and molecular docking techniques for the discovery of potent monoamine oxidase B inhibitors: Computer-aided generation of new rasagiline bioisosteres. *Curr. Top. Med. Chem.* **2012**, *12*, 1734–1747.

(53) Baskin, I. I.; Skvortsova, M. I.; Stankevich, I. V.; Zefirov, N. S. On the basis of invariants of labeled molecular graphs. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 527–531.

(54) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships I. Partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565–577.

(55) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21–35.

(56) Ghose, A. K.; Pritchett, A.; Crippen, G. M. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships III: Modeling hydrophobic interactions. *J. Comput. Chem.* **1988**, *9*, 80–90.

(57) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.

(58) Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced adversarial neural computer for de novo molecular design. *J. Chem. Inf. Model.* **2018**, *58*, 1194–1204.

(59) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.

(60) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **2018**, *4*, 120–131.

(61) Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. druGAN: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol. Pharm.* **2017**, *14*, 3098–3104.

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

(63) Speck Planche, A.; Cordeiro, M. N. D. S. Speeding up the virtual design and screening of therapeutic peptides: Simultaneous prediction of anticancer activity and cytotoxicity. In *Multi-Scale Approaches in Drug Discovery: From Empirical Knowledge to In Silico Experiments and Back*, 1st ed.; Speck-Planche, A., Ed.; Elsevier: Oxford, U.K., 2017; pp 127–147.

(64) Kleandrova, V. V.; Speck Planche, A. Multitasking model for computer-aided design and virtual screening of compounds with high anti-HIV activity and desirable ADMET properties. In *Multi-Scale Approaches in Drug Discovery: From Empirical Knowledge to In Silico Experiments and Back*, 1st ed.; Speck-Planche, A., Ed.; Elsevier: Oxford, U.K., 2017; pp 55–81.

(65) 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.

(66) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

(67) Irwin, J. J.; Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

(68) Williams, A. J. Chemspider: a platform for crowdsourced collaboration to curate data derived from public compound databases. In *Collaborative Computational Technologies for Biomedical Research*; Ekins, S., Hupcey, M. A. Z., Williams, A. J., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, 2011; pp 363–386.

(69) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1PII of

original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3-25. 1. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.

(70) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* **1999**, *1*, 55−68.

(71) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.

(72) *CambridgeSoft ChemDraw Ultra. v8.0*, Cambridge, MA, 2003.

(73) *ChemAxon Marvin Sketch, JChem. v15.11.16.0*, Budapest, Hungary, 1998−2016.

(74) Iorio, F.; Knijnenburg, T. A.; Vis, D. J.; Bignell, G. R.; Menden, M. P.; Schubert, M.; Aben, N.; Gonçalves, E.; Barthorpe, S.; Lightfoot, H.; Cokelaer, T.; Greninger, P.; van Dyk, E.; Chang, H.; de Silva, H.; Heyn, H.; Deng, X.; Egan, R. K.; Liu, Q.; Mironenko, T.; Mitropoulos, X.; Richardson, L.; Wang, J.; Zhang, T.; Moran, S.; Sayols, S.; Soleimani, M.; Tamborero, D.; Lopez-Bigas, N.; Ross-Macdonald, P.; Esteller, M.; Gray, N. S.; Haber, D. A.; Stratton, M. R.; Benes, C. H.; Wessels, L. F. A.; Saez-Rodriguez, J.; McDermott, U.; Garnett, M. J. A landscape of pharmacogenomic interactions in cancer. *Cell* **2016**, *166*, 740−754.

(75) Yang, W.; Soares, J.; Greninger, P.; Edelman, E. J.; Lightfoot, H.; Forbes, S.; Bindal, N.; Beare, D.; Smith, J. A.; Thompson, I. R.; Ramaswamy, S.; Futreal, P. A.; Haber, D. A.; Stratton, M. R.; Benes, C.; McDermott, U.; Garnett, M. J. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* **2013**, *41*, D955−D961.

(76) Garnett, M. J.; Edelman, E. J.; Heidorn, S. J.; Greenman, C. D.; Dastur, A.; Lau, K. W.; Greninger, P.; Thompson, I. R.; Luo, X.; Soares, J.; Liu, Q.; Iorio, F.; Surdez, D.; Chen, L.; Milano, R. J.; Bignell, G. R.; Tam, A. T.; Davies, H.; Stevenson, J. A.; Barthorpe, S.; Lutz, S. R.; Kogera, F.; Lawrence, K.; McLaren-Douglas, A.; Mitropoulos, X.; Mironenko, T.; Thi, H.; Richardson, L.; Zhou, W.; Jewitt, F.; Zhang, T.; O'Brien, P.; Boisvert, J. L.; Price, S.; Hur, W.; Yang, W.; Deng, X.; Butler, A.; Choi, H. G.; Chang, J. W.; Baselga, J.; Stamenkovic, I.; Engelman, J. A.; Sharma, S. V.; Delattre, O.; Saez-Rodriguez, J.; Gray, N. S.; Settleman, J.; Futreal, P. A.; Haber, D. A.; Stratton, M. R.; Ramaswamy, S.; McDermott, U.; Benes, C. H. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **2012**, *483*, 570−575.

