European Journal of Pharmaceutical Sciences 62 (2014) 171–179

Contents lists available at ScienceDirect

European Journal of Pharmaceutical Sciences

journal homepage: www.elsevier.com/locate/ejps

QSAR models for thiophene and imidazopyridine derivatives inhibitors of the Polo-Like Kinase 1

Nieves C. Comelli [a,*], Pablo R. Duchowicz [b], Eduardo A. Castro [b]

[a] *Facultad de Ciencias Agrarias, Universidad Nacional de Catamarca, Av. Belgrano y Maestro Quiroga, 4700 Catamarca, Argentina*

[b] *Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (UNLP, CCT La Plata-CONICET), Diag. 113 y 64, C.C. 16, Sucursal 4, 1900 La Plata, Argentina*

ARTICLE INFO

Article history:
Received 28 January 2014
Received in revised form 27 May 2014
Accepted 28 May 2014
Available online 6 June 2014

Keywords:
Chemoinformatics
Multivariate linear regression analysis
Polo-Like Kinase 1 (PLK1) inhibitors
Molecular modeling
Thiophene and imidazopyridines derivatives

ABSTRACT

The inhibitory activity of 103 thiophene and 33 imidazopyridine derivatives against Polo-Like Kinase 1 (PLK1) expressed as pIC_{50} ($-\log IC_{50}$) was predicted by QSAR modeling. Multivariate linear regression (MLR) was employed to model the relationship between 0D and 3D molecular descriptors and biological activities of molecules using the replacement method (MR) as variable selection tool.

The 136 compounds were separated into several training and test sets. Two splitting approaches, distribution of biological data and structural diversity, and the statistical experimental design procedure D-optimal distance were applied to the dataset. The significance of the training set models was confirmed by statistically higher values of the internal leave one out cross-validated coefficient of determination (Q^2) and external predictive coefficient of determination for the test set (R^2_{test}).

The model developed from a training set, obtained with the D-optimal distance protocol and using 3D descriptor space along with activity values, separated chemical features that allowed to distinguish high and low pIC_{50} values reasonably well. Then, we verified that such model was sufficient to reliably and accurately predict the activity of external diverse structures.

The model robustness was properly characterized by means of standard procedures and their applicability domain (AD) was analyzed by leverage method.

1. Introduction

Polo-Like Kinases (PLKs) are important regulators of cell cycle progression and mitosis. They are an evolutionary conserved family of serine/threonine kinases characterized by an amino-terminal kinase domain and a C-terminal region composed of polo boxes. There are four identified Polo-Like Kinases, PLK1, PLK2 (SNK), PLK3 (PRK/FNK) and PLK4 (SAK), which have similar but non-overlapping functions in the cell cycle progression (Harris et al., 2012; Song et al., 2012). Particularly, PLK1 is the most investigated member of the family as an anticancer drug target because it is overexpressed in several human tumor types and inhibition of its activity results in a potent antitumor effect both in vitro and in vivo (Chopra et al., 2010; Degenhardt and Lampkin, 2010; Rudolph et al., 2009; Zhang et al., 2009).

In recent years much attention was focused to the research of small molecule PLK1 inhibitors. To date, eleven kinase inhibitors were approved by the Food and Drug Administration (FDA) for the treatment of cancer, and there are continuous efforts to get more candidates (Bohari et al., 2011; Chahrour et al., 2012; Liu and Gray, 2006; Strebhardt and Ullrich, 2006).

Most kinase inhibitors discovered to date are ATP competitive. Other kinase inhibitors were identified by interacting with a hydrophobic pocket directly adjacent to the ATP binding site or by binding to a hydrophobic pocket remote from the ATP binding site (Zhang et al., 2009).

In recent publications, two series of ATP-competitive PLK1 inhibitors formed by 103 thiophene benzimidazole (TP) and 33 imidazopyridine (IP) analogs were described (Emmitte et al., 2009a,b; Rheault et al., 2010; Sato et al., 2009). These compounds were developed from the compound in Fig. 1 which was characterized as a potent inhibitor of PLK1 and a useful tool for further investigation on the biological functions of PLK1 (Lansing et al., 2007).

From 5-(5,6-dimetoxy-1H-benzimidazol-1-yl)-3-{(2-trifluoromethylbenzyl)oxy} thiophene-2-carboxamide, the new 136 PLK1 inhibitors were prepared by introducing of chemical modifications such as the methylation of the benzylic carbon with (R)-configuration and the incorporation of functional groups on the aryl and heteroaryl groups with a relatively wide range of size, polarizability, hydrophobicity, and lipophilicity.

* Corresponding author. Tel.: +54 0383154698975.
E-mail address: ncomelli75@gmail.com (N.C. Comelli).

http://dx.doi.org/10.1016/j.ejps.2014.05.029
0928-0987/© 2014 Elsevier B.V. All rights reserved.

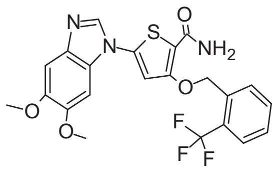


Fig. 1. 5-(5,6-Dimetoxy-1H-benzimidazol-1-yl)-3-{[2-trifluoromethylbenzyl]oxy} thiophene-2-carboxamide molecule.

Structure–activity relationship (SAR) studies assisted by docking simulations have provided useful insight into the structural requirements for inhibitory activity of TP and IP against PLK1 (Emmitte et al., 2009a,b; Rheault et al., 2010; Sato et al., 2009). Moreover, some quantitative structure–activity relationship (QSAR) methods have been applied in order to get a better understanding of the chemical features that influenced their activity and improve further design of new PLK1 inhibitors. Particularly, in silico models were developed using the comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) (Cao, 2012).

QSAR studies consist of relating the biological activities of a series of compounds with appropriate molecular descriptors (Bohari et al., 2011; Cronin, 2010). Such relationships may be used to predict the activity/property of new compounds and to design virtual compound libraries. By their widely accepted predictive and diagnostic power, QSAR models promise considerable savings in time, money, and a reduction in use of experimental animals when compared with conventional testing strategies.

Since the application of QSAR methodology on PLK1 inhibitors has received little attention and, to the best of our knowledge, only Cao's paper published QSAR models for the inhibitory activity of 103 TP analogs and 33 IP derivatives, herein we report QSAR analysis based on the curated dataset of 103 TP and 33 IP derivatives since we detected some errors in previously published data.

