International Journal of
Molecular Sciences

MDPI

Article

Development of Multi-Target Chemometric Models for the Inhibition of Class I PI3K Enzyme Isoforms: A Case Study Using QSAR-Co Tool

Amit Kumar Halder and M. Natália Dias Soeiro Cordeiro *

Department of Chemistry and Biochemistry, University of Porto, 4169-007 Porto, Portugal
* Correspondence: ncordeir@fc.up.pt

Received: 14 July 2019; Accepted: 24 August 2019; Published: 27 August 2019

Abstract: The present work aims at establishing multi-target chemometric models using the recently launched quantitative structure–activity relationship (QSAR)-Co tool for predicting the activity of inhibitor compounds against different isoforms of phosphoinositide 3-kinase (PI3K) under various experimental conditions. The inhibitors of class I phosphoinositide 3-kinase (PI3K) isoforms have emerged as potential therapeutic agents for the treatment of various disorders, especially cancer. The cell-based enzyme inhibition assay results of PI3K inhibitors were curated from the CHEMBL database. Factors such as the nature and mutation of cell lines that may significantly alter the assay outcomes were considered as important experimental elements for mt-QSAR model development. The models, in turn, were developed using two machine learning techniques as implemented in QSAR-Co: linear discriminant analysis (LDA) and random forest (RF). Both techniques led to models with high accuracy (ca. 90%). Several molecular fragments were extracted from the current dataset, and their quantitative contributions to the inhibitory activity against all the proteins and experimental conditions under study were calculated. This case study also demonstrates the utility of QSAR-Co tool in solving multi-factorial and complex chemometric problems. Additionally, the combination of different in silico methods employed in this work can serve as a valuable guideline to speed up early discovery of PI3K inhibitors.

Keywords: PI3K inhibitors; cancer; QSAR; multi-target models; linear discriminant analysis; random forest

1. Introduction

Machine learning-based chemometric modelling is widely applied to the design and discovery of new therapeutic agents with superior biological activities. It is now being realised that ligand–protein interaction-based traditional drug discovery methods may no longer be sufficient to satisfy clinical drug safety criteria. 'Data fusion' techniques that incorporate multiple structural, genetic and pharmacological data types and sources are now becoming essential for the discovery of safe and effective therapeutic agents [1,2]. Quantitative structure–activity relationship (QSAR), which relates the variations in the observed activity to numerical descriptors, is extensively used to find structural requirements for higher active molecules and/or to predict activity of novel compounds [3–5]. Conventional QSAR models are developed based on the experimental activity of the compounds tested against a single biological target following a specific experimental condition. However, multi-target QSAR models have drawn considerable attention very recently. It is worth mentioning also that the 'multi-target' term is used here to define those chemometric models that truly integrate the data for simultaneous prediction of response parameters against multiple biological targets under different experimental conditions [6–10]. Our group has launched QSAR-Co [11], a publicly available

Java-based tool to support Box–Jenkins based mt-QSAR model development, validation and screening (https://sites.google.com/view/qsar-co). In order to understand how this tool may perform with diverse and complex datasets containing multi-factorial response variables, we performed a case study with this QSAR-Co tool considering the inhibitors of different class I phosphoinositide 3-kinases (PI3Ks) isoforms as potential anticancer agents. As per our literature search, the current case study is the first report on the multi-target chemometric modelling on the inhibition of class I PI3K enzyme isoforms.

The phosphoinositide 3-kinases (PI3Ks) belong to a large lipid enzyme family involved in the phosphorylation of the 3′-OH group of phosphatidylinositols (PI) present in the plasma membranes [12]. In response to various external stimuli such as growth factors, hormones and environmental variations, PI3Ks generate intracellular signals to regulate diverse cellular processes such as cell proliferation, survival, differentiation, migration, inflammation and metabolism [13]. Based on their structural- and enzymatic-kinetic differences, PI3Ks are classified into three classes, i.e., class I, class II and class III [14]. Among all these, only class I PI3Ks are able to phosphorylate membrane substrate phosphorylate phosphatidylinositol-4,5-bisphosphate (PIP_2) to produce a second messenger named phosphatidylinositol-3,4,5-trisphosphate (PIP_3), which subsequently initiates a signalling cascade for activation of the downstream effectors to trigger increased cell growth, metabolism and cell-cycle progression [15,16]. The class I PI3Ks are further categorised into class IA and IB enzymes. The class IA includes three enzyme isoforms, namely: PI3Kα, PI3Kβ and PI3Kδ; each isoform consists of a p110 catalytic unit (p110α, p110β and p110δ) that forms heterodimers with a regulatory subunit. On the other hand, the class IB includes only one enzyme isoform PI3Kγ in which the heterodimer is formed between catalytic subunit p110γ and a regulatory subunit [12,17]. Each PI3K isoform is found to possess distinct tissue specificity as well as unique physiological functionalities. The PI3Kα and PI3Kβ are expressed ubiquitously whereas the expressions of PI3Kδ and PI3Kγ are limited only to hematopoietic system [18,19]. PI3Kα is involved in regulating glucose homeostasis, insulin signalling and mitochondrial growth. PI3Kβ promotes platelet adhesion and aggregation, playing significant roles in the progression of thrombotic diseases [15,20]. Additionally, PI3Kβ participates in several immune responses. Both PI3Kγ and PI3Kδ play significant roles in inflammation and immunisation [21,22]. Due to their crucial involvement in various physiological functions, it is not surprising that class I PI3K inhibitors are implicated in the treatment of a range of diseases such as arthritis [22], thrombosis, asthma [23], inflammation, cardiovascular diseases, PIK3CA-related overgrowth syndromes, etc. Nevertheless, class I PI3K inhibitors are above all well-recognised as potential anticancer agents. Dysregulation of the PI3K pathway leads to severe abnormality in the cell cycle, cell growth survival, metabolism and motility, which are some common hallmarks of cancer [17,18,24]. A number of PI3K inhibitors, including pan-PI3K inhibitors, isoform-selective PI3K inhibitors and dual PI3K/mTOR inhibitors are currently under clinical trials for cancer treatment [24]. Idelalisib, a PI3Kδ-specific inhibitor, has been approved by the United States Food and Drug Administration (USFDA) for the treatment of follicular non-Hodgkin's B-cell lymphoma, small-cell lymphocytic lymphoma and chronic lymphoid leukemia [25]. Two other PI3K inhibitors, that is duvelisib (PI3Kγ/δ inhibitor) and copanlisib (pan-PI3K inhibitor), have also found approval for the treatment of leukemia and lymphoma [26]. The p110α encoding gene *PIK3CA* is one of the most frequently mutated genes in most human cancers, especially breast, ovarian and colorectal cancers [17,18,27]. Although p110β catalytic subunit mutations are not commonly found, its tumorigenic potential is associated with the loss of the enzyme phosphatase and tensin homolog (known as PTEN), a negative regulator of PI3K pathway that dephosphorylates PIP3 [24,28]. While PI3Kδ signalling is associated with the activation, proliferation and survival of the B cells [29], PI3Kγ is activated in response to tissue hypoxia and plays crucial roles in the development of tumour microenvironment [30]. Due to their immunomodulatory potencies, PI3Kδ and PI3Kγ inhibitors are potential candidates for cancer immunotherapy [29]. Nevertheless, inhibition of PI3Kδ and PI3Kγ may give rise to moderate to severe adverse immune responses. It is worth mentioning here that despite promising therapeutic results, similar to many targeted cancer therapies, different types of PI3K inhibitors (pan-/isoform-specific) may give rise to serious adverse effects that restrict their

clinical applications [31]. Very recently, Curigliano and Shah reported a detailed review of the adverse effects obtained from different clinically tested class I PI3K inhibitors [29]. It is generally postulated that pan-PI3K inhibitors may exhibit higher efficacy against cancer cells whereas isoform-specific PI3K inhibitors may have less overall toxicity [15,31]. While overall efficacy and relative benefit of pan- and isoform-specific PI3K inhibitors will be decided through clinical trials, a detailed understanding of the structural and physicochemical factors responsible for higher potency and selectivity towards different class I PI3K inhibitors will definitely help in the development of the next generation of PI3K inhibitors.

