

Multiple Linear Regression QSAR Studies of CDK4 and FLT3 Inhibitors for the Identification of Potential Dual Kinase Inhibitors

Suvarna N. Vakare ^{1*}, Rajesh B. Patil ²

^{1,2}Department of Pharmaceutical Chemistry, Sinhgad Technical Education Society's, Sinhgad College of Pharmacy, Vadgaon (Bk), Pune 411041, Maharashtra, India
Email ID:- vakare.suvarna@gmail.com

ABSTRACT

The escalation of chronic, non-communicable gastrointestinal disorders in contemporary society is directly tied to aberrant Dual kinase inhibitors hold promising therapeutic applications in various cancers, including acute myeloid leukaemia (AML). Cyclin-dependent kinase 4 (CDK4) and FMS-like tyrosine kinase 3 (FLT3) are relevant therapeutic targets in haematological cancers and AML. However, the design of dual kinase inhibitors of CDK4/FLT3 is challenging, and quantitative approaches for identifying promising compounds with dual kinase activity are limited. In the present study, multiple linear regression (MLR) quantitative structure-activity relationship (QSAR) models were developed for CDK4 and FLT3 inhibitors using QSARINS with genetic algorithm-based variable selection. The IC₅₀ data and structures of these inhibitors were retrieved from ChEMBL. Data were further curated with the DataWarrior tool for salt removal, structure transformation, and conversion of IC₅₀ to pIC₅₀. The final datasets comprised 195 CDK4 inhibitors (pIC₅₀ ranging from 5.75 to 9.00) and 54 FLT3 inhibitors (pIC₅₀ ranging from 6.84 to 9.70). PaDEL descriptors representing various structural features were calculated for both datasets, and constant and highly correlated descriptors were removed, followed by elastic net feature selection to obtain 144 unique descriptors. The CDK4 model, developed using 158 training and 37 test compounds with ten descriptors, showed good statistical fit ($R^2 = 0.774$; $R^2_{\text{adj}} = 0.758$; $F = 50.3$) and good internal predictive performance ($Q^2_{\text{LOO}} = 0.735$; $Q^2_{\text{LMO}} = 0.727$). Y-scrambling analysis confirmed there is a low risk of chance correlation ($R^2_{\text{Yscr}} = 0.063$). The external validation showed good statistical fit with $R^2_{\text{ext}} = 0.501$, $Q^2_{\text{F1}} = 0.518$, $Q^2_{\text{F3}} = 0.739$, and $\text{RMSE}_{\text{ext}} = 0.395$, and test compounds within the model's applicability domain. For the FLT3 inhibitor dataset, the eight-descriptor model showed good predictive performance ($R^2 = 0.893$; $Q^2_{\text{LOO}} = 0.848$; $Q^2_{\text{LMO}} = 0.828$). The QSAR models highlighted the contribution of binary fingerprint, Burden eigenvalue, and radial distribution function descriptors to CDK4 and FLT3 inhibitory activity. The developed QSAR models could be useful for prioritising compounds with potential dual CDK4/FLT3 activity.

Keywords:-QSAR, CDK4 inhibitors, FLT3 inhibitors, Dual kinase inhibitors, MLR

How to cite this article Suvarna N. Vakare , Rajesh B. Patil Multiple Linear Regression Qsar Studies of Cdk4 and Flt3 Inhibitors for the Identification of Potential Dual Kinase Inhibitors Int J Drug Deliv Technol. 2026;2173-2183(7): DOI: 10.25258/ijddt.16.7.222

Source of support: Nil.

Conflict of interest: None

1.INTRODUCTION

Cyclin-dependent kinase 4 (CDK4) is involved in the transition of cells from the G₁ to the S phase of the cell cycle. It binds to D-type cyclins and, upon binding, phosphorylates the retinoblastoma protein, which drives the cell cycle to DNA synthesis and division. Aberration in this process results in several type of cancers. Thus, CDK4/6 driven pathway is one of the targets of cancer intervention, and currently several CDK4/6 inhibitors are clinically used, including palbociclib, ribociclib, and abemaciclib [1,2]. Hence, there is a continued need for investigation of CDK4 as a target for anticancer drug discovery. FMS-like tyrosine kinase 3 (FLT3) is another important target in acute myeloid leukaemia (AML). FLT3 mutations, particularly internal tandem duplications (FLT3-ITD), are responsible for leukemic cell proliferation in the majority of AML patients. The FLT3 inhibitors such as midostaurin and gilteritinib are clinically used for patients with FLT3-mutated AML [3,4]. However, resistance and haematological toxicities remain a challenge. Hence,

investigations and the search for newer FLT3 inhibitors, and inhibitors that target more than one pathway involved in cancer, are current research needs.[5,6]. CDK4 regulates cell-cycle progression, whereas FLT3 supports growth and survival signalling in AML. Hence, targeting both CDK4 and FLT3 provides a useful basis for exploring dual target inhibition. Previous reports suggest efforts in the development of dual inhibitors of CDK4/FLT3. AMG 925, a dual CDK4/FLT3 inhibitor, is in clinical trials with potent activity against both targets [5]. Several other CDK4/FLT3 dual inhibitors, including FN-1501, are in clinical trials [1,7]. Thus, the literature suggests dual inhibition as a potential drug discovery strategy, especially in AML by targeting CDK4 and FLT3. Quantitative structure-activity relationship (QSAR) modelling can help identify links between molecular structure and biological activity [8]. It uses molecular descriptors of the compounds to build mathematical models from the experimentally determined activity of the compounds. Databases such as ChEMBL provide bioactivity data along with the structures and

*Author for Correspondence: vakare.suvarna@gmail.com

experimental conditions [9,10]. Multiple linear regression (MLR) is useful for developing interpretable QSAR models. The descriptor selection method, which identifies key descriptors relevant to the activity, is important in MLR-based QSAR models. The QSAR models, in turn, can be validated using suitable statistical measures, internal and external validation, and the applicability domain of the model [11–13]. For simple molecular descriptors, PaDEL [14] descriptors, which include 1D-3D descriptors and several fingerprint descriptors, are useful in QSAR studies. QSARINS software is useful in QSAR model development with MLR methodology, along with several in-built options for training set and test set splitting, selecting a genetic algorithm for model equation calculation, and options for internal and external model validations [15]. In this study, curated CDK4 and FLT3 inhibitor datasets from ChEMBL were analysed separately to develop MLR-based QSAR models. The objective of this QSAR study was to identify molecular descriptors associated with activity against CDK4 and FLT3 targets, and assess the developed models'

statistical performance with appropriate validation steps. The dataset's compound activity data for both targets were evaluated to establish the pattern relevant to potential dual CDK4/FLT3 inhibition. The study is expected to support the prioritisation of dual inhibitors of CDK4/FLT3 and help in the design of such dual inhibitors.