More specifically, we construct models for each class of compounds and hybrid models including compounds from two classes. Conditions evaluated to create hybrid models are: (i) Dataset is comprised of structurally-related molecules with a common 3-(benzyloxy)thiophene scaffold hopping and shared action mechanism; (ii) even though, biological data of TP and IP analogs were determined by different protocols, we found that both techniques measure the similarly inhibitory effects. The consistency of data was observed comparing the pIC_{50} values of 5-(1H-benzimidazol-1-yl)-3-{2-trifluoromethylbenzyloxy}-thiophene-2-carboxamide identified as compound **16** (Emmitte et al., 2009b) and compound **1** in (Sato et al., 2009). Since the pIC_{50} difference among the compounds (0.368 log unit) is within the mean experimental error ($0.5340 \leqslant \varepsilon \leqslant 0.6366$ log unit), we recognize that the basic paradigm of QSAR methods can be applied without difficulty to the curated dataset of 136 compounds in (Cao, 2012).

Statistical parameters of the models based on different theoretical molecular descriptors were sensitive to the selection of the training and test set using a rational splitting method. The model developed from a training set obtained with the D-optimal distance protocol and using 3D descriptor space along with activity values gave good predictions and afforded rationale for the search of novel leads as PLK1 inhibitors.

2. Materials and methods

2.1. Experimental dataset

The experimental PLK1 inhibitory activity measured as $pIC_{50} = -\log_{10}IC_{50}$ of 136 PLK1 kinase inhibitors were obtained from the literature (Emmitte et al., 2009a,b; Rheault et al., 2010; Sato et al., 2009). These compounds were categorized into two structural classes: a dataset of 103 thiophene (TP) analogs and 33 imidazopyridine (IP) derivatives (see Table S1 in the Supporting Information). The inhibitory activities were determined in enzymatic and tumor cell lines assays as were described in (Lansing et al., 2007; Sato et al., 2009).

2.2. Modeling and molecular descriptors calculation

We retained the R-configuration for the benzylic carbon atom on all the molecular structures except in the compound **4** where the chirality of the sp^3 C-atom was changed to S-configuration (see Table 1S). The initial conformations of the compounds were drawn with the aid of the "Model Build" modulus of the HyperChem 7.5 program for Windows (HyperChem Release 7.5 for Windows, 2002) and pre-optimized using molecular mechanics force fields (MM+). Then, the molecular geometries were refined using semi-empirical quantum method Austin Method 1 (AM1) and the Polack–Ribiere algorithm also implemented in Hyperchem until an energy gradient smaller than 0.1 kcal mol^{-1} Å^{-1}. Finally, the .hin files of the found geometries were converted into .smile by using the Open Babel 2.3.1 software (O'Boyle et al., 2011). This molecular format was employed as input for the generation of all molecular descriptors.

As modeling input variables, we used a wide set of theoretical molecular descriptors that takes into account different structural features -constitutional (0D), mono-dimensional (1D), bi-dimensional (2D) and three-dimensional (3D)- for capturing and magnifying distinct aspects of chemical structures. Thus, a total of 2594 molecular descriptors including molecular properties (such as $\log P$, molar refractivity, number of rotatable bonds, H-donors, H-acceptors, and topological surface area), constitutional descriptors, topological descriptors, connectivity indices, autocorrelation descriptors, charge descriptors and molecular fragments were calculated. The descriptor typology is as follows: (a) 248 transferable atom equivalent (TAE) descriptors using RECON 5.5 (Recon 5.5, 2002), (b) 694 3D-descriptors available in the Online Chemical Modeling Environment (OCHEM): Inductive, MERA, MERSY, Spectrophores and MOPAC descriptors (Sushko et al., 2011), and (c) 931 conformation-independent molecular plus 721 3D-descriptors using E-Dragon (E-Dragon 1.0, 2005). The list of these molecular descriptors, their meaning, and the calculation procedures were provided with literature references by the RECON 5.5, E-DRAGON 1.0 packages and the online web OCHEM platform. For more details, see the related publications (Breneman and Rhem, 1997; Breneman et al., 1995; Bultinck et al., 2003; Consonni and Todeschini, 2010; Cherkasov, 2005; Todeschini and Consonni, 2009).

Prior to the development of QSAR models, descriptors were normalized using the following formula,

$$X_{ij}^{n} = \frac{X_{ij} - X_{j,\min}}{X_{j,\max} - X_{j,\min}}$$

where X_{ij} and X_{ij}^{n} are the non-normalized and normalized j-th ($j = 1, \ldots, k$) descriptor values for compound i ($i = 1, \ldots, n$), and $X_{j,\min}$ and $X_{j,\max}$ are the minimum and maximum values for the j-th descriptor. Thus, for all descriptors, $\min(X_{ij}^{n}) = 0$ and $\max(X_{ij}^{n}) = 1$. This procedure is a good practice, especially when the predictor variables have different scale ranges. In this way, all the variables are treated as if they were of equal importance, regardless of their scale of measurement.

2.3. Model development

2.3.1. Molecular descriptors selection

Here the replacement method (RM) (Duchowicz et al., 2005, 2006) was used to search, via multivariate linear regressions (MLRs), the molecular descriptors responsible for the activity of 136 PLK1 inhibitors. This method, successfully applied elsewhere (Duchowicz et al., 2008; Goodarzi et al., 2009; Pasquale et al., 2012; Pomilio et al., 2010; Talevi et al., 2011), is considered an efficient optimization tool which generates models on a training set by searching in a set having D descriptors for an optimal subset having $d \ll D$ ones with smallest training set standard deviation (S_{train}) or smallest root mean square deviation ($RMSD_{train}$). The quality of the results achieved with this technique approaches that obtained by performing an exact (combinatorial) full search of molecular descriptors although with a much shorter CPU time (Mercader et al., 2010, 2011).

For the selection of the most relevant descriptors, we considered independently seven descriptors pools containing (**a**) 248 Recon descriptors; (**b**) 694 3D OCHEM descriptors; (**c**) 721 3D-Dragon descriptors; (**d**) 931 (0D–2D)-Dragon descriptors; and (**f**) their combinations: (i) 1179 descriptors by merging the **a** and **d** classes, (ii) 1415 descriptors relating the **b** and **c** classes and, (iii) the pool formed by 2594 descriptors. In this way, we try to capture the most relevant variables in modeling the response of the training set since we have no a priori knowledge of which descriptors, and which particular combinations with others are related to the studied response and are able to be used in models for prediction aims.

2.3.2. Training and test set selection

One of the most important characteristics of a QSAR model is its predictive power. This can be defined as the ability of a model to accurately predict the biological activity of compounds that were not used for model development.

A QSAR model's predictive ability depends largely on the nature of the training set used to establish the structure–activity relationship (Martin et al., 2012). In order to obtain a model of high statistical rigor and predictive power, a dataset should be adequately split into the training and test sets. This division must satisfy the following conditions (Golbraikh et al., 2003; Martin et al., 2012): (i) the training set chemicals must be structurally diverse enough to cover the whole descriptor space of the overall data set, and (ii) the compounds in the training and test sets should be close to each other.