So far, a number of different ligand-based in silico methods and protocols have been employed by different researchers for the discovery of PI3K inhibitors, and these include 2D-quantitative structure–activity relationship (2D-QSAR) modelling, 3D-QSAR, 3D-pharmacophore mapping, etc. [32–49]. However, all these methods suffer from at least one of the following shortcomings. From one side, only a limited number of data samples have been considered for modelling and therefore, the overall applicability of these models may be limited. On the other hand, only one PI3K enzyme isoform has been considered as the biological target for modelling. Considering the multi-factorial nature of the diseases, these single target models may also pose serious limitations to the applicability of these in silico models. Herein we report, for the first time, a multi-target in silico chemometric modelling with different isoforms of class I PI3K inhibitors. Even though Liew et al. earlier performed a consensus QSAR modelling with an integrated dataset containing inhibition data for all four class I PI3K inhibitors, the final models could not differentiate among inhibitory activities against different PI3K isoforms [36]. The mt-QSAR models developed for the current case study, apart from being highly predictive in nature, are also capable of differentiating among different biological targets as well as experimental assay conditions. Based on the mt-QSAR model, contributions of some structural fragments for higher/lower inhibitory potency against PI3K enzymes are also discussed.

2. Results and Discussion

2.1. Linear Mt-QSAR Model Development

The major concepts and design strategies of QSAR-Co have been reported elsewhere [11]. A step-by-step instruction manual for the use of this tool is also available online (https://sites.google.com/view/qsar-co). The multi-target modelling of the QSAR-Co tool is based on the Box–Jenkins moving average approach [10,11,50]. The tool currently allows for the development of a linear mt-QSAR model by genetic algorithm based linear discriminant analysis (GA-LDA) technique [11,51–53]. A dataset comprising 726 compounds were collected from CHEMBL (https://www.ebi.ac.uk/chembl/). As can be observed from the dataset (provided in supplementary material, SM1.xlsx), each compound was tested against class I PI3K enzymes by a specific cell-based assay. In cell-based assays, the enzyme inhibitory potency of the inhibitors is measured in a specific cell line, and the outcomes of these assays are thus affected not only by the type of biological target (specific class I PI3K enzyme isoform, in this case) but also by the complex multi-factorial conditions that exist inside a specific cellular system [54]. The rationale and methodology of resorting to the Box–Jenkins moving average approach for setting up the present mt-QSAR modelling have been documented in detail previously [10,11,50], thus an overview highlighting only the most principal technical aspects will be given here. Prior to model development, the experimental elements for each data point are settled by the Box–Jenkins approach [11]. Any experimental element may simply be defined as the specific condition of the assay which is likely to alter the assay results of the compound under study. The first element considered in the current work is the biological enzyme target or *bt*. Each dataset compound was assayed against at least one of the four above-mentioned class I human PI3K enzymes (*bt*). Moreover, 34 different types of cell lines were found in the retrieved dataset (SM1.xlsx). Some of these cell lines are wild types whereas the others are found to have at least one mutation. Furthermore, in some assays PTEN-deficient or PTEN-null human cell lines were used. Therefore, it may be inferred that depending on the cell type, the experimental outcomes should also vary to a considerable extent. Depending on the nature of the

cell lines, we considered two more experimental elements, namely *cl* (type of cell line) and *mt* (wild or mutated) for the multi-target modelling. Overall, each combination of these three elements *bt*, *cl* and *mt* defines a specific experimental condition, which may be expressed as an ontology of the form $cj \rightarrow (bt, cl, mt)$. It is worth mentioning here that some compounds of the dataset have been assayed against more than one experimental element. Depending on the biological response, each dataset sample *i* was annotated as active [$IAi(cj) = 1$] or inactive [$IAi(cj) = -1$]. The response variable $IAi(cj)$ is a binary variable characterising the inhibitory potency of the *i*th sample under a specific experimental condition *cj*. Any compound with $IC_{50}/K_i/K_d$ values ≤ 600 nM was assigned as active whereas the remaining data samples considered as inactive. Generally speaking, any compound exhibiting an inhibitory activity in the micromolar range is considered as a '*hit*' molecule in the context of drug discovery [55]. In this work, the selected cut-off value appears in the sub-micromolar range to make the models more rigorous for the selection of more potent inhibitors. This cut-off also prevented excessive imbalance between active and inactive data points.

Details about data curation, dataset division, molecular descriptor calculation, descriptor modification, data-pre-treatment and model development are described in the Materials and Methods section. In the Box–Jenkins approach, each molecular descriptor calculated for a dataset compound is modified through a systemic procedure so that contributions of each experimental element may be easily incorporated (see Section 3.2). The mt-QSAR model was established with a modelling dataset, which is a combination of a sub-training set (n = 453) on which the model was developed, and a test set (n = 113), which was used to validate the model.

The resulting best-fit mt-QSAR-LDA model found (a ten-variable equation) is given below along with the statistical parameters of the LDA.

$$\begin{aligned} IAi(cj) = {} & +1.070 + 16.121\,\Delta[SpMAD\_A]b_t + 5.979\,\Delta[HATS2i]b_t + 5.582\,\Delta[nROCON]c_l \\ & + 0.475\,\Delta[F07[N\text{-}N]]c_l + 0.131\,\Delta[HTm]m_t + 0.069\,\Delta[SM15\_EA(\mathrm{dm})]c_l - 2.898\,\Delta[nCONN]b_t \\ & -2.662\,\Delta[R1m]c_l - 0.905\,\Delta[Mor18m]b_t - 0.850\,\Delta[HATS8s]c_l \\ & N = 453,\ \lambda = 0.234,\ \chi^2 = 648.27,\ D^2 = 11.522,\ F\,(10, 442) = 144.89,\ p\text{-value} < 10^{-16} \end{aligned} \quad (1)$$

As can be seen, this mt-QSAR-LDA model is satisfactory in both statistical significance, goodness of fit and robustness. The low value found for the Wilks λ statistic (0.234) [56] shows also that the model displays an adequate discriminatory power. The classification results obtained for the sub-training and test sets are presented in Table 1, outlining the overall performance of the present mt-QSAR LDA model.

Table 1. Overall performance of the final multitarget-quantitative structure–activity relationship (mt-QSAR) linear discriminant analysis (LDA) model.

Classification [a]	Sub-Training Set	Test Set
ND_{Total}	453	113
ND_{active}	324	89
CCD_{active}	310	84
Sensitivity (%)	95.68	94.38
$ND_{inactive}$	129	24
$CCD_{inactive}$	122	22
Specificity (%)	94.57	91.67
F-measure	0.967	0.960
Accuracy (%)	95.36	93.80
MCC	0.889	0.825

[a] ND_{Total}—Number of total data-points; ND_{active}—Number of active data-points, $ND_{inactive}$—Number of inactive data-points CCD—Correctly classified data, MCC—Matthews correlation coefficient.

As can be observed from Table 1, the QSAR model achieved an accuracy of 95.36% and 93.80% for the sub-training and the test sets, respectively. Furthermore, the model correctly classified 95.68% of the active samples and 94.57% the inactive ones of the sub-training set. Similarly, the model could

correctly classify 94.38% of active and 91.67% of inactive samples of the test set, respectively. These findings corroborate that this mt-QSAR-LDA model is highly capable of discriminating active PI3K inhibitors from inactive ones. The high values found for the Matthews correlation coefficients (i.e., 0.889 for the sub-training and 0.825 for the test set) further confirms the statistical robustness of the model [57].

Figure 1 shows the receiver operating characteristic (ROC) plots for the sub-training (10-fold) and the test sets. The area under the ROC curve (AUROC) is another important well-known statistical parameter to evaluate the statistical significance of a classifier model. For a random classifier, an area under ROC (AUROC) value of 0.5 is obtained [58]. In the current case, AUROC values of 0.977 and 0.968 are obtained for the sub-training and the test sets. Therefore, these values further emphasised the high statistical significance of the developed mt-QSAR model.

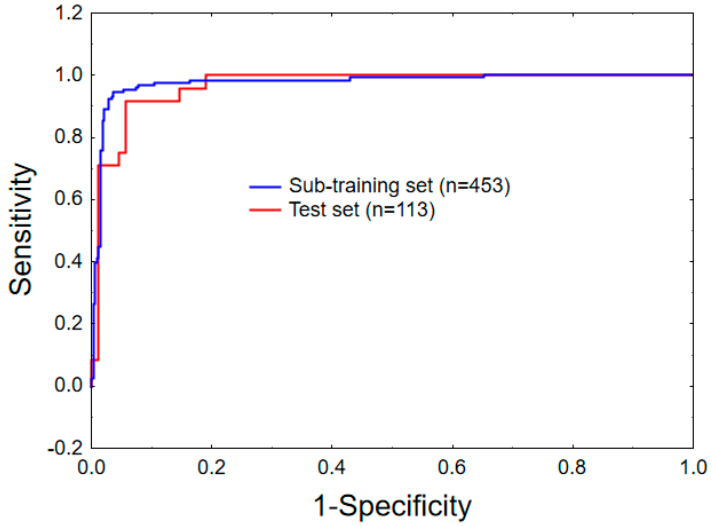

Figure 1. Receiver operating characteristic (ROC) curves for the sub-training (10-fold) and the test sets.