METHODS

Datasets of CDK4 and FLT3 and dataset curation

The dataset of CDK4 and FLT3 inhibitors was retrieved from ChEMBL. The dataset was filtered to a common structural core in the DataWarrior program, such that the dataset represents a congeneric series, excluding structurally unrelated compounds. The common shared scaffold was a pyrido[4,3-b]pyrrolo[2,3-d]pyrimidine tricycle bearing a cycloalkyl group at the pyrrole nitrogen and an arylamino substituent at C2; the three core variants present across the two series are the N-cyclopentyl, N-cyclohexyl and N-(trans-4-substituted cyclohexyl) analogues (Figure 1)...

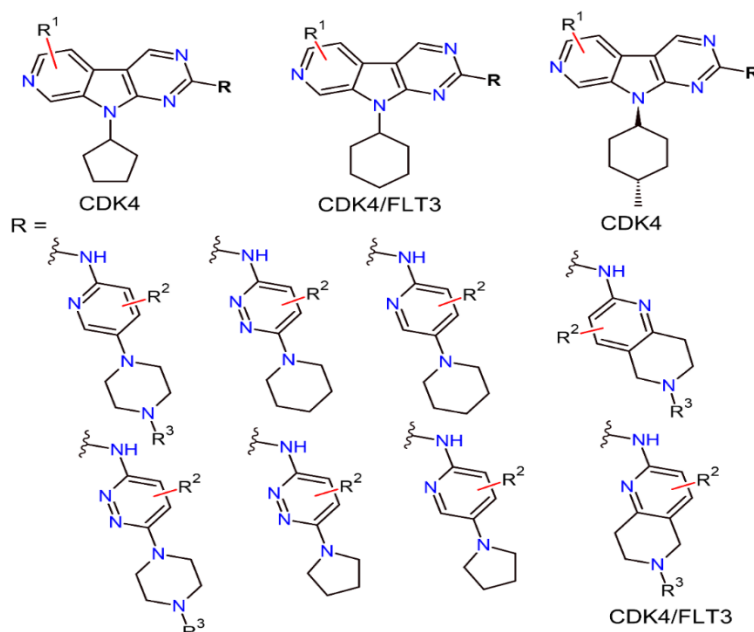


Figure 1. Common structural scaffolds of the CDK4 and FLT3 dataset and substituents.

The CDK4 dataset details and FLT3 dataset details are provided in Tables 1 and 2, respectively

Table 1: CDK4 dataset compound details

CHEMBL ID	pIC ₅₀	CHEMBL ID	pIC ₅₀	CHEMBL ID	pIC ₅₀
CHEMBL3698649	9.0000	CHEMBL3642024	8.2441	CHEMBL3702099	6.2218
CHEMBL3702053	9.0000	CHEMBL3641996	8.2365	CHEMBL3641963	6.1249
CHEMBL3702060	9.0000	CHEMBL3698742	8.2291	CHEMBL3641962	6.0861
CHEMBL3642037	9.0000	CHEMBL3642038	8.2076	CHEMBL3684988	6.0721
CHEMBL3702046	8.9208	CHEMBL3702095	8.1938	CHEMBL3642039	6.0315
CHEMBL3237702	8.9208	CHEMBL3698735	8.1739	CHEMBL3684987	5.7471