In this work the search of validated and predictive QSAR models was initiated by ordering the chemicals as in (Cao, 2012) and with the partition into a training and test set proposed there. Then, the model development process for each class of inhibitors was followed by using 78 TP and 26 IP as training set and the developed models were validated using the remaining compounds (25 TP and 7 IP) as test set. Meanwhile, hybrid models from two classes compounds were built and validated with 104 (78 P and 26 IP) and 32 (25 TP and 7 IP) compounds respectively.

As regards the above-mentioned partition, such division cannot guarantee that the training set compounds represent the entire calculated descriptor space in this work. Then we use the statistical Experimental Design procedure, D-optimal distance (Marengo and Todeschini, 1992; Todeschini et al., 2012), to develop several training and test sets.

Briefly, D-optimal distance protocol selects samples through a fast exchange algorithm where, in each cycle, substitution provides the maximum increase in the minimum distance between currently selected compounds. This algorithm provides a final distribution of the most dissimilar compounds selected from the set of allowed candidates which are used as a training set to develop models. Once the models are established, predictions are made for the remaining molecules under study (test set).

The D-optimal distance methodology was widely used for developing predictive QSAR model (Gramatica, 2013). The resultant classification ensure that the similarity principle can be employed for the activity prediction of the test set and satisfies the condition of closeness between the test set points to the training set points.

For the purpose of this paper, the sampling was performed with each descriptor set described in Section 2.3.1 including the activity values. Consequently, from the original dataset, seven partitions in training and test set -each one with 104 (training set) and 32 (test set) molecules- were selected. The compounds selected as test compounds are listed in Table 2S.

2.3.3. Model validation

The statistical qualities and validity of the built MLR equations were judged by means of (Roy, 2007; Todeschini et al., 2009): (a) internal validation or cross-validation (CV) by leave-one-out (LOO) and leave-some-out (LSO) procedure; (b) using the test set; (c) data randomization or Y-scrambling and; (d) examining if the following conditions are satisfied (Golbraikh and Tropsha, 2002; Martin et al., 2012):

$$Q^2 > 0.5$$

$$R^2_{test} > 0.6$$

$$\frac{(R^2_{test} - R^2_0)}{R^2_{test}} < 0.1 \quad \text{and} \quad 0.9 \leqslant k \leqslant 1.1$$

$$\text{or} \quad \frac{(R^2_{test} - R'^2_0)}{R^2_{test}} < 0.1 \quad \text{and} \quad 0.9 \leqslant k' \leqslant 1.1$$

$$|R^2_0 - R'^2_0| < 0.3$$

and

$$R^2_m > 0.5$$

where Q^2 is the leave one out correlation coefficient for the training set, R^2_{test} is the square of the Pearson's correlation coefficient (coefficient of determination) between the observed and predicted inhibitory activity for the test set, R^2_0 and R'^2_0 are coefficients of determination for trend lines through the origin between the observed and predicted and predicted and observed pIC_{50} values for the test sets and k and k' are the corresponding slopes. Moreover, we examined the difference between R^2_{test} and R^2_0 denoted as R^2_m in (Roy and Roy, 2007).

Regarding the application of Y-randomization, after the analysis of 100 cases of repetitive randomization of the response data (Y) of the compounds in the several training sets without making any change in the descriptor matrix, the statistical reliability was corroborated by noting that random models have significantly higher standard deviation (S^{rand}) than the original models (S).

All parameters taken as indicators of predictive capacity of the models reported in this work appear defined along with their correspondent equations in Table 3S.

2.3.4. Applicability domain analysis

Since a structure–activity model is defined and limited by the nature and quality of the data used in the training set for model development, a QSAR model is only valid within its respective response and chemical structure space (applicability domain, DA). This region is defined by the nature of the chemicals in the training set, and can be characterized in various ways (Netzeva et al., 2005).

In this work, we use the leverage approach (Atkinson, 1991; Gramatica, 2007, 2013) for the applicability domain definition. The Williams plot of standardized cross-validated residuals (r_i)

vs. leverage values (h) was used for an immediate and simple graphical detection of both the response outliers -i.e., compounds with cross-validated standardized residuals greater than 3 standard deviation units, $r_i > 3s$- and structurally influential chemicals in a model ($h > h^*$). In this plot, the applicability domain is established inside a squared area with $r_i = \pm 3s$ and a leverage threshold h^* (with $h^* = 3p/n$, where p is the number of model variables plus one, and n is the number of the molecules used to calculate the model).

From the leverage values, a compound is considered outside the applicability domain (outlier) when $h > h^*$ and $r_i > 3s$. Conversely, when a compound has a leverage value lower than the critical one, the probability of accordance between the predicted and the actual values is as high as that for the training set chemicals. The same is true for a chemical with $h > h^*$ in the training set and $r_i < 3s$. Such compound is a structurally influential chemical in the model.

2.3.5. Degree of contribution of selected descriptors

In order to find out the relative importance of the j-th descriptor in the linear model, we standardized its regression coefficient (b_j^s, see Table 3S). The larger the absolute value of b_j^s, the greater the importance of such descriptor (Draper and Smith, 1981).

For all calculations in this work we used algorithms written in the language of technical computing Matlab 7.12 (Matlab, 2011).

3. Results and discussion

3.1. QSAR models from training set by considering distribution of biological data and structural diversity

From the training and test sets proposed previously (Cao, 2012), we have constructed twenty-one types of models using RM approach: seven separate models for each class of inhibitors (fourteen models in total) and seven hybrid models including compounds from two classes.

In Table 1A–C we present the linear regressions by group of descriptors with minimum $S_{train}/RMSD_{train}$ and $S_{test}/RMSD_{test}$ values obtained by removing 1–8 variables. Following the common practice of keeping a relatively small number of descriptors in the model, we consider it to be an acceptable model if it is a solution with $R^2_{test} > 0.6$ and $Q^2 > 0.5$.

According to our calculations (see Table 1A–C), the best MLR models have two to five parameters. For the models of TP analogs in Table 1A, while the internal validation is always within an acceptable statistical range ($Q^2 > 0.5$), the test set selected does not produce good external validation statistics ($R^2_{test} < 0.6$). In contrast, the models generated for the IP molecules provided consistently high values of Q^2 and R^2_{test} (i.e. $0.70 \leqslant Q^2 \leqslant 0.86$ and $0.66 \leqslant R^2_{test} < 0.85$). However, we consider that the predictive capacity of such models is questionable due to the small dataset used to obtain the models and the remarkable difference in the $S_{train}/RMSD_{train}$ and $S_{test}/RMSD_{test}$ values. As regards to the hybrid QSAR models in Table 1C, these do not display satisfactory predictive ability despite the higher diversity and larger number of compounds included in the calculations. In fact, R^2_{test} is not always acceptable in all the models ($R^2_{test} < 0.6$ in most models) and the $S_{train}/RMSD_{train}$ and $S_{test}/RMSD_{test}$ values are not of the same quality.