Further analysis of this classification model should only be carried out after checking the degree of collinearity among the independent variables of the model, and that may be easily diagnosed by analysing the cross-correlation matrix. As can be seen in Table 2, the maximum Pearson correlation coefficient (i.e., *r*) obtained is 0.74, leading us to infer that the developed model is non-redundant in nature.

Table 2. Degree of collinearity among the variables of the mt-QSAR LDA model.

Descriptors	Δ[*nCONN*]b_t	Δ[*Mor18m*]b_t	Δ[*HTm*]m_t	Δ[*nROCON*]c_l	Δ[SpMAD_A]b_t	Δ[*HATS2i*]b_t	Δ[*HATS8s*]c_l	Δ[*R1m*]c_l	Δ[*F07*[*N-N*]]c_l	Δ[*SM15_EA*(*dm*)]c_l
Δ[*nCONN*]b_t	1.00	0.14	0.08	−0.20	−0.18	−0.29	−0.23	−0.17	0.00	0.07
Δ[*Mor18m*]b_t	0.14	1.00	−0.37	−0.18	0.15	0.11	0.21	0.14	−0.07	0.11
Δ[*HTm*]m_t	0.08	-0.37	1.00	0.09	−0.14	−0.29	−0.21	−0.07	0.14	0.13
D[*nROCON*]c_l	−0.20	-0.18	0.09	1.00	0.01	0.17	0.00	0.02	0.23	0.12
D[SpMAD_A]b_t	−0.18	0.15	−0.14	0.01	1.00	−0.06	−0.11	0.01	0.15	−0.26
D[*HATS2i*]b_t	−0.29	0.11	−0.29	0.17	−0.06	1.00	0.18	0.02	−0.04	0.01
D[*HATS8s*]c_l	−0.23	0.21	−0.21	0.00	−0.11	0.18	1.00	0.74	−0.09	0.29
D[*R1m*]c_l	−0.17	0.14	−0.07	0.02	0.01	0.02	0.74	1.00	−0.02	0.36
D[*F07*[*N-N*]]c_l	0.00	−0.07	0.14	0.23	0.15	−0.04	−0.09	−0.02	1.00	0.17
D[*SM15_EA*(*dm*)]c_l	0.07	0.11	0.13	0.12	−0.26	0.01	0.29	0.36	0.17	1.00

Then, the Y-randomization test [59,60] was performed on the sub-training set to confirm that the model was not developed by chance. An average λ value of 0.978 was obtained from 100 randomized models in which the dependent parameter was randomly scrambled. As this average value is far greater than the original λ value (i.e., 0.234), one may conclude that the mt-QSAR-LDA model is unique in nature.

Finally, the applicability domain (AD) of the model was determined by the standardisation approach proposed by Roy et al [61] with the help of QSAR-Co [11]. Fifteen data-samples of the sub-training and two samples of the test set were found to remain outside the AD of the model (SM1.xlsx). Overall, it may thus be concluded that the developed mt-QSAR-LDA model satisfies each required criterion for being a robust classifier model.

In addition to model development, the QSAR-Co tool also allows the screening of large datasets. Using this screening facility, the mt-QSAR-LDA model was used to screen an external validation set (n = 160) to further confirm the external predictivity. Details of this external dataset and calculated descriptors, as well as the results of the predictions are provided in supplementary materials (SM2.xlsx). The GA-LDA model could correctly predict 118 out of 123 active datapoints achieving a sensitivity of 95.93%. On the other hand, a specificity of 89.19% was obtained as 33 out of 37 inactive data points were predicted correctly. Therefore, as far as the screening of external validation set is concerned, the model has an accuracy of 94.38% and a Matthews correlation coefficient (MCC) value of 0.843. Moreover, when the AD was calculated, only five samples of the external validation set were found as outliers. Altogether, these diverse statistics demonstrate the high internal quality as well as predictive power of the derived mt-QSAR-LDA model.

2.2. Physicochemical and Structural Interpretation of the Molecular Descriptors

Let us now scrutinise the physicochemical/structural information of the ten variables appearing in the derived mt-QSAR-LDA model (see Equation (1)) using the QSAR-Co software [11]. Evidently, each model variable not only expresses the contribution of a specific molecular descriptor (i.e., core descriptor) but also highlights the significance of a specific experiment element (i.e., *cj*). Therefore, in order to understand the significance of these variables, both these factors (i.e., core descriptor and experimental element) should be taken into consideration. As it is evident, all three experimental elements considered in the current study (i.e., b_t, c_l and m_t) found a place in the final mt-QSAR-LDA model (Equation (1)). The c_l element, which incorporates the information due to cell-type changes, appeared five times in the model whereas the element representing the biological enzyme target (i.e, b_t) was coupled to four independent variables of the model, meanwhile the third element *mt*, representing the mutation status of the cell, appeared only once. The meaning of these variables is described in Table 3. To understand the relative importance of these model variables, one needs to inspect the absolute values of their standardised coefficients (see Figure 2), that is, those are as follows: $\Delta[nCONN]b_t > \Delta[R1m]c_l > \Delta[SpMAD\_A]b_t > \Delta[HTm]m_t > \Delta[F07[N\text{-}N]]c_l > \Delta[HATS8s]c_l > \Delta[SM15\_EA(dm)]c_l > \Delta[HATS2i]b_t > \Delta[nROCON]c_l > \Delta[Mor18m]b_t$.

The most significant variable of the model is $\Delta[nCONN]b_t$, where *nCONN* is a functional group count descriptor that annotates the number of urea/thiourea fragments present in the structure of the compound. This variable, which is also sensitive towards the biological enzyme target element (b_t), indicates the significance of urea/thiourea molecular fragment for determining the potency of the PI3K inhibitors. The core descriptor of variable $\Delta[R1m]c_l$ (i.e., *R1m*) stands for the R-autocorrelation of lag-1, obtained through the leverage/geometry matrix of the corresponding molecular graphs. *R1m* is a 3D GETAWAY (GEometry, Topology and Atom Weights AssemblY) type of descriptor [62,63], which attempts to match molecular geometry with chemical information obtained from atomic masses. Furthermore, $\Delta[R1m]c_l$ also incorporates the changes due the variation in the cell type in which the assay was performed. Both variables $\Delta[nCONN]b_t$ and $\Delta[R1m]c_l$ have negative coefficients, whereas the third most significant descriptor $\Delta[SpMAD\_A]b_t$ has a positive coefficient. $\Delta[SpMAD\_A]b_t$ is based on the 2D matrix-based descriptor *SpMAD_A*, which stands for the spectral mean absolute deviation obtained from the adjacency matrix [64,65]. $\Delta[HTm]m_t$ is the only variable which is modified on the basis of mutation status of the cell, and the core descriptor *HTm* is a 3D GETAWAY descriptor weighted by atomic mass, similar to the *R1m* described above. However, unlike *R1m*, which is obtained from the leverage/geometry matrix, the calculation *HTm* is based on the molecular influence matrix [62,63]. The $\Delta[F07[N\text{–}N]]c_l$ is dependent on the 2D-atom pair descriptor *F07*[*N–N*], which is the frequency of

N−N at a topological distance of seven [64]. Presence of this descriptor in the model indicates that the distances between two nitrogen atoms in the compound structure may contribute to its inhibitory activity, depending at the same time on the cell type of the enzyme inhibition assay. Both $\Delta[HTm]m_t$ and $\Delta[F07[N–N]]c_l$ are positively correlated with the response variable.

Table 3. Ten variables selected in the multi-target LDA model.