CHEMBL3698698	8.8860	CHEMBL3702044	8.1611	CHEMBL5624526	8.1549
CHEMBL3642002	8.8860	CHEMBL3702189	8.1426	CHEMBL6162969	8.1549
CHEMBL3702120	8.8569	CHEMBL3642022	8.1426	CHEMBL6175521	8.0000
CHEMBL3698706	8.8538	CHEMBL3642034	8.0969	CHEMBL5900891	7.2218
CHEMBL3698642	8.8239	CHEMBL3642036	8.0969	CHEMBL3642021	8.3279
CHEMBL3641998	8.8239	CHEMBL3698737	8.0757	CHEMBL3641999	8.3187
CHEMBL3698654	8.7958	CHEMBL3702211	8.0655	CHEMBL3642014	8.3187
CHEMBL3698744	8.7958	CHEMBL3698648	8.0177	CHEMBL3642035	8.3098
CHEMBL3698692	8.7695	CHEMBL3642028	8.0132	CHEMBL3702082	8.3010
CHEMBL3641997	8.7695	CHEMBL3702109	8.0000	CHEMBL3702083	8.3010
CHEMBL3642031	8.7695	CHEMBL3684978	7.9788	CHEMBL3702160	8.3010
CHEMBL3702050	8.7447	CHEMBL3698679	7.9586	CHEMBL3702185	8.3010
CHEMBL3702110	8.7447	CHEMBL3702045	7.9586	CHEMBL3702210	8.3010
CHEMBL3698739	8.7212	CHEMBL3702074	7.9586	CHEMBL3698749	8.2924
CHEMBL3702091	8.7212	CHEMBL3702184	7.9586	CHEMBL3698689	8.2839
CHEMBL3642000	8.7212	CHEMBL3642025	7.9586	CHEMBL3642010	8.2839
CHEMBL3698693	8.6989	CHEMBL3237451	7.9586	CHEMBL3702073	8.2596
CHEMBL3698746	8.6989	CHEMBL3698687	7.9208	CHEMBL3698733	8.2518
CHEMBL3702085	8.6989	CHEMBL3698738	7.9208	CHEMBL3702079	8.2518
CHEMBL3702092	8.6989	CHEMBL3642015	7.9208	CHEMBL3641960	6.7569
CHEMBL3702154	8.6989	CHEMBL3698638	7.8860	CHEMBL3641914	6.7520
CHEMBL3642019	8.6989	CHEMBL3702100	7.8860	CHEMBL3641922	6.6757
CHEMBL3642032	8.6989	CHEMBL3641992	7.8860	CHEMBL3698677	6.6252
CHEMBL3642033	8.6989	CHEMBL3702188	7.8538	CHEMBL3641923	6.5883
CHEMBL3237443	8.6989	CHEMBL3702056	7.8239	CHEMBL3641951	6.5228
CHEMBL3237444	8.6989	CHEMBL3702047	7.7958	CHEMBL3641958	6.5214
CHEMBL3680379	8.6777	CHEMBL3684984	7.7825	CHEMBL3641925	6.4814
CHEMBL3698646	8.6575	CHEMBL3698704	7.7695	CHEMBL3641956	6.4814
CHEMBL3642001	8.6575	CHEMBL3702057	7.7695	CHEMBL3641961	6.4559
CHEMBL3698703	8.6382	CHEMBL3640010	7.7447	CHEMBL3641964	6.4436
CHEMBL3237719	8.6197	CHEMBL3684986	7.7258	CHEMBL3684994	6.3344
CHEMBL3680385	8.6197	CHEMBL3702028	7.7212	CHEMBL3641913	6.2365
CHEMBL3702101	8.6197	CHEMBL3702078	7.7212	CHEMBL3641924	6.7986
CHEMBL3971812	8.6020	CHEMBL3642013	7.7212	CHEMBL3641921	6.7904
CHEMBL3702077	8.6020	CHEMBL3684985	7.7166	CHEMBL3641959	6.7695
CHEMBL3680381	8.5850	CHEMBL3702061	7.6989	CHEMBL3698731	8.3979
CHEMBL3702186	8.5850	CHEMBL3702094	7.6989	CHEMBL3702062	8.3872
CHEMBL3642030	8.5850	CHEMBL3237445	7.6989	CHEMBL3642026	8.3872
CHEMBL3702183	8.5528	CHEMBL3684983	7.6925	CHEMBL3698743	8.3767
CHEMBL3702093	8.5376	CHEMBL3642027	7.6197	CHEMBL3698748	8.3767
CHEMBL5207533	8.5228	CHEMBL3702097	7.5816	CHEMBL3702102	8.3767

CHEMBL3684993	8.5228	CHEMBL3237447	7.5686	CHEMBL3641994	8.3767
CHEMBL3698708	8.5228	CHEMBL3684981	7.5044	CHEMBL3702029	8.3665
CHEMBL3702059	8.5228	CHEMBL3702080	7.4814	CHEMBL3684982	8.3565
CHEMBL3698647	8.5086	CHEMBL3641965	7.4559	CHEMBL3698725	8.3565
CHEMBL3698732	8.5086	CHEMBL3702081	7.4317	CHEMBL3698741	8.3565
CHEMBL3237442	8.4948	CHEMBL3702055	7.3872	CHEMBL3702096	8.3565
CHEMBL3698734	8.4814	CHEMBL3641920	7.3565	CHEMBL3680380	8.3467
CHEMBL3698740	8.4685	CHEMBL3641966	7.3565	CHEMBL3702104	8.3467
CHEMBL3702052	8.4685	CHEMBL3702054	7.3010	CHEMBL3642016	8.3467
CHEMBL3702084	8.4685	CHEMBL3680384	7.2806	CHEMBL3680382	8.3372
CHEMBL3237704	8.4559	CHEMBL3698705	7.2518	CHEMBL3702103	8.3372
CHEMBL3702119	8.4436	CHEMBL3640009	7.2518	CHEMBL3702179	8.3372
CHEMBL3641995	8.4317	CHEMBL3641957	7.2441	CHEMBL3642009	8.3372
CHEMBL3698645	8.4202	CHEMBL3641919	7.1804	CHEMBL3698691	8.3979
CHEMBL3702051	8.4202	CHEMBL3702098	7.1674	CHEMBL3680383	6.9208
CHEMBL3698640	8.4089	CHEMBL3641918	7.1487	CHEMBL3698655	8.3979
CHEMBL3702155	8.4089	CHEMBL3680386	7.1378	CHEMBL3642012	6.9586

Table 2: FLT3 dataset details

CHEMBL ID	pIC ₅₀	CHEMBL ID	pIC ₅₀	CHEMBL ID	pIC ₅₀
CHEMBL3237712	9.6989	CHEMBL3680383	8.4814	CHEMBL3237717	8.5228
CHEMBL3237706	9.5228	CHEMBL3680386	8.4685	CHEMBL3237705	8.5228
CHEMBL3237718	9.2218	CHEMBL3237710	8.3979	CHEMBL3684983	8.5086
CHEMBL3237711	9.2218	CHEMBL3237716	8.3010	CHEMBL3684982	8.4948
CHEMBL3684989	9.1870	CHEMBL3237444	8.3010	CHEMBL3680380	8.4948
CHEMBL3237719	9.0000	CHEMBL3237450	8.3010	CHEMBL6191695	9.3187
CHEMBL5207533	9.0000	CHEMBL3237707	8.2218	CHEMBL6175521	8.0969
CHEMBL3237708	9.0000	CHEMBL3237713	8.2218	CHEMBL5624526	7.9586
CHEMBL3680379	8.8860	CHEMBL3237714	8.0969	CHEMBL6162969	7.9586
CHEMBL3684984	8.8538	CHEMBL3237715	8.0969	CHEMBL1288582	7.2218
CHEMBL3684986	8.7958	CHEMBL3237443	7.9586	CHEMBL3684994	8.6382
CHEMBL3684978	8.7695	CHEMBL3237442	7.8538	CHEMBL3684980	8.5686
CHEMBL3680381	8.7447	CHEMBL3237704	7.7695	CHEMBL3684985	8.5376
CHEMBL3684981	8.7447	CHEMBL3237446	7.7695	CHEMBL3237447	7.3372
CHEMBL3237703	8.6989	CHEMBL3680384	7.6695	CHEMBL3684979	7.1325
CHEMBL3237702	8.6989	CHEMBL3237445	7.6575	CHEMBL3684988	6.8416
CHEMBL3237709	8.6989	CHEMBL3684987	7.4814	CHEMBL3237451	8.6989
CHEMBL3680382	8.6989	CHEMBL3680385	7.4424	CHEMBL3237449	7.3665