The results in Table 1A–C emphasize that the data in the training and test set are non-homogeneous. In fact, we understand that the predictive capacity of the reported linear models is fairly low due to the use of test data with members which are not well represented in the training set. Likewise, QSAR models in Table 1 have limitations associated with chemical diversity of the training set: such selection does not ensure that all chemical classes in the original dataset are well represented. So, despite of the certain predictive capacity over compounds into of the training set ($R^2_{train}, Q^2 > 0.50$), selection of insufficiently representative test set reveals that the resulting models may not be satisfactorily practical for the search of related structures with similar biological activity from large chemical databases.

One way to accomplish a representative training set is through multiple designs including selection on the basis of relevant

Table 1
Statistical quality of the developed models using training and test set proposed in (Cao, 2012) for: (A) thiophene analogs, (B) imidazopyridines analogs and (C) from two classes compounds including 104 compounds (78 thiophene and 26 imidazopyridine) as training set. Here, **D** represents the set of descriptors explored during the modeling and **d** is the number of descriptors in the selected model.

D	d	R^2_{train}	S_{train}	$RMSD_{train}$	R^2_{test}	S_{test}	$RMSD_{test}$	$R^2_{ij,\max}$	Q^2	S_{loo}
(A)										
248	5	0.60	0.53	0.51	0.38	0.66	0.57	0.50	0.53	0.58
694	5	0.75	0.42	0.40	0.41	0.86	0.75	0.29	0.69	0.46
721	5	0.70	0.46	0.44	0.43	0.69	0.60	0.23	0.62	0.51
931	2	0.54	0.56	0.54	0.35	0.65	0.61	0.07	0.71	0.58
1179	5	0.72	0.45	0.43	0.47	0.71	0.62	0.80	0.66	0.49
1415	5	0.76	0.41	0.39	0.45	0.62	0.54	0.31	0.72	0.44
2594	5	0.78	0.39	0.37	0.47	0.71	0.62	0.18	0.74	0.42
(B)										
248	3	0.79	0.33	0.31	0.81	0.90	0.59	0.32	0.7	0.39
694	4	0.89	0.24	0.22	0.80	0.81	0.43	0.65	0.83	0.31
721	3	0.89	0.24	0.22	0.75	0.80	0.52	0.26	0.83	0.31
931	3	0.88	0.25	0.23	0.85	0.49	0.32	0.70	0.77	0.36
1179	3	0.89	0.24	0.22	0.69	0.87	0.57	0.29	0.81	0.31
1415	3	0.92	0.20	0.18	0.66	0.90	0.59	0.27	0.86	0.26
2594	3	0.92	0.19	0.18	0.58	1.02	0.67	0.45	0.89	0.25
(C)										
248	4	0.50	0.60	0.58	0.46	0.67	0.62	0.37	0.45	0.63
694	5	0.66	0.49	0.48	0.58	0.63	0.57	0.06	0.62	0.53
721	4	0.63	0.52	0.50	0.62	0.55	0.50	0.46	0.60	0.54
931	5	0.68	0.48	0.47	0.45	0.70	0.63	0.52	0.62	0.52
1179	5	0.71	0.46	0.43	0.42	0.72	0.52	0.28	0.66	0.50
1415	5	0.73	0.45	0.44	0.59	0.58	0.65	0.34	0.69	0.48
2594	5	0.76	0.42	0.41	0.49	0.65	0.59	0.15	0.72	0.45

molecular descriptors and structural similarity. So, the dataset under study was divided into a training set with 104 compounds and a test set with 32 compounds using all the available structural information as described in Section 2.3.2. The compounds selected as test set are listed in Table 2S.

3.2. QSAR models from training and test sets obtained with the D-optimal distance protocol

The best models built with the diverse training and test sets and by removing 1–8 variables from each D set are summarized in Table 2. Such models were defined from four to five descriptors. A higher or lower number of variables did not have any significant effect on the accuracy of the developed models.

According to our calculations, the model generated from the training set select in descriptor space including 694 3D OCHEM and 721 3D-Dragon had consistently high values of R^2_{test} and Q^2 (see Table 2, $D = 1415, R^2_{test} = 0.71, Q^2 = 0.69$). Others slightly inferior validated models were obtained with training and test sets selected from 3D OCHEM space and from the set with all available structural information (i.e. $D = 694$, $R^2_{test} = 0.63$, $Q^2 = 0.58$ and $D = 2594$, $R^2_{test} = 0.68$, $Q^2 = 0.67$).

Direct comparison of the models in Tables 1 and 2 to corroborate our initial assumption about the limited structural variability of the compounds included in the training set in (Cao, 2012) with a direct impact in the predictive capacity of the final models. In this context, better prediction of the pIC_{50} values with $D = 1415$ in both training and test set in Table 2 reveals the importance of the 3D data to formulate informative QSAR equations. Other structural features important for the inhibitory activity appear in the model combining all types of available descriptors ($D = 2594$). Conversely, it is worth noting the relatively low statistical significance and predictive power of the QSAR formulations applying 0D–2D descriptors (i.e. $D = 248$, 931, 1179). In such case, we consider that linear regression is not sufficient to describe the data, and that training non-linear models may be useful.

From Table 2, the model that exhibits the best balance between the statistical parameters obtained on the training and the test set was,

$$pIC_{50} = 3.598RDF0.5m - 1.558Mor02p - 2.179E2v - 3.115ASyV2 + 3.954Small\_RSI\_Mol - 2.334 \tag{1}$$

$$N_{train} = 104, \quad d = 5, \quad N_{train}/d = 21, \quad R^2_{train} = 0.73,$$
$$S_{train} = 0.45, \quad F = 53.23, \quad R^2_{ij\max} = 0.56,$$
$$R^2_{loo} = 0.69, \quad S_{loo} = 0.48, \quad R^2_{l20\%o} = 0.63, \quad S_{l20\%o} = 0.53,$$
$$S^{rand} = 0.79,$$
$$N_{test} = 32, \quad R^2_{test} = 0.71, \quad S_{test} = 0.47.$$

where F and $R^2_{ij\max}$ are the Fisher parameters and the maximum correlation coefficient between descriptor pairs. The large F value and the small $R^2_{ij\max}$ are indicative of the model statistical significance and the trivial degree of multicollinearity among the chosen subset of descriptors.