(77) Varbanov, H. P.; Kuttler, F.; Banfi, D.; Turcatti, G.; Dyson, P. J. Repositioning approved drugs for the treatment of problematic cancers using a screening approach. *PLoS One* **2017**, *12*, No. e0171052.

(78) *ChemAxon Standardizer Tool for Structure Canonicalization and Transformation, JChem. v15.11.16.0*, Budapest, Hungary, 1998−2016.

(79) Medina Marrero, R.; Marrero-Ponce, Y.; Barigye, S. J.; Echeverría Díaz, Y.; Acevedo-Barrios, R.; Casañola-Martín, G. M.; García Bernal, M.; Torrens, F.; Pérez-Giménez, F. QuBiLs-MAS method in early drug discovery and rational drug identification of antifungal agents. *SAR QSAR Environ. Res.* **2015**, *26*, 943−958.

(80) Brito-Sánchez, Y.; Castillo-Garit, J. A.; Le-Thi-Thu, H.; González-Madariaga, Y.; Torrens, F.; Marrero-Ponce, Y.; Rodríguez-Borges, J. E. Comparative study to predict toxic modes of action of phenols from molecular structures. *SAR QSAR Environ. Res.* **2013**, *24*, 235−251.

(81) Rescigno, A.; Casañola-Martin, G. M.; Sanjust, E.; Zucca, P.; Marrero-Ponce, Y. Vanilloid derivatives as tyrosinase inhibitors driven by virtual screening-based QSAR models. *Drug Test. Anal.* **2011**, *3*, 176−181.

(82) Valdés-Martini, J. R.; García-Jacas, C. R.; Marrero-Ponce, Y.; Silveira Vaz d' Almeida, Y.; Morell, C. *QUBILs-MAS: Free software for Molecular Descriptors calculator from Quadratic, Bilinear and Linear Maps Based on Graph-Theoretic Electronic-Density Matrices and Atomic Weightings. v1.0*; CEDA registration number: 2373-2012; CAMD-BIR Unit: Villa Clara. http://tomocomd.com/, 2012.

(83) Valdes-Martini, J. R.; Marrero-Ponce, Y.; Garcia-Jacas, C. R.; Martinez-Mayorga, K.; Barigye, S. J.; Vaz d'Almeida, Y. S.; Pham-The, H.; Perez-Gimenez, F.; Morell, C. A. QuBiLS-MAS, open source multi-platform software for atom- and bond-based topological (2D) and chiral (2.5D) algebraic molecular descriptors computations. *J. Cheminf.* **2017**, *9*, 35.

(84) *Talete-srl DRAGON (Software for Molecular Descriptor Calculation). v6.0.* http://www.talete.mi.it/, 2015.

(85) Luan, F.; Cordeiro, M. N. D. S.; Alonso, N.; García-Mera, X.; Caamaño, O.; Romero-Duran, F. J.; Yañez, M.; González-Díaz, H. TOPS-MODE model of multiplexing neuroprotective effects of drugs and experimental-theoretic study of new 1,3-rasagiline derivatives potentially useful in neurodegenerative diseases. *Bioorg. Med. Chem.* **2013**, *21*, 1870−1879.

(86) Speck-Planche, A.; Cordeiro, M. N. D. S. Chemoinformatics for medicinal chemistry: in silico model to enable the discovery of potent and safer anti-cocci agents. *Future Med. Chem.* **2014**, *6*, 2013−2028.

(87) Urias, R. W. P.; Barigye, S. J.; Marrero-Ponce, Y.; García-Jacas, C. R.; Valdes-Martiní, J. R.; Perez-Gimenez, F. IMMAN: free software for information theory-based chemometric analysis. *Mol. Diversity* **2015**, *19*, 305−319.

(88) Pearson, K. Notes on regression and inheritance in the case of two parents. *Proc. R. Soc. London* **1895**, *58*, 240−242.

(89) Statsoft-Team. *STATISTICA. Data Analysis Software System. v6.0*, Tulsa, 2001.

(90) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.