The dataset was curated for salts; duplicates were removed, the compounds with definite IC₅₀ values were retained, and then converted to pIC₅₀ values using the following equation.

$$pIC_{50} = -\log_{10}(IC_{50} \times 10^9)$$

The curated datasets comprised 195 compounds for CDK4 and 54 compounds for FLT3. The dataset comprises 31 compounds common to both sets. The pIC₅₀ values range from 5.75 to 9.00 for CDK4, and 6.84 to 9.70 for FLT3.

Structure preparation and descriptor calculation

The 2D structures of the respective compounds from each dataset were converted to 3D structures after energy minimization using RDKit, and the single lowest energy conformation was saved in SDF format. Molecular descriptors were calculated with PaDEL Descriptor [Version 2.21] [16]. The descriptors included CDK fingerprints (FP), CDK graph-only fingerprints (GraphFP), CDK extended fingerprints (ExtFP), and Klekota Roth fingerprints. After removing constant and highly correlating descriptors the final descriptor sets included autocorrelation (ATSC2i, GATS5e, GATS2i), Burden-modified eigenvalue (SpMax2_Bhs, SpMin5_Bhv, VE2_Dzs), connectivity and path-count (VCH-6, piPC6, piPC8), molecular distance edge (MDEC-13), radial distribution function and fingerprint bit classes.

Training and test sets and model development

Each dataset was divided into a training set and an external prediction set by random splitting in QSARINS [17]. The CDK4 dataset was split into 156 training and 39 prediction compounds using an 80:20 split ratio. The FLT3 dataset was split into 44 training and 10 prediction compounds. Random splitting was employed to avoid user bias. Models were developed by genetic algorithm multiple linear regression (GA-MLR) implemented in QSARINS. The genetic algorithm was run with the leave one out (LOO) cross validation (Q²LOO), as the fitness function. Model performance was evaluated by examining average R² and Q² as a function of the number of variables from one to ten.

Model validation

Models were validated according to the OECD principles for the regulatory acceptance of QSAR models [18]. Goodness of fit was assessed by the coefficient of

determination R², the adjusted coefficient R²_{adj}, the Fisher ratio F, the root-mean-square error and mean absolute error of the training set (RMSE_{tr}, MAE_{tr}), the concordance correlation coefficient (CCC_{tr}) and the Friedman lack-of-fit function (LOF). The difference R² – R²_{adj} was observed as a measure of redundant descriptors. Internal prediction was assessed by LOO cross-validation (Q²LOO, RMSE_{cv}, MAE_{cv}, PRESS_{cv}, CCC_{cv}) and by leave-many-out cross-validation (Q²LMO). Any chance correlation was analysed by Y-randomisation, where the activity data were permuted, and the model parameters were determined. External predictive ability was assessed on the test set compounds, and the R²_{ext} was determined. Further, three more external prediction measures were determined namely correlation coefficient Q²-F1, Q²-F2 and Q²-F3.

Applicability domain analysis

The applicability domain was defined by the leverage approach and visualised as a Williams plot of standardised residuals against leverage. Leverage h_i is the diagonal element of the hat matrix for compound i, and the warning leverage h* was taken as

$$h^* = 3(p + 1) / n$$

where p is the number of descriptors in the model and n is the number of training compounds, giving h* = 0.2115 for the CDK4 model (p = 10, n = 156) and h* = 0.614 for the FLT3 model (p = 8, n = 44). Compounds with h_i > h* were considered as structural outliers.

Results and Discussion

Two independent MLR-based QSAR models were built for CDK4 and FLT3 datasets. Each model was developed, interpreted, and validated independently.

Dataset composition and chemical space

The CDK4 dataset contained 195 compounds, and the pIC₅₀ values spanned 3.25 log units in the range 5.75 to 9.00. The FLT3 dataset contained 54 compounds with pIC₅₀ values spanning 2.86 log units in the range 6.84 to 9.70 (Table 3). The 2 log unit minimum difference in activity or pIC₅₀ values is generally required for a meaningful regression [19].

Table 3: CDK4 and FLT3 dataset details

Property	CDK4 set	FLT3 set
Compounds	195	54
Training/test set	156 / 39	44 / 10
Compounds common in both datasets	31	
pIC ₅₀ range	5.75 - 9.00	6.84 - 9.70
Molecular weight range	400 - 570	264 - 547
Heavy atoms	30 - 41	20 - 40
Halogen-bearing compounds	46	9
Descriptors in final model	10	8