The model given by Eq. (1) involves five descriptors that explain 73% of the experimental variance (R^2_{train}). This is considered a valid structure–activity relationship since it approves the internal validation process of cross-validation through the exclusion of one molecule at a time and also by excluding 20% of the observations (21 molecules). Other facts that suggest that the reported model is properly validated are:

- $S_{train} < S^{rand}$
- $\frac{(R^2_{test} - R^2_0)}{R^2_{test}} = 0.008$
- $\frac{(R^2_{test} - R'^2_0)}{R^2_{test}} = 0.3$
- $k = 1.01$
- $k' = 0.99$
- $|R^2_0 - R'^2_0| = 0.21$
- $R^2_m = 0.65$

The meaning of the variables in Eq. (1) and of all models reported in this work appears in Table 4S. Meanwhile, Table 5S displays the activities predicted by Eq. (1) along with the numerical values for the best molecular descriptors of pIC_{50} inhibitory activities.

More considerations on the predictive power of the chosen model were summarized in Fig. 2 and Fig. 1S in Supplementary Material.

Particularly, Fig. 2 shows the regression line contrasting the predicted and experimental pIC_{50} values and Fig. 1S reveals the dispersion plot of residuals. The alignment of the points in a straight line in Fig. 2 and the fact that residuals tend to follow a random pattern around the zero line in Fig. 1S reveals that the assumption of the MLR technique is fulfilled (i.e. confirm the fine correlation between selected descriptors and the modeled property).

On the other hand, Fig. 3 shows the Williams plot of the model for TP and IP as PLK1 inhibitors. Here, no compound has been identified as outlier and just two compounds in the training set (molecules **61** and **68**) are predicted as slightly influential. Such molecules, structurally somewhat distant from the other chemicals (with substituted amide and ester group in the 2-position of the thiophene), are considered "good leverage points" since the information that they encode contribute to make the model more precise.

The aforementioned concepts arise after checking the structural similarity of the test chemical **67** outside the applicability domain with the influential chemicals. At that point, we determined that the predicted response for this molecule is reliable and that splitting by D-optimal design allows to obtain a model with completely interpolated predictions.

As regards the introduced descriptors in Eq. (1), these were easily accessible from the x, y, z coordinates of the molecule atoms and other quantities derived from the coordinates (interatomic distances or distances from a specified origin, molecular volume, electronegativity and covalent radii). They reflect aspects related to the molecular size, shape and the steric interactions quantification. Particularly:

Table 2
Best QSAR found with the replacement method and representative training and test sets selected in different descriptors spaces.

D	**d**	R^2_{train}	S_{train}	$RMSD_{train}$	R^2_{test}	S_{test}	$RMSD_{test}$	$R^2_{ij,\max}$	Q^2	S_{loo}	Descriptors
248	5	0.50	0.64	0.62	0.63	0.44	0.40	0.98	0.43	0.68	*Del(Rho)NA10, SIEPMax, PIP3, FDRNA10, FPIP5*
694	5	0.64	0.53	0.51	0.63	0.47	0.42	0.64	0.58	0.57	*Spectrosph.Part_4, ASYV2, SYMC6X, SYMS4Z, SmallestRsIMol*
721	4	0.66	0.51	0.49	0.57	0.59	0.54	0.46	0.63	0.53	*RDF110v, RDF035p, Mor17e, HATS8u*
931	4	0.63	0.52	0.51	0.46	0.65	0.59	0.66	0.59	0.55	*IC3, BEHm7, GGI9, C-042*
1179	5	0.67	0.51	0.49	0.33	0.60	0.54	0.61	0.59	0.56	*piPC08, GATS2m, BELv1,* $nArCONH_2$*, Energy*
1415	**5**	**0.73**	**0.45**	**0.44**	**0.71**	**0.47**	**0.42**	**0.31**	**0.69**	**0.48**	*RDF035m, Mor02pE2v, ASYV2, SmallestRsIMol*
2594	5	0.71	0.46	0.52	0.61	0.66	0.64	0.16	0.68	0.48	*BIC5, SIEPMin, RDF075m, RDF035p, ASYV2*

The final model chosen for QSAR analysis is highlighted in bold font.

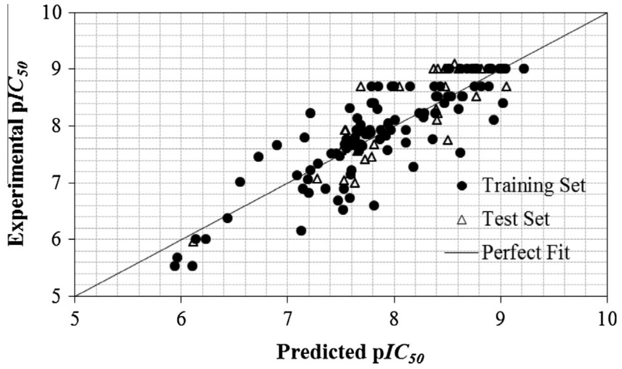


Fig. 2. Experimental pIC_{50} for PLK1 as function of predicted pIC_{50} according to Eq. (1).

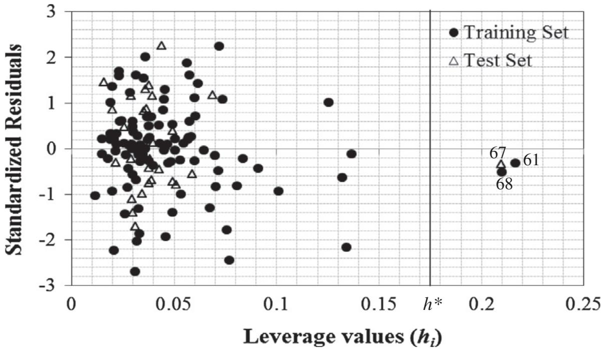


Fig. 3. Williams plot for the externally validated QSAR model of TP and IP inhibitors against PLK1.

- *RDF035m* is the radial distribution function on a spherical volume of radius 3.5 angstroms weighted by atomic masses. This descriptor, based on the geometrical interatomic distance, provides valuable information about interatomic distances in a molecule.
- *ASyV*2, is a measure of the asymmetry of molecular volume along the second principal inertia axis. It is built, using 3D representation of molecules in the framework of Model of Effective Radii of Atoms (Sushko et al., 2011) in such a way as to capture molecular information regarding atom distribution in invariant reference frames.
- *Small_RSI_Mol* inform the smallest value of atomic steric influence in a molecule. It is an inductive reactivity index which quantifies steric effects of a single atom onto a group of atoms through the basic and readily accessible parameters of bound atoms: electronegativities, covalent radii, and intramolecular distances.
- E2v is the 2nd component accessibility directional WHIM index weighted by atomic van der Waals volumes. This and other similar descriptors are statistical indices calculated on the projections of the atoms along the principal inertial axes. They estimate dispersion and distribution of molecular properties around the geometric center. Particularly, $E2v$ is related to the quantity of unfilled space per projected atom and has been called density (or emptiness).
- 3D-MoRSE signal 2 weighted by atomic polarizabilities (*Mor*02*p*) is one of the 3D-MoRSE descriptors that can be calculated by summing atomic properties viewed by a different angular scattering function at 32 equidistant values in the range 1–31 Å^{-1}. These codes have great potential for representation of molecular structure since they reflect the three-dimensional arrangement of the atoms, molecular size and shape.

