International Journal of
Molecular Sciences

MDPI

Article

AKT Inhibitors: The Road Ahead to Computational Modeling-Guided Discovery

Amit Kumar Halder and M. Natália D. S. Cordeiro *

LAQV@REQUIMTE/Faculty of Sciences, University of Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal; amit.halder@fc.up.pt
* Correspondence: ncordeir@fc.up.pt

Abstract: AKT, is a serine/threonine protein kinase comprising three isoforms—namely: AKT1, AKT2 and AKT3, whose inhibitors have been recognized as promising therapeutic targets for various human disorders, especially cancer. In this work, we report a systematic evaluation of multi-target Quantitative Structure-Activity Relationship (mt-QSAR) models to probe AKT' inhibitory activity, based on different feature selection algorithms and machine learning tools. The best predictive linear and non-linear mt-QSAR models were found by the genetic algorithm-based linear discriminant analysis (GA-LDA) and gradient boosting (Xgboost) techniques, respectively, using a dataset containing 5523 inhibitors of the AKT isoforms assayed under various experimental conditions. The linear model highlighted the key structural attributes responsible for higher inhibitory activity whereas the non-linear model displayed an overall accuracy higher than 90%. Both these predictive models, generated through internal and external validation methods, were then used for screening the Asinex kinase inhibitor library to identify the most potential virtual hits as pan-AKT inhibitors. The virtual hits identified were then filtered by stepwise analyses based on reverse pharmacophore-mapping based prediction. Finally, results of molecular dynamics simulations were used to estimate the theoretical binding affinity of the selected virtual hits towards the three isoforms of enzyme AKT. Our computational findings thus provide important guidelines to facilitate the discovery of novel AKT inhibitors.

Keywords: AKT inhibitors; multi-target QSAR models; pharmacophore-based mapping; molecular docking; molecular dynamics simulations

Citation: Halder, A.K.; Cordeiro, M.N.D.S. AKT Inhibitors: The Road Ahead to Computational Modeling-Guided Discovery. *Int. J. Mol. Sci.* **2021**, *22*, 3944. https://doi.org/10.3390/ijms22083944

Academic Editor: Hanoch Senderowitz

Received: 14 February 2021
Accepted: 8 April 2021
Published: 11 April 2021

1. Introduction

AKT, also known as protein kinase B (PKB), is a serine/threonine-specific protein kinase that belongs to the AGC family of kinases. Three closely related isoforms of AKT occur in mammals, namely: AKT1 (PKBα), AKT2 (PKBβ) and AKT3 (PKBγ) [1,2]. All these AKT isoforms have a common structure comprised of three domains, i.e., a pleckstrin homology (PH) domain at the N terminus that binds to phosphatidylinositol-3-kinase (PI3K), a catalytic kinase domain with an ATP-binding site and a hydrophobic motif (HM) at the C-terminus [3]. A high degree of sequence homology is observed at the catalytic domain of AKT isoforms though the other two domains vary to a certain extent. AKT1 and AKT2 are expressed ubiquitously whereas AKT3 is found primarily in the brain, kidney, and heart. Being a key enzyme of the PI3K cascade, AKT plays a crucial role in the regulation of diverse cellular functions. Some major functions of AKT include cell proliferation, cell cycle progression, cell survival and inhibition of apoptosis through inactivation of more than 20 pro-apoptotic proteins [1,4–6]. AKT signaling is regularly impaired in several types of cancers and increased AKT activity has been detected in a number of aggressive malignancies. Therefore, these enzymes are considered as promising targets for the development of novel anticancer agents [7–9]. However, the scope of AKT inhibitors action is not just limited to the treatment of malignant diseases. Recent

investigations suggested that AKT inhibitors might also be applied in the treatment of neurological diseases, diabetes, obesity, cardiovascular diseases, idiopathic pulmonary fibrosis, inflammatory and autoimmune diseases [6,10,11]. Multiple attempts have so far been made to develop AKT inhibitors as anticancer agents. GSK690693 was the first clinically tested AKT inhibitor that was followed by AZD5363, ipatasertib, afuresertib, uprosertib, MK-2206, etc. [12]. Currently, at least seven AKT inhibitors are in different stages of clinical trials [13]. Most of these clinically tested agents bind to the catalytic domain of the enzymes, and therefore these simultaneously inhibit all three isoforms of the AKT (i.e., they are pan-AKT-inhibitors). However, allosteric inhibitors were also developed in order to obtain more selectivity towards one or more AKT isoforms [14]. As an example, BAY1125976 is an allosteric inhibitor with higher specificity towards AKT1 and AKT2 isoforms. However, the advantage of isoform specific AKT inhibitors against pan-AKT inhibitors are yet to be established clinically [13,15].

Nevertheless, it has already been confirmed that bioactivity against all three isoforms should be taken into consideration while developing novel AKT inhibitors.

Machine learning-based (ML) tools have thus far been proved to be an extremely useful strategy for the design and discovery of therapeutically active agents [16–18]. In particular, these have been frequently employed in Quantitative Structure–Activity Relationships (QSAR) modeling to find structural requirements for higher active molecules and/or to predict activity of novel hit molecules [19–21]. ML tools have also been applied in multi-target QSAR (mt-QSAR) modeling for jointly predicting the bioactivity of compounds optimized under multiple biological targets and assay conditions [16,20,22–26]. Recently, we have launched the software code QSAR-Co for easy tackling of multi-target classification-based QSAR modeling efforts [22]. In QSAR-Co (available in https://sites.google.com/view/qsar-co (accessed on 8 February 2021)), linear mt-QSAR models are developed by the genetic algorithm based linear discriminant analysis (GA-LDA) whereas non-linear models are generated by the random forest (RF) technique. With the desire to extend its functionalities that can further help in understanding the scope and reliability of computational modeling-guided approaches, we examined here various feature selection algorithms and machine learning tools to build reliable mt-QSAR models for probing the inhibitory action of AKT enzyme isoforms. The best predictive linear and non-linear models were then used to screen a focused kinase library to obtain the most potential virtual hits that were further investigated by structure-based methods, such as pharmacophore-based prediction, docking and molecular dynamics (MD) simulation techniques. Even though several computational modeling works targeting AKT inhibitors have been reported so far, these were always focused only on one subtype of AKT pertaining to one experimental assay condition [27–33]. To the best of our knowledge, the current work is the first one to report multi-target computational modeling-guided discovery of inhibitors for all three AKT isoforms assayed under multiple experimental assay conditions.

2. Results and Discussion

2.1. Dataset Collection and Preparation

Dataset compounds were collected from the ChEMBL database (https://www.ebi.ac.uk/chembl/ (accessed on 1 June 2020)). Details of the dataset can be found in Supplementary Materials (SM1.xlsx). Each of these compounds has been tested against at least one of the three isoforms of AKT (i.e., AKT1, AKT2 and AKT3, BAO label: Single protein) and the corresponding activity evaluated according to either half-maximal inhibitory concentration (IC_{50}) or binding affinity (Ki). Moreover, at least one of two assay techniques has been applied, i.e., either a binding assay (B) or a functional assay (F). After removal of duplicate data-points, a dataset containing 5523 samples was used for modeling the inhibitory activity, in which the latter was converted into a binary categorical response variable, $IAi(c_j)$, with values +1 (active) and −1 (inactive). Samples with IC_{50}/K_i values $\leq$ 500 nM were considered as active [$IAi(c_j) = +1$], otherwise they were considered as inactive [$IAi(c_j) = -1$] [20,34]. Further, we adopted the Box-Jenkins moving average

approach for handling the mt-QSAR modeling. Details of this computational modeling approach have been thoroughly discussed over the past and so we limit ourselves here to a brief outline [16,22,25,35]. According to the Box-Jenkins moving average approach, the input structural descriptors (D_i) of each compound are converted to deviation descriptors ($\Delta(D_i)c_j$ based on the experimental conditions c_j (or ontology) these have been tested for. As referred to above, three different experimental elements are considered here for mt-QSAR modeling, i.e., the biological target (b_t: AKT1, AKT2, or AKT3), measure of effect (m_e: IC_{50} or Ki), and assay type (a_t: B or F). The final deviation descriptors $\Delta(D_i)c_j$ not only encode structural aspects of the compounds but also information related to the experimental conditions under which these have been assayed (i.e., c_j) [22,36,37]. Details regarding the calculation of input descriptors (D_i), data curation and dataset division schemes, as well as model development strategies are discussed in the Materials and Methods section. The mt-QSAR model development strategy is outlined in Figure 1.

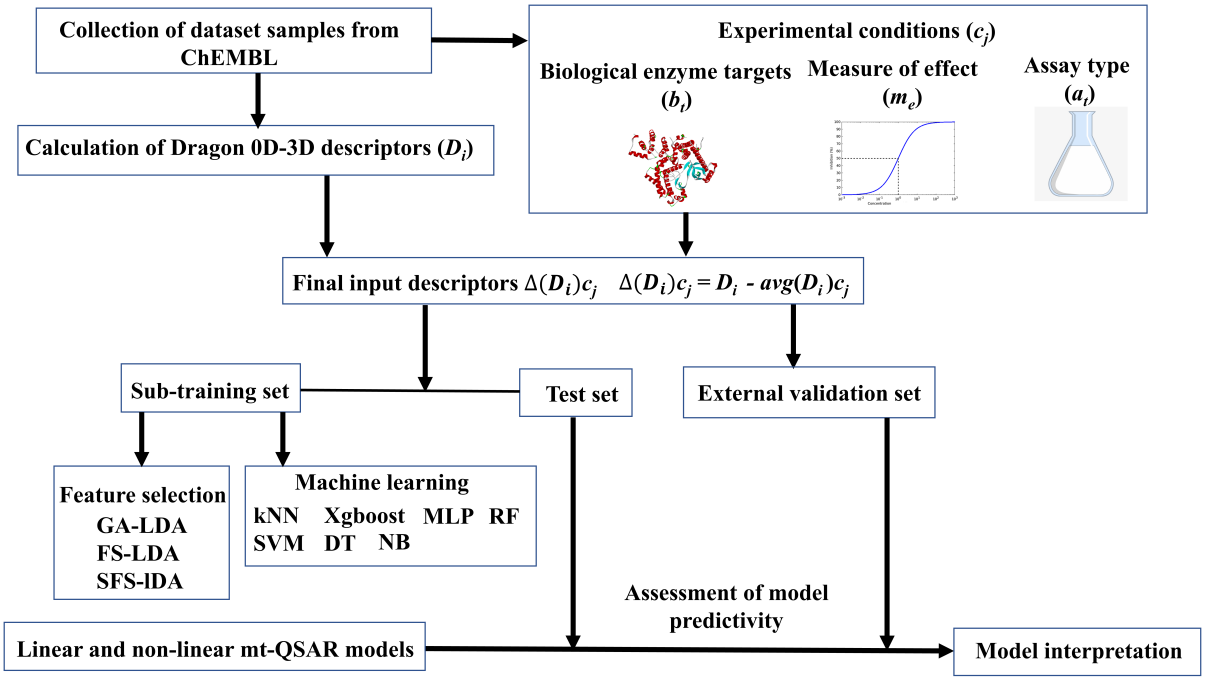

Figure 1. Flowchart showing the multi-target Quantitative Structure-Activity Relationships (mt-QSAR) work performed in the current work.