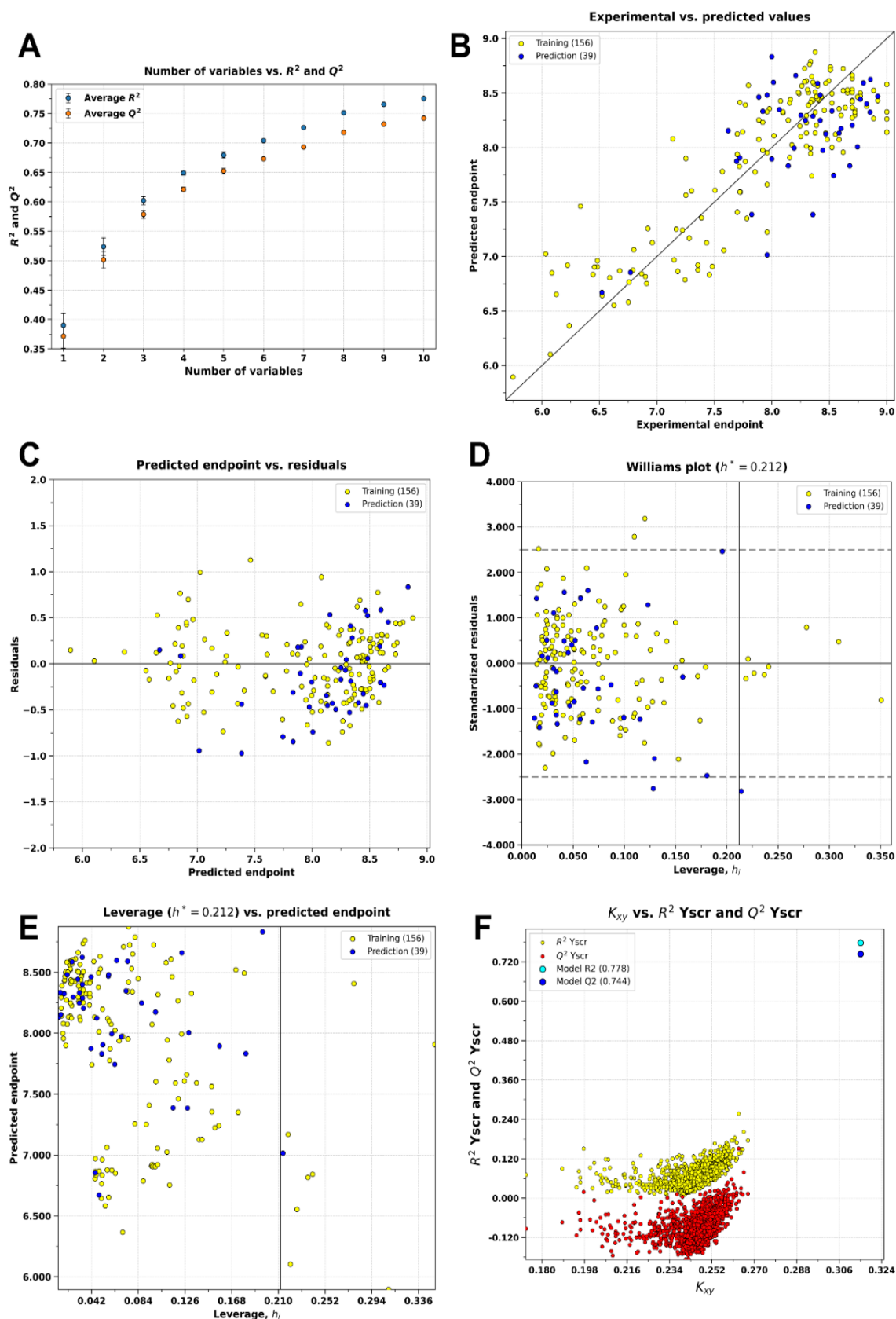


Figure 2: CDK4 best QSAR model statistics. A) Plot showing number of variables vs R^2/Q^2 , B) Plot showing experimental vs predicted values, C) Plot showing predicted values vs residuals, D) Williams plot, E) Plot showing applicability domain, F) Plot showing K_{xy} vs R^2 y-scramble values and Q^2 y-scramble values

Briefly, the dataset comprises the pyridopyrrolopyrimidine series analogues of dual CDK4/FLT3 inhibitor AMG 925, a pyrido[4,3-b]pyrrolo[2,3-d]pyrimidine tricyclic compounds with different N-substituents such as cyclopentyl, cyclohexyl or trans-4-substituted cyclohexyl, and different arylamino groups such as aminopyridine, aminopyridazine or tetrahydronaphthyridine, each bearing a terminal basic amine ring such as piperazine, piperidine or pyrrolidine (Figure 1). Thirty one

compounds were common in both datasets, enabling dual target analysis.

CDK4 MLR-based QSAR model development and model validation

The model selection dependence was analysed from the average R^2 and Q^2 , where the 10-descriptor model showed reasonably good predictive statistics and reached almost a plateau.

Table 4: Five best 10-descriptor models along with their statistical performance

Model	Descriptors unique to the model	R^2	R^2_{adj}	Q^2_{LOO}	$RMSE_{tr}$	$RMSE_{cv}$	R^2_{ext}	CCC_{ext}
1	ATSC2i, SpMin5_Bh v	0.779	0.763	0.744	0.364	0.391	0.359	0.579
2	GATS2i, SpMin5_Bh v	0.777	0.762	0.742	0.365	0.392	0.372	0.588
3	KRFP435, SpMin5_Bh v	0.777	0.761	0.744	0.365	0.391	0.417	0.63
4	GATS2i, GraphFP782	0.776	0.761	0.741	0.366	0.394	0.368	0.586
5	KRFPC435, SpMin5_Bh v	0.776	0.761	0.744	0.366	0.392	0.425	0.634

The five best 10-descriptor models differed in their external predictive performance and show slightly different unique descriptors contributing to the models (Table 4). All five models share a common eight-descriptor pool consisting the descriptors FP1013, FP874, VCH-6, SpMax2_Bhs, GraphFP255, RDF40v, GraphFP87 and MDEC-13. Model 1, with good internal and external predictive performance, was taken further as the best model for further interpretations. Models 3 and 5, where the unique descriptors KRFP435 and KRFPC435, instead of the Burden descriptor, showed good external statistics (R^2_{ext} 0.417 and 0.425) and are corroborating models. The Model 1 statistical results are shown in Figure 2B-F. The Model 1 fitting statistics were R^2 0.779, R^2_{adj} 0.763, F value 50.95 ($p < 0.0001$), $RMSE_{tr}$ 0.364, and CCC_{tr} 0.875. The difference in R^2 and R^2_{adj} is small (0.015), indicating that all the descriptors are important and no descriptor is redundant. The residual standard deviation of 0.377 is within an acceptable limit for a heterogeneous dataset.