Standardized coefficient values (b_j^s) in decreasing order of significance are reported below:

$$RDF035m(0.79), ASyV2(0.54), Small\_RSI\_Mol(0.43), \mathrm{E2v}(0.35), Mor02p(0.20)$$

This trend reveals that the *RDF*035*m*, *ASyV*2 and *Small_RSI_Mol* variables were the most informative to the model.

It is worth to highlight that only *RDF*035*m* and *Small_RSI_Mol* contribute positively on the pIC_{50} values. Thus, higher values for *RDF*035*m*, *Small_RSI_Mol* and lower values for *ASyV*2, would lead to more potent compounds (higher predicted pIC_{50} activities), that is to say, bulky substituents in lateral positions of 3-(benzyloxy) thiophene scaffold. In fact, this is observed in compounds with the substitution pattern: pyridine, pyrimidine and pyrazole ring with straight-chain alkyl and hydrophilic groups.

In contrast, *RDF*035*m* and *Small_RSI_Mol* diminish, while *ASyV*2 increases in compounds with imidazole and aminothiazole substituents replacing the benzimidazole moiety. Likewise, this trend was observed in derivatives with small substituents and substituted amides, ester, cyano and thioamide groups in the 2-position of the thiophene.

We understand that these results are consistent with the preliminary SAR observations described in the literature (Emmitte et al., 2009a,b; Rheault et al., 2010; Sato et al., 2009). Then, we propose that a potent inhibitor of PLK1 is that whose structure presents a subtle balance between steric, hydrophobic and electrostatic factors which play an important role in stacking interactions with PLK1.

3.3. Prediction of anticancer activity for an external data set

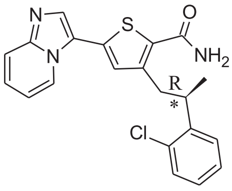
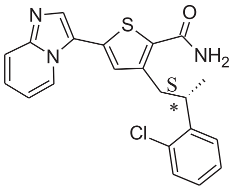
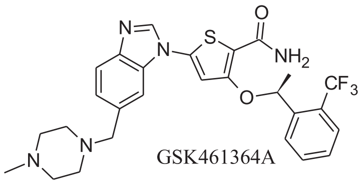
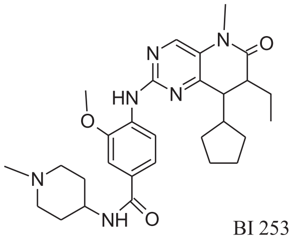
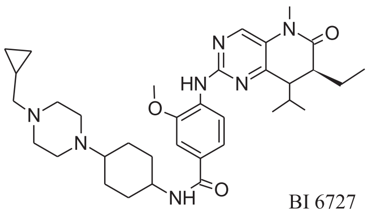
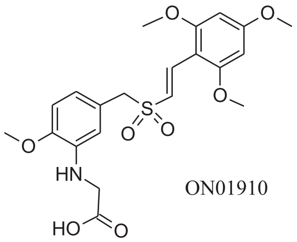
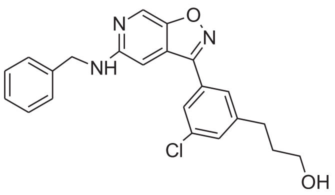
A reliable and truly predictive QSAR model should be able to accurately predict activities of new compounds in external sets. To this end, the model in Eq. (1) validated with the test sets was used to predict the activity of seven compounds (Table 3) (Murugan et al., 2011; Sato et al., 2009), which were not available prior to our QSAR study of the 136 derivatives dataset.

Thus, we have decided to test the developed model with two imidazopyridines cited in (Sato et al., 2009) and omitted without explanation in (Cao, 2012), a thiophene benzimidazole (GSK 461364A), two dihydropteridinone derivative (BI 2536 and BI 6727), a benzy styryl sulfone (ON01910) and an isoxazolopyridines. Such external set, formed by three compounds with similar core structure to the training set (congeneric, imidazopyridines and GSK161364A) and four novel compounds structurally diverse to the training set (BI 2536, BI 6727, ON01910 and isoxazolopyridines), displayed moderate to high activity ($6.5229 \leqslant pIC_{50} \leqslant 9.0970$; Table 3). These compounds were intentionally selected since they have a range of activity paralleling those used in the training and test set during the model building and because we recognized in them electrostatic, steric and hydrophobic factors that can give approximately the same contribution in the interaction with PLK1 as in the analysis for TP and IP derivatives.

Table 3 lists the predicted activity values for the external data set along with their leverage values (h_i) and standardized residuals (r_i). According to our calculations, Eq. (1) predicted high and low pIC_{50} values reasonably well and all compounds (congeneric and novels) were inside the applicability domain of the training set. In fact, the leverage values of the compounds are lower than the critical value ($h^* = 0.173$) and $r_i < 3$. Then, all chemicals are correctly predicted and the capability of detecting novel structures confirms our assumption about the structural factors that mediate the inhibitory activity over PLK1.

These results demonstrate that the developed model for inhibitory activity was adequate and can be considered an effective tool for new ‘in silico’ inhibitors discovery.

Table 3
Plk1 kinase inhibitors selected as external test set.

Molecule	Exp. pIC_{50}	Pred. pIC_{50}	h_i	r_i
	8.1550	7.5465	0.036	1.36
	6.5229	7.5234	0.036	−2.23
 GSK461364A	8.6990	8.3011	0.053	0.89
 BI 2536	9.0970	8.5112	0.085	1.30
 BI 6727	9.0605	8.4220	0.151	1.42
 ON01910	8.0222	8.3749	0.079	−0.79
	7.2924	6.7508	0.176	1.20

4. Conclusion

The inhibitory activity of 103 thiophene benzimidazole (TP) and 33 imidazopyridine (IP) derivatives against Polo-Like Kinase 1 (PLK1) was quantitatively analyzed in terms of a wide set of chemometric descriptors. Before modeling, a procedure of chemical data curation was applied to enable the development of reliable and predictive QSAR models. Our analysis suggested that QSAR methods can be applied any difficulty on the entire dataset of 136 compounds.

Two methodologies were used to split the original chemical data set into training and test sets, and the consequences on the models performance were analyzed.

A model with an optimistic predictive performance was obtained with the training set selected using D-optimal design and the 3D descriptor space along with activity values.

The model was applied to predict pIC_{50} for seven related compounds which were not included in model building.