Name	Description	Descriptor Type
$\Delta[nCONN]b_t$	Number of urea (-thio) fragment, depending on the chemical structure and enzyme target	Functional group counts
$\Delta[R1m]c_l$	R autocorrelation of lag 1/weighted by mass, depending on the chemical structure and cell type	GETAWAY indices
$\Delta[SpMAD\_A]b_t$	Spectral mean absolute deviation from the adjacency matrix, depending on the chemical structure and biological target enzyme	2D matrix-based adjacency matrix descriptors
$\Delta[HTm]m_t$	H total index/weighted by mass, depending on the cell mutation and chemical structure	GETAWAY H-indices
$\Delta[F07[N–N]]c_l$	Frequency of N-N at topological distance 7, depending on the chemical structure and cell type	2D Atom Pairs
$\Delta[HATS8s]c_l$	Leverage-weighted autocorrelation of lag 8/weighted by I-state, depending on the chemical structure and cell type	GETAWAY H-indices
$\Delta[SM15\_EA(dm)]c_l$	Spectral moment of order 15 from edge adjacency matrix weighted by dipole moment, depending on the chemical structure and cell type	Edge adjacency indices
$\Delta[HATS2i]b_t$	Leverage-weighted autocorrelation of lag 2/weighted by ionization potential, depending on the chemical structure and biological target enzyme	GETAWAY H-indices
$\Delta[nROCON]c_l$	Number of (thio-) carbamates (aliphatic), depending on the chemical structure and cell type	Functional group counts
$\Delta[Mor18m]b_t$	Signal 18/weighted by mass, depending on the chemical structure and biological target enzyme	3D-MoRSE, weighted by mass

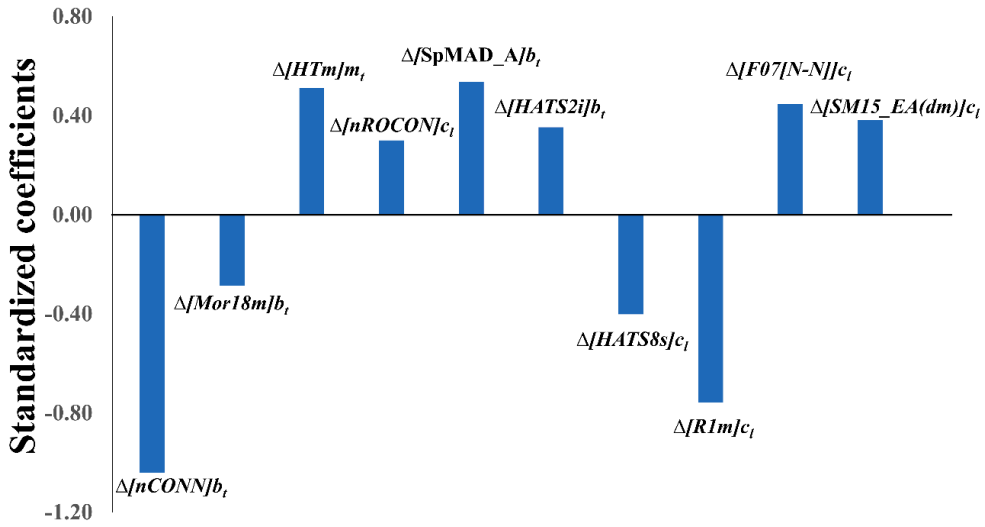

Figure 2. Standardised coefficients *vs.* variables in the mt-QSAR LDA model.

The mt-QSAR model showed the importance of two other 3D-GETAWAY based descriptors, which are $\Delta[HATS8s]c_l$ and $\Delta[HATS2i]b_t$ [62,63]. While $\Delta[HATS8s]c_l$ has a negative coefficient, $\Delta[HATS2i]b_t$ is found to be associated with a positive one. Significantly, both these descriptors are calculated on the basis of the leverage weighted autocorrelation of molecular graphs. As atomic weights, intrinsic state (*I*-state) and ionization potential are used for the calculation of *HATS8s* and *HATS2i* descriptors, respectively [62,63]. Another descriptor $\Delta[SM15\_EA(dm)]c_l$ is based on the 2D-edge adjacency spectral moment descriptor *SM15_EA*(*dm*). Similar to 2D matrix-based adjacency matrix descriptors, edge adjacency matrices are also calculated considering the chemical structures as weighted graphs. In the case of edge adjacency indices however, the elements of *edges* are substituted by the bond orders between connected atoms in the molecule [66]. Like $\Delta[nCONN]b_t$, another variable of the model $\Delta[nROCON]c_l$ is based on a simple functional group count descriptor. The core descriptor of the latter, *nROCON*, annotates the number of aliphatic carbamates or thiocarbamates in the compounds. Unlike $\Delta[nCONN]b_t$, $\Delta[nROCON]c_l$ is however dependent on the cell type element (c_l). Both $\Delta[SM15\_EA(dm)]c_l$ and $\Delta[nROCON]c_l$ are sensitive to changes of cell types and both are associated with positive coefficients. The last variable of the model $\Delta[Mor18m]b_t$ is based on the 3D-Morse (molecular representation of structures based on electronic diffraction) descriptor *Mor18m*. 3D-Morse are descriptors calculated by summing atomic weights observed from different angular scattering functions (or signals) [67]. For the calculation of *Mor18m*, the mass is used as atomic weight. $\Delta[Mor18m]b_t$, which is dependent on the biological enzyme target, is negatively correlated with the response variable.

2.3. Non-Linear Mt-QSAR Model Development

In addition to GA-LDA based linear model development, QSAR-Co allows for the generating of non-linear mt-QSAR models using the random forest (RF) strategy [11,68,69]. It is often observed (but not always) that non-linear models developed with all calculated descriptors produce more predictive models than the linear models developed with a limited number of descriptors [70,71]. Of course, overall interpretability of such non-linear models is inferior to that of linear models [70–73]. RF is basically an ensemble classification method that makes predictions by averaging over the predictions of multiple independent decision trees. Due to its high accuracy and superiority, RF has received considerable attention in recent years [68,69]. One of the major advantages of RF is that it is less susceptible to produce overfitted models. As such, RF may be preferred over many other non-linear machine learning methods to produce highly accurate mt-QSAR models [74–76].

In the current case study, we applied the random forest (RF) based classification approach to develop non-linear models [68,69]. All the descriptors calculated by AlvaDesc were applied for the development of RF model with the help of the QSAR-Co tool [11]. The obtained results for the best RF model found are depicted in Table 4.

Table 4. Overall performance of the final mt-QSAR random forest (RF) model.

Classification [a]	Sub-training Set (10-fold CV) [b]	Test Set
ND_{Total}	453	113
ND_{active}	324	89
CCD_{active}	313	87
Sensitivity (%)	96.6	97.75
$ND_{inactive}$	129	24
$CCD_{inactive}$	117	21
Specificity (%)	90.7	87.5
F-measure	0.965	0.972
Accuracy (%)	94.92	95.57
MCC	0.875	0.866

[a] ND_{Total}—Number of total data-points; ND_{active}—Number of active data-points, $ND_{inactive}$—Number of inactive data-points, CCD—Correctly classified data, MCC—Matthews correlation coefficient. [b] 10-fold cross-validation statistics.

The internal predictivity of the RF model was judged by 10-fold cross-validation statistics (as implemented in QSAR-Co tool) and in so doing, an overall accuracy of 94.92% was obtained. On the other hand, the RF model depicted an accuracy of 95.57% for the test set. However, when the developed RF model was used to predict the external validation set to check its true external predictivity, the model could correctly classify 119 out of 123 active and 28 out of 37 inactive data points. Therefore, the RF model reveals a sensitivity, specificity, accuracy and MCC of 96.75%, 75.68%, 91.88% and 0.763, respectively. Overall, the LDA model could correctly classify 689 out of 726 data-points whereas the RF model correctly classified 685 out of 726 data-points. Evidently, in terms of predictive power, the performance of the LDA model is slightly better than that of the RF model. The attribute importance (based on the average impurity decrease per attribute over the trees) are provided in the supplementary materials (SM3.xlsx).

2.4. Quantitative Contributions of the Fragments Towards Inhibitory Activity

Since the mt-QSAR LDA model was shown to be slightly more predictive than the non-linear mt-QSAR RF model, we attempted to use it as a tool to calculate the quantitative contributions of some important cyclic/ring fragments present in the dataset. To identify such fragments, the Bemis–Murcko scaffolds [77] were calculated with the entire dataset containing 726 data-points with the help of the OCHEM web server [78]. Thirty-five small alicyclic/ring fragments that were found in more than 10 dataset compounds were selected and the inhibitory activities of these fragments were subsequently calculated against all the experimental conditions (i.e., combinations of the elements b_t, m_t and c_l) reported in this work (a total of 40). After that, the variables appearing in Equation (1) were calculated for these fragments, following the same procedure by which these were calculated for the external validation set molecules, and considering all 40 different experimental conditions. A total of 1400 (= 35 × 40) scores were obtained by putting the calculated variables into Equation (1). These scores are, however, non-standardised and in order to obtain standardised scores the following procedure was adopted. The arithmetic means and the standard deviation of all these non-standardised scores were calculated. Subsequently, a standardisation procedure was employed where the arithmetic scores were subtracted from each non-standardised score and these subtracted values were then divided by the standard deviation [10,50]. These standardised scores represent the quantitative contributions of the fragments for the inhibitory potentials of these fragments and various experimental conditions. The chemical structures and the average standardised scores (or contributions) calculated for each fragment are presented in Figure 3, whereas all these scores are provided as supplementary materials (SM4.xlsx).

As can be deduced from Figure 3, fragments such as **F12, F22, F10, F26, F23, F11, F20, F28, F21** and **F29** depicted highly positive average standardised scores (>0.5) signifying that these fragments had positive contributions against all experimental conditions. These fragments may, therefore, be considered as the most suitable fragments for the design of novel class I PI3K inhibitors. Similarly, some fragments such as **F35, F34, F8, F7, F6, F2, F4, F9, F5, F3, F1, F31** and **F32** displayed highly negative average standardised scores (<−0.5). Thus, the later fragments should be avoided while designing novel PI3K inhibitors. Moderate positive contributions are obtained for **F18, F27, F13, F19, F25, F30** and **F17**, whereas slightly negative contributions are observed for fragments such as **F33, F15, F16, F14** and **F24.** These fragments are likely to have mixed responses against different experimental conditions. It should be noted here that some compounds of the current dataset may be found with positive response even though these contain fragments with negative scores. The opposite case is also true. It is not unnatural because combinations and connections of these fragments ultimately determine the overall inhibitory potential of the compounds, rather than their mere presence in the compound structures.

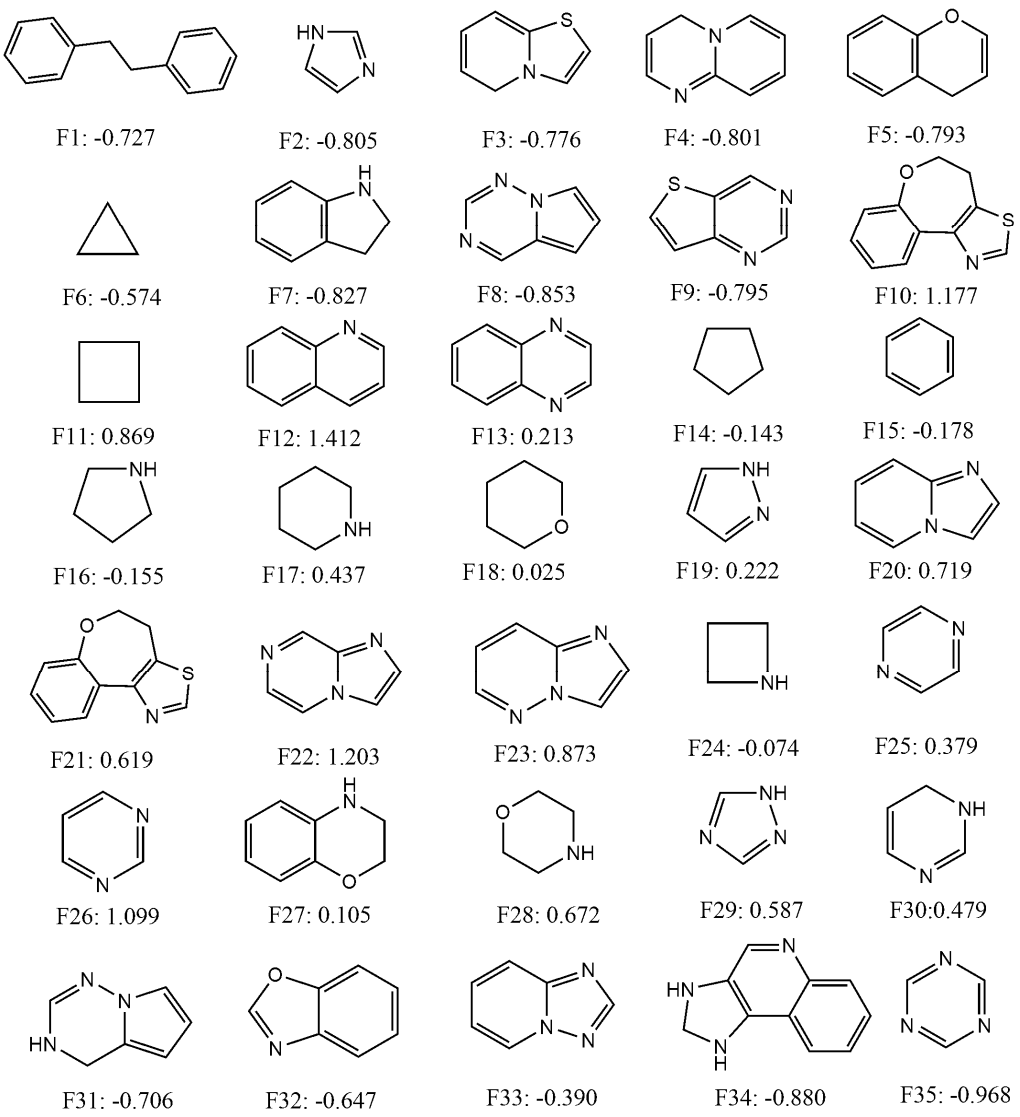

Figure 3. Chemical structures and average standardized scores of selected fragments.

3. Materials and Methods

3.1. Dataset Curation and Descriptor Calculation

After collecting the dataset compounds from CHEMBL (https://www.ebi.ac.uk/chembl/), the dataset was curated by removing duplicate data-points. The SMILES formats of the molecules obtained from the CHEMBL were converted into SDF formats by the MarvinView v18.18.0 software (https://docs.chemaxon.com/display/docs/MarvinView). Molecular descriptors were calculated by the AlvaDesc v.1.0.8 software (https://www.alvascience.com/alvadesc/) with the help of OCHEM [78], a freely available web server for QSAR descriptor calculation, model development and data storage. Before the calculation of these descriptors, the compounds were pre-processed (structures were standardised, neutralised, cleaned and salts were removed) in the OCHEM platform. Furthermore, during descriptor calculation, each dataset compound was geometrically optimised in OCHEM by the Corina software [79].

3.2. Box–Jenkins Approach

The global descriptors calculated by AlvaDesc v.1.0.8 only consider the chemical structures of the compounds, and these descriptors are thus incapable of discriminating the influence on the chemical

structure when a specific molecule is assayed under more than one experimental condition (i.e., *cj*). To solve this problem, we adopted the Box–Jenkins moving average approach [80], the details of which have been discussed previously in detail [10,11,50,60]. Briefly, Box–Jenkins operators are used to calculate successive average values of a defined property of a defined system at different intervals of time. In Box–Jenkins based mt-QSAR modelling, the time domain is not considered. Rather, the arithmetic average of any molecular descriptors for a specific experimental condition is calculated as follows [11]:

$$avg(D_i)c_j = \sum_{i=1}^{n(c_j)} D_i \quad (2)$$

Here, D_i is the calculated descriptor of the individual compound '*i*' and $n(c_j)$ is the number of actives in the modelling dataset (= sub-training set + test set) assayed under the same element of the experimental condition c_j. The $avg(D_i)c_j$ is thus the arithmetic mean of the descriptors (D_i) for a specific experimental condition (c_j). After generating the $avg(D_i)c_j$ values, the final modified descriptors ($\Delta(D_i)cj$) are subsequently generated using the following formula:

$$\Delta(D_i)c_j = D_i - avg(D_i)c_j \quad (3)$$

In this equation, $\Delta(D_i)c_j$ is a deviation descriptor that actually measures to what extent a chemical structurally deviates from a set of compounds assigned as active and tested against the same experimental condition [11,81,82]. In the QSAR-Co software, the calculated D_i descriptors are provided as inputs as these descriptors are automatically converted into $\Delta(D_i)c_j$ by this tool for the development of mt-QSAR models [11].

3.3. Model Development and Validation

In the current work, we employed the QSAR-Co tool to develop mt-QSAR models by GA-LDA and RF methods [11]. Before setting up both these models, the dataset was divided into a modelling set and an external validation set by the *k*-means cluster analysis (*k*-MCA) technique [83] with the help of the STATISTICA software package [84]. The purpose of *k-MCA* is to ensure that the validation set covers the same chemical-biological space as the training set, on which the model is built with [60]. Constitutional descriptors calculated by AlvaDesc were used along with the *IAi*(*cj*) values for generating five clusters based on the Euclidian distances from 500 iterations. From each cluster, external validation set samples were randomly collected to form an external validation set of 160 data samples. It is worth mentioning here that, the mt-QSAR models were developed only with the modelling set of 566 samples. Once the models were developed with the modified descriptors as described in Equations (2) and (3), they were then used to screen the external validation set in order to estimate their true predictivity. The best predictive model was selected based on the predictivity obtained for the external validation set. For setting up the models, however, the modelling dataset was further randomly divided into a sub-training (80% of the training data) and a test set (20% of the training data) with the help of the QSAR-Co tool [11].

The parameter settings used for the GA-LDA technique in QSAR-Co were: (a) total number of iteration/generation: 100, (b) equation length: 10 (fixed), (c) mutation probability: 0.3, (d) initial number of equation generated: 100, (e) number of equation selected in each generation: 30. Although no data pre-treatment strategy was used during the development of the models, inter-collinearity of the model descriptors was checked by examining the cross-correlation matrix and models generated with highly correlated descriptors ($-0.8 < r < 0.8$, r being the Pearson's correlation coefficient) were discarded. With the help of QSAR-Co [11] and STATISTICA [84], the internal predictivity of the models developed based on the sub-training set was evaluated by standard statistical indices such as the Wilks' lambda (λ), chi-squared (χ^2), the square of Mahalanobis distance (D^2), Fisher's statistic index (F) and the corresponding p-value (p)[85]. The goodness of prediction for the sub-training, test and external

validation sets was evaluated by computing the following statistical measures: sensitivity (correct classification of the active cases), specificity (correct classification of inactive cases), accuracy (overall correct classification), *F*-measure and Matthews correlation coefficient (MCC) [57,85]. Moreover, a *Y*-randomization test was used on the sub-training set to judge the uniqueness of the statistical model [59,60]. Therefore, the values of the dependent variable were randomly scrambled 100 times, and the Wilk's lambda (λ) of the original model was then compared with the average Wilk's lambda (λ_{rand}) of the randomized models. For determining the applicability domain, the standardization approach [61] was employed with the help of the QSAR-Co tool [11].

For settling the RF model, important parameters settings of QSAR-Co were considered, namely: (a) each bag size: 100, (b) maximum depth: 0 (unlimited), (c) number of randomly chosen features: 0 (i.e., n = int(log2[#Predictors]+1)), (d) number of iterations: 100. It should be noted that the change of these parameter settings failed to improve the predictivity of the modelling dataset to a considerable extent.

4. Conclusions

Multi-target QSAR modelling based on the Box–Jenkins approach was successfully utilised in recent years to establish a number of validated predictive chemometric models for various targets [9–11,50,60,82,86]. In the current work, we carried out such kind of mt-QSAR modelling on inhibitors of four different isoforms of class I PI3K enzyme using the recently introduced QSAR-Co tool [11]. Cell-based enzyme inhibition assay results obtained from chemical databases like CHEMBL are rarely utilised for the development of chemometric models. However, these results represent a more complicated bio-functional scenario that exists inside living cells, and therefore they may be more challenging to predict by using chemometric models. In the current investigation, our aim was to focus on the cell-based assays performed for PI3K enzyme inhibition. From one side, the developed mt-QSAR models help in understanding specific structural and physicochemical contributions responsible for higher potency. At the same time, the highly predictive models make them suitable for the screening of chemical libraries towards the search of novel hit molecules. The high statistical quality of the developed mt-QSAR-LDA model allowed us to calculate the relative contributions of some important cyclic/ring fragments towards higher PI3K inhibitory potential at various experimental conditions. Similarly, the model may also be employed to understand the contribution of aliphatic and other small structural fragments. The combination of the different in silico techniques employed in this work can serve as valuable guidelines to speed up early discovery of PI3K inhibitors. Finally, the current case study emphasises that the QSAR-Co tool may be utilised in the future to develop predictive models when response variables are dependent on multiple experimental conditions.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/20/17/4191/s1.

Author Contributions: Conceptualization, A.K.H. and M.N.D.S.C.; methodology, A.K.H. and M.N.D.S.C.; software, A.K.H.; validation, A.K.H. and M.N.D.S.C.; formal analysis, A.K.H; investigation, A.K.H. and M.N.D.S.C.; resources, M.N.D.S.C.; data curation, A.K.H.; writing—original draft preparation, A.K.H. and M.N.D.S.C.; writing—review and editing: A.K.H. and M.N.D.S.C.

Funding: This work was supported by UID/QUI/50006/2019 with funding from FCT/MCTES through national funds.

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

References

1. Searls, D.B. Data integration: Challenges for drug discovery. *Nat. Rev. Drug Discov.* **2005**, *4*, 45–58. [CrossRef] [PubMed]
2. Lo, Y.C.; Rensi, S.E.; Torng, W.; Altman, R.B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **2018**, *23*, 1538–1546. [CrossRef] [PubMed]

3. Sun, G.H.; Fan, T.J.; Zhang, N.; Ren, T.; Zhao, L.J.; Zhong, R.G. Identification of the Structural Features of Guanine Derivatives as MGMT Inhibitors Using 3D-QSAR Modeling Combined with Molecular Docking. *Molecules* **2016**, *21*, 823. [CrossRef] [PubMed]
4. Fan, T.J.; Sun, G.H.; Zhao, L.J.; Cui, X.; Zhong, R.G. QSAR and Classification Study on Prediction of Acute Oral Toxicity of N-Nitroso Compounds. *Int. J. Mol. Sci.* **2018**, *19*, 3015. [CrossRef] [PubMed]
5. Sun, G.H.; Fan, T.J.; Sun, X.D.; Hao, Y.X.; Cui, X.; Zhao, L.J.; Ren, T.; Zhou, Y.; Zhong, R.G.; Peng, Y.Z. In Silico Prediction of O-6-Methylguanine-DNA Methyltransferase Inhibitory Potency of Base Analogs with QSAR and Machine Learning Methods. *Molecules* **2018**, *23*, 2892. [CrossRef]
6. Speck-Planche, A. Recent advances in fragment-based computational drug design: Tackling simultaneous targets/biological effects. *Future Med. Chem.* **2018**, *10*, 2021–2024. [CrossRef]
7. 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. [CrossRef]
8. 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. [CrossRef]
9. 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. [CrossRef]
10. Speck-Planche, A.; Cordeiro, M.N.D.S. Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins. *Mol. Divers.* **2017**, *21*.
11. Ambure, P.; Halder, A.K.; Gonzalez Diaz, H.; Cordeiro, M. QSAR-Co: An Open Source Software for Developing Robust Multitasking or Multitarget Classification-Based QSAR Models. *J. Chem. Inf. Model.* **2019**, *59*, 2538–2544. [CrossRef]
12. Burke, J.E. Structural Basis for Regulation of Phosphoinositide Kinases and Their Involvement in Human Disease. *Mol. Cell* **2018**, *71*, 653–673. [CrossRef]
13. Porta, C.; Paglino, C.; Mosca, A. Targeting PI3K/Akt/mTOR Signaling in Cancer. *Front. Oncol.* **2014**, *4*, 64. [CrossRef]
14. Yip, P.Y. Phosphatidylinositol 3-kinase-AKT-mammalian target of rapamycin (PI3K-Akt-mTOR) signaling pathway in non-small cell lung cancer. *Transl. Lung Cancer Res.* **2015**, *4*, 165–176. [CrossRef]
15. Miller, M.S.; Thompson, P.E.; Gabelli, S.B. Structural Determinants of Isoform Selectivity in PI3K Inhibitors. *Biomolecules* **2019**, *9*, 82. [CrossRef]
16. Maheshwari, S.; Miller, M.S.; O'Meally, R.; Cole, R.N.; Amzel, L.M.; Gabelli, S.B. Kinetic and structural analyses reveal residues in phosphoinositide 3-kinase alpha that are critical for catalysis and substrate recognition. *J. Biol. Chem.* **2017**, *292*, 13541–13550. [CrossRef]
17. De Santis, M.C.; Gulluni, F.; Campa, C.C.; Martini, M.; Hirsch, E. Targeting PI3K signaling in cancer: Challenges and advances. *Biochim. Biophys. Acta Rev. Cancer* **2019**, *1871*, 361–366. [CrossRef]
18. Thorpe, L.M.; Yuzugullu, H.; Zhao, J.J. PI3K in cancer: Divergent roles of isoforms, modes of activation and therapeutic targeting. *Nat. Rev. Cancer* **2015**, *15*, 7–24. [CrossRef]
19. Piddock, R.E.; Bowles, K.M.; Rushworth, S.A. The Role of PI3K Isoforms in Regulating Bone Marrow Microenvironment Signaling Focusing on Acute Myeloid Leukemia and Multiple Myeloma. *Cancers* **2017**, *9*, 29. [CrossRef]
20. Jackson, S.P.; Schoenwaelder, S.M.; Goncalves, I.; Nesbitt, W.S.; Yap, C.L.; Wright, C.E.; Kenche, V.; Anderson, K.E.; Dopheide, S.M.; Yuan, Y.; et al. PI 3-kinase p110beta: A new target for antithrombotic therapy. *Nat. Med.* **2005**, *11*, 507–514. [CrossRef]
21. Zhang, J.; Grubor, V.; Love, C.L.; Banerjee, A.; Richards, K.L.; Mieczkowski, P.A.; Dunphy, C.; Choi, W.; Au, W.Y.; Srivastava, G.; et al. Genetic heterogeneity of diffuse large B-cell lymphoma. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 1398–1403. [CrossRef]
22. Cushing, T.D.; Metz, D.P.; Whittington, D.A.; McGee, L.R. PI3Kdelta and PI3Kgamma as targets for autoimmune and inflammatory diseases. *J. Med. Chem.* **2012**, *55*, 8559–8581. [CrossRef]
23. Yoo, E.J.; Ojiaku, C.A.; Sunder, K.; Panettieri, R.A., Jr. Phosphoinositide 3-Kinase in Asthma: Novel Roles and Therapeutic Approaches. *Am. J. Respir. Cell Mol. Biol.* **2017**, *56*, 700–707. [CrossRef]
24. Yang, J.; Nie, J.; Ma, X.; Wei, Y.; Peng, Y.; Wei, X. Targeting PI3K in cancer: Mechanisms and advances in clinical trials. *Mol. Cancer* **2019**, *18*, 26. [CrossRef]