CDK4 descriptor interpretation

The regression equation for model 1 is:

$$pIC_{50} = 8.5057 - 1.3533 (FP1013) - 0.8047 (FP874) - 7.3993 (VCH - 6) - 1.3101 (SpMax2_{Bhs}) - 0.0088 (ATSC2i) - 0.3035 (GraphFP255) + 0.0861 (RDF40v) + 3.9556 (SpMin5_{Bhv}) + 0.3981 (GraphFP87) - 0.1787 (MDEC13)$$

These ten descriptors belong to different classes such as CDK fingerprint, Burden eigenvalue, molecular distance edge, valence chain connectivity, radial distribution, graph-only fingerprint, and centered Broto-Moreau autocorrelation. FP874 is a CDK fingerprint class descriptor and showed a negative coefficient in the model equation, suggesting its contribution is detrimental to the model equation. SpMax2_Bhs belongs to a burden eigenvalue (I-state weighted) class descriptor and the negative coefficient suggested electro negative substituents at the extremes of structure reduce the

potency. FP1013 a CDK fingerprint class descriptor with negative coefficient suggested disfavoured substructure pattern. MDEC-13 is a molecular distance edge class descriptor, and the negative coefficient suggested that transformation of a primary carbon to a tertiary carbon makes the structures less active. VCH-6 is a valence chain connectivity class of descriptor and the negative coefficient suggested six membered ring induces unfavourable or lower activity. RDF40v is a radial distribution class of descriptor, and the positive coefficient suggests that bulky substituents away from 4 Å from the molecular centre are favourable. GraphFP255 is a graph-only fingerprint class of descriptor, and the negative coefficient suggests an unfavourable topological pattern in the molecules. GraphFP87 is another graph-only fingerprint descriptor, and the positive coefficient suggests it contributed favourably based on a unique topological pattern in molecules. ATSC2i is a centred Broto-Moreau autocorrelation class of descriptor and the negative coefficient suggested electron rich atom pairs two bonds apart are unfavourable. SpMin5_Bhv is a burden eigenvalue, volume-weighted class of descriptor, and the positive coefficient suggested a compact and evenly distributed volume in molecules is favourable. Overall, molecules with electronic and steric features evenly distributed have better CDK4 inhibitory activity.

CDK4 QSAR model robustness, Y-randomisation and applicability domain

Internal validation confirms that model 1 is stable. Leave-one-out (LOO) cross-validation with Q^2_{LOO} 0.744, $RMSE_{cv}$ 0.391 and CCC_{cv} 0.856; the difference in R^2 and Q^2_{LOO} of 0.034 is below the threshold 0.3, suggesting no overfitting. Leave-many-out (LMO) cross-validation, with Q^2_{LMO} 0.728, showed model 1 is robust and model is equally applicable to almost all compounds. Y-randomisation showed R^2_{Yscr} 0.065 and Q^2_{Yscr} -0.088, suggesting there is no chance correlation. The Williams plot with $h^* = 0.2115$ showed that most of the compounds remain inside the applicability domain and within ± 2.5 standardised residuals. However, five compounds fall outside the limit of standardised residuals, which include CHEMBL3684994, CHEMBL3642039, CHEMBL3702074, CHEMBL3684982, and CHEMBL3680386. However, these compounds have lower pIC_{50} values and were not considered further.

FLT3 MLR-based QSAR model development and model validation

The FLT3 dataset with 54 compounds supported an eight-descriptor model. The model selection dependence was analysed from the average R^2 and Q^2 , where the 8-descriptor model showed reasonably good predictive statistics and reached almost a plateau (Figure 3).

Table 5: Five best 8-descriptor models along with their statistical performance

Model	Descriptors unique to the model	R^2	R^2_{adj}	Q^2_{LOO}	$RMSE_{tr}$	$RMSE_{cv}$	CCC_{tr}
1	piPC6, RDF90v, RDF145e	0.893	0.869	0.848	0.204	0.244	0.944
2	AATSC5c, piPC6, RDF90e	0.891	0.867	0.844	0.206	0.247	0.943
3	piPC8, RDF90v, RDF145e	0.891	0.866	0.844	0.207	0.247	0.942
4	piPC6, RDF145e, RDF90p	0.891	0.865	0.843	0.207	0.248	0.942
5	piPC6, RDF90e, ExtFP879	0.89	0.865	0.847	0.208	0.245	0.942