The predictive ability of the combination of variables *RDF*035*m*, *ASyV*2, *Small_RSI_Mol*, *E*2*v*, and *Mor*02*p*, ($R^2_{test} = 0.71; Q^2 = 0.69$), highlighted the importance of the molecular size and shape, and the steric interactions in modeling the studied property.

From a detailed examination of the correlations found, we recognized that a potent inhibitor of PLK1 is described as a structure with a subtle balance between steric, hydrophobic and electrostatic factors which could play an important role in the binding affinity to the PLK1 active site. Similar results were also identified as guidelines for optimization strategies in previous SAR studies. Then, the proposed model can be utilized to design and predict new potent compounds as PLK1 inhibitor candidates, and to discover compounds with novel scaffolds that can act as PLK1 inhibitors via similar mechanisms.

Acknowledgements

We thank the financial support provided by the National Research Council of Argentina (CONICET) PIP11220100100151 project and to Ministerio de Ciencia, Tecnología e Innovación Productiva for the electronic library facilities. N.C.C. acknowledges a fellowship from CONICET and Facultad de Ciencias Agrarias, Universidad Nacional de Catamarca.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ejps.2014.05.029.

References

Atkinson, A.C., 1991. Plots, Transformation, Regression. Clarendon Press, Oxford, UK.

Bohari, M.H., Srivastava, H.K., Sastry, G.N., 2011. Analogue-based approaches in anti-cancer compound modelling: the relevance of QSAR models. Org. Med. Chem. Lett. 1, 1–12.

Breneman, C.M., Rhem, M., 1997. A QSPR analysis of HPLC column capacity factors for a set of high-energy materials using electronic Van der waals surface property descriptors computed by the transferable atom equivalent method. J. Comput. Chem. 18, 182–197.

Breneman, C.M., Thompson, R., Rhem, M., Dung, M., 1995. Electron density modeling of large systems using the transferable atom equivalent method. Comput. Chem. 19, 161–179.

Bultinck, P., Langenaeker, W., Carbó-Dorca, R., Tollenaere, J.P., 2003. Fast calculation of quantum chemical molecular descriptors from the electronegativity equalization method. J. Chem. Inf. Comput. Sci. 43, 422–428.

Cao, S., 2012. QSAR, molecular docking studies of thiophene and imidazopyridine derivatives as polo-like kinase 1 inhibitors. J. Mol. Struct. 1020, 167–176.

Chahrour, O., Cairns, D., Omran, Z., 2012. Small molecule kinase inhibitors as anti-cancer therapeutics. Mini Rev. Med. Chem. 12, 399–411.

Cherkasov, A., 2005. Inductive QSAR descriptors. Distinguishing compounds with antibacterial activity by artificial neural networks. Int. J. Mol. Sci. 6, 63–86.

Chopra, P., Sethi, G., Dastidar, S.G., Ray, A., 2010. Polo-like kinase inhibitors: an emerging opportunity for cancer therapeutics. Expert Opin. Investig. Drugs 19, 27–43.

Consonni, V., Todeschini, R., 2010. Molecular Descriptors. In: Puzyn, T., Leszczynski, J., Cronin, M.T.D. (Eds.), Recent Advances in QSAR Studies. Methods and Applications. Springer Science+Business Media, New York.

Cronin, M.T.D., 2010. Quantitative Structure–Activity Relationships (QSARs) – Applications and Methodology. In: Puzyn, T., Leszczynski, J., Cronin, M.T.D. (Eds.), Recent Advances in QSAR Studies: Methods and Applications. Springer Science & Business Media, New York.

Degenhardt, Y., Lampkin, T., 2010. Targeting polo-like kinase in cancer therapy. Clin. Cancer Res. 16, 384–389.

Draper, N.R., Smith, H., 1981. Applied Regression Analysis. John Wiley & Sons, New York.

Duchowicz, P.R., Castro, E.A., Fernández, F.M., González, M.P., 2005. A new search algorithm of QSPR/QSAR theories: normal boiling points of some organic molecules. Chem. Phys. Lett. 412, 376–380.

Duchowicz, P.R., Castro, E.A., Fernández, F.M., 2006. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. MATCH Commun. Math. Comput. Chem. 55, 179–192.

Duchowicz, P.R., Talevi, A., Bruno-Blanch, L.E., Castro, E.A., 2008. New QSPR study for the prediction of aqueous solubility of drug-like compounds. Bioorg. Med. Chem. Lett. 16, 7944–7955.

E-Dragon 1.0, 2005. Milano Chemometrics and QSAR Research Group. In: VCCLAB, Virtual Computational Chemistry Laboratory, http://michem.disat.unimib.it/chm.

Emmitte, K.A., Adjebang, G.M., Andrews, C.W., Badiang Alberti, J.G., Bambal, R., Chamberlain, S.D., Davis-Ward, R.G., Dickson, H.D., Hassler, D.F., Hornberger, K.R., Jackson, J.R., Kuntz, K.W., Lansing, T.J., Jr., M.R.A., Nailor, K.E., Pobanz, M.A., Smith, S.C., Sung, C.M., Cheung, M., 2009a. Design of potent thiophene inhibitors of polo-like kinase 1 with improved solubility and reduced protein binding. Bioorg. Med. Chem. Lett. 19, 1694–1697.

Emmitte, K.A., Andrews, C.W., Badiang, J.G., Davis-Ward, R.G., Dickson, H.D., Drewry, D.H., Emerson, H.K., Epperly, A.H., Hassler, D.F., Knick, V.B., Kuntz, K.W., Lansing, T.J., Linn, J.A., Mook, R.A.J., Nailor, K.E., Salovich, J.M., Spehar, G.M., Cheung, M., 2009b. Discovery of thiophene inhibitors of polo-like kinase. Bioorg. Med. Chem. Lett. 19, 1018–1021.

Golbraikh, A., Tropsha, A., 2002. Beware of q2! J. Mol. Graph. Model. 20, 269–276.

Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.D., Lee, K.H., Tropsha, A., 2003. Rational selection of training and test sets for the development of validated QSAR models. J. Comput. Aided Mol. Des. 17, 241–253.

Goodarzi, M., Duchowicz, P.R., Wu, C.H., Fernández, F.M., Castro, E.A., 2009. New hybrid genetic based Support Vector Regression as QSAR approach for analyzing flavonoids-GABA(A) complexes. J. Chem. Inf. Model. 49, 1475–1485.

Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 26, 694–701.

Gramatica, P., 2013. On the Development and Validation of QSAR Models. In: Reisfeld, B., Mayeno, A.N. (Eds.), Computational Toxicology. Springer Science & Business Media, New York, U.S.A.