25. Miller, B.W.; Przepiorka, D.; de Claro, R.A.; Lee, K.; Nie, L.; Simpson, N.; Gudi, R.; Saber, H.; Shord, S.; Bullock, J.; et al. FDA approval: Idelalisib monotherapy for the treatment of patients with follicular lymphoma and small lymphocytic lymphoma. *Clin. Cancer Res.* **2015**, *21*, 1525–1529. [CrossRef]
26. Sanchez, V.E.; Nichols, C.; Kim, H.N.; Gang, E.J.; Kim, Y.M. Targeting PI3K Signaling in Acute Lymphoblastic Leukemia. *Int. J. Mol. Sci.* **2019**, *20*, 412. [CrossRef]
27. Fruman, D.A.; Chiu, H.; Hopkins, B.D.; Bagrodia, S.; Cantley, L.C.; Abraham, R.T. The PI3K Pathway in Human Disease. *Cell* **2017**, *170*, 605–635. [CrossRef]
28. Wee, S.; Wiederschain, D.; Maira, S.M.; Loo, A.; Miller, C.; deBeaumont, R.; Stegmeier, F.; Yao, Y.M.; Lengauer, C. PTEN-deficient cancers depend on PIK3CB. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13057–13062. [CrossRef]
29. Curigliano, G.; Shah, R.R. Safety and Tolerability of Phosphatidylinositol-3-Kinase (PI3K) Inhibitors in Oncology. *Drug Saf.* **2019**, *42*, 247–262. [CrossRef]
30. Evans, C.A.; Liu, T.; Lescarbeau, A.; Nair, S.J.; Grenier, L.; Pradeilles, J.A.; Glenadel, Q.; Tibbitts, T.; Rowley, A.M.; DiNitto, J.P.; et al. Discovery of a Selective Phosphoinositide-3-Kinase (PI3K)-gamma Inhibitor (IPI-549) as an Immuno-Oncology Clinical Candidate. *ACS Med. Chem. Lett.* **2016**, *7*, 862–867. [CrossRef]
31. Greenwell, I.B.; Ip, A.; Cohen, J.B. PI3K Inhibitors: Understanding Toxicity Mechanisms and Management. *Oncol.* **2017**, *31*, 821–828.
32. Bharate, S.B.; Singh, B.; Bharate, J.B.; Jain, S.K.; Meena, S.; Vishwakarma, R.A. QSAR and pharmacophore modeling of N-acetyl-2-aminobenzothiazole class of phosphoinositide-3-kinase-alpha inhibitors. *Med. Chem. Res.* **2013**, *22*, 890–899. [CrossRef]
33. Chadha, N.; Jasuja, H.; Kaur, M.; Bahia, M.S.; Silakari, O. Imidazo[1,2-a]pyrazine inhibitors of phosphoinositide 3-kinase alpha (PI3K alpha): 3D-QSAR analysis utilizing the Hybrid Monte Carlo algorithm to refine receptor-ligand complexes for molecular alignment. *Sar Qsar Env. Res.* **2014**, *25*, 221–247. [CrossRef] [PubMed]
34. Kaur, M.; Silakari, O. Identification of new dual spleen tyrosine kinase (Syk) and phosphoionositide-3-kinase (PI3K) inhibitors using ligand and structure-based integrated ideal pharmacophore models. *Sar Qsar Env. Res.* **2016**, *27*, 469–499. [CrossRef] [PubMed]
35. Li, Y.; Wang, Y.; Zhang, F. Pharmacophore modeling and 3D-QSAR analysis of phosphoinositide 3-kinase p110alpha inhibitors. *J. Mol. Model.* **2010**, *16*, 1449–1460. [CrossRef] [PubMed]
36. Liew, C.Y.; Ma, X.H.; Yap, C.W. Consensus model for identification of novel PI3K inhibitors in large chemical library. *J. Comput. Aided Mol. Des.* **2010**, *24*, 131–141. [CrossRef] [PubMed]
37. Oluic, J.; Nikolic, K.; Vucicevic, J.; Gagic, Z.; Filipic, S.; Agbaba, D. 3D-QSAR, Virtual Screening, Docking and Design of Dual PI3K/mTOR Inhibitors with Enhanced Antiproliferative Activity. *Comb. Chem. High. Throughput Screen.* **2017**, *20*, 292–303. [CrossRef] [PubMed]
38. Peddi, S.R.; Sivan, S.K.; Manga, V. Discovery and design of new PI3K inhibitors through pharmacophore-based virtual screening, molecular docking, and binding free energy analysis. *Struct. Chem.* **2018**, *29*, 1753–1766. [CrossRef]
39. Peng, X.X.; Feng, K.R.; Ren, Y.J. Molecular modeling studies of quinazolinone derivatives as novel PI3K delta selective inhibitors. *RSC Adv.* **2017**, *7*, 56344–56358. [CrossRef]
40. Ran, T.; Lu, T.; Yuan, H.; Liu, H.; Wang, J.; Zhang, W.; Leng, Y.; Lin, G.; Zhuang, S.; Chen, Y. A selectivity study on mTOR/PI3Kalpha inhibitors by homology modeling and 3D-QSAR. *J. Mol. Model.* **2012**, *18*, 171–186. [CrossRef]
41. Safavi-Sohi, R.; Ghasemi, J.B. Quasi 4D-QSAR and 3D-QSAR study of the pan class I phosphoinositide-3-kinase (PI3K) inhibitors. *Med. Chem. Res.* **2013**, *22*, 1587–1596. [CrossRef]
42. Sharma, P.; Shukla, A.; Kalani, K.; Dubey, V.; Luqman, S.; Srivastava, S.K.; Khan, F. In-silico & In-vitro Identification of Structure-Activity Relationship Pattern of Serpentine & Gallic Acid Targeting PI3Kgamma as Potential Anticancer Target. *Curr. Cancer Drug Targets* **2017**, *17*, 722–734. [CrossRef] [PubMed]
43. Taha, M.O.; Al-Sha'er, M.A.; Khanfar, M.A.; Al-Nadaf, A.H. Discovery of nanomolar phosphoinositide 3-kinase gamma (PI3K gamma) inhibitors using ligand-based modeling and virtual screening followed by in vitro analysis. *Eur. J. Med. Chem.* **2014**, *84*, 454–465. [CrossRef] [PubMed]
44. Takeda, T.; Wang, Y.; Bryant, S.H. Structural insights of a PI3K/mTOR dual inhibitor with the morpholino-triazine scaffold. *J. Comput. Aided Mol. Des.* **2016**, *30*, 323–330. [CrossRef] [PubMed]

45. Wang, F.X.; Chen, Y.D. Pharmacophore models generation by catalyst and phase consensus-based virtual screening protocol against PI3K alpha inhibitors. *Mol. Simul.* **2013**, *39*, 529–544. [CrossRef]
46. Wang, J.; Wang, F.; Xiao, Z.; Sheng, G.; Li, Y.; Wang, Y. Molecular simulation of a series of benzothiazole PI3Kalpha inhibitors: Probing the relationship between structural features, anti-tumor potency and selectivity. *J. Mol. Model.* **2012**, *18*, 2943–2958. [CrossRef]
47. Wu, F.; Hou, X.Y.; Luo, H.; Zhou, M.; Zhang, W.J.; Ding, Z.Y.; Li, R. Exploring the selectivity of PI3K alpha and mTOR inhibitors by 3D-QSAR, molecular dynamics simulations and MM/GBSA binding free energy decomposition. *Medchemcomm* **2013**, *4*, 1482–1496. [CrossRef]
48. Yang, W.J.; Shu, M.; Wang, Y.Q.; Wang, R.; Hu, Y.; Meng, L.X.; Lin, Z.H. 3D-QSAR and docking studies of 3-Pyridine heterocyclic derivatives as potent PI3K/mTOR inhibitors. *J. Mol. Struct.* **2013**, *1054*, 107–116. [CrossRef]
49. Yu, M.; Gu, Q.; Xu, J. Discovering new PI3K alpha inhibitors with a strategy of combining ligand-based and structure-based virtual screening. *J. Comput. Aided Mol. Des.* **2018**, *32*, 347–361. [CrossRef]
50. 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. [CrossRef]
51. Rogers, D.; Hopfinger, A.J. Application of Genetic Function Approximation to Quantitative Structure-Activity-Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866. [CrossRef]
52. Snedecor, G.W.; Cochran, W.G. *Statistical methods*, 8th ed.; Iowa State University Press: Ames, IA, USA, 1989; p. 503.
53. Ambure, P.; Aher, R.B.; Gajewicz, A.; Puzyn, T.; Roy, K. "NanoBRIDGES" software: Open access tools to perform QSAR and nano-QSAR modeling. *Chemom. Intell. Lab. Syst.* **2015**, *147*, 1–13. [CrossRef]
54. Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J.P. Activity, assay and target data curation and quality in the ChEMBL database. *J. Comput. Aided Mol. Des.* **2015**, *29*, 885–896. [CrossRef]
55. Anderson, A.C. The process of structure-based drug design. *Chem. Biol.* **2003**, *10*, 787–797. [CrossRef]
56. Wilks, S.S. Certain generalizations in the analysis of variance. *Biometrika* **1932**, *24*, 471–494. [CrossRef]
57. Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS ONE* **2017**, *12*. [CrossRef]
58. Hanczar, B.; Hua, J.; Sima, C.; Weinstein, J.; Bittner, M.; Dougherty, E.R. Small-sample precision of ROC-related estimates. *Bioinformatics* **2010**, *26*, 822–830. [CrossRef]
59. Rucker, C.; Rucker, G.; Meringer, M. y-Randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357. [CrossRef]
60. Halder, A.K.; Natalia, M.; Cordeiro, D.S. Probing the Environmental Toxicity of Deep Eutectic Solvents and Their Components: An In Silico Modeling Approach. *Acs Sustain. Chem. Eng.* **2019**, *7*, 10649–10660. [CrossRef]
61. Roy, K.; Kar, S.; Ambure, P. On a simple approach for determining applicability domain of QSAR models. *Chemom. Intell. Lab. Syst.* **2015**, *145*, 22–29. [CrossRef]
62. Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693–705. [CrossRef] [PubMed]
63. Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682–692. [CrossRef] [PubMed]
64. Todeschini, R.; Consonni, V.; Mannhold, R.; Kubinyi, H.; Folkers, G.; Wiley Online Library. Molecular Descriptors for Chemoinformatics. Vol. I & II. In *Methods and Principles in Medicinal Chemistry Ser 82*; Wiley-VCH Imprint. John Wiley & Sons, Incorporated: Hoboken, NJ, USA, 2010; p 1 online resource.
65. Consonni, V.; Todeschini, R.; Wiley Online Library. Handbook of molecular descriptors. In *Methods and principles in medicinal chemistry 11*; Wiley-VCH: Weinheim, Germany; New York, NY, USA, 2000; p 1 online resource.
66. Estrada, E.; Ramirez, A. Edge adjacency relationships and molecular topographic descriptors. Definition and QSAR applications. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 837–843. [CrossRef]

67. Devinyak, O.; Havrylyuk, D.; Lesyk, R. 3D-MoRSE descriptors explained. *J. Mol. Graph. Model.* **2014**, *54*, 194–203. [CrossRef] [PubMed]
68. Fawagreh, K.; Gaber, M.M.; Elyan, E. Random forests: From early developments to recent advancements. *Syst. Sci. Control. Eng.* **2014**, *2*, 602–609. [CrossRef]
69. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
70. Nizami, B.; Tetko, I.V.; Koorbanally, N.A.; Honarparvar, B. QSAR models and scaffold-based analysis of non-nucleoside HIV RT inhibitors. *Chemom. Intell Lab.* **2015**, *148*, 134–144. [CrossRef]
71. Halder, A.K. Finding the structural requirements of diverse HIV-1 protease inhibitors using multiple QSAR modelling for lead identification. *Sar Qsar Env. Res.* **2018**, *29*, 911–933. [CrossRef]
72. Marchese Robinson, R.L.; Palczewska, A.; Palczewski, J.; Kidley, N. Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets. *J. Chem. Inf. Model.* **2017**, *57*, 1773–1792. [CrossRef]
73. Guha, R. On the interpretation and interpretability of quantitative structure-activity relationship models. *J. Comput. Aid. Mol. Des.* **2008**, *22*, 857–871. [CrossRef]
74. Ishwaran, H. Variable importance in binary regression trees and forests. *Electron. J. Stat.* **2007**, *1*, 519–537. [CrossRef]
75. Nawar, S.; Mouazen, A.M. Comparison between Random Forests, Artificial Neural Networks and Gradient Boosted Machines Methods of On-Line Vis-NIR Spectroscopy Measurements of Soil Total Nitrogen and Total Carbon. *Sensors* **2017**, *17*, 2428. [CrossRef]
76. Lee, K.; Lee, M.; Kim, D. Utilizing random Forest QSAR models with optimized parameters for target identification and its application to target-fishing server. *BMC Bioinform.* **2017**, *18*. [CrossRef]
77. Bemis, G.W.; Murcko, M.A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893. [CrossRef]
78. Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A.K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V.V.; Tanchuk, V.Y.; et al. Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.* **2011**, *25*, 533–554. [CrossRef]
79. Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008. [CrossRef]
80. Gonzalez-Diaz, H.; Herrera-Ibata, D.M.; Duardo-Sanchez, A.; Munteanu, C.R.; Orbegozo-Medina, R.A.; Pazos, A. ANN multiscale model of anti-HIV drugs activity vs AIDS prevalence in the US at county level based on information indices of molecular graphs and social networks. *J. Chem. Inf. Model.* **2014**, *54*, 744–755. [CrossRef]
81. Alonso, N.; Caamano, O.; Romero-Duran, F.J.; Luan, F.; Cordeiro, M.N.D.S.; Yanez, M.; Gonzalez-Diaz, H.; Garcia-Mera, X. Model for high-throughput screening of multitarget drugs in chemical neurosciences: Synthesis, assay, and theoretic study of rasagiline carbamates. *ACS Chem. Neurosci.* **2013**, *4*, 1393–1403. [CrossRef]
82. Speck-Planche, A.; Scotti, M.T. BET bromodomain inhibitors: Fragment-based in silico design using multi-target QSAR models. *Mol. Divers.* **2018**. [CrossRef]
83. Gore, P.A. 11 - Cluster Analysis. In *Handbook of Applied Multivariate Statistics and Mathematical Modeling*; Tinsley, H.E.A., Brown, S.D., Eds.; Academic Press: San Diego, CA, USA, 2000; pp. 297–321. [CrossRef]
84. Statsoft-Team (2001) STATISTICA. Data analysis software system. v6.0, Tulsa. Available online: http://www.statsoft.com/Products/STATISTICA-Features (accessed on 22 May 2019).
85. Brown, M.T.; Wicker, L.R. 8 - Discriminant Analysis. In *Handbook of Applied Multivariate Statistics and Mathematical Modeling*; Tinsley, H.E.A., Brown, S.D., Eds.; Academic Press: San Diego, CA, USA, 2000; pp. 209–235. [CrossRef]
86. 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. [CrossRef]