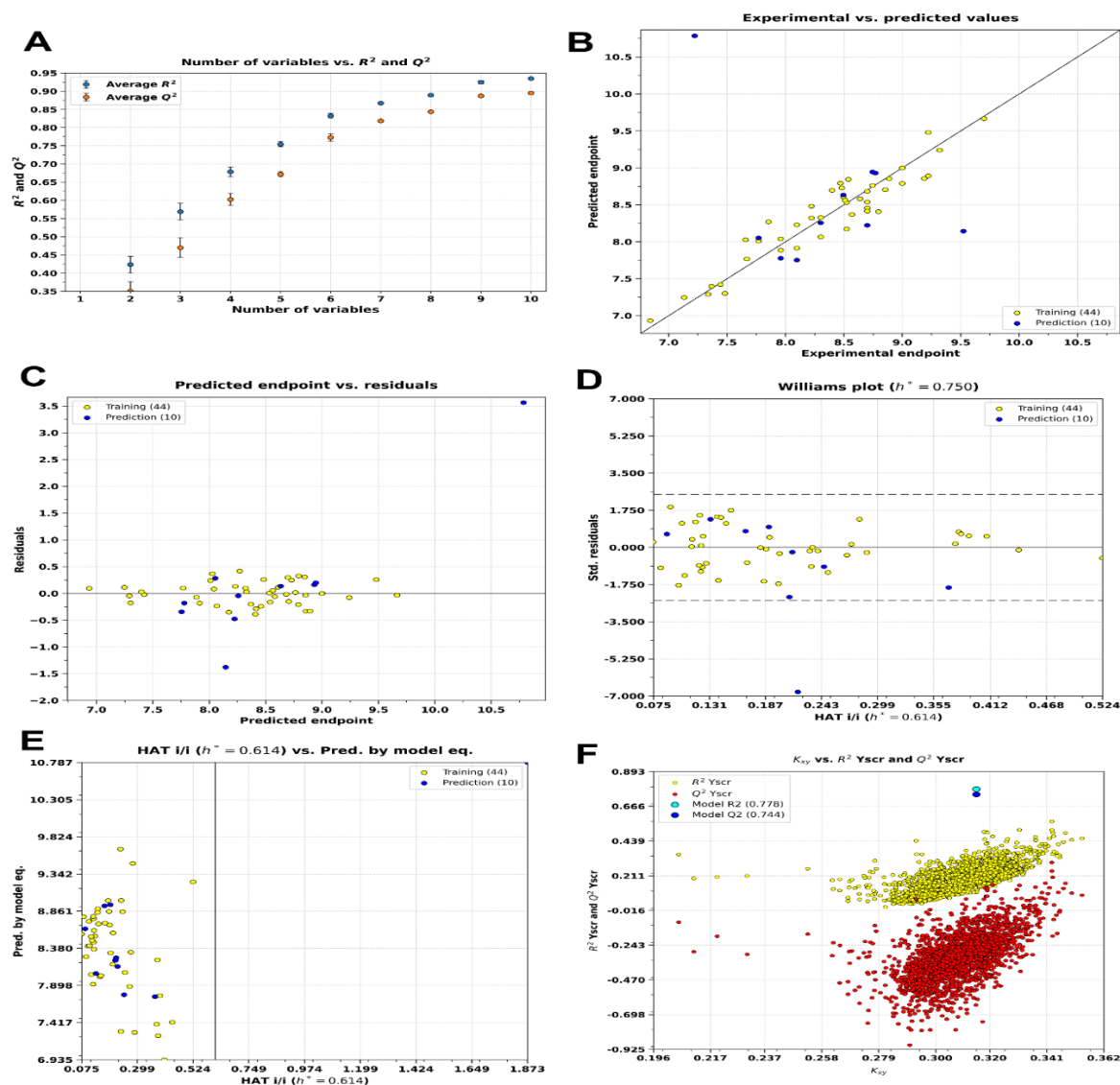


Figure 3: FLT3 best QSAR model statistics. A) Plot showing number of variables vs R^2/Q^2 , B) Plot showing experimental vs predicted values, C) Plot showing predicted values vs residuals, D) Williams plot, E) Plot showing applicability domain, F) Plot showing K_{xy} vs R^2 y-scramble values and Q^2 y-scramble values

All five models share GATS5e, VE2_Dzs and the three fingerprint bits FP206, FP640 and FP817, and differ only in the choice of path count (piPC6 or piPC8) and in which weighting of the radial distribution function is used (v, e or p) (Table 5). Internal validation was correspondingly strong: Q2LOO 0.848, CCCcv 0.919, Q2LMO 0.827 and an $R^2 - Q^2$ LOO difference of 0.045. Y-randomisation showed R^2 Yscr 0.184 and Q^2 Yscr -0.319. Overall, the MLR-based QSAR model with 8 descriptors showed good statistical measures.

FLT3 descriptor interpretation

The regression equation for FLT3 model 1 is:

$$\text{pIC}_{50} = 29.9215 + 1.1764 (\text{GATS5e}) - (\text{VE2_Dzs}) - 3.3183 (\text{piPC6}) - 0.0493 (\text{RDF90v}) - 0.0224 (\text{RDF145e}) + 0.8326 (\text{FP206}) - 1.1332 (\text{FP640}) - 0.9019 (\text{FP817})$$

The descriptor FP640 with a negative coefficient showed that this fingerprint is unfavourable for FLT3 activity.

Similarly, FP817 with a negative coefficient showed an unfavourable contribution to Model 1. The descriptor FP206, with a positive coefficient, is favourable for the activity. VE2_Dzs, an averaged Randic-like eigenvector, with negative coefficient is unfavourable which suggested delocalized electron system is unfavourable for the FLT3 activity. piPC6, a conventional bond-order path count descriptor with negative coefficient showed six bond conjugated system is unfavourable. GATS5e, a Geary autocorrelation descriptor with a positive coefficient, showed that electronegativity between atoms five bonds away is favourable for the activity. RDF145e, a radial distribution descriptor with a negative coefficient, showed that electronegative atoms too far away from the center of the molecule are unfavourable. RDF90v, another radial distribution descriptor with a negative coefficient, showed that bulky substituents 9 Å away from the center of the compound are unfavourable.

In external validation, model 1 showed a small number of compounds outside the applicability domain. For FLT3 QSAR models, the external statistics were slightly poor, with R^2_{ext} 0.149, Q^2_{F1} -3.115, CCExt -0.354, and RMSE_{ext} 1.231. Due to the small sample size, the external prediction might be poor. The compound CHEMBL1288582 was the outlier, having the leverage of 1.873 against the threshold h^* of 0.614.

CDK4/FLT3 QSAR models in design of dual inhibitors
The CDK4 and FLT3 MLR-based QSAR models were developed independently and can be applied together to predict the consequent CDK4 or FLT3 activity. Thirty-one compounds common to both datasets further confirmed good predictive statistics for both models. From this common dataset across both targets, the CDK4 and FLT3 activities showed weak correlation with $r = 0.365$ and $r^2 = 0.133$, suggesting these two activities are largely independent. Hence, dual target activity prediction might be feasible if these model QSARs are applied together to identify both CDK4 and FLT3 activities. It was observed that 15 of the 31 compounds predicted activity falls within ± 0.5 log units of the activity against both targets. These 10 compounds are hence key structural starting points for further design, and they hold the features of a dual pharmacophore. Overall, the cyclohexyl-substituted tricyclic core is necessary for dual CDK4/FLT3 activity, while

cyclopentyl and trans-4-substituted cyclohexyl cores are associated with CDK4 activity. Further, the N-cycloalkyl group is good for CDK4 activity, but slightly detrimental for FLT3 activity. The developed QSAR models utilized only a few 2D and fingerprint descriptors during regressions. Further studies on binding mode, docking, or molecular dynamics might be necessary to establish the importance of these distance-dependent descriptors.