Harris, P.S., Venkataraman, S., Alimova, I., Birks, D.K., Donson, A.M., Knipstein, J., Dubuc, A., Taylor, M.D., Handler, M.H., Foreman, N.K., Vibhakar, R., 2012. Polo-like kinase 1 (PLK1) inhibition suppresses cell growth and enhances radiation sensitivity in medulloblastoma cells. BMC Cancer 12, 1–13.

HyperChem Release 7.5 for Windows, Hypercube, Inc., http://www.hyper.com, 2002.

Lansing, T.J., McConnell, R.T., Duckett, D.R., Spehar, G.M., Knick, V.B., Hassler, D.F., Noro, N., Furuta, M., Emmitte, K.A., Gilmer, T.M., Mook, R.A.J., Cheung, M., 2007. In vitro biological activity of a novel small-molecule inhibitor of polo-like kinase 1. Mol. Cancer Ther. 6, 450–459.

Liu, Y., Gray, N.S., 2006. Rational design of inhibitors that bind to inactive kinase conformations. Nat. Chem. Biol. 2, 358–364.

Marengo, E., Todeschini, R.A., 1992. New algorithm for optimal distance- based experimental design. Chemom. Int. Lab. Syst. 16, 37–44.

Martin, T.M., Harten, P., Young, D.M., Muratov, E.N., Golbraikh, A., Zhu, H., Tropsha, A., 2012. Does rational selection of training and test sets improve the outcome of QSAR modeling? J. Chem. Inf. Model. 52, 2570–2578.

Matlab, version 7.12., The MathWorks, Inc. http://www.mathworks.com, 2011.

Mercader, A.G., Duchowicz, P.R., Fernandez, F.M., Castro, E.A., 2010. Replacement method and enhanced replacement method versus the genetic algorithm approach for the selection of molecular descriptors in QSPR/QSAR theories. J. Chem. Inf. Model. 50, 1542–1548.

Mercader, A.G., Duchowicz, P.R., Fernandez, F.M., Castro, E.A., 2011. Advances in the replacement and enhanced replacement method in QSAR and QSPR theories. J. Chem. Inf. Model. 51, 1575–1581.

Murugan, R.N., Park, J.E., Kim, E.H., Shin, S.Y., Cheong, C., Lee, K.S., Bang, J.K., 2011. Plk1-targeted small molecule inhibitors: molecular basis for their potency and specificity. Mol. Cells 32, 209–220.

Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T.D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D.W., Schultz, T.W., Stanton, D.T., van de Sandt, J.J.M., Tong, W., Veith, G., Yang, C., 2005. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. ATLA 33, 155–173.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R., 2011. Open Babel: an open chemical toolbox. J. Cheminf. 3, 1–14.

Pasquale, G., Romanelli, G.P., Autino, J.C., García, J., Ortiz, E.V., Duchowicz, P.R., 2012. Quantitative structure–activity relationships on chalcone derivatives: mosquito larvicidal studies. J. Agric. Food Chem. 60, 692–697.

Pomilio, A.B., Giraudo, M.A., Duchowicz, P.R., Castro, E.A., 2010. QSPR analyses for aminograms in food: citrus juices and concentrates. Food Chem. 123, 917–927.

Recon 5.5, 2002. Rensselaer Polytechnic Institute, T., http://www.drugmining.com, New York, USA.

Rheault, T.R., Donaldson, K.H., Badiang-Alberti, J.G., Davis-Ward, R.G., Andrews, C.W., Bambal, R., Jackson, J.R., Cheung, M., 2010. Heteroaryl-linked 5-(1H-benzimidazol-1-yl)-2-thiophenecarboxamides: potent inhibitors of polo-like kinase 1 (PLK1) with improved drug-like properties. Bioorg. Med. Chem. Lett. 20, 4587–4592.

Roy, K., 2007. On some aspects of validation of predictive quantitative structure–activity relationship models. Expert Opin. Drug Discov. 2, 1567–1577.

Roy, P.P., Roy, K., 2007. On some aspects of variable selection for partial least squares regression models. QSAR Comb. Sci. 27, 302–313.

Rudolph, D., Steegmaier, M., Hoffmann, M., Grauert, M., BAum, A., Quant, J., Haslinger, C., Garin-Chesa, P., Adolf, G.R., 2009. BI 6727, A polo-like kinase inhibitor with improved pharmacokinetic profile and broad antitumor activity. Clin. Cancer Res. 15, 3094–3102.

Sato, Y., Onozaki, Y., Sugimoto, T., Kurihara, H., Kamijo, K., Kadowaki, C., Tsujino, T., Watanabe, A., Otsuki, S., Mitsuya, M., Iida, M., Haze, K., Machida, T., Nakatsuru, Y., Komatani, H., Kotani, H., Iwasawa, Y., 2009. Imidazopyridine derivatives as potent and selective polo-like kinase (PLK) inhibitors. Bioorg. Med. Chem. Lett. 19, 4673–4678.

Song, B., Liu, X.S., Liu, X., 2012. Polo-like kinase 1 (Plk1): an un expected player in DNA replication. Cell. Div. 7, 1–7.

Strebhardt, K., Ullrich, A., 2006. Targeting polo-like kinase 1 for cancer therapy. Nat. Rev. Cancer 6, 321–330.

Sushko, I., Novotarskyi, S., Kürner, R., Kumar Pandey, A., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V.V., Tanchuk, V.Y., Todeschini, R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J., Schwab, C., Baskin, I.I., Palyulin, V.A., Radchenko, E.V., Welsh, W.J., Kholodovych, V., Chekmarev, D., Cherkasov, A., Aires-de-Sousa, J., Zhang, J., Bender, A., Nigsch, F., Patiny, Williams, A., Tkachenko, V., Tetko, I.V., 2011. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J. Comput. Aided Mol. Des. 25, 533–554.

Talevi, A., Goodarzi, M., Ortiz, E.V., Duchowicz, P.R., Bellera, C.L., Pesce, G., Castro, E.A., Bruno-Blanch, L.E., 2011. Prediction of drug intestinal absorption by new linear and non-linear QSPR. Eur. J. Med. Chem. 46, 218–228.

Todeschini, R., Consonni, V., 2009. Molecular Descriptors for Chemoinformatics. Wiley-VCH, Weinheim, Germany.

Todeschini, R., Consonni, V., Gramatica, P., 2009. Chemometrics in QSAR. In: Brown, S., Tauler, R., Walczak, R. (Eds.), Comprehensive Chemometrics. Elsevier, Oxford, pp. 129–172.

Todeschini, R., Ballabio, D., Consonni, V., Mauri, A., Cassotti, M., 2012. Distance-based optimal design. Milano Chemometrics & QSAR Research Group. University of Milano-Bicocca Milano, Italy.

Zhang, J., Yang, P.Y., Gray, N.S., 2009. Targeting cancer with small molecule kinase inhibitors. Nat. Rev. Cancer 9, 28–39.