2.2. Linear Interpretable Mt-QSAR Models

The dataset was first randomly divided into a training set containing 3867 data points (70% of the data) and a validation set containing 1656 data-points (30% of the data) using the random division technique of QSAR-Co tool [22]. A total of 5305 input descriptors (D_i) were calculated for the training set by employing the alvaDesc tool [38], and these descriptors were subsequently converted to 15,915 deviation descriptors with the help of QSAR-Co tool [22]. It must be noted here that all models were set up solely on the basis of the training set, which was further split into a sub-training set containing 2707 data points (70% of the training set) and a test set containing 1160 data points (30% of the training set) by applying the random division technique of QSAR-Co tool [22] for models' development purposes. The predictivity of the built models was finally tested with the validation set. Therefore, the difference between the test set and validation set is that the test set datapoints participated in the calculation of deviation descriptors but datapoints of the validation set had no role on that.

Then, 2881 descriptors were identified to have an intercorrelation greater than 0.999 with other descriptors and after removing these descriptors the remaining 13,034 descrip-

tors were considered for development of the linear models. However, the intercorrelations among the selected descriptors of the final models were critically examined.

Three different feature selection techniques, namely: genetic algorithm (GA), forward stepwise (FS) and sequential forward selection (SFS), were used one-by-one for the development of linear interpretable models. For each of the following linear discriminant analysis (LDA) models—i.e., GA-LDA, FS-LDA and SFS-LDA, a maximum of ten descriptors was allowed. The best linear models derived from the sub-training set, with descriptors selected by these feature selection techniques, are depicted in Table 1 along with the LDA statistical parameters.

As seen, the low Wilk's lambda (λ) values and high chi-square (χ^2), squared Mahalanobis distance (D^2) and F values are indicative of the statistical significance of all three models developed. Among these models, the FS-LDA model is found to have the lowest λ value. Significantly, the goodness-of-fit of GA-LDA is very similar to that of the FS-LDA model. The degree of collinearity among the selected variables was also inspected, and the resultant cross-correlation matrices can be found in the Supplementary Materials (Tables S1–S3). The highest Pearson correlation coefficients (r) observed between two independent variables were 0.838, 0.614 and 0.742 for the GA-LDA, FS-LDA and SFS-LDA models, respectively. It must be pointed out here that we discarded all models generated with highly intercorrelated independent variables ($r \geq 0.85$). That was the case, for example, of two initial SFS-LDA models that had to be discarded and then re-generated after removing one of the descriptors with $r > 0.85$.

The next step was to verify the uniqueness of the derived models, which can easily be done by applying the Y-based randomization technique [39]. In our previous works [16,20], the Y-based randomization was performed only by scrambling the response variable but here, we slightly modified this technique and named the new technique as Yc randomization. Generally, the Y-based randomization allows one to check if the linear model was not developed by chance. In conventional computational-guided modeling, the response variable is randomly shuffled n times to generate n number of randomized models, the statistical parameters of which are then compared to that of the original model [22,40]. However, in the Box-Jenkins based mt-QSAR, the experimental elements (c_j) participate also in the calculation of the final deviation descriptors. Therefore, these experimental elements should be randomized along with the response variables to assess the robustness of the models. In order to fulfil such criteria, both the responses $IA_i(c_j)$ and elements c_j were shuffled 100 times to generate 100 different randomized datasets along with their deviation descriptors. The models developed subsequently using the same feature selection techniques were evaluated by computing the corresponding λ (λ_r) values. The average of the latter values (λ_{rm}) was then compared with the λ values obtained for the original models. The λ_{rm} values obtained for the GA-LDA, FS-LDA and SFS-LDA randomized models (0.994, 0.996 and 0.992, respectively) were found to be much higher than the λ values obtained for the original models (0.414, 0.408 and 0.507, respectively), thus confirming the unique nature of the later models.

Table 1. Goodness-of-fit of the linear models produced by different feature selection algorithms.

Method	Model	λ	χ^2	D^2	p	F (10,2696)
GA-LDA	$IA_i\left(c_j\right) = +1.766\Delta[\mathrm{VE1\_Dz(Z)}]_{b_t} + 1.399\Delta(\mathrm{Mor32m})_{b_t} + 0.173\Delta(\mathrm{L2m})_{b_t} - 0.464\Delta(\mathrm{C-032})_{m_e} - 0.280\Delta(\mathrm{F02[N-O]})_{m_e} - 0.004\Delta(\mathrm{Wi\_D/Dt})_{m_e} + 1.186\Delta(\mathrm{nRNH2})_{a_t} + 0.768\Delta(\mathrm{Mor27m})_{a_t} + 0.549\Delta(\mathrm{Mor21u})_{a_t} - 0.323\Delta(\mathrm{nArNHR})_{a_t} + 1.532$	0.414	2381.34	5.89	$<10^{-16}$	374.53
FS-LDA	$IA_i\left(c_j\right) = +2.043\Delta(\mathrm{C-030})_{b_t} + 0.953\Delta(\mathrm{nCt})_{b_t} + 0.309\Delta(\mathrm{L2m})_{b_t} + 0.013\Delta(\mathrm{D/Dtr05})_{b_t} - 0.589\Delta(\mathrm{CATS3D\_18\_DL})_{b_t} - 0.460\Delta(\mathrm{CATS3D\_10\_PL})_{b_t} + 1.016\Delta(\mathrm{nPyridines})_{m_e} + 0.481\Delta(\mathrm{CATS3D\_07\_DA})_{m_e} - 0.007\Delta[\mathrm{(T(N..O)}]_{m_e} - 3.545\Delta(\mathrm{nRHNH2})_{a_t} + 2.545$	0.408	2420.04	6.156	$<10^{-16}$	391.07
SFS-LDA	$IA_i\left(c_j\right) = +0.954\Delta(\mathrm{F08[N-S]})_{b_t} + 0.851\Delta(\mathrm{Mor31u})_{b_t} - 1.081\Delta(\mathrm{CATS2D\_02\_DD})_{b_t} - 3.280\Delta(\mathrm{B03[S-Br]})_{b_t} + 2.494\Delta(\mathrm{nRNH2})_{m_e} + 0.177\Delta(\mathrm{H-05})_{m_e} + 0.567\Delta(\mathrm{F07[N-Cl]})_{a_t} + 0.099\Delta(\mathrm{SsNH2})_{a_t} + 0.005\Delta[T\mathrm{(N..N)}]_{a_t} - 1.419\Delta(\mathrm{CATS2D\_06\_DD})_{a_t} + 1.543$	0.507	1831.98	4.120	$<10^{-16}$	261.77

Let us now check the overall predictive ability of these linear models. To do so, statistical parameters such as the sensitivity, specificity, *F*-measure, accuracy and the Matthews correlation coefficient [41,42] values were carefully examined not only for the sub-training subset but also, to infer their external predictivity, firstly for the test set (n = 1160) and finally for the validation set (n = 1656). As seen in Table 2, all models display a high predictivity against the sub-training, test and validation sets. The overall predictivity of the GA-LDA model however supersedes that of both FS-LDA and SFS-LDA models, judging from the obtained accuracy values for such sets (88.2%, 89.6%, 88.2%, respectively). Interestingly, the overall predictivity of SFS-LDA model is similar to that of the GA-LDA model. Even though FS-LDA model had the highest goodness-of-fit (lowest λ value), it afforded a lower overall predictive power compared to that of the other two models.

Table 2. Overall performance of the final linear models.

Classification [a]	GA-LDA			FS-LDA			SFS-LDA		
	Sub-Training	Test	Validation	Sub-Training	Test	Validation	Sub-Training	Test	Validation
ND_{Total} [b]	2707	1160	1656	2707	1160	1656	2707	1160	1656
ND_{active} [b]	1027	459	620	1027	459	620	1027	459	620
CCD_{active} [c]	916	413	553	901	399	541	905	409	546
Sensitivity (%)	89.2	90.0	88.6	87.4	88.6	88.5	88.2	89.9	88.7
$ND_{inactive}$ [b]	1680	701	1036	1680	701	1036	1680	1680	1036
$CCD_{inactive}$ [c]	1472	626	918	1468	621	917	1481	630	919
Specificity (%)	87.6	89.3	89.2	87.7	86.9	87.3	88.1	89.1	88.1
F-measure	0.852	0.872	0.857	0.842	0.851	0.845	0.849	0.871	0.851
Accuracy (%)	88.2	89.6	88.8	87.5	87.9	88.0	88.1	89.6	88.5
MCC [d]	0.756	0.785	0.767	0.741	0.75	0.749	0.753	0.784	0.758

[a] LDA classification statistical parameters. [b] ND: Number of datapoints. [c] Correctly classified datapoints. [d] Matthews correlation coefficient.

Another way of confirming the classification ability of any LDA model is by means of the receiver operating characteristics (ROC) curve [43]. Figure 2 shows the ROC plots for the GA-LDA and SFS-LDA models. One can see that both these models are not random, but truly statistically significant classifiers, since the area under the ROC curves (ROC AUC scores calculated with Scikit-learn [44]) for all the sets are statistically higher (>0.88) than that of a random classifier model (ROC AUC score = 0.5).

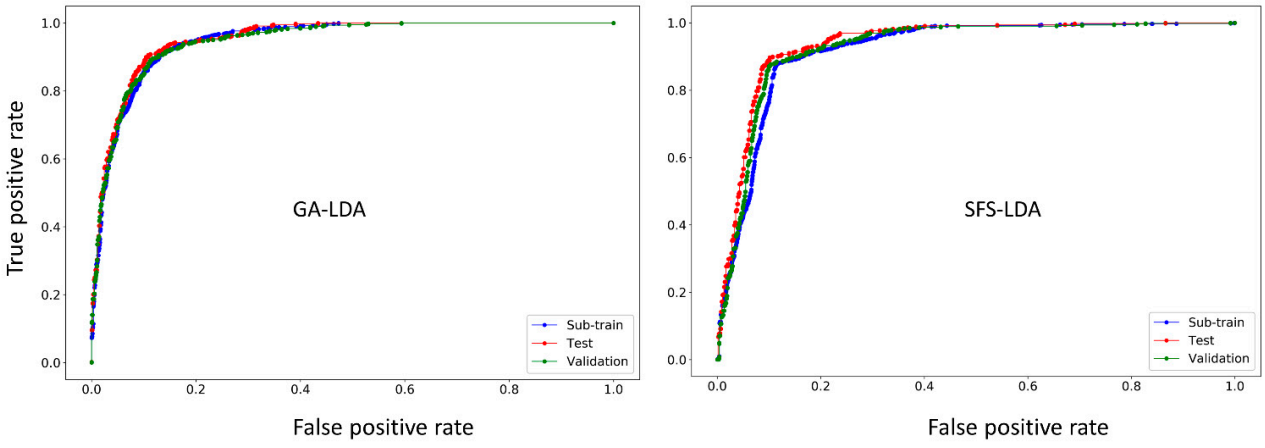

Figure 2. ROC curves for the two best linear models (GA-LDA and SFS-LDA).

To sum up, it can be inferred that the GA-LDA is the best linear model considering its goodness-of-fit as well as internal and external predictive ability. Moreover, when the sub-training set of GA-LDA model was subjected to a 10-fold cross-validation, an accuracy of 87.99% and MCC value of 0.752 were obtained indicating high internal predictivity of this model.

Yet, to establish the overall reliability of any mt-QSAR model, one should also access its applicability domain (AD). When the GA-LDA model was examined by means of the standardization-based AD approach [45], 64 data-points of the sub-training set, 20 data-points of test set and 49 data-points of the validation set are found to be possible structural outliers, meaning that for those predictions might not be reliable.

Further approval of this GA-LDA classification model should only be carried out after verifying if a consensus modeling approach might not yield a new model with higher predictivity. To accomplish this, the predicted response variables from three different LDA models were collected and the outcomes occurring more frequently were regarded as the consensus predicted activities. By comparing results from both models (Table 3), it was observed that the predictive power of the consensus LDA model is particularly similar but not higher than that of the original GA-LDA model. Therefore, the GA-LDA model is to be considered as the best linear interpretable model for the current dataset. However, considering the performance of the other LDA models, one can reach to the conclusion that the feature selection algorithms used here may simultaneously be applied for future mt-QSAR modeling of other datasets.

Table 3. Predictive ability by different LDA models.

Classification [a]	GA-LDA Model			Consensus LDA Model		
	Sub-Training	**Test**	**Validation**	**Sub-Training**	**Test**	**Validation**
ND_{Total}	2707	1160	1656	2707	1160	1656
ND_{active}	1027	459	620	1027	459	620
CCD_{active}	916	413	553	902	406	541
Sensitivity (%)	89.19	89.98	88.61	87.83	88.45	87.26
$ND_{inactive}$	1680	701	1036	1680	701	1036
$CCD_{inactive}$	1472	626	918	1486	633	927
Specificity (%)	87.62	89.30	89.19	88.45	90.30	89.48
F-measure	0.852	0.872	0.857	0.850	0.870	0.852
Accuracy (%)	88.21	89.57	88.83	88.22	89.57	88.65
MCC	0.756	0.785	0.767	0.754	0.783	0.760

[a] LDA classification statistical parameters (For information see footnotes on Table 1).

2.3. Interpretation of Molecular Descriptors

Undeniably, one of the primary advantages of linear models is the possibility of identifying the most crucial structural and physicochemical factors responsible for the higher activity of the compounds [46]. A description of all the descriptors appearing in the three linear models is given in Table 4. However, considering the higher statistical quality of both the GA-LDA and SFS-LDA models, our discussion will focus mainly on the descriptors appearing in these models. Since the relative contributions of the descriptors can only be understood by analyzing the absolute values of their standardized coefficients, these are shown in Figure 3 for the GA-LDA and SFS-LDA models.

Table 4. Deviation descriptors of the mt-QSAR LDA models and their respective definitions.

Model	Deviation Descriptors	c_j	Core Descriptor	Description	Descriptor Type [a]
GA-LDA	$\Delta[\mathrm{VE\_1Dz(Z)}]_{b_t}$	biological target	VE1_Dz(Z)	Coefficient sum of the last eigenvector (absolute values) from Barysz matrix weighted by atomic number	2D matrix-based
	$\Delta(\mathrm{Mor32m})_{b_t}$	biological target	Mor32m	Signal 32/weighted by mass	3D-MoRSE
	$\Delta(\mathrm{L2m})_{b_t}$	biological target	L2m	2nd component size/weighted by mass	WHIM directional
	$\Delta(\mathrm{C-032})_{m_e}$	measure of effect	C−032	X–CX–X	Atom-centered fragments
	$\Delta(\mathrm{F02[N-O]})_{m_e}$	measure of effect	F02[N−O]	Frequency of N−O at topological distance 2	2D Atom Pairs
	$\Delta(\mathrm{Wi\_D/Dt})_{m_e}$	measure of effect	Wi_D/Dt	Wiener-like index from distance/detour matrix	2D matrix-based
	$\Delta(\mathrm{nRNH2})_{a_t}$	assay type	nRNH2	Number of primary amines (aliphatic)	Functional group counts
	$\Delta(\mathrm{nArNHR})_{a_t}$	assay type	nArNHR	Number of secondary amines (aromatic)	Functional group counts
	$\Delta(\mathrm{Mor27}m)_{a_t}$	assay type	Mor27m	Signal 27/weighted by mass	3D-MoRSE
	$\Delta(\mathrm{Mor21u})_{a_t}$	assay type	Mor21u	Signal 21/unweighted	3D-MoRSE
FS-LDA	$\Delta(\mathrm{D/Dtr05})_{b_t}$	biological target	D/Dtr05	Distance/detour ring index of order 5	Ring
	$\Delta(\mathrm{C-030})_{b_t}$	biological target	C−030	X–CH–X	Atom-centered fragments
	$\Delta(\mathrm{nCt})_{b_t}$	biological target	nCt	Number of total tertiary C(sp3)	Functional group counts
	$\Delta(\mathrm{L2m})_{b_t}$	biological target	L2m	2nd component size/weighted by mass	WHIM directional
	$\Delta(\mathrm{CATS3D\_18\_DL})_{b_t}$	biological target	CATS3D_18_DL	Donor-Lipophilic BIN 18 (18–19 Å)	3D-CATS
	$\Delta(\mathrm{CATS3D\_10\_PL})_{b_t}$	biological target	CATS3D_10_PL	Positive-Lipophilic BIN 10 (10–11 Å)	3D-CATS
	$\Delta(\mathrm{nPyridines})_{m_e}$	measure of effect	nPyridines	Number of Pyridines	Functional group counts
	$\Delta[\mathrm{T(N..O)}]_{m_e}$	measure of effect	T(N..O)	Sum of topological distances between N..O	2D Atom Pairs
	$\Delta(\mathrm{CATS3D\_07\_DA})_{m_e}$	measure of effect	CATS3D_07_DA	Donor-Acceptor BIN 7 (7–8 Å)	3D-CATS
	$\Delta(\mathrm{nRNH2})_{a_t}$	assay type	nRNH2	Number of primary amines (aliphatic)	Functional group counts
SFS-LDA	$\Delta(\mathrm{F08[N-S]})_{b_t}$	biological target	F08[N−S]	Frequency of N−S at topological distance 8	2D Atom Pairs
	$\Delta(\mathrm{B03[S-Br]})_{b_t}$	biological target	B03[S−Br]	Presence/absence of S−Br at topological distance 3	2D Atom Pairs
	$\Delta(\mathrm{Mor31u})_{b_t}$	biological target	Mor31u	Signal 31/unweighted	3D-MoRSE
	$\Delta(\mathrm{CATS2D\_02\_DD})_{b_t}$	biological target	CATS2D_02_DD	Donor-Donor at lag 2	2D-CATS
	$\Delta(\mathrm{H-052})_{m_e}$	measure of effect	H−052	H attached to C0(sp3) with 1X attached to next C	Atom-centered fragments
	$\Delta(\mathrm{nRNH2})_{m_e}$	measure of effect	nRNH2	Number of primary amines (aliphatic)	Functional group counts
	$\Delta(\mathrm{CATS2D\_02\_DD})_{b_t}$	assay type	T(N..N)	Sum of topological distances between N..N	2D Atom Pairs
	$\Delta(\mathrm{F07[N-Cl]})_{a_t}$	assay type	F07[N−Cl]	Frequency of N-Cl at topological distance 7	2D Atom Pairs
	$\Delta(\mathrm{SsNH2})_{a_t}$	assay type	SsNH2	Sum of sNH2 E-states	Atom-type E-state indices
	$\Delta(\mathrm{CATS2D\_06\_DD})_{a_t}$	assay type	CATS2D_06_DD	Donor-Donor at lag 6	2D-CATS

[a] MoRSE: Molecular Representation of Structures based on Electronic diffraction; WHIM: Weighted Holistic Invariant Molecular [47]; CATS: Chemically Advanced Template Search [48].

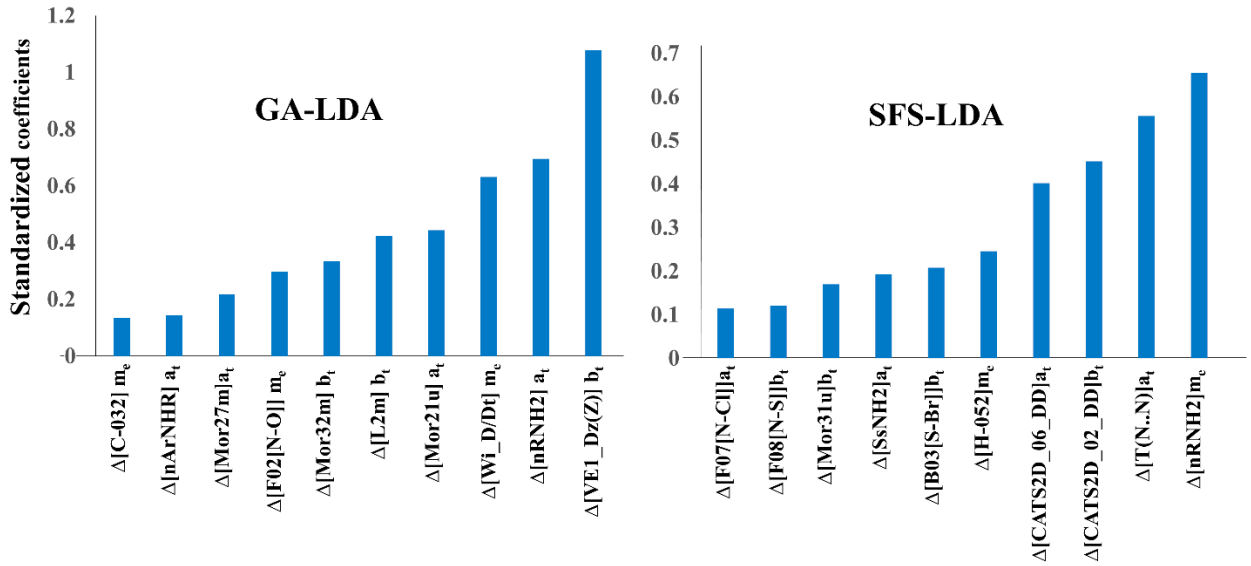

Figure 3. Absolute standardized coefficients vs. variables in the mt-QSAR models.

Firstly, it is noteworthy that all the experimental elements considered in this work (i.e., a_t, m_e, and b_t) consistently appeared in the final LDA models demonstrating their importance (Table 4). To set up the models, 30 distinct categories of descriptors (available in alvaDesc [38]) were considered but only 10 persisted. The latter pertained more frequently to 2D atom pairs (pairs of atoms at a given topological distance) and functional group counts descriptors, as well as to atom-centered fragments [49]. Importantly, these three categories of descriptors are easily interpretable. Nevertheless, 3D descriptors, such as 3D-Morse (3D-Molecular Representation of Structures based on Electronic diffraction) [50], WHIM (Weighted Holistic Invariant Molecular) [47], and CATS3D (Chemically Advanced Template Search 3D) [48], also appeared in the final models.

As can be seen in Figure 3, the three most important descriptors of the GA-LDA model pertain to two topological indices (i.e., the graph-based descriptors $\Delta[\text{VE\_1Dz(Z)}]_{b_t}$ and $\Delta(\text{Wi\_D/Dt})_{m_e}$) and one constitutional descriptor ($\Delta(\text{nRNH2})_{a_t}$: counts of the number of primary amines). Other contributing descriptors of the model are geometrical descriptors either weighted by atomic masses or unweighted, these obviously containing information about the whole 3D-molecular structure of the compounds [47,49]. Interestingly, similar to the number of aliphatic amines, the number of aromatic amines ($\Delta(\text{nArNHR})_{a_t}$) is found also to be important in the GA-LDA model for triggering a higher biological activity, both being dependent on the experimental element assay type (a_t). Except one 3D-MoRSE descriptor, all SFS-LDA descriptors are found as 2D descriptors. For instance, its most significant descriptor is $\Delta(\text{nRNH2})_{m_e}$, which along with the $\Delta(\text{nRNH2})_{a_t}$ descriptor of GA-LDA (which also appears in the FS-LDA model), reiterates the importance of aliphatic primary amines for achieving high activity against the AKT enzyme isoforms. Other important descriptors in this model are the frequency of atom pairs at particular topological distances, e.g., between two nitrogen atoms or sulfur and bromine atoms of the compounds [49]. Two mentions also are the two CATS2D descriptors [48] of the SFS-LDA model, i.e., descriptors $\Delta(\text{CATS2D\_02\_DD})_{b_t}$ and $\Delta(\text{CATS2D\_06\_DD})_{a_t}$, which embody a potential pharmacophore point on pairs of atoms. Both these involve two hydrogen bond donor groups (DD) in the compounds located at different topological distances.

Overall, it may thus be concluded that the presence as well as topological orientation of primary amines, amides and hydrogen bond donor groups in the molecules play crucial roles in promoting activity against the AKT isoforms. The topological distance between two nitrogen atoms, nitrogen and chlorine atoms, nitrogen and oxygen atoms, as well as sulfur and bromine atoms in the compounds appear to be the key factors for a higher AKT inhibition. At the same time, the importance of carbon atoms (e.g., $\Delta(\text{C}-032)_{m_e}$ and

$\Delta(H-052)_{m_e}$), topological graph-based descriptors and geometrical descriptors (particularly 3D-MoRSE ones) is revealed. All these descriptors are found to be significant in ascertaining the biological activity of the compounds against AKT enzymes.

2.4. Non-Linear Predictive Mt-QSAR Models

Herein, descriptors with variances less than 0.001 and inter-correlations higher than 0.95 were removed, leaving 3342 deviation descriptors for setting up various non-linear mt-QSAR models, based on the ML tools: Xgboost [51], (RF) [52], *k*NN [53], RBF-SVC [54], MLP [55], DT [56], and NB [57]. Since the statistical quality of the resultant non-linear models substantially depends on the ML parameter settings [58], the latter were optimized by hyperparameter tuning [59]. A 10-fold cross-validation (CV) grid search scheme was employed to optimize the parameters and to achieve the best possible model based upon the sub-training set.

Table 5 shows the parameter values that were optimized through hyperparameter tuning as well as the final optimized values of these parameters for each of the machine learning tools applied. Furthermore, Table 5 also depicts the 10-fold CV accuracies obtained when the machine learning tools with optimized parameter values were fitted with the sub-training set.

Table 5. Hyper-parameters values explored in the machine learning tools applied for development of the non-linear mt-QSAR models.

Method	Parameters Tuned	Parameters Selected	10-Fold CV Accuracy (%) [a]
RF	Bootstrap: True/False Criterion: Gini, Entropy, Maximum depth: 10, 30, 50, 70, 90, 100, None Maximum features: Auto, Sqrt Minimum samples leaf: 1, 2, 4 Minimum samples split: 2, 5, 10 Number of estimators: 50, 100, 200,500	False Gini 90 Sqrt 1 5 200	91.02
k*NN*	Number of neighbors: 1–31 Weight options: Uniform, Distance Algorithms: Auto, Ball tree, kd_tree, brute	20 Distance Auto	79.20
Xgboost	Minimum child weight: 1,5,10 Gamma: 0, 0.5, 1, 1.5, 2, 5 Sum sample: 0.6, 0.8, 1.0 Number of estimators: 50, 100, 200,300 Maximum depth: 3, 4, 5	1 0 0.8 100 5	91.54
RBF-SVC	C: 0.1, 1, 10, 100, 1000 Gamma: 1, 0.1, 0.01, 0.001	1 1	62.30
MLP	Hidden layer sizes:(50,50,50), (50,100,50), (100,) Activation: Identity, Logistic, Tanh, Relu Solver: SGD, Adam Alpha: 0.0001, 0.001, 0.01,1 Learning rate: Constant, Adaptive, Inverse scaling	(100,) Relu Adam 0.0001 Adaptive	82.97
DT	Criterion: Gini, EntropyMaximum depth: 10,30,50,70,90,100, NoneMaximum features: Auto, SqrtMinimum samples leaf: 1,2,4Minimum samples split: 2–50	Entropy100Sqrt113	84.33
NB	Alpha: 1,0.5,0.1Fit prior: True, False	0.1True	69.40

[a] The cross-validation accuracy was estimated only on the sub-training set.

It is evident that the classification ability of these machine learning tools varies to a considerable extent, but those achieving the better ones are the DT, RF and Xgboost tools. In fact, the MLP model demonstrates moderate internal predictivity, and the classifications produced by *k*NN, SVC and NB suggest an extremely poor internal predictivity of these tools.

The RF and Xgboost tools produced a 10-fold CV accuracy on the sub-training set of 91.02% and 91.54%, respectively, indicating that both resultant models display high and almost similar internal predictivity. These two models were then chosen to further investigate their external predictivity on the test and validation sets. Table 6 shows the overall predictivity of the RF and Xgboost models, whereas their ROC plots are displayed in Figure S1 of the Supplementary Materials.

Table 6. Overall predictivity of the RF and Xgboost non-linear models.

Classification [a]	RF			Xgboost		
	Sub-Training (10-CV)	**Test**	**Validation**	**Sub-Training (10-CV)**	**Test**	**Validation**
ND_{Total}	2707	1160	1656	2707	1160	1656
ND_{active}	1027	459	620	1027	459	620
CCD_{active}	919	417	573	932	422	578
Sensitivity (%)	89.48	92.58	93.53	90.75	91.87	93.24
$ND_{inactive}$	1680	701	1036	1680	701	1036
$CCD_{inactive}$	1545	649	969	1546	644	966
Specificity (%)	91.96	90.85	92.42	92.02	91.94	93.23
F-measure	0.883	0.899	0.909	0.891	0.900	0.912
Accuracy (%)	91.02	91.90	93.11	91.54	91.90	93.24
MCC	0.810	0.831	0.854	0.822	0.832	0.857

[a] Non-linear classification statistical parameters (For information see footnotes on Table 1).

As expected, the internal and external predictivities of the RF and Xgboost based non-linear models are noticeably higher than those of the linear models (see Tables 2 and 3). On the basis of their overall predictivity (cf. accuracy values and MCC scores for all sets in Table 6), the Xgboost model stands for the best non-linear model.

One aspect that warrants explicit attention is to inspect how the two best-derived models (i.e., GA-LDA and Xgboost) manage to classify the datapoints pertaining to different experimental elements c_j (Table 7). In general, the datasets applied in mt-QSAR computational modeling encompass a large variation in the number of data-points vis-à-vis the various experimental elements. As expected, the same situation happens in the current dataset. Still, the non-linear Xgboost model is unaffected by that since it affords high accuracies irrespectively of the experimental element or validation set. The GA-LDA model, with less overall predictivity than the Xgboost model, shows also high accuracies in case of the test set. Nevertheless, it reaches low accuracy values against some experimental conditions (e.g., for c_j = 4 and 7). Yet, if both these models are considered simultaneously, there is apparently a greater chance of finding more accurate predictions.

Table 7. The predictive accuracies of GA-LDA and Xgboost models with respect to the different experimental elements c_j.

c_j	m_e	a_t	b_t	Test Set					External Validation Set				
				N_{Sample} [a]	Xgboost		GA-LDA		N_{Sample} [a]	Xgboost		GA-LDA	
					# Incorrect [b]	% Accuracy	# Incorrect [b]	% Accuracy		# Incorrect [b]	% Accuracy	# Incorrect [b]	% Accuracy
1	**IC50**	**B**	**AKT**	**509**	80	84.28	95	81.34	707	95	86.56	142	79.92
9	Ki	F	AKT2	163	2	98.77	2	98.77	238	0	100.00	5	97.90
10	Ki	F	AKT3	148	1	99.32	3	97.97	236	2	99.15	2	99.15
8	Ki	F	AKT	156	3	98.08	2	98.72	203	1	99.51	2	99.01
2	IC50	B	AKT2	121	2	98.35	14	88.43	167	8	95.21	23	86.23
3	IC50	B	AKT3	35	4	88.57	4	88.57	53	2	96.23	3	94.34
5	Ki	B	AKT	17	2	88.24	1	94.12	27	2	92.59	3	88.89
6	Ki	B	AKT2	4	0	100.00	0	100.00	12	1	91.67	1	91.67
4	IC50	F	AKT	4	0	100.00	0	100.00	9	1	88.89	3	66.67
7	Ki	B	AKT3	3	0	100.00	0	100.00	4	0	100.00	1	75.00

[a] Number of data samples. [b] Number of data samples (#) incorrectly predicted by the model.

2.5. Virtual Screening

In order to describe how the developed mt-QSAR models perform on identifying virtual hits, we employed the GA-LDA and Xgboost models for screening the focused library ***Asinex kinase*** (http://www.asinex.com/focus_kinases/, accessed on 17 August 2020), which comprises 6538 compounds. Details about this dataset can be found in Supplementary Materials (SM2.xlsx). Similarly, the descriptors of all such compounds were calculated by the alvaDesc tool [38]. In the modeling dataset used here, we found 10 unique experimental elements c_j depending on the m_e, a_t and b_t conditions (cf. Table 7). Each compound of the Asinex kinase inhibitor library was also assigned to those conditions, and a virtual dataset containing 65,380 samples was then prepared. Afterwards, the deviation descriptors of each of these samples were calculated using the QSAR-Co tool [22]. Initially the GA-LDA model was used for screening this virtual dataset, and 28 compounds were predicted as active ($IA_i(c_j)$ = +1) against at least 7 out of 10 experimental elements. The predictivity of these 28 compounds was then tested with the best non-linear model (i.e., the Xgboost model) and seven compounds were predicted to be active against at least 6 experimental elements c_j. These seven compounds (Figure 4) are thus to be considered as the most potential virtual hits according to the current computational multi-target modeling. Note that, for screening the dataset, we mainly relied on the linear model since it is likely to be less susceptible to overfitting than the non-linear model.

Table 8 displays the number of experimental conditions pertaining to each virtual hit picked by the GA-LDA and Xgboost models, whereas details of these experimental conditions are provided in Supplementary Materials (Table S4). It is observed that each of these virtual hits are predicted to be active against all three AKT isoforms in one or more experimental assay conditions (defined by the combinations of a_t and b_t).

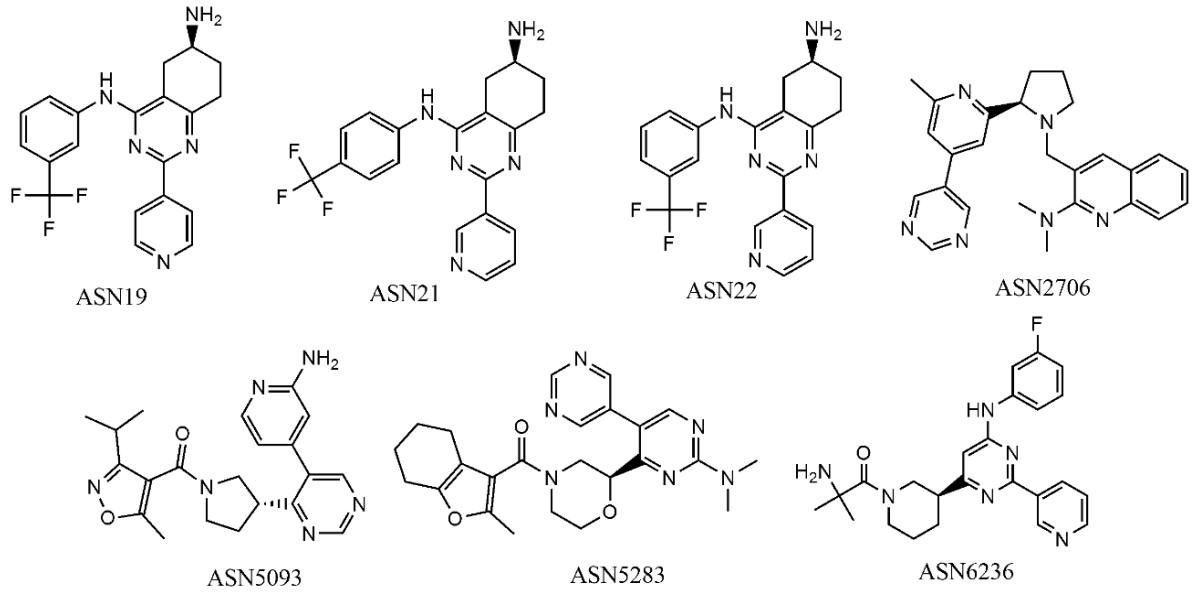

Figure 4. Structures of the most potential virtual hits obtained with the GA-LDA and Xgboost models.

Table 8. Number of experimental conditions pertaining to the most potential virtual hits gathered by the GA-LDA and Xgboost models.

Compound	No of Experimental Conditions (GA-LDA)	No of Experimental Conditions (Xgboost)
Asn0019	10	6
Asn0021	10	6
Asn0022	10	6
Asn2706	7	6
Asn5093	8	6
Asn5283	7	6
Asn6236	7	6

2.6. Pharmacophore Based Biological Target Identification

After identifying seven virtual hits from multi-target computational-guided modeling, we have decided to implement a filtering scheme based on reverse pharmacophore mapping. The latter will allow pre-filtering such hits that best fulfil simple geometric and chemical functionality requirements to trigger a biological response, before embarking on more computationally demanding approaches. For this, each of these virtual hits were investigated with the PharmMapper webserver (http://www.lilab-ecust.cn/pharmmapper/index.html, accessed on 4 January 2021)) [60,61] to identify their possible human macromolecular targets. The results of PharmMapeer based predictions are outlined in Table 9. Details about the fitting of the virtual hits with the obtained structure-based pharmacophores are provided in the Supplementary Materials (Figures S2 and S3).

Interestingly, except Asn5283 and Asn2706, all virtual hits are mapped with the structure-based pharmacophores coming from both human AKT1 (PDB ID: 3CQU) and human AKT2 (PDB ID: 2UW9). Note that the PharmMapper predictions are based on the protein structures available in the Protein Data Bank (PDB) and so far, no complete structure of AKT3 is available. Therefore, these results mean that multiple AKT isoforms may be possible biological targets for these five virtual hits, which are Asn0019, Asn0021, Asn0022, Asn5093 and Asn6236. Furthermore, all these five virtual hits are fitted with structure-based pharmacophores generated at the catalytic sites of these enzyme isoforms. We randomly selected five Asinex kinase inhibitor library compounds (names and structures are provided in SM2.xlsx of Supplementary Materials) that were predicted to be inactive in the QSAR based virtual screening and none of these compounds was mapped with structure-based pharmacophores pertaining to both AKT1 and AKT2. Based upon the current pharmacophore analysis, we removed Asn5283 and Asn2706, and decide to further investigate the remaining six virtual hits by other more complicated structure-based analysis techniques.

Table 9. Fitting of the virtual hits to the structure-based pharmacophores predicted by the PharmMapper webserver.

Compound	PDB ID	Target Name	Feature Type [a]	No of Features	Fit Score
Asn0019	3CQU	AKT1	2H,A,D	4	2.925
Asn0019	2UW9	AKT2	3H,P,A,D	6	3.078
Asn0021	3CQU	AKT1	2H,A,D	4	2.274
Asn0021	2UW9	AKT2	3H,P,A,D	6	3.385
Asn0022	3CQU	AKT1	2H,A,D	4	2.724
Asn0022	2UW9	AKT2	3H,P,A,D	6	3.000
Asn5093	3CQU	AKT1	2H,A,D	4	2.637
Asn5093	2UW9	AKT2	3H,P,A,D	6	3.135
Asn6236	3CQU	AKT1	2H,A,D	4	3.126
Asn6236	2WU9	AKT2	3H,P,A,D	6	2.955

[a] A: Acceptor, D: Donor, H: Hydrophobic, P: Positive ionizable.

2.7. Structure-Based Prediction of the Virtual Hits

Even though pharmacophore mapping insinuated that selected virtual hits may interact with multiple AKT enzyme isoforms, considering high flexibilities of kinase enzymes, it may not be able to predict the stabilities ligand-receptor complexes. Therefore, we decided to go a step forward and carry out molecular dynamics (MD) simulations to understand the dynamic behavior of the virtual hits within the AKT enzyme isoforms. The five virtual hits predicted by both the mt-QSAR modeling and the pharmacophore-based mapping (PharmMapper) were initially docked into the isoforms AKT1 (PDB ID:4GV1) and AKT2 (PDB ID: 1O6K), as well as into an AKT3 homology model. For this, two molecular docking tools were employed, namely: (a) Autodock Vina [62] and (b) Autodock 4.2 [63]. The homology modeling of AKT3 was carried out using the SWISS-MODEL webserver (swiss-model.expasy.org) [64,65]. Details about the procedures of homology modeling used as well as the validation of the model are described in the Materials and Methods section.

Since multiple binding sites exist in these AKT enzyme isoforms, we initially conducted a blind docking experiment for the virtual hits with the help of Autodock Vina. To do so, a grid box of 100 Å × 100 Å × 100 Å dimensions was centered on each of such macromolecules. The blind docking (performed with an exhaustiveness of 45) indicated that all the virtual hits may preferably bind to the catalytic domain of the enzyme isoforms. Thereafter, an Autodock based docking was performed for all the virtual hits by setting a grid box of 50 Å × 50 Å × 50 Å dimensions, with a grid spacing of 0.375 Å, centered on the catalytic residue of Asp292, Asp293, and Asp289 for AKT1, AKT2, and AKT3, respectively. The pan-AKT inhibitor GSK690693 was employed as a reference compound in this Autodock docking experiment. The results of both docking experiments performed with Autodock Vina and Autodock 4.2 are provided in Table S5 of Supplementary Materials.

These MD simulations were carried out with docked (Autodock-based) complexes of these compounds for 50 ns. The docked complexes of some selected virtual hits (i.e., the starting ligand-protein complexes used in the MD simulations) are provided in the Supplementary Materials. The root-mean-square-deviation (RMSD) plots of the backbone atoms of the receptor-ligand complexes, ligand RMSD plots, the root-mean-square-fluctuation (RMSF) plots of these protein structures, and the radius of gyration' plots are presented in the Supplementary Materials (Figures S4–S6). A close look at these plots reveals an adequate dynamic stability as well as compactness of the ligand-macromolecule complexes. However, our major goal was to estimate the binding free energies of these ligand-protein complexes, which was computed by the molecular mechanics-generalized Born surface area (MM-GBSA) approach.

As observed from the results in Table 10, Asn0019 depicts maximum theoretical binding energy values against three different enzyme isoforms of AKT among five virtual hits. The theoretical binding free energy values of Asn0019 against AKT2 and AKT3 enzymes were comparable to those of the reference compound GSK690693. Even though Asn5093 depicted slightly increased theoretical ΔG_{bind} values against AKT1 and AKT2 as compared to Asn0019 (as well as Asn0021 and Asn0022), its binding energy was found to be reduced against AKT3 enzyme. The Asn0019 may be projected as the most promising virtual hits based on its average theoretical ΔG_{bind} value.

Table 10. Calculated binding free energies (ΔG_{bind} in kcal/mol) for the AKT1, ATK2, and ATK3 bound ligands.

Compound	AKT1	AKT2	AKT3
Asn0019	−32.54	−27.61	−43.61
Asn0021	−25.81	−29.03	−39.04
Asn0022	−27.75	−23.56	−37.29
Asn5093	−36.34	−29.44	−22.67
Asn6236	−18.08	−19.72	−26.82
GSK690693	−46.88	−29.78	−43.17

Per residue decomposition analysis of Asn0019 was performed to understand the binding interactions of this virtual hit against X-ray crystal structures AKT isoforms (i.e., AKT1 and AKT2). Total energy values obtained from important binding site residues are depicted in Figure 5. It is observed that interactions of Asn0019 with Leu156/158, Val164/166, Ala177/179, Lys179/181, Tyr229/231, Ala230/232, Glu278/279, Asn279/280, Met281/282, Thr291/292 and Phe438/439 of AKT1 and AKT2 enzymes may play crucial roles in determining the binding affinities of this molecule in these proteins.

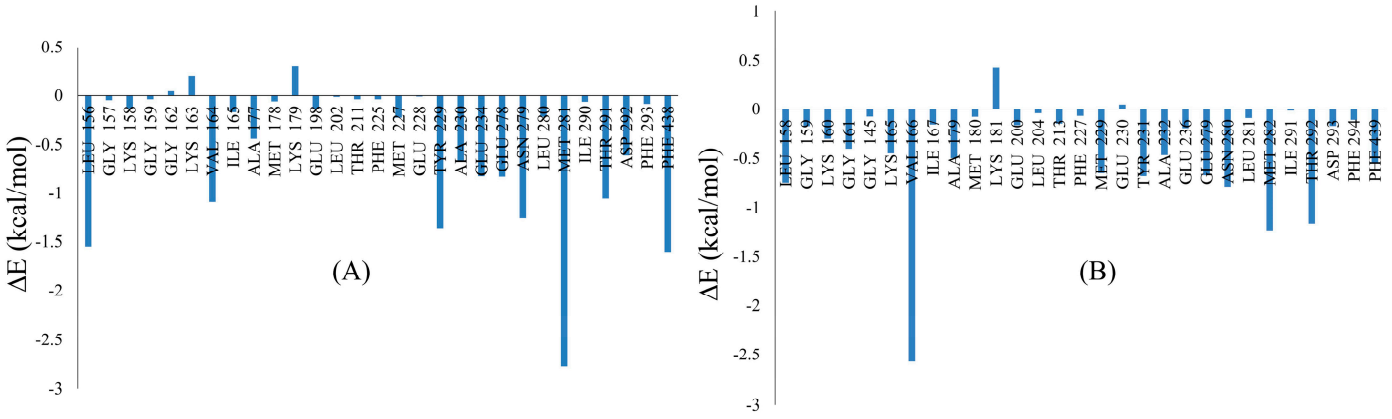

Figure 5. Total energy obtained from per-residue decomposition analysis of Asn0019 in (**A**) AKT1 and (**B**) AKT2 enzymes.

3. Materials and Methods

3.1. Descriptor Calculation

The SMILES structures of the compounds were converted to 2D structures using the MarvinView software (https://docs.chemaxon.com/display/docs/MarvinView, accessed on 6 July 2020). Such structures were subsequently standardized by the ChemAxon Standardizer tool using the following options: strip salts, aromatize, clean 3D, tautomerize, neutralize and add explicit hydrogens [66]. The initial descriptors were calculated by alvaDesc tool (https://www.alvascience.com/alvadesc/, accessed on 6 July 2020) [38] in the OCHEM web-platform [67]. Geometry optimizations of the compounds were performed with the Corina software [68] under OCHEM [67]. Calculation of the deviation descriptors ($\Delta(D_i)c_j$) from the initial descriptors (D_i) and experimental elements (c_j) were carried out using the QSAR-Co tool [22].

3.2. Development of Linear Interpretable Models

The linear models were based on linear discriminant analysis (LDA) models and three different feature selection algorithms, i.e., (a) genetic algorithm (GA), (b) forward stepwise (FS), and (c) sequential forward selection (SFS). For the first one, the resultant GA-LDA model was set up with the help of the QSAR-Co tool [22]. Details of the GA selection procedure have been extensively reported in the past [69]. Briefly, GA is based on the evolution of a population of randomly generated models. Firstly, parent models are chosen, and these are then subjected to random "cross-over" and "mutation" processes to produce child models, which are then used to check their fitness scores. The new generated models with the highest fitness scores are then forwarded to a next iteration. The algorithm terminates either when the maximum number of allowed population models is reached or when no improvement regarding the fitness score is observed for the subsequent 10 population generations. The parameters employed here for setting up the GA-LDA model with QSAR-Co were: (i) total number of iterations/generations: 100, (ii) equation length: 10 (fixed), (iii) mutation probability: 0.3, (iv) initial number of equations generated: 100, and (v) number of equations selected in each generation: 30. Forward stepwise (FS) is a very popular feature selection algorithm in which the independent descriptors are included in the model stepwise depending on a specific statistical parameter. In this work, the FS-LDA model was set up with a python program where the features are selected and included in the model stepwise by the corresponding *p*-values of the Fischer statistic [69]. Initially, criteria for the forward selection (i.e., *p*-value to enter) as well as for the backward elimination (*p*-value to remove) have to be set. The descriptor with the lowest *p*-value is first included and subsequently other descriptors are included in the model based on their *p*-values only if the criterion for forward selection is met. However, if the *p*-value of a descriptor included in the model is found to be greater than the "*p*-value to remove", it is eliminated from the model. In the current work, both *p*-values to enter and to remove

were fixed at 0.05. The final LDA models were developed and were subsequently validated using the *LinearDiscriminantAnalysis* function of Scikit-learn [44]. The python code used for FS-LDA model development is provided in the Supplementary Materials (file SM1.py). The last feature selection algorithm is based on a sequential forward selection which adds features into an empty set until the performance of the model is not improved either by addition of another feature or by reaching the maximum allowed number of features [70]. Similar to FS, this procedure is also a greedy search algorithm where the best subsets of descriptors are selected stepwise and the model performance is justified by the user-specific statistical measure. In this work, the python based SequentialFeatureSelector algorithm of the "mlxtend" library (http://rasbt.github.io/mlxtend/, accessed on 27 July 2020) was applied to setup the resultant SFS-LDA model. The parameters used for that purpose were: (i) maximum number of features: 10, (ii) forward: True, (iii) floating: True, (iv) scoring: accuracy, and (v) cross-validation (cv) = 0. Thereby, accuracy was employed as the user-specific statistical measure for feature selection, in contrast to FS where the *p*-value was used, and no cross-validation was performed during feature selection.

Several diagnostic statistical indices were employed for assessing our model equations, in terms of the criteria goodness-of-fit and goodness-of-prediction. Measures of goodness-of-fit of the LDA models based on the sub-training set were estimated by standard indices such as the Wilks' lambda (λ) [71], chi-squared (χ^2), the square of Mahalanobis distance (D^2), the Fisher's statistic index (*F*), and the corresponding *p*-value (*p*) [41]. Measures of goodness-of-prediction for both linear and non-linear models were estimated by computing the following statistical measures for the sub-training, test and external validation sets: sensitivity—correct classification of the active cases, specificity—correct classification of inactive cases, accuracy—overall correct classification, *F*-measure, and Matthews correlation coefficient (MCC) [41,42].

3.3. Non-Linear Model Development

Seven different machine learning techniques were used for setting up the non-linear models, namely: (a) gradient boost classifier (Xgboost) [51], (b) Random Forests (RF) [52], (c) *k*-Nearest Neighbors (*k*NN) [53], (d) Radial Basis Function based Support Vector Classifier (RBF-SVC) [54], (e) Multilayer Perceptron (MLP) neural-networks [55], (f) Decision Trees (DT) [56], and (g) Bernoulli Naïve Bayes (NB) [57]. The python 2 based Xgboost (version 1.0.0) algorithm (https://xgboost.readthedocs.io/en/latest/, accessed on 26 October 2020) was used for developing the Xgboost models. Regarding the RF ML, that was implemented with the help of the QSAR-Co tool [22], using the following parameters: (i) each bag size: 100, (ii) maximum depth: 0 (unlimited), (iii) number of randomly chosen features: 0 [i.e., n = int(log2(#Predictors) + 1)], and (iv) number of iterations: 100.

All other machine learning techniques were applied by resorting to the respective Scikit-learn machine learning packages [44]. The model development parameters for each of these techniques were determined by hyperparameter tuning as implemented in GridSearchCV of Scikit-learn [58]. During hyperparameter optimization, a 10-fold cross-validation (10-fold CV) was performed with the sub-training set to identify the best model estimators. The 10-fold CV accuracies obtained from the different machine learning techniques were compared to find the highly predictive classifiers. Finally, the external predictivity of these classifiers were estimated with the test and the validation sets.

3.4. PharmMapper Based Prediction of Biological Targets

The freely-accessed PharmMapper (version 2017) web server (http://www.lilab-ecust.cn/pharmmapper/, accessed on 4 January 2021) searches for the best mapping poses of the given molecules against structure-based pharmacophore models generated with all targets of PharmTargetDB [60,61]. PharmMapper applies a large database of receptor-based pharmacophores (>7000 pharmacophore models based on 1627 drug targets, 459 of which are human protein targets.) to find possible macromolecular targets for the input ligands. It implements Cavity (version 1.1) program in order to identify the binding sites on the

surface of a protein structure and to rank these subsequently as per their druggability scores. The receptor-based pharmacophore modelling was then performed using Pocket (version 4.0) program for extracting the pharmacophore features within these druggable cavities.

In this work, the virtual hits obtained from multi-target QSAR computational modeling were subjected to PharmMapper based predictions of biological targets. A maximum number of 1000 conformers were generated during the search of pharmacophore fitting on human biological targets and the top 200 targets based on the fit scores were analyzed.

3.5. Homology Modeling

The homology model of AKT3 was set up with the help of the SWISS-MODEL webserver [64,65]. The FASTA sequence of human AKT3 was retrieved from Uniprot (Uniprot ID: Q9Y243) and we used the sequence 141–479 for homology modeling. The FASTA sequence was uploaded to the SWISS-MODEL server to search for templates and then, ranging from 50 templates to 10 templates were selected for generating the model on the basis of higher GMQE values and also on their diversity. Each of the homology models was analyzed by the 'structural assessment' option available in this webserver. Among these models, one model developed with the template of 6CCY.1.A (sequence identity: 83.18%) was found to have the lowest MolProbity value (=1.02), calculated from http://molprobity.biochem.duke.edu/, accessed on 11 January 2021). This model, which showed a Qualitative Model Energy Analysis (QMEAN) value of −1.84, was selected for further structural modifications. From the low MolProbity value, it is ensured that the homology model is of good structural quality, however, the model showed some complications mainly due to its Clashscore (all atoms) of 0.37, the presence of 5 Ramachandran outliers, as well as the presence of 26 bad angles. Therefore, the UCSF Chimera software [72] was utilized for structural modifications of this homology model. The latter included step-by-step loop refinement, rotamer adjustment and structural minimization of the selected residues. Then, for each modification step, the quality of the model was checked in the Molprobilty webserver (http://molprobity.biochem.duke.edu/, accessed on 11 January 2021) [64]. Ultimately, the modified homology model was found to have an even lower Molprobilty score of 0.87, indicating that the initially developed model was clear structurally improved. In this final model, the Clashscore (all atoms) reduced to zero and the number of outliers to one, as well as the number of bad angles lowered to two. The final model also depicted an improved Global Model Quality Estimation (GMQE) value of −1.61. Alignment of Q9Y243 with 6CCY.1.A, validation parameters as well as Ramachandran plots of the final homology model are shown in the Supplementary Materials (Figures S7 and S8).

3.6. Molecular Docking

The docking experiments with the X-ray crystal structures and homology model were performed with Autodock Vina (version 1.1.2., The Scripps Research Institute, La Jolla, CA, USA) [62] and Autodock 4.2 (The Scripps Research Institute, La Jolla, CA, USA) [63]. During preparation of the macromolecules, water molecules and the peptide substrate obtained from the PDB were removed. Other necessary details about the methodology followed on such docking experiments are the same as described earlier [20].

3.7. Molecular Dynamics Simulations

All MD simulations were carried out using the software package AMBER 12 [73,74]. Initially, the protonation states of the amino acid residues of each protein were fixed at pH = 7.0. These protonation states were attained by the PDB2PQR server (http://server.poissonboltzmann.org/pdb2pqr, accessed on 12 January 2021), using the AMBER forcefield and output naming scheme [75]. The ff99SB and general AMBER forcefield (GAFF) were applied for describing the protein-inhibitor and inhibitor-solvent interactions, respectively [76,77]. The optimization of ligands was carried out with the help of Leap in Antechamber, by performing MD simulations of the hydrated complexes centered in

a cubic box of edge length of 8 Å, and applying the forcefields GAFF, ff99SB, and TIP3P for water molecules [78]. Subsequently, the negative charges of the complex systems were neutralized. The SHAKE algorithm was used to constrain all bonds related to hydrogen atoms, whereas the Partial Mesh Ewald (PME) method was employed to handle long range electrostatic forces (using a cutoff of 12 Å). Energy minimization of the complexes was performed in two stages. In the first stage, only the ions and water molecules were allowed to relax during 1000 steps of the steepest descent method and during 1000 steps of the conjugate gradient algorithm using a restrained force of 500 kcal/mol on the solute. In the second stage, the whole system was relaxed during 5000 minimization steps, i.e., 2500 steps of steepest decent minimization followed by 2500 of conjugated gradient. The minimized systems were then gradually heated up from 0 to 300 K (50 K in each step) with a weak harmonic restraint of 10 kcal/mol to keep the solute fixed for 200 ps. Subsequently, a 2-ns equilibration on the *NpT* ensemble was performed, with pressure kept fixed at 1 bar and temperature at 300 K. Finally, the 50-ns MD simulations without restrictions were run with constant temperature (T = 300 K) and constant pressure (p = 1 bar). After completion of such simulations, various post-dynamic analyses were carried out with PTRAJ and CPPTRAJ implemented in the AMBER package [79]. The graphs were plotted using the QTGRACE tool (https://sourceforge.net/projects/qtgrace/files/, accessed on 1 February 2021). Molecular Mechanics Generalized Born Surface Area (MM-GBSA) based binding free energies of the ligands were calculated using MMPBSA.py program in AMBER [80,81]. One hundred snapshots were taken from the last 10 ns of MD trajectory.

The binding energy calculation is represented as the following equation:

$$\Delta G = \Delta E_{ele} + \Delta E_{vdW} + \Delta G_{pol} + \Delta G_{nonpolar} - T\Delta S \quad (1)$$

ΔE_{ele} and ΔE_{vdW} are electrostatic and van der Waals interactions between the ligands and the proteins (in gas phase), respectively. The polar solvation free energy (ΔG_{pol}) accounts for the polar interactions with the solvent molecules whereas the $\Delta G_{nonpolar}$ term represents non-polar solvation free energy, which is obtained from the equation $\Delta G_{nonpolar} = \gamma SASA + \beta$. The the solvent accessible surface area is represented by the term SASA. The surface tension proportionality constant (γ) and the free energy of nonpolar solvation of a point solute (β), were set to 0.00542 kcal mol^{-1} Å^{-2} and 0 kcal mol^{-1}, respectively. The entropic contribution ($T\Delta S$) is not calculated because apart from being computationally expensive (especially for large macromolecular complexes), it has been reported to be less accurate [80,81].

The energy contributions of the close contact amino acid residues into the total binding free energies were computed using MM-GBSA per residue free energy decomposition method with Amber 12 MM-GBSA module [73,74,82,83]. All energy components (van der Waals, electrostatic, polar solvation, and nonpolar solvation contributions) were calculated using 200 snapshots extracted from the last 10 ns MD trajectories.

4. Conclusions

The latest advances in machine learning tools, coupled with the availability of ever-larger data sets, brought about a fresh wave for faster and less complicated computational-guided drug discovery efforts [84–86]. In this work, we could assemble a large dataset containing 5523 inhibitors of the three AKT isoforms, assayed under a variety of experimental conditions. With the desire to build reliable predictive multi-target QSAR classification models for probing the inhibitory activity from such data, we examined the use of various machine learning tools along with several feature selection algorithms. Considering that machine learning is a powerful mean for finding drug like leads [18], the best linear and non-linear mt-QSAR models were finally used for screening a focused kinase library to identify virtual hits as potential pan-AKT inhibitors. The results obtained were finally post-processed by structure-based approaches.

With regard to the mt-QSAR modeling, the combination of LDA with feature selection algorithms such as FS, SFS, or GA was found to produce classification models exhibiting

very good accuracy (>87%), as well as internal and external predictivity. Nevertheless, the GA feature selection algorithm yielded the best predictive linear model, even though in a not so straightforward and less time-consuming way as the other two algorithms. More significantly, these linear models aided us in understanding the most crucial structural and/or physicochemical properties required for higher AKT inhibition. At the same time, the classification ability of the seven different ML-based mt-QSAR models were found to vary to a considerable extent. The Xgboost technique produced the most predictive non-linear mt-QSAR model (accuracy > 90%), but its overall predictivity was similar to that of the RF model. This leads us to suppose that tree-based modeling techniques are superior to other machine learning ones for multi-target computational modeling. Yet, more investigations are needed to confirm that supposition.

To judge how the best linear and nonlinear mt-QSAR models perform on identifying virtual hits with activity against AKT inhibitors, we used them to screen the Asinex kinase inhibitor library. The obtained virtual hits were further evaluated by structure-based pharmacophore modeling, molecular docking and MD simulations studies. Worth mentioning here that in case of multi-target modeling, structure-based approaches become more problematic when one or more biological targets are not sufficiently characterized. Indeed, this was the case for the AKT3 isoform, the X-ray crystal structure of which is yet to be reported, and thereby a reliable homology model had to be derived. The pharmacophore-mapping target-identification search led to results reinforcing the former mt-QSAR based predictions. Further, the results obtained in the following MD simulations allowed us to put forward Asn0019 as the most potent virtual hit for the inhibition of all AKT isoforms.

To conclude, the information gathered and the derived mt-QSAR computational models provide important guidelines for the discovery of novel AKT inhibitors. What is more, such models are not limited only to pan-inhibitors but can also be applied to identify inhibitors that have selectivity towards one or two AKT isoforms.

Supplementary Materials: The following are available online https://www.mdpi.com/article/10.3390/ijms22083944/s1, Table S1: Degree of collinearity among the variables of the GA-LDA model; Table S2: Degree of collinearity among the variables of the FS-LDA model; Table S3: Degree of collinearity among the variables of the SFS-LDA model; Figure S1: ROC curves for the two best non-linear models (Xgboost–ROC-AUC score (test): 0.919, ROC-AUC score (validation): 0.932, and RF–AUROC: ROC-AUC score (test): 0.917, ROC-AUC score (validation): 0.930); Table S4: Experimental conditions under which the virtual hits were predicted to be active; Table S5: Molecular docking results with binding energy values (in kcal/mol) of the virtual hits; Figure S2: Fitting of virtual hits to the structure-based pharmacophores of AKT1 (PDB: 3CQU) and AKT2 (PDB: 2UW9) enzymes; Figure S3: Fitting of virtual hits to the structure-based pharmacophores of AKT1 (PDB: 3CQU) and AKT2 (PDB: 2UW9) enzymes.; Figure S4: Protein backbone RMSD, ligand RMSD, RMSF and radius of gyration plots for the AKT1 complexes.; Figure S5: Protein backbone RMSD, ligand RMSD, RMSF and radius of gyration plots for the AKT2 complexes; Figure S6: Protein backbone RMSD, ligand RMSD, RMSF and radius of gyration plots for the AKT3 complexes; Figure S7: (A) Alignment of the AKT3 target sequence with potential template; (B) Z-score estimation of the AKT3 homology model; (C) Local QMEAN estimates after manual refinement; (D) AKT3 homology model built using the Swiss-Model server and the 6CCy.1.A template 3D structure; Figure S8: Ramachandran plots for the homology model of AKT3 enzyme. Starting_protein_ligand_complexes.zip: PDB files containing the complexes of virtual hits with the enzymes, used in the MD simulations.

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. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by UID/QUI/50006/2020 with funding from FCT/MCTES through national funds.

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

References

1. Santi, S.A.; Douglas, A.C.; Lee, H. The AKT isoforms, their unique functions and potential as anticancer therapeutic targets. *Biomol. Concepts* **2010**, *1*, 389–401. [CrossRef]
2. Hinz, N.; Jucker, M. Distinct functions of AKT isoforms in breast cancer: A comprehensive review. *Cell Commun. Signal.* **2019**, *17*, 154. [CrossRef]
3. Barile, E.; De, S.K.; Carlson, C.B.; Chen, V.; Knutzen, C.; Riel-Mehan, M.; Yang, L.; Dahl, R.; Chiang, G.; Pellecchia, M. Design, Synthesis, and structure-activity relationships of 3-Ethynyl-1H-indazoles as Inhibitors of the phosphatidylintositol 3-kinase signaling pathway. *J. Med. Chem.* **2010**, *53*, 8368–8375. [CrossRef] [PubMed]
4. Altomare, D.A.; Testa, J.R. Perturbations of the AKT signaling pathway in human cancer. *Oncogene* **2005**, *24*, 7455–7464. [CrossRef]
5. Manning, G.; Whyte, D.B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298*, 1912–1934. [CrossRef]
6. Nitulescu, G.M.; Van De Venter, M.; Nitulescu, G.; Ungurianu, A.; Juzenas, P.; Peng, Q.; Olaru, O.T.; Gradinaru, D.; Tsatsakis, A.; Tsoukalas, D.; et al. The AKT pathway in oncology therapy and beyond (Review). *Int. J. Oncol.* **2018**, *53*, 2319–2331. [CrossRef]
7. Kumar, A.; Rajendran, V.; Sethumadhavan, R.; Purohit, R. AKT kinase pathway: A leading target in cancer research. *Sci. World J.* **2013**, *2013*, 756134. [CrossRef]
8. Dumble, M.; Crouthamel, M.C.; Zhang, S.Y.; Schaber, M.; Levy, D.; Robell, K.; Liu, Q.; Figueroa, D.J.; Minthorn, E.A.; Seefeld, M.A.; et al. Discovery of novel AKT inhibitors with enhanced anti-tumor effects in combination with the MEK inhibitor. *PLoS ONE* **2014**, *9*, e100880. [CrossRef]
9. Mundi, P.S.; Sachdev, J.; McCourt, C.; Kalinsky, K. AKT in cancer: New molecular insights and advances in drug development. *Br. J. Clin. Pharmacol.* **2016**, *82*, 943–956. [CrossRef]
10. Hers, I.; Vincent, E.E.; Tavare, J.M. AKT signalling in health and disease. *Cell. Signal.* **2011**, *23*, 1515–1527. [CrossRef]
11. Huang, X.; Liu, G.; Guo, J.; Su, Z. The PI3K/AKT pathway in obesity and type 2 diabetes. *Int. J. Biol. Sci.* **2018**, *14*, 1483–1496. [CrossRef]
12. Nitulescu, G.M.; Margina, D.; Juzenas, P.; Peng, Q.; Olaru, O.T.; Saloustros, E.; Fenga, C.; Spandidos, D.A.; Libra, M.; Tsatsakis, A.M. AKT inhibitors in cancer treatment: The long journey from drug discovery to clinical use. *Int. J. Oncol.* **2016**, *48*, 869–885. [CrossRef]
13. Song, M.Q.; Bode, A.M.; Dong, Z.G.; Lee, M.H. AKT as a therapeutic target for cancer. *Cancer Res.* **2019**, *79*, 1019–1031. [CrossRef]
14. Narayan, R.S.; Fedrigo, C.A.; Brands, E.; Dik, R.; Stalpers, L.J.A.; Baumert, B.G.; Slotman, B.J.; Westerman, B.A.; Peters, G.J.; Sminia, P. The allosteric AKT inhibitor MK2206 shows a synergistic interaction with chemotherapy and radiotherapy in glioblastoma spheroid cultures. *BMC Cancer* **2017**, *17*, 204. [CrossRef]
15. Brown, J.S.; Banerji, U. Maximising the potential of AKT inhibitors as anti-cancer treatments. *Pharmacol. Ther.* **2017**, *172*, 101–115. [CrossRef]
16. Halder, A.K.; Cordeiro, M.N.D.S. Development of multi-target chemometric models for the inhibition of class i PI3K enzyme isoforms: A case study using qsar-co tool. *Int. J. Mol. Sci.* **2019**, *20*, 4191. [CrossRef]
17. Lima, A.N.; Philot, E.A.; Trossini, G.H.G.; Scott, L.P.B.; Maltarollo, V.G.; Honorio, K.M. Use of machine learning approaches for novel drug discovery. *Expert Opin. Drug Discov.* **2016**, *11*, 225–239. [CrossRef]
18. 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]
19. Kausar, S.; Falcao, A.O. An automated framework for QSAR model building. *J. Cheminform.* **2018**, *10*, 1. [CrossRef]
20. Halder, A.K.; Giri, A.K.; Cordeiro, M.N.D.S. Multi-target chemometric modelling, fragment analysis and virtual screening with erk inhibitors as potential anticancer agents. *Molecules* **2019**, *24*, 3909. [CrossRef]
21. Lewis, R.A.; Wood, D. Modern 2D QSAR for drug discovery. *Wires Comput. Mol. Sci.* **2014**, *4*, 505–522. [CrossRef]
22. 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] [PubMed]
23. Speck-Planche, A. Multi-Scale modeling in drug discovery against infectious diseases. *Mini Rev. Med. Chem.* **2019**, *19*, 1560–1563. [CrossRef]
24. Speck-Planche, A. Multiple perspectives in anti-cancer drug discovery: From old targets and natural products to innovative computational approaches. *Anticancer Agents Med. Chem.* **2019**, *19*, 146–147. [CrossRef]
25. Speck-Planche, A.; Scotti, M.T. BET bromodomain inhibitors: Fragment-based in silico design using multi-target QSAR models. *Mol. Divers.* **2019**, *23*, 555–572. [CrossRef]
26. Muratov, E.N.; Bajorath, J.; Sheridan, R.P.; Tetko, I.V.; Filimonov, D.; Poroikov, V.; Oprea, T.I.; Baskin, I.I.; Varnek, A.; Roitberg, A.; et al. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49*, 3716. [CrossRef]
27. Vyas, V.K.; Ghate, M.; Gupta, N. 3D QSAR and HQSAR analysis of protein kinase B (PKB/AKT) inhibitors using various alignment methods. *Arab. J. Chem.* **2017**, *10*, S2182–S2195. [CrossRef]
28. Dong, X.W.; Jiang, C.Y.; Hu, H.Y.; Yan, J.Y.; Chen, J.; Hu, Y.Z. QSAR study of AKT/protein kinase B (PKB) inhibitors using support vector machine. *Eur. J. Med. Chem.* **2009**, *44*, 4090–4097. [CrossRef]
29. Fei, J.; Zhou, L.; Liu, T.; Tang, X.Y. Pharmacophore modeling, virtual screening, and molecular docking studies for discovery of novel AKT2 inhibitors. *Int. J. Med. Sci.* **2013**, *10*, 265–275. [CrossRef]

30. Akhtar, N.; Jabeen, I. A 2D-QSAR and grid-independent molecular descriptor (grind) analysis of quinoline-type inhibitors of AKT2: Exploration of the binding mode in the pleckstrin homology (ph) domain. *PLoS ONE* **2016**, *11*, e0168806. [CrossRef] [PubMed]
31. Al-Sha'er, M.A.; Taha, M.O. Ligand-based modeling of AKT3 lead to potent dual AKT1/AKT3 inhibitor. *J. Mol. Graph. Model.* **2018**, *83*, 153–166. [CrossRef] [PubMed]
32. Ajmani, S.; Agrawal, A.; Kulkarni, S.A. A comprehensive structure-activity analysis of protein kinase B-alpha (AKT1) inhibitors. *J. Mol. Graph. Model.* **2010**, *28*, 683–694. [CrossRef] [PubMed]
33. Muddassar, M.; Pasha, F.A.; Neaz, M.M.; Saleem, Y.; Cho, S.J. Elucidation of binding mode and three dimensional quantitative structure-activity relationship studies of a novel series of protein kinase B/AKT inhibitors. *J. Mol. Model.* **2009**, *15*, 183–192. [CrossRef] [PubMed]
34. Anderson, A.C. The process of structure-based drug design. *Chem. Biol.* **2003**, *10*, 787–797. [CrossRef]
35. Halder, A.K.; Cordeiro, M.N.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]
36. 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. [CrossRef] [PubMed]
37. 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*, 511–523. [CrossRef] [PubMed]
38. Mauri, A. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. In *Ecotoxicological QSARs*; Roy, K., Ed.; Springer: New York, NY, USA, 2020; pp. 801–820.
39. Rucker, C.; Rucker, G.; Meringer, M. Y-Randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357. [CrossRef]
40. Ojha, P.K.; Roy, K. Comparative QSARs for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection. *Chemom. Intellig. Lab. Syst.* **2011**, *109*, 146–161. [CrossRef]
41. 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.
42. Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE* **2017**, *12*, e0177678. [CrossRef]
43. 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]
44. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
45. Roy, K.; Kar, S.; Ambure, P. On a simple approach for determining applicability domain of QSAR models. *Chemom. Intellig. Lab. Syst.* **2015**, *145*, 22–29. [CrossRef]
46. Speck-Planche, A.; Cordeiro, M.N.D.S. Speeding up early drug discovery in antiviral research: A fragment-based in silico approach for the design of virtual anti-hepatitis C leads. *ACS Comb. Sci.* **2017**, *19*, 501–512. [CrossRef]
47. Gramatica, P. WHIM descriptors of shape. *QSAR Comb. Sci.* **2006**, *25*, 327–332. [CrossRef]
48. Reutlinger, M.; Koch, C.P.; Reker, D.; Todoroff, N.; Schneider, P.; Rodrigues, T.; Schneider, G. Chemically Advanced Template Search (CATS) for scaffold-hopping and prospective target prediction for "Orphan' molecules. *Mol. Inform.* **2013**, *32*, 133–138. [CrossRef]
49. Todeschini, R.; Consonni, V.; Todeschini, R. *Molecular Descriptors for Chemoinformatics*, 2nd ed.; John Wiley Distributor: Weinheim, Germany; Chichester, UK, 2009; p. 1.
50. Devinyak, O.; Havrylyuk, D.; Lesyk, R. 3D-MoRSE descriptors explained. *J. Mol. Graph. Model.* **2014**, *54*, 194–203. [CrossRef] [PubMed]
51. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
52. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
53. Cover, T.M.; Hart, P.E. Nearest neighbor pattern classification. *IEEE T. Inform. Theory* **1967**, *13*, 21–29. [CrossRef]
54. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the 5th annual workshop on Computational learning theory, Pittsburgh, PA, USA, 27–29 July 1992; pp. 144–152.
55. Huang, G.B.; Babri, H.A. Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. *IEEE Trans. Neural Netw.* **1998**, *9*, 224–229. [CrossRef] [PubMed]
56. Hammann, F.; Drewe, J. Decision tree models for data mining in hit discovery. *Expert Opin. Drug Discov.* **2012**, *7*, 341–352. [CrossRef]
57. McCallum, A.; Nigam, K. A comparison of event models for naive bayes text classification. *Work Learn. Text Categ.* **2001**, *752*, 41–48.
58. Koutsoukas, A.; Monaghan, K.J.; Li, X.L.; Huan, J. Deep-learning: Investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J. Cheminform.* **2017**, *9*, 42. [CrossRef]
59. Stalring, J.C.; Carlsson, L.A.; Almeida, P.; Boyer, S. AZOrange—High performance open source machine learning for QSAR modeling in a graphical programming environment. *J. Cheminform.* **2011**, *3*, 28. [CrossRef]

60. Wang, X.; Shen, Y.H.; Wang, S.W.; Li, S.L.; Zhang, W.L.; Liu, X.F.; Lai, L.H.; Pei, J.F.; Li, H.L. PharmMapper 2017 update: A web server for potential drug target identification with a comprehensive target pharmacophore database. *Nucleic Acids Res.* **2017**, *45*, W356–W360. [CrossRef] [PubMed]
61. Liu, X.F.; Ouyang, S.S.; Yu, B.A.; Liu, Y.B.; Huang, K.; Gong, J.Y.; Zheng, S.Y.; Li, Z.H.; Li, H.L.; Jiang, H.L. PharmMapper server: A web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Res.* **2010**, *38*, W609–W614. [CrossRef]
62. Trott, O.; Olson, A.J. Software news and update AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461.
63. Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791. [CrossRef] [PubMed]
64. Haddad, Y.; Adam, V.; Heger, Z. Ten quick tips for homology modeling of high-resolution protein 3D structures. *PLoS Comput. Biol.* **2020**, *16*, e1007449. [CrossRef]
65. Guex, N.; Peitsch, M.C. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **1997**, *18*, 2714–2723. [CrossRef]
66. ChemAxon. *Standardizer, Version 15.9.14.0 Software*; ChemAxon: Budapest, Hungary, 2010.
67. 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]
68. 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]
69. Ambure, P.; Aher, R.B.; Gajewicz, A.; Puzyn, T.; Roy, K. "NanoBRIDGES" software: Open access tools to perform QSAR and nano-QSAR modeling. *Chemom. Intellig. Lab. Syst.* **2015**, *147*, 1–13. [CrossRef]
70. Menzies, T.; Kocagüneli, E.; Minku, L.; Peters, F.; Turhan, B. Chapter 22—Complexity: Using assemblies of multiple models. In *Sharing Data and Models in Software Engineering*; Menzies, T., Kocagüneli, E., Minku, L., Peters, F., Turhan, B., Eds.; Morgan Kaufmann: Boston, MA, USA, 2015; pp. 291–304.
71. Wilks, S.S. Certain generalizations in the analysis of variance. *Biometrika* **1932**, *24*, 471–494. [CrossRef]
72. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. [CrossRef]
73. Case, D.A.; Cheatham, T.E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K.M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R.J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688. [CrossRef] [PubMed]
74. Salomon-Ferrer, R.; Case, D.A.; Walker, R.C. An overview of the Amber biomolecular simulation package. *WIRES Comput. Mol. Sci.* **2013**, *3*, 198–210. [CrossRef]
75. Dolinsky, T.J.; Nielsen, J.E.; McCammon, J.A.; Baker, N.A. PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665–W667. [CrossRef]
76. Wang, J.M.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A. Development and testing of a general amber force field. *J. Comput. Chem.* **2005**, *26*, 114. [CrossRef]
77. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712–725. [CrossRef] [PubMed]
78. Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935. [CrossRef]
79. Roe, D.R.; Cheatham, T.E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095. [CrossRef]
80. Wang, L.; Chen, L.; Yu, M.; Xu, L.H.; Cheng, B.; Lin, Y.S.; Gu, Q.; He, X.H.; Xu, J. Discovering new mTOR inhibitors for cancer treatment through virtual screening methods and in vitro assays. *Sci. Rep.* **2016**, *6*, 18987. [CrossRef]
81. Berishvili, V.P.; Kuimov, A.N.; Voronkov, A.E.; Radchenko, E.V.; Kumar, P.; Choonara, Y.E.; Pillay, V.; Kamal, A.; Palyulin, V.A. Discovery of novel tankyrase inhibitors through molecular docking-based virtual screening and molecular dynamics simulation studies. *Molecules* **2020**, *25*, 3171. [CrossRef] [PubMed]
82. Srinivasan, J.; Cheatham, T.E.; Cieplak, P.; Kollman, P.A.; Case, D.A. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate—DNA helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409. [CrossRef]
83. Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **2015**, *10*, 449–461. [CrossRef]
84. Schoning, V.; Hammann, F. How far have decision tree models come for data mining in drug discovery? *Expert Opin. Drug Discov.* **2018**, *13*, 1067–1069. [CrossRef]
85. Yang, X.; Wang, Y.F.; Byrne, R.; Schneider, G.; Yang, S.Y. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev.* **2019**, *119*, 10520–10594. [CrossRef]
86. Zhao, L.L.; Ciallella, H.L.; Aleksunes, L.M.; Zhu, H. Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. *Drug Discov. Today* **2020**, *25*, 1624–1638. [CrossRef]