CONCLUSION

Two MLR-based QSAR models were developed independently on congeneric series sharing a pyrido[4,3-b]pyrrolo[2,3-d]pyrimidine core containing compounds. The 10 descriptor model for CDK4 inhibition ($n = 195$; R^2 0.779, Q^2_{LOO} 0.744, Q^2_{LMO} 0.728) and the 8 descriptor model for FLT3 inhibition ($n = 54$; R^2 0.893, Q^2_{LOO} 0.848, Q^2_{LMO} 0.827). Both the QSAR models fulfil internal and, to some extent, external prediction. The Y-randomisation showed no chance correlation. The QSAR model for the CDK4 dataset showed the importance of steric and electronic character, and radial distance from the center of molecules. The QSAR model for FLT3 showed the importance of some key substructure patterns and fingerprint descriptors. The two QSAR models together could be useful for dual CDK4/FLT3 inhibitor design

REFERENCE

1. Zhang Y, Hsu C-P, Lu J-F, Kuchimanchi M, Sun Y-N, Ma J, et al. FLT3 and CDK4/6 inhibitors: Signaling mechanisms and tumor burden in subcutaneous and orthotopic mouse models of acute myeloid leukemia. *J Pharmacokinet Pharmacodyn*. 2014;41:675–91. <https://doi.org/10.1007/s10928-014-9393-x>
2. Uras IZ, Walter GJ, Scheicher R, Bellutti F, Prchal-Murphy M, Tigan AS, et al. Palbociclib treatment of FLT3-ITD+ AML cells uncovers a kinase-dependent transcriptional regulation of FLT3 and PIM1 by CDK6. *Blood*. 2016;127:2890–902. <https://doi.org/10.1182/blood-2015-11-683581>
3. Stone RM, Mandrekar SJ, Sanford BL, Laumann K, Geyer S, Bloomfield CD, et al. Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation. *N Engl J Med*. 2017;377:454–64. <https://doi.org/10.1056/NEJMoa1614359>
4. Perl AE, Martinelli G, Cortes JE, Neubauer A, Berman E, Paolini S, et al. Gilteritinib or Chemotherapy for Relapsed or Refractory FLT3 - Mutated AML. *N Engl J Med*. 2019;381:1728–40. <https://doi.org/10.1056/NEJMoa1902688>
5. Li Z, Wang X, Eksterowicz J, Gribble MW, Alba GQ, Ayres M, et al. Discovery of AMG 925, a FLT3 and CDK4 Dual Kinase Inhibitor with Preferential Affinity for the Activated State of FLT3. *J Med Chem*. 2014;57:3430–49. <https://doi.org/10.1021/jm500118j>
6. Wang Y, Zhi Y, Jin Q, Lu S, Lin G, Yuan H, et al. Discovery of 4-((7 H -Pyrrolo[2,3- d]pyrimidin-4-yl)amino)- N -(4-((4-methylpiperazin-1-yl)methyl)phenyl)-1 H -pyrazole-3-carboxamide (FN-1501), an FLT3- and CDK-Kinase Inhibitor with Potentially High Efficiency against Acute Myelocytic Leukemia. *J Med Chem*. 2018;61:1499–518. <https://doi.org/10.1021/acs.jmedchem.7b01261>
7. Li X, Yang T, Hu M, Yang Y, Tang M, Deng D, et al. Synthesis and biological evaluation of 6-(pyrimidin-4-yl)-1H-pyrazolo[4,3-b]pyridine derivatives as novel dual FLT3/CDK4 inhibitors. *Bioorganic Chemistry*. 2022;121:105669. <https://doi.org/10.1016/j.bioorg.2022.105669>
8. Kubinyi H. Quantitative structure-activity relationships (QSAR) and molecular modelling in cancer research. *J Cancer Res Clin Oncol*. 1990;116:529–37. <https://doi.org/10.1007/BF01637071>
9. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*. 2012;40:D1100–7. <https://doi.org/10.1093/nar/gkr777>
10. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*. 2024;52:D1180–92. <https://doi.org/10.1093/nar/gkad1004>
11. Gramatica P, Pilutti P, Papa E. Approaches for externally validated QSAR modelling of Nitrated

- Polycyclic Aromatic Hydrocarbon mutagenicity. SAR and QSAR in Environmental Research. 2007;18:169–78.
<https://doi.org/10.1080/10629360601054388>
12. Chirico N, Gramatica P. Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. *J Chem Inf Model*. 2011;51:2320–35.
<https://doi.org/10.1021/ci200211n>
 13. Roy K, Mitra I. On the Use of the Metric rm^2 as an Effective Tool for Validation of QSAR Models in Computational Drug Design and Predictive Toxicology. *MRMC*. 2012;12:491–504.
<https://doi.org/10.2174/138955712800493861>
 14. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*. 2011;32:1466–74.
<https://doi.org/10.1002/jcc.21707>
 15. Gramatica P. Principles of QSAR Modeling: Comments and Suggestions From Personal Experience. *International Journal of Quantitative Structure-Property Relationships*. 2020;5:61–97.
<https://doi.org/10.4018/IJQSPR.20200701.oa1>
 16. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*. 2011;32:1466–74.
<https://doi.org/10.1002/jcc.21707>
 17. Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S. QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. *J Comput Chem*. 2013;34:2121–32.
<https://doi.org/10.1002/jcc.23361>
 18. Gómez-Jiménez G, Gonzalez-Ponce K, Castillo-Pazos DJ, Madariaga-Mazon A, Barroso-Flores J, Cortes-Guzman F, et al. The OECD Principles for (Q)SAR Models in the Context of Knowledge Discovery in Databases (KDD). *Advances in Protein Chemistry and Structural Biology* [Internet]. Elsevier; 2018 [cited 2026 Oct 10]. p. 85–117.
<https://doi.org/10.1016/bs.apcsb.2018.04.001>
 19. Cortés-Ciriano I, Škuta C, Bender A, Svozil D. QSAR-derived affinity fingerprints (part 2): modeling performance for potency prediction. *J Cheminform*. 2020;12:41.
<https://doi.org/10.1186/s13321-020-00